CASE STUDY: ANALYSIS OF A DOUBLE-BLIND DATASET

Oliver Benning[1], Amani Chikh[1], Shenxin Gao[1], Xinyi Miao[1], Adam Rhyndress[1], Victoria Silverman[1], Ashley Shi[1], Nicole Vingerhoeds[1], Jingyuan Wang[1], Zijie Xia[1], Patrick Boily[1,2,3]

with additional contributions by Shintaro Hagiwara[3], Yue Huang[3], Andrew Macfie[3], and Jen Schellinck[2,4]

Abstract

Case studies and practical examples are often used as pedagogical tools to illustrate how the concepts and methods of data analysis, data science, and machine learning can be used in real-world situations (see [1, 8, 10, 19] for a number of impressive examples). This pedagogical picture is somewhat simplistic, and the contrast with real-world data can be jarring, especially for data scientists tackling their first project. In this case study, we showcase how data visualization can be used to better understand a fictitious (albeit realistic) transportation scenario. The objective remains firmly grounded in practical considerations: can we convey the intricacies and nuances of a complex dataset while still asking and answering interesting questions and providing useful insights?

Dataset and Code

The artificial dataset shares conceptual characteristics with *Canadian Air Transport Security Authority* data, but its content is emphatically **not** representative of actual Canadian airport pre-board screening data, and the questions that are asked, the answers that are provided, and the insights that arise may not necessarily be transferable to the Canadian context. The dataset (and some of the code) is available from GitHub at github.com/obenn/IQC4376-F20-DATA.git ; the data slicing tool can be found at basa.strikethrough.net .

Contributors

The bulk of the work was conducted by the students in the Fall 2020 edition of Introduction to Quantitative Consulting (MAT4376G) offered through the University of Ottawa's Department of Mathematics and Statistics. Contributions were also made by Idlewyld Analytics and Sysabee consultants.

Keywords

Case study, applications, data analysis, data science, quantitative methods, data cleaning, data validation, data processing, data visualization, data representation.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[2]Data Action Lab, Ottawa
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
[4]Sysabee, Ottawa
Email: pboily@uottawa.ca

Contents

1. Introduction

Data analysis case studies often play out like the edited and condensed replay of the top performances at an archery competition somehow finding it way to the nightly news:

- there's the introduction of the specific event that is being covered and a zoom on the target;
- a focus on one or two major athletes;
- highlights of some of the best attempts;
- an on-screen table showing the final scores;
- an excerpt of the medal ceremony, and
- a mention that this year's winner is the event's youngest ever, say, and so forth,
- all of which is packaged for an audience with (at

best) a burgeoning understanding of the context and relevance of what is being related.

In practice, insightful data analysis is closer in spirit to a dynamic hybrid of hockey and decathlon, played over a 6-month period by a handful of die-hards, with:

- numerous disconnected events;
- an evolving cast and a revolving set of rules and referees who contradict one another;
- a series of false starts and do-overs;
- stormy weather,
- own-goals galore,
- all of which is done in front of a tiny (and quite often hostile) crowd.

Reports of the first kind have the marked advantage of following a simpler narrative although they are not usually as 'true-to-life', while the complicated path of the second kind of study may have the unintended consequence of scaring away would-be data scientists, consultants, and clients.

Depending on the context, a case can be made for either approach, but the contrast between the idyllic vision and its rough-and-tumble brethren can be jarring when a student is in the midst of their first non-academic project and it suddenly starts to feel as though they accidentally wandered onto a leaking ship that is seconds away from hitting an iceberg in shark-infested waters just as a flying saucer is crashing into it from starboard.[1]

Consider this case study, then, to be an attempt to remedy the situation by preparing data analysts to face the reality of ... well, of data analysis.

Background We dive into a world almost identical to ours, but where people travel the skies on blimps and dirigibles instead of airplanes. Even though no project can ever be dissociated entirely from the context in which it arises, we will assume that the history of that world is largely irrelevant to the tasks at hand, and will forego the traditional attempts at justifying this alternative timeline.

Borealia is a large country in Vespuchia; four of its major cities have class-A airfields (see Figure 1). Borealia is bordered by several countries; consequently, a bevy of domestic and international travellers enter, leave, or pass through the country on a daily basis.

The nation's airspace security is assured by the *Borealian Aeronatics Security Agency* (BASA); the agency runs preboard screening of passengers and crew for all flights departing its class-A airfields. BASA has collected three years worth of data (20X6-20X8) about the passengers' wait time experiences at the 4 class-A airfields.

Figure 1. The nations of Vespuchia, with Borealia's 4 class-A airfields.

2. Objectives and Scope

This project's aims are to help BASA understand their data as it relates to its pre-boarding system, and to unleash its potential to lead to actionable insights.

In the absence of any real-life information about BASA and Borealia (which remain fictions, as a client and as a location, after all), we will assume that the various processes are similar to those occurring in Canadian airports, with the important caveat that the dataset and airfields are in **no way** to be considered as being part of a *roman-à-clef* about BASA's Canadian counterparts, the *Canadian Air Transport Security Agency* (CATSA).

We start with the client's presentation.

3. The Client and the Problem(s)

By providing efficient and effective pre-board screening (PBS), the Borealian Aeronautic Security Agency (BASA) ensures the safety of all passengers and crew aboard flights departing Borealian airfields while maintaining an appropriate balance between staffing and the wait time experienced by passengers.

PBS process The screening process is structurally similar at each airfield: passengers arriving at the beginning of the main queue may have their boarding passes scanned at the S_1 position, but they are always scanned at the S_2 position (see Figure 2).

[1] That feeling never really goes away, to be honest.

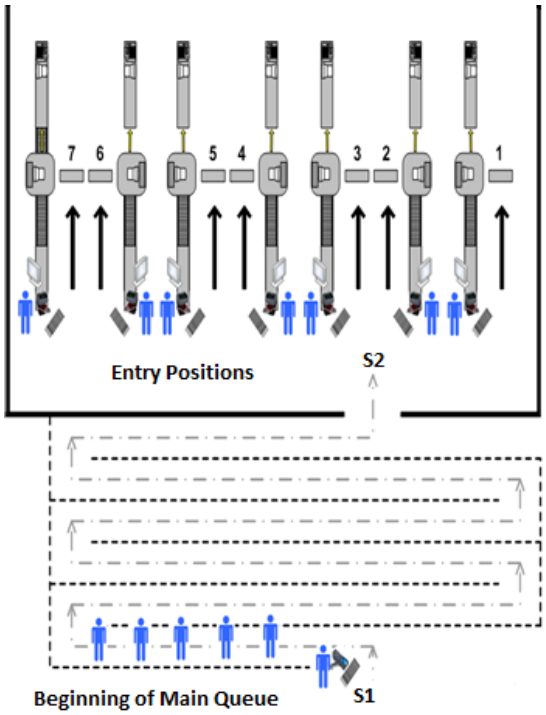


Figure 2. Schematics of pre-board screening (PBS). Passengers enter the main queue, where their boarding pass may be screened at S_1. Once they reach the end of the main queue, their boarding pass is screened at S_2 and they are sent to one of the active lines for processing.

Available Data For each of 4 airfields, BASA can provide the following data elements, dating from 20X6 to 20X8:

- **Airfield:** Auckland [AUC], Chebucto [CWL], Queenston [QUE], Saint-François [SAF];
- **Passenger ID:** a unique identifier for each passenger which exited the main queue;
- **Scan at** S_2: the date and time at which passengers exited the main queue, recorded to the nearest minute;
- **Wait Time**: the interval of time spent in the main queue, for each passenger which was also scanned upon entering it, rounded up to the minute;
- C_{start}: the reported number of active servers when a passenger entered the main queue, for each passenger with a recorded wait time (integer);
- C_0: the reported number of active servers when a passenger exited the main queue (integer);
- C_{avg}: the average reported number of active servers during the period spent in the main queue, for each passenger with a recorded wait time;
- **Scheduled Departure**: the scheduled departure time of each passenger's flight;
- **Actual Departure:** the actual departure time of each passenger's flight;
- **Destination City and Country**: the final airfield and country destination for each passenger exiting the main queue (may not be the flight destination).

Notes As a passenger may not have been scanned upon entering the main queue, the fields for the wait time, C_{start} and C_{avg} are sometimes empty. There are occasional blips with the other fields as well.

The number of active servers is reported in 15-minute blocks in a dataset to which we do not have access. The server vacation policy is fluid and may not necessarily use the same time blocks; since C_0 and C_{start} are integers, they are by definition estimates. All that can be said on the topic is that an external validation of these estimates has been conducted to show that the estimates are consistent and fairly representative of the real situation, but BASA still has reservations.

The Questions Numerous factors clearly influence the PBS wait time: the schedule intensity of departing flights, the volume of passengers on these flights, the number of servers and processing rates at a given airfield, etc.

But there might also be yearly, seasonal, time-of-day, day-of-week, and various interaction effects, depending on the specific airfield, on the flight destination, or any other factor. There could be trend level shifts in the number of passengers, flights, destinations, etc.

Ultimately, BASA is seeking an in-depth understanding of their data to help make Borealian airfields as efficient and secure as possible. For instance, BASA would ultimately like answers to the following questions (or if answers cannot be provided, an evidence-based argument to suggest what other information would be needed):

1. What does the dataset "look" like? What insights could be gleaned by visualizing the data?
2. What do anomalous observations look like at the passenger, flight, and active server levels?
3. In what circumstances are passengers not scanned at S_1?
4. When do passengers typically arrive to be scanned at S_1?
5. On average, how long do passengers wait in the main queue? What factors affect the waiting time?
6. Does server performance change according to traffic patterns?
7. Is it possible to forecast passenger arrival patterns based on flight schedule?
8. Is it possible to predict main queue waiting times given specific arrival patterns, flight schedule, and server vacation policy?
9. Is it possible to set a server vacation policy to control waiting times based on predicted arrival patterns?
10. Do passengers ever miss flights because of the waiting time? Can we find factors that are linked with missed flights?
11. Based on the size, schedule, and final destination of the passengers, what flights are most similar? Most dissimilar?

12. Do the number of flights and number of passengers exhibit seasonal patterns or trend level shifts?
13. Is there any way to detect if servers or airfields are not reporting their data correctly, either through fraud, or incompetence?
14. Can we predict the effects that temporarily shutting down an airfield or modifying the number of flights between airfields could have on the Borealian network?
15. Can anything else insightful be said about the data?

Outline In this chapter, we will mostly focus on a subset of the client's questions:

- we will start by describing our attempts to **Understand the System**, including some of the issues that were encountered in the later stages;
- this will be followed by section on **Data Preparation**, which includes: preliminary **Data Exploration**, **Data Cleaning**, and **Data Visualization**.

Due to time constraints, the data cleaning will be restricted to identifying and understanding **holes** in the data.

4. Understanding the BASA System

In order to understand how various aspects of the world (whether the BASA world or our own world) interact with one another, we need to **carve out chunks** corresponding to various system aspects and define their **boundaries**.

Thinking in Systems Terms Working in teams requires a **shared understanding** of what is being studied.

A **system** is made up of **objects** with **properties** that potentially change over time. Within the system we perceive **actions** and **evolving properties** leading us to think of the situation under study in terms of **processes**.

The objects themselves have various properties. Natural processes **generate** (or **destroy**) objects, and may **modify** the properties of these objects over time.

We **observe**, **quantify**, and **record** specific values of these properties at particular points in time. This generates data points, capturing the **underlying reality** to some degree of **accuracy**, but the process always yields errors and can at best create approximations.

Identifying Gaps in Knowledge A **gap in knowledge** is identified when we realize that what we thought we knew about a system proves incomplete or false. This might occur repeatedly, at any moment in the process:

- data preparation;
- data exploration, and
- data analysis.

In the BASA system, for instance, it was first assumed that observations existed for each passenger in the transportation network, but data exploration lead to a series of questions that smashed that assumption: only those passengers that are screened prior to boarding a flight are scanned – layovers are not included in the dataset.

When faced with a knowledge gap (and we promise that this will happen), the suggested approach is to be flexible:

- **revisit** your (explicit and implicit) assumptions;
- **return** to the client and **request** clarification and **ask** questions, and
- modify the system representation as required.

This process needs to be repeated in order to create an **explicit conceptual model**; work can eventually forge ahead, but the assumptions under which the team is labouring need to be stated explicitly.

Relating the Data to the System Is the data going to be of any use when it comes to understanding the system? This question can only be answered if we understand:

- how the data is **collected**;
- the **approximate nature** of data and system, and
- what the data represents (observations and features).

Whether the combination of system and data is **sufficient** to understand the aspects of the world under consideration is crucial: if the data, the system, and the world are out of alignment, insights might prove useless (see Figure 3).

4.1 The BASA System

4.1.1 Conceptual Models

We have created a three-tier system diagram to show how individuals interact with the Borealian Air Transport Network (BATN). Each tier represents a more granular view within the BASA system, with the third tier being the most granular.

The "External Interactions with the BATN" diagram of Figure 4 shows how local and international passengers, as well as employees, enter and exit the BATN.

Within the BATN, one can see how arriving passengers (both international and domestic) may, or may not, go through Pre-Board Screening (PBS) while all new passengers must go through PBS before boarding their departing flight. Furthermore, arriving international passengers need to go through customs before either entering PBS, boarding their flight, or exiting the airfield (see Figure 5).

PBS is a simple **first-in, first-out** (FIFO) **queuing system** with passengers entering the queue at S_1 and exiting at S_2, simultaneously beginning the scanning process (see Figure 2). Queues and queueing models will be revisited in a later chapter.

4.1.2 Relating the Data to the System

The dataset for Borealia's four class-A airfields contains about ten million records with twenty features each, spanning approximately three years. Importantly, the collected data represents **passengers** (and possibly BASA employees) who have gone through PBS in one of the four checkpoints.

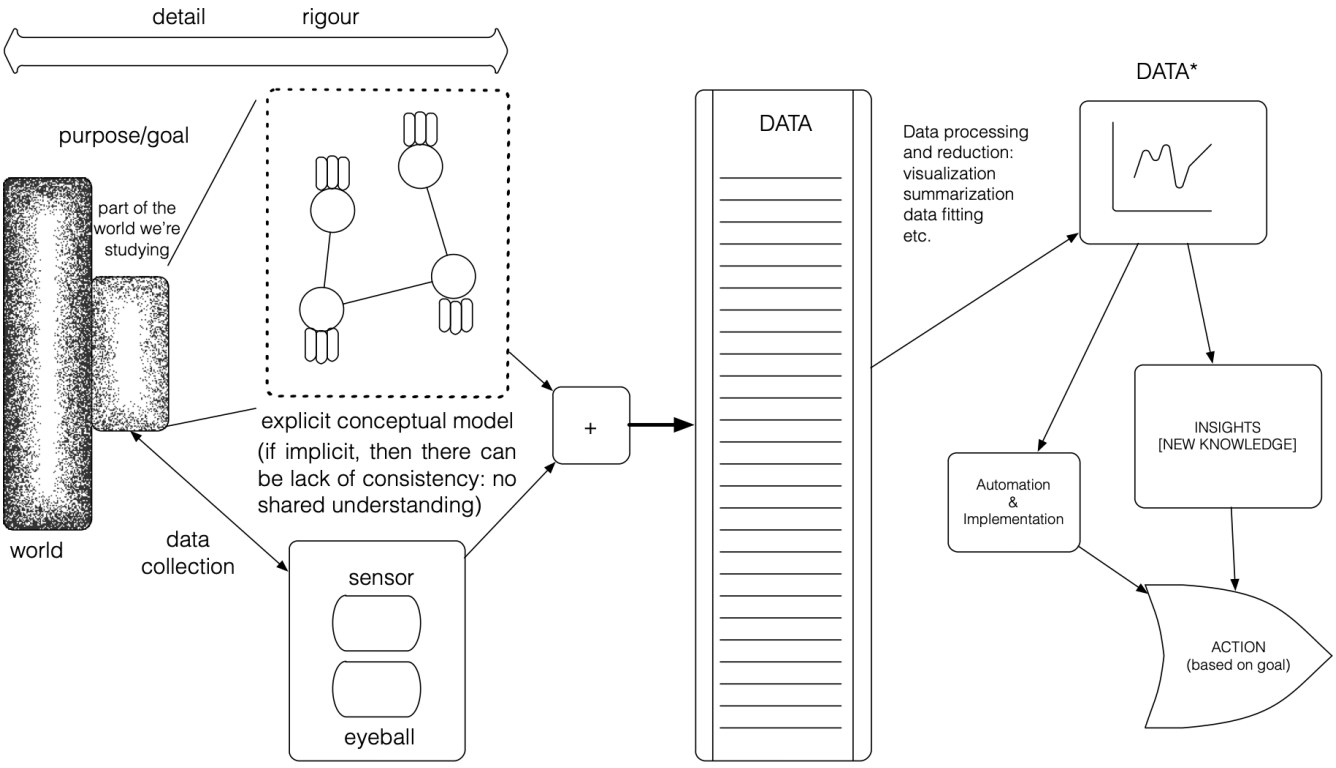


Figure 3. Schematic diagram of system understanding [29].

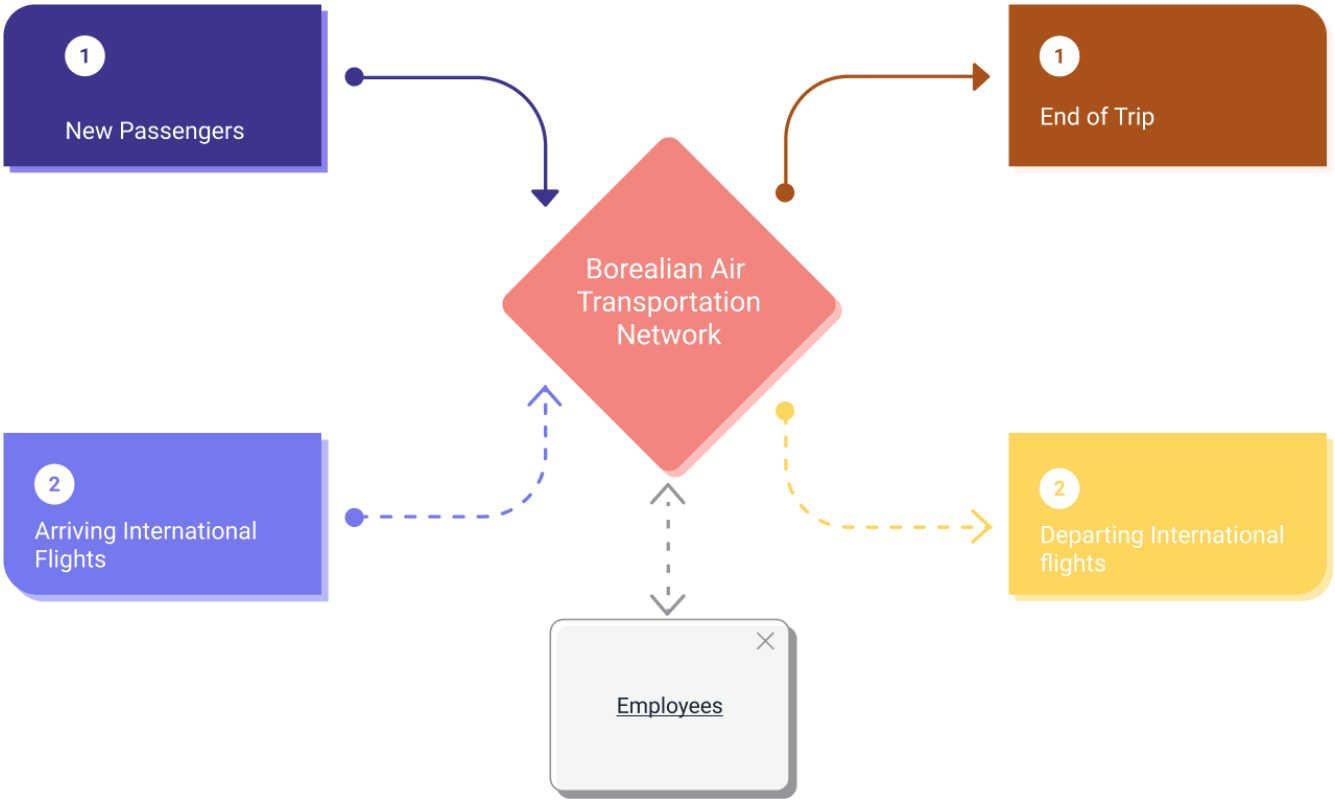


Figure 4. BASA system understanding – Tier 1: external interactions with the Borealian air transport network.

The majority of the available data is captured at the S_2 scan in PBS where a security agent scans the passengers' boarding pass and passport. The remaining data is collected at S_1 (if scanned) and at the time of flight departure, respectively.

Note that the dataset only contains records on passengers who have gone through PBS in the BATN; as such, it is a proper subset of the data for the entire system (if passengers have already gone through PBS at a previous airfield, domestic or international, they are not included in the wait time data).

4.2 Data Infrastructure

4.2.1 Data Structure

The BASA dataset contains 20 variables, divided into 3 aspects:

- **flight-related**
  - `Sch_Departure`
  - `Act_Departure`
  - `Departure_Date`
  - `Departure_Time`
  - `S2`
  - `Time_of_Day`
  - `Period_of_Week`
  - `Day_of_Week`
  - `Month`
  - `Season`
  - `Year`
- **passenger-related**
  - `Pass_ID`
  - `Order`
  - `Wait_Time`
  - `C_start`
  - `C0`
  - `C_Avg`)
- **geography-related**
  - `Airfield`
  - `BFO_Dest_City`
  - `BFO_Dest_Country_Code`

More information on the variables is available through the data dictionary (Section 5.1.2).

Variables can be selected and observations filtered out depending on the analysis of interest.

In order to answer questions related to the wait time and to the differences between scheduled and actual departure times by airfield, for instance, we could select

- `Wait_Time`,
- `S2`,
- `Sch_Departure`,
- `Act_Departure`, and
- `Airfield`;

to answer questions related to busy times, we could select

- `Season`,
- `Month`,
- `Time_of_Day`,
- `Period_of_Week`,
- `S2`, and
- `Airfield`;

for questions related to servers, we could select

- `Pass_ID`,
- `C0`,
- `C_start`,
- `Airfield`,
- `Season`;

to answer flight network or fligh capacity questions, we could select

- `Pass_ID`,
- `Departure_Date`,
- `Sch_Departure`,
- `Act_Departure`,
- `BFO_Dest_City`,
- `BFO_Dest_Country_Code`,
- `Airfield`,
- `Season`.

These only cover a small subset of questions that could be asked, of course, and derived variables might be required as well.

4.2.2 Loading a "Big" Dataset

Working with large data-sets is quite challenging for data scientists. Being able to efficiently process a large dataset often boils down to whether or not it can fit in memory, and, if not, to whether the operations under consideration can be split into sub-operations on multiple subsets of the data, and their results combined – a technique known as MapReduce [4].

The raw data is contained in a 2GB Comma Separated Value (CSV) file, where rows of data are plain-text rows in a file with column attributes separated by a comma character. This affords readability and straightforward parsing of the data, but the performance of the format leaves a lot to be desired. Early approaches that loaded the entire CSV into an R data frame or a Panda object in Python required the entire data frame to be loaded into memory, which not all users could do with their hardware.

The de-facto standard for operations on data that do not lie entirely in memory is to abstract the data into a database format to be queried via Structured Query Language (SQL). Ultimately, we converted the CSV file into an SQLite file (an optimised binary encoding designed for SQL operations). This allowed for improved operational performance and eliminated the pain of loading the entire CSV into memory.

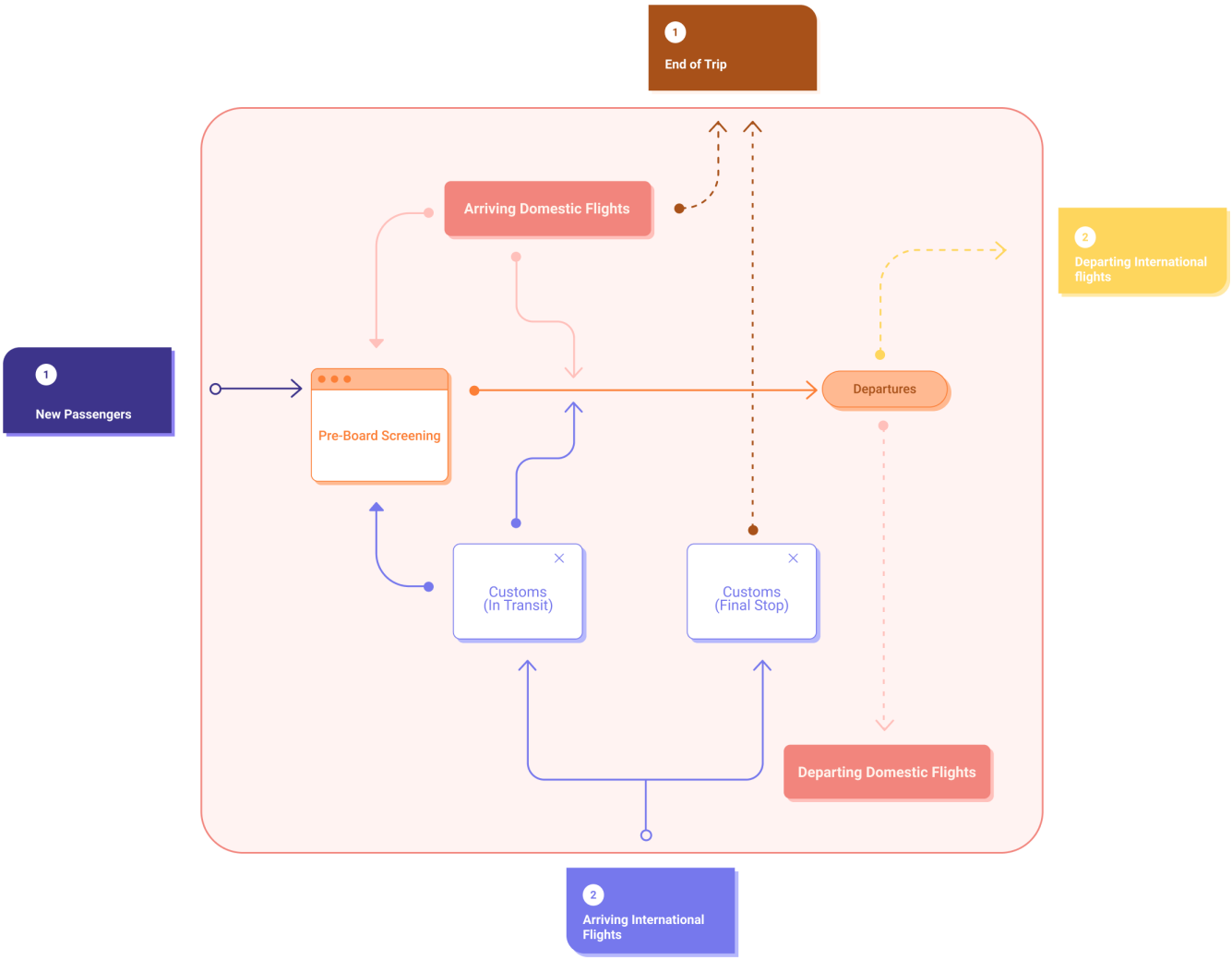


Figure 5. BASA system understanding – Tier 2: Inside the Borealian air transport network.

4.2.3 Code Infrastructure

Cloud services are extremely helpful for remote development collaborations. To host the data and associated scripts, we initially used GitHub in conjunction with Git-LFS, a Git module that enables the storage of large files.

Knowing the computing limitations some were experiencing, we leveraged free access to Azure through the University of Ottawa domain to stage an Azure Machine Learning Lab.

Instead of downloading the data to bring it to local machines, the data and computation resources reside in the Azure cloud and are exposed via a Jupyter-style interface. The solution was impressive and worked extremely well, providing enough RAM for users to simply load the entire dataset into their notebooks.

4.2.4 API Availability

To make subsets of the data more readily available, we staged an HTTP application programming interface (API) at https://basa-data.strikethrough.net , accessible through the front-end at https://basa.strikethrough.net .

The API allows users to access subsets of the data based on columns, filters and computed columns to prevent the need to download the entire data-set when only a subset is needed. It provides a shareable URL to retrieve each query so that subsets can easily be shared to others, or retrieved via requests in code (see Figure 6).

The back-end is structured as a Python program that loads the entire data into a Pandas dataframe on first start, and serves requested subsets via Flask-HTTP.

5. Data Preparation

Many procedures must be implemented prior to analyses being conducted. For instance, data should be examined for **missing observations**, as many quantitative methods cannot handle non-responses. **Anomalous entries** may influence the outcome of an analysis, leading to invalid conclusions. Data should also be confirmed to be **logically consistent** so that two any variables do not contradict one another.

Similarly, it is important to verify the **quality of the data**. Understanding the data structure (see Sections 4.2.1, 5.1.2), as well as its content, is key to successful and insightful analyses. Thorough data preparation is an important predecessor to data analysis.

Data preparation is an iterative process, involving both **data cleaning** and **data exploration**. In this section, we discuss the results of the initial iteration. We start by providing an update on data cleaning – the verification of data quality – then move on to data exploration and visualization, which helped us understand the information contained in the data, the structure of the data, potential questions to be answered with available data, as well as additional information required to answer specific questions.

5.1 Data Cleaning

Data is said to be **consistent** if it is technically correct and has also been cleaned to a point that analysis can be undertaken in earnest. Obtaining data that is prepared for analysis requires detection, selection, and correction. In the case of non-missing data, the data must be checked for values that are logically inconsistent or anomalous in some way. [28]

5.1.1 Technically Correct Data

As noted in [28], **technically correct** data contains values with the correct data type, provided in a consistent format. An example of technically correct variable is `S2`. As it represents the date and time at which a passenger exit the queue, we expect all observations to have consistent format of `YYYY-MM-DD HH:MM:SS`, which is indeed the case.

The BASA data was technically correct in all but one variable: `BFO_Dest_City`, which had four airfields of interest (i.e., AUC, CWL, QUE, and SAF) not following the convention of AAA###.

5.1.2 Invalid Entries

The allowable range of a data set's variable must abide by defined project constraints, and be in line with common sense. **Invalid entries** are those which fail on either of those two fronts and setting domains for the variables in the data set. According to the analysis for the whole dataset, it is obvious that invalid entries exist in both numerical and categorical variables.

The categorical variable `Year`, for instance, is a variable that contains invalid entries. The data has been collected for the years 20X6 to 20X9 (the "X" in the dataset is a "2" in the file to make it resolve to a date). Yet there are entries outside of the allowable range, such as "1989" or "1900". Invalid years appear in the data 330 times. Additionally, other variables with a year component (`S2`, `Sch_Departure`, `Act_Departure`, and `Departure_Date`) share the same also contain invalid entries.

	V1	Airfield	S2	Wait_Time
1	1688301	CWL	2028-09-12 08:08:00	1755
2	1690682	CWL	2028-09-26 07:37:00	1687
3	2195205	CWL	2027-11-25 14:24:00	1578
4	3920345	CWL	2028-10-29 10:04:00	1748
5	4174160	CWL	2028-10-15 14:29:00	1776
6	4927270	QUE	2028-12-17 14:58:00	1693
7	4971733	QUE	2029-05-14 05:51:00	1496
8	5070787	QUE	2027-10-05 17:59:00	1930

Table 3. Table of invalid waiting times.

Some other data consistency checks are shown in Table 2. For the waiting time, as an example, the maximum observation is 1930 minutes (32h10min) which seems an unlikely value. For logistical purposes, we assume that the upper bound on the waiting time is 24 hours (1440min). There are only 8 observations with a wait time greater than 1440 minutes (see Table 3), which suggests that these are uncommon occurrences and are in line with our assumption they are invalid.

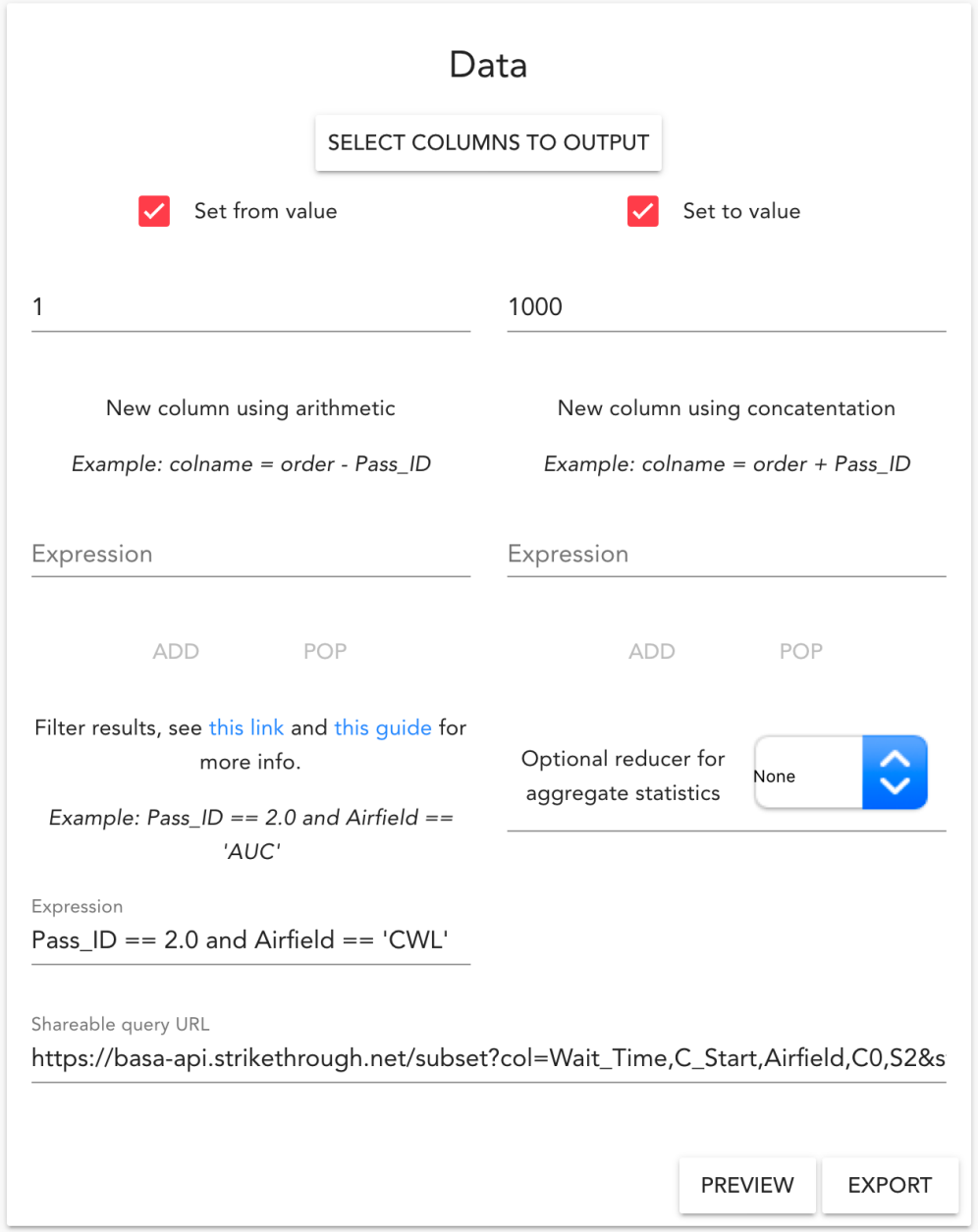
Furthermore, it was found that 95% of passengers go through the queue within 17 minutes, 99% within 24 minutes, 99.8% within 30 minutes, and 99.99% within 1 hour. It seems safe to say that any wait time longer that 2 hours is unreasonable, and should probably be handled separately.[2]

For categorical variables, it was found that *Period_of_Week* was always inconsistent with *Day_of_Week*, as Saturdays and Sundays were recorded as weekdays, and Monday through Friday as the weekend. Seasonal values were also sometimes incorrect (months did not correctly match up with seasons). These types of data issues are readily correctable, however.

As another example, the minimum value of `Pass_ID` is 1, and its maximum value is 9,904,000. But there are only 10,330 entries with `Pass_ID` below 99,999; this seems to indicate that `Pass_ID` is not a regular count. Further investigation is required to understand how `Pass_ID` is allocated before determining if these are invalid entries.

Apart from these **scope constraints** (which require values to lie in a certain range), there are also **regular expression constraints** (which require values to have compatible formats); for instance, the `Departure_Time` does not meet the standard time format "`xx:xx:xx`", and there are two formats for destination cities: "`ABC123`", or "`AUC`", "`CWL`", "`QUE`", and "`SAF`".

[2] Although it should be noted that the majority of the authors have waited more than 2 hours before boarding an aircraft at least once, and that this is not an usual situation to occur.



	Wait_Time	C_Start	Airfield	C0	S2
1098463	NaN	NaN	CWL	NaN	2027-03-01 22:25:00
1098464	NaN	NaN	CWL	NaN	2027-03-03 15:18:00
1098465	NaN	NaN	CWL	NaN	2027-03-04 19:29:00
1098466	NaN	NaN	CWL	NaN	2027-03-05 16:30:00
1098467	NaN	NaN	CWL	NaN	2027-03-06 16:25:00
1098468	NaN	NaN	CWL	NaN	2027-03-07 16:30:00
1098469	NaN	NaN	CWL	NaN	2027-03-08 17:14:00

Figure 6. The BASA database API front-end; this query will return all records between the 1st and the 1000th for which `Pass_ID` is 2 and `Airfield` is CWL.

Field Name	Data Type	Data Format	Description	Example
NA (row number)	Integer	#####	Gives each row a distinct ID number starting with 1,2,3,...	9984587
Airfield	TLA	ABC	The airfield from switch the passenger departs form	SAF
S2	Date & Time	YYYY-MM-DD HH:MM:SS (AM/PM)	The date and time at which the passengers boarding pass was scanned at S2	2027-11-25 11:46:00 AM
Wait_Time	Integer	#####	The time it takes a passenger to go form S1 to S2	6
C_Start	Integer	#####	The number of security agents at S1 when the passenger arrives at S1 (collected in 15 second intervals)	3
C0	Integer	#####	The number of security agents at S2 when the passenger arrives at S2 (collected in 15 second intervals)	2
C_avg	Real Number	#####	The average of number of security agents while the passenger is in the PBS queue (collected in 15 second intervals)	2.667
Sch_Departure	Date & Time	YYYY-MM-DD HH:MM:SS	The date and time that the flight was scheduled for depature	2028-08-13 17:39:00
Act_Departure	Date & Time	YYYY-MM-DD HH:MM:SS	The date and time at which the flight actually departed	2028-08-13 17:57:00
BFO_Dest_City	TLA / 3 Digit Integer	ABC### / ABC	The destination city of the flight (if a Borealian airfield only a TLA)	VES077 & CWL
BFO_Destination_Country_Code	TLA	ABC	The destination country of the flight	VES
order	Integer	#####	A unique number corresponding to the order at which the passenger was scanned at S2	5287858
Pass_ID	Integer	#####	A unique number assigned to each passenger when scanned at S2	5187391
Departure_Date	Date	YYYY-MM-DD	The date that the flight departed	2028-08-13
Departure_Time	Integer	#####	The time of departure converted to a number in seconds by the following formula: int(hh:mm:ss*24*60*60). For example, 17:57:00 -> int(time*24*60*60)= 64620	64620
Time_of_Day	Integer & Text	# - Text	The time of day that the flight departed	2 - MORNING
Period_of_Week	Integer & Text	# - Text	The period of week that the flight departed	2 - WEEKEND
Day_of_Week	Integer & Text	# - Text	The day of week that the flight departed	5 - FRI
Month	Integer & Text	# - Text	The month that the flight departed	08-Aug
Season	Integer & Text	# - Text	The season tha the flight departed	0 - NODATA
Year	Integer	YYYY	The year that the flight departed	2028

Table 1. Data dictionary for the BASA dataset at the passenger level; missing value codes depend on the variable – they include (blank), NA, '.', '0 - NO DATA', '0 - NOD', and '00 - NOD'.

5.1.3 Missing Values

Missing values must be identified and inconsistent approaches to indicating missing values standardized. Table 7 provides an overview of missing values: there are three variables with high proportions of missing observations: `Wait_Time` (45%), `C_Start` (54%), and `C_avg` (54%).

Potential reasons to explain why `Wait_Time` is missing are discussed in Section 5.2. But the measurements for `Wait_Time` are at least required to obtain information for `C_Start` and `C_avg`, it is reasonable that these variables have greater proportions of missing observations.

We notice that there are two types of missing values: those that are coded as missing (see caption, Table 1) and those for which no value is present. Apart from the variables discussed previously, it seems that missing values, while present, are not prevalent in the dataset in general.

Original data

(Numerical) Variable	Min	1st Q.	Median	3rd Q.	Max
S2	2027/01/01 2:19	2027/10/29 13:55	2028/07/24 15:26	2029/04/12 18:03	2030/01/07 17:36
Wait_Time	0	2	4	9	1930
C_Start	0	2	3	4	7
C0	0	2	3	4	7
C_avg	0	2	3	4	7
Sch_Departure	2026/12/08 8:28	2027/10/27 19:21	2028/07/21 20:07	2029/04/11 7:43	2030/01/07 21:14
Act_Departure	1899/12/31 8:00	2027/10/27 18:59	2028/07/21 19:54	2029/04/11 7:48	2030/01/07 20:52
order	1	2,496,000	4,982,000	7,478,000	9,974,000
Pass_ID	1	2,495,000	4,980,000	7,424,000	9,904,000
Departure_Date	1899/12/08	2027/10/27	2028/07/21	2029/04/11	2030/01/07
Departure_Time	0:00	8:38	13:32	17:37	23:59
Year	1899	2027	2028	2029	2030

Corrected data (based on within variable consistency check)

(Numerical) Variable	Min	1st Q.	Median	3rd Q.	Max
S2	2027/01/01 2:19	2027/10/29 13:55	2028/07/24 15:26	2029/04/12 18:03	2030/01/07 17:36
Wait_Time	0	2	4	9	1440
C_Start	0	2	3	4	7
C0	0	2	3	4	7
C_avg	0	2	3	4	7
Sch_Departure	2027/01/01 5:08	2027/10/27 19:21	2028/07/21 20:07	2029/04/11 7:43	2030/01/07 21:14
Act_Departure	2027/1/1 5:08	2027/10/27 19:26	2028/07/21 20:17	2029/04/11 8:02	2030/01/07 20:52
order	1	2,496,000	4,982,000	7,478,000	9,974,000
Pass_ID	1	2,495,000	4,980,000	7,424,000	9,904,000
Departure_Date	2027/01/01	2027/10/27	2028/07/21	2029/04/11	2030/01/07
Departure_Time	0:00	8:38	13:32	17:37	23:59
Year	2027	2027	2028	2029	2030

Table 2. Data consistency checks.

We can also look at the prevalence of missing values in **observations**.The vast majority of observations have either no missing values (≈ 46%) or they have exactly 3 missing values – `Wait_Time`, `C_Start`, and `C_avg` (≈ 49%).

The proportion of observations with 9 or more missing values is fairly small (≈ 2%), but keep in mind that the dataset is quite large (so it corresponds to ≈ 200,000 observations in total).

Depending on the task under consideration (and on which variables have been retained to tackle it), it might be an acceptable solution to remove those observations with too many missing values. On the other hand, there might be a need to impute the missing values, in which case a strategy will have to be developed to speed up the process. Automation may help, but regular audits are recommended.

5.1.4 Duplicate Values

As expected, some variables in the dataset have duplicate values.[3] There are variables (`Wait_Time`, `C_avg`, etc.) where this is not problematic, but others for which this should not occur, or, at the very least, for which specific combinations of variables should not occur (`Pass_ID`, `S_2`×`Airfield`, etc.).

There are 9,903,426 unique entries based on `Pass_ID`, and 9,906,787 unique records based on the combination of `Pass_ID` and `S_2`.[4]

The combination of `Pass_ID`, `Airfield` and `S_2` gives rise to a unique identifier for each observation, [5] once the 77,900 duplicates for the `Pass_ID` and `S_2` were removed.

5.1.5 Variable Syntax

Another potential issue with the data is that the same words could be spelled in different ways in the data set.

To find the different spellings of the character variables, the unique entries for each variable were extracted. A list of the chosen spellings for each of the levels was curated, as seen in Table 4.

In the destination city variable, the standard for all the entries except for the four class A Borealian airfields is the three letter acronym for the country code and and a three digit number that represents the city. The 4 Borealian airfield cities were converted to the form `BORXYZ`, the first three digits being the country code for Borealia and the last three digits corresponding to the city code (`AUC`, `CWL`, `QUE`, `SAF`).

[3] Since 99.99% of the 5,450,590 observations with a `Wait_Time` measurement have waited less than 1 hour, and since that variable is measured in minutes, the pigeon-hole principle guarantees duplicates.

[4] Most duplicate `Pass_ID` occurred 2 or 3 times, but some were duplicated a substantial number of times (304 times for `Pass_ID`= 2, for instance.

[5] This has been incorporated as a derived variable, see below.

Field	Number of non-missing entries	Number of missing entries	%Missing
Airfield	9,984,687	-	0%
S2	9,984,687	-	0%
Wait_Time	5,450,590	4,534,097	45%
C_Start	4,588,266	5,396,421	54%
C0	8,792,155	1,192,532	12%
C_avg	4,588,266	5,396,421	54%
Sch_Departure	9,792,456	192,231	2%
Act_Departure	9,792,456	192,231	2%
BFO_Dest_City	9,858,613	126,074	1%
BFO_Destination_Country_Code	9,858,613	126,074	1%
order	9,984,687	-	0%
Pass_ID	9,984,210	477	0%
Departure_Date	9,792,456	192,231	2%
Departure_Time	9,792,456	192,231	2%
Time_of_Day	9,792,456	192,231	2%
Period_of_Week	9,792,456	192,231	2%
Day_of_Week	9,792,456	192,231	2%
Month	9,792,456	192,231	2%
Season	9,792,456	192,231	2%
Year	9,792,456	192,231	2%

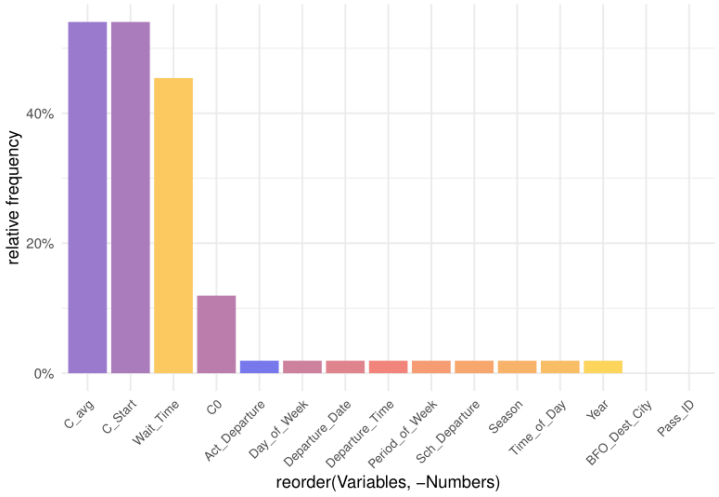


Figure 7. Proportions of missing observations in the BASA dataset.

Variable	Range
Month	NA,[1:12]
Period of Week	NA,[1,2] (1=Weekday, 2=Weekend)
Time of Day	NA, [1,4] (1=Night, 2=Morning, 3=Afternoon, 4=Evening
Day of Week	NA, [1:7]
Season	NA, [1,4] (1=Winter, 2=Spring, 3=Summer, 4=Autumn)

Table 4. New levels of the dataset's categorical variables.

	Derived variable	Example
1	Airport_Wait_Time	180
2	Flight_Delay_Time	0
3	Flight_Dest_Type	FALSE
4	Flight_ID	SAF1850621280VES
5	S_1	2029/12/6 8:01:00 AM
6	Unique_Record_ID	SAF50448271850621280

Table 5. Examples of derived variables.

5.1.6 Derived Variables

We can also derive new variables that might prove useful in better understanding the BASA system and for eventual analyses, such as

- `Airport_Wait_Time`: the amount of time passengers spend in the airfield after they exit the security checkpoint for the last time (`Act_Departure` − `S_2`);
- `Flight_Delay_Time`: which could be negative if the flight leaves before it is scheduled to do so (`Act_Departure` − `Sch_Departure`);
- `Flight_ID`: concatenation of strings `Airfield` and `Sch_Departure`;
- `Most_Frequent_Dest`: the most frequent level of `BFO_Dest_City` for all passengers assigned to a `Flight_ID`;
- `Flight_Dest_Type`: depending on the value of `Most_Frequent_Dest` (domestic – going to another Borealian airfield – or international);
- `S_1`: datetime of entrance into the PBS queue for passengers with a `Wait_Time` (`S_2` − `Wait_Time`);
- `Unique_Record_ID`: a unique identifier for passengers going through PBS, the concatenation of `Airfield`, `Pass_ID` and `S_2`.

Some examples are shown in Table 5.

5.1.7 Data Summarization

As mentioned in Section 4, the dataset consists of individuals going through PBS; it excludes those passengers who board a flight at a given airfield but were scanned at another airfield, earlier in their travels. The focus of the data is on **PBS waiting times**, not necessarily on passengers.

To get a better sense of the system, we can summarize the data in various ways.

First, we consider the distributions of arrival times (`S_1`) at each airfield for those passengers for which we have wait time data. The histograms of Figure 8 show that passenger arrival patterns at PBS vary from one airfield to the next. Note the inferred (approximated) periods of PBS operation for each airfield:

- `AUC` – 6:00AM to 9:15PM, with peaks 7AM and 5PM;
- `CWL` – 3:00AM to 11:00PM, with various peaks during the day;
- `QUE` – 1:45AM to 9:00PM, with peaks around 5AM, 11:30AM, and 3:00PM-5:00PM;
- `SAF` – 5:15AM to 5:30PM, with lulls between 9:30AM and 11:30AM, and between 1:30PM and 3:30PM.

These histograms make it clear that arrival patterns do exist, but the relative traffic volumes are difficult to compare at a

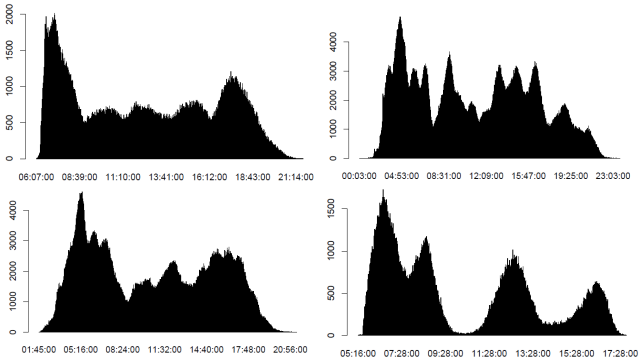


Figure 8. Distribution of arrival times (at S_1) for passengers with wait time information: from left to right, top to bottom – AUC, CWL, QUE, SAF (20X7-20X9).

glance, especially since only those passengers with a scan at S_1 are included.

In Figure 9, we see the traffic volume for all PBS passengers (with a bin-size of 1 hour): the histogram shapes are more or less preserved, which provides some evidence to suggest that the missing wait times are **missing completely at random** (or that the distribution of available wait times might be representative of the overall distribution of wait times in each airfield).

We clearly see that QUE is the busiest airfield, followed closely by CWL, and that both AUC and SAF are substantially less busy. The shorter hours of operation at SAF are easy to spot as well.[6]

A boxplot of the Wait_Time variable (see Figure 10) clearly shows that the vast majority of passengers wait less than 2 hours at PBS.

A small multiple boxplot of wait times by airfield shows that SAF tends to have the shortest waiting times, while the wait times at CWL and QUE are typically much longer, highlighting the strong degree of correlation between airfield and PBS wait times.

5.1.8 Output Datasets for Team Projects

A **clean dataset** requires the data to be standardised and documentation of the issues that have been found, so that analysts can make informed decisions as to whether certain observations should be included or excluded.

Given the large number of missing values in the dataset, and since the data needs will change based on the specific of each analysis project, we have decided to leave it up to the analysts to impute missing values and correct invalid entries.

[6]The shape and volume of traffic at each airfield seem to indicate that QUE, which is located inland, is likely to be the Borealian hub for domestic and Vespuchian (continental) travels, whereas CWL, which is located on the country's East Coast, is likely to be the hub for international flights to Europe and Africa (although these assertions would have to be validated by looking at the most likely destinations of flights leaving the two airfields).

- For **categorical date/time** variables, we suggest coordinating with (and comparing to) S_2.
- Imputing for **waiting times** requires several assumptions to be made about when people arrive at the PBS queue and how the queue its ordered; as there are several unknown factors affecting the queue's performance (server vacation policy, protocol to deal with hazardous passengers, etc.), and since we have some evidence to suggest that the available wait times are representative of all wait times, we have decided against imputing for missing wait time entries.
- For **location** variables, we suggest looking at the derived Most_Frequent_Dest variable and make the (reasonable?) assumption that this is the flight's direct destination. If the actual and scheduled departure times are also unavailable, however, there is more uncertainty regarding the final destination as we would first have to predict what flight the passenger was meant to board. This could presumably be done by looking for strong Flight_ID signals in neighbouring PBS arrival observations.

5.2 Data Exploration and Visualization

As an initial step in working with and understanding the BASA data, we have carried out a first iteration of data cleaning and data preparation. The data is generally correct, both technically and logically; however, some logical issues will need to be resolved in order for the data to be fully cleaned and ready for in-depth analyses. Also, it was found that some variables had a relatively high proportion of missing data, which may require some follow-up investigation.

In what follows, we further explore the dataset, focusing in particular on airfield usage and queuing patterns for the four airfields, as well as the popularity of both domestic and international destinations.

5.2.1 Airfield Use Patterns

There are more than 18,000 people flying from the four class-A airfields every day. Figure 9 and the accompanying discussion provide traffic information for each airfield airfield throughout the day, from 20X7 to 20X9.

Are there seasonal patterns? Figure 12 summarize the traffic for each airfield, per year, per month.

We observe that Auckland is relatively quiet between May and September and has stable traffic during the rest of year. This seasonal trend continues from 20X7 to 20X9; however, it is interesting to note that the number of passengers increases every year for all twelve months.

Chebucto follows a very different pattern: it is consistently busier during the warmer seasons, and traffic goes down during winter months. Also, there is practically no traffic between January and February 20X7, which may be due to

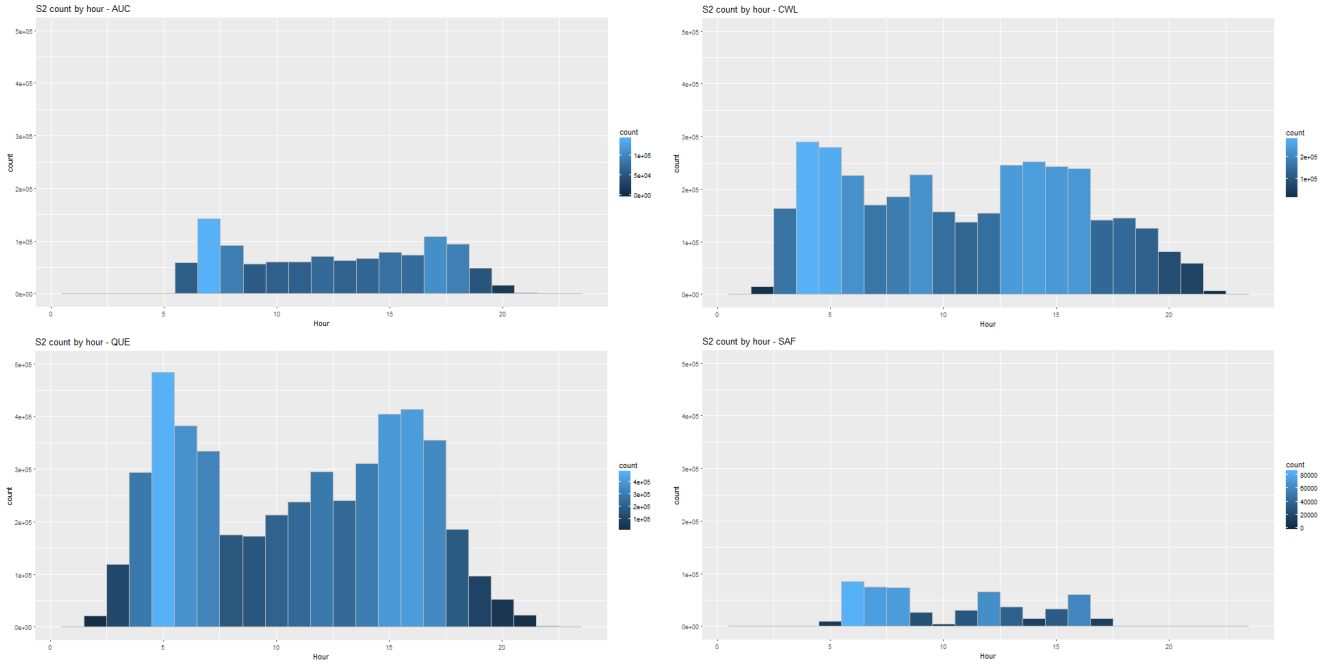


Figure 9. Distribution of total number of passengers (`S_2`) going through PBS throughout the day (2027-2029).

closure of the airfield, or a defect in the data collection/reporting system.

This does not seem to apply to Queenston, where the yearly trend is much more apparent than the monthly trend.

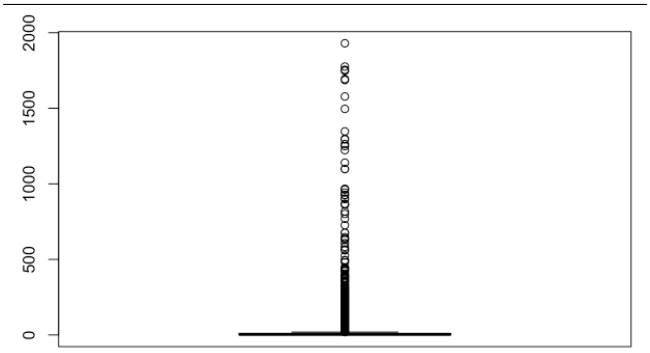


Figure 10. Boxplot of PBS wait times (combined).

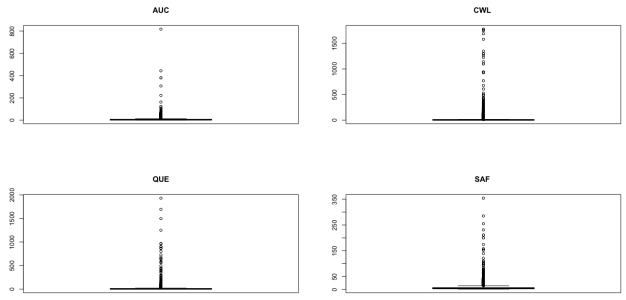


Figure 11. Boxplots of PBS wait times, by airfield.

On the other hand, trends in Saint-François are similar to the Chebucto patterns: higher traffic is observed during warmer seasons, with relatively consistent performance for all three years. It also has no traffic records between January to February of 20X7.

5.2.2 Queuing information

As seen in Table 7, 45% of PBS passengers did not get scanned at S_1. We have discussed how problematic this is as information about `Wait_Time`, `C_Start`, and `C_avg` cannot be obtained for those passengers.

Therefore, we begin our exploration by looking at the patterns of missing values for the `Wait_Time` variable.

As noted previously, there are some differences in traffic depending on time of day, month, year, and airfield.

We first take a look at the effect of peak hours on the number of passengers scanned at S_1. Figure 14 gives the average number of passengers scanned at S_2 for each hour over the 3-year period and the corresponding proportion of passengers scanned at S_1 for each airfield. Across airfields, it is clear that the proportion of scans at S_1 increases as average hourly traffic increases. This trend is particularly true for the two small airfields, `AUC` and `SAF`. For larger airfields, this upward trend plateaus at a certain point. At `CWL`, for instance, the proportion rapidly increases until the average hourly traffic at S_2 reaches about 100, after which S_1 scan rate hovers around 70% for busier hours. A similar pattern is observed at `QUE`, with a lower plateau between 40% and 50%.

Intuitively, we expect that it is unlikely for people to be scanned at S_1 if traffic and wait times are small. As the

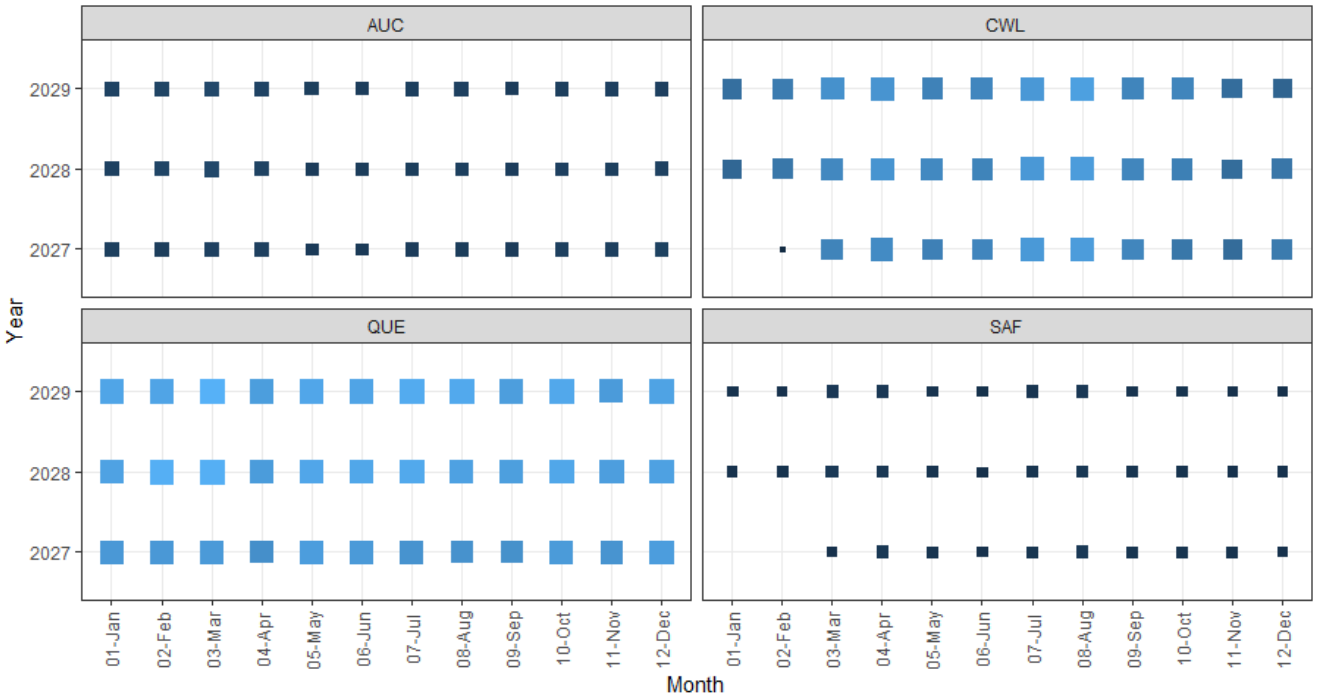


Figure 12. Traffic density per year, per month, per airfield. Note that traffic data for Chebucto and Saint-François seems not to be collected in earnest before March 20X7. Size and colour are correlated to number of PBS passengers.

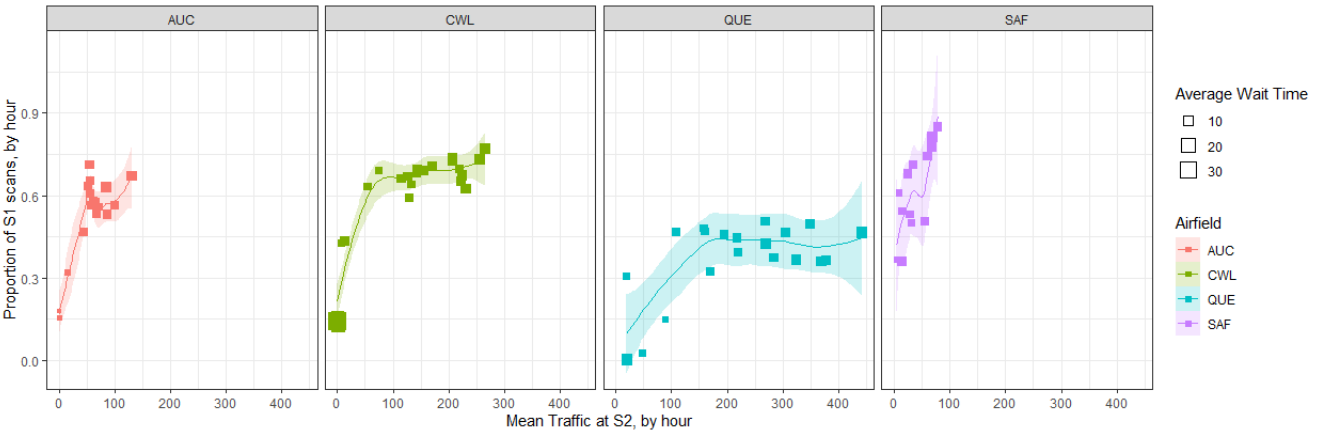


Figure 13. Proportion of passengers scanned at S_1 and hourly average traffic at S_2, by airfield.

traffic increases, passengers will wait in the queue and it becomes important for larger proportions of passengers to be scanned in order to collect wait time information; as passengers experience similar wait times in these cases, it would make sense that only a portion of them would receive the non-mandatory scan.

We switch our focus to the distribution of S_2 scans relative to the scheduled departure time. In general, we would expect passengers to enter the queue a considerable amount of time prior to the scheduled departure time.[7] Figure 14 provides a visual for the difference between scan times at `S_2` and `Sch_Departure`.[8]

Ignoring some extreme values, we observe that, overall, the distribution of time difference is **bimodal**. This feature breaks down when we look at each airfield individually.

In the middle plot of Figure 14 (`QUE`), passengers arrive at S_2 with plenty of time to spare (the same remark applies `AUC` and `SAF`). For `CWL`, however, many passengers are scanned around the scheduled departure time, and some passengers are scanned **after** their scheduled departure.

[7] Although this may depend on their destination and the size of the airfield from which their dirigible leaves.

[8] A negative value represents passengers arriving at S_2 prior to their flight departure.

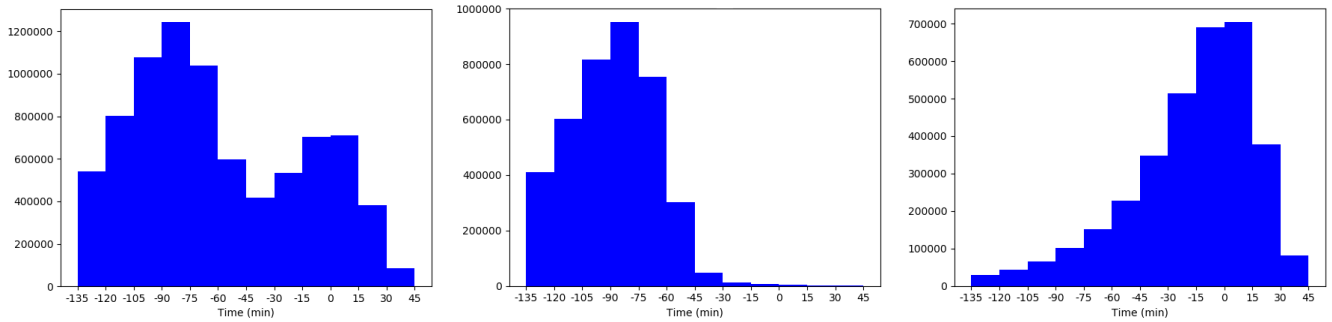


Figure 14. Distribution of S_2 values relative to **scheduled** departure time. The plot on the left combines data from all 4 airfields; the middle and rightmost plots show the QUE and CWL data, respectively.

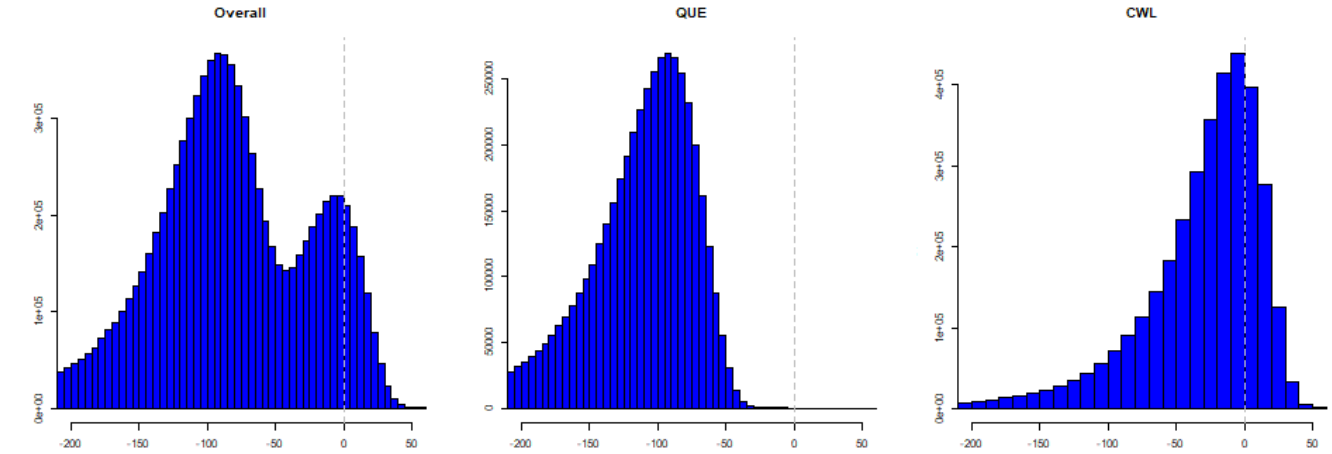


Figure 15. Distribution of S_2 values relative to **actual** departure time. The plot on the left combines data from all 4 airfields; the middle and rightmost plots show the QUE and CWL data, respectively.

Figure 15 presents three histograms charting the time differences between S_2 and Act_Departure. This difference shows similar patterns as those observed in Figure 14, but the proportion of passengers scanned at S_2 **after** the time of departure is smaller. Indeed, while 35% of passengers were scanned at S_2 after their scheduled departure, only 25% were scanned after their actual departure.

That is still an unreasonably high proportion of passengers who miss their flight, however – what is going on? This would need to be revisited with the client.

5.2.3 Patterns in Destinations

Understanding the popularity of different international and domestic flight destinations can be a useful component in the estimation of the potential numbers of queuing passengers at airfield security. We can measure destination popularity either based on number of flights to a destination, or on the number of travellers to a destination.

A Preliminary Exploration of Destination Popularity

For the purposes of this preliminary analysis we will be using the number of screened travellers as a representation of popularity. Using the number of flights as a popularity measure could be considered as well.

We assume that each record (i.e., each row) represents a screened traveller going to a destination. If the same traveller is scanned multiple times, this may result in an overestimate of the number of travellers; however, we suspect this occurrence is relatively rare and consequently any overestimate will be small (see, however, Section 5.2.5 for a further discussion of this assumption and Pass_ID).

We also assume that, although this analysis necessarily leaves out records with missing values for destination data, the patterns found are nonetheless broadly representative, and that unique scheduled departure times within each airfield are a good approximation for distinct flights.

This may result in an underestimate of the number actual flights, since it may be possible for multiple flights to be scheduled to depart at exactly the same time in the same airfield (perhaps for different airlines), but we suggest that this underestimate will be relatively minor in the context of our exploratory analysis.

We only consider a number of dimensions at this stage:

- popularity measure: travellers, flights
- airfields: ALL, AUC, CWL, QUE, SAF
- location: sphere (dom., int.) - country - city
- season - Spring, Summer, Autumn, Winter

```
            DOM        INT
AUC 0.11356707 0.10064935
CWL 0.37957317 0.29545455
QUE 0.47789634 0.48214286
SAF 0.02896341 0.12175325
            DOM        INT
AUC 0.11287879 0.10112360
CWL 0.37878788 0.29695024
QUE 0.47954545 0.48154093
SAF 0.02878788 0.12038523
```

Figure 16. % of travellers (domestic/international): flight/destination combinations (above); # of screened passengers (below).

Questions of interest may include:

- is there a meaningful difference between the two proposed popularity measures (# of travellers vs. # of flights)?
- what are the most popular destinations over all (domestic and international)?
- what is the popularity of different international and domestic destinations for each airfield?
- Do different airfields have different patterns for international and domestic flights?
- which destination is most popular in each season (across all airfields)?

Different Popularity Measures

We begin with a comparison of the proposed popularity measures: the number of screened travellers and the number of flights. If the two measures provide essentially the same picture, we will dispense with one of them for the remainder of the exploration.

As a means of comparison, Figure 16 shows the percentage of travel (domestic and international) based on flights/destinations count and on screened travellers count. As can be seen, values are compatible with each other. Consequently, we will use the number of screened travellers as the popularity measure for exploration.

Destination Popularity: The Big Picture

In order to gain a better understanding of popularity and travel patterns, it will be useful to get an overall sense of which destinations are most popular.

The top and bottom 10 destination lists shown in Tables 6,7, 8, and 9 give an initial overview of the most popular destinations, both domestically and internationally, from the 4 class-A Borealian airfields. Figures 17 and 18 show at a glance the number of travellers to each foreign country (and cities within these countries) served by the four airfields.

Borealian City	Traveller Count
BOR050	2,858,020
AUC	712,145
BOR040	638,887
BOR041	619,837
SAF	579,388
QUE	567,602
CWL	356,892
BOR030	335,775
BOR047	41,141
BOR045	28,020

Table 6. Top 10 most popular Borealian cities to visit, across all four airfields.

Borealian City	Traveller Count
BOR005	12
BOR043	10
BOR010	9
BOR021	8
BOR011	6
BOR002	5
BOR012	3
BOR019	3
BOR003	2
BOR004	2

Table 7. Bottom 10 popular Borealian cities to visit, across all four airfields.

Destinations: A Comparison between Airfields

Table 19 shows the screened passenger destination frequency for each airfield, broken into domestic and international destinations, as well as the corresponding relative frequencies.

We can see from this that the pattern of of popularity with respect to international and domestic destinations is not the same across airfields. In particular, SAF sends more screened passengers to international destinations than to domestic ones, unlike the other three airfields.

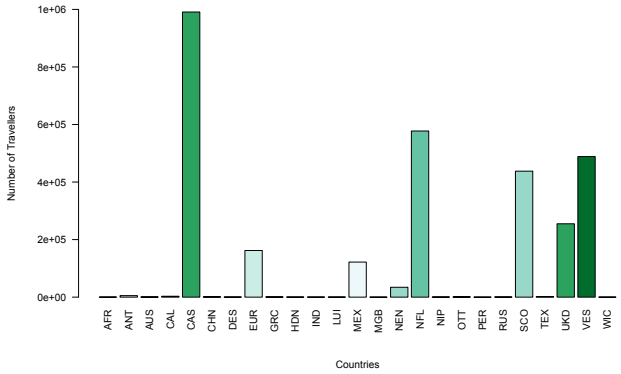


Figure 17. Number of travellers to foreign countries.

Country Code	Traveller Count
CAS	990,954
NFL	577,413
VES	488,560
SCO	437,809
UKD	254,928
EUR	162,096
MEX	121,861
NEN	34,301
ANT	5,282
CAL	3,104

Table 8. Top 10 most popular international destinations (countries) to visit, across all four airfields.

Country Code	Traveller Count
RUS	557
NIP	496
DES	424
HDN	359
LUI	296
AFR	264
IND	236
WIC	115
PER	42
MGB	13

Table 9. Bottom 10 least popular international destinations (countries) to visit, across all four airfields.

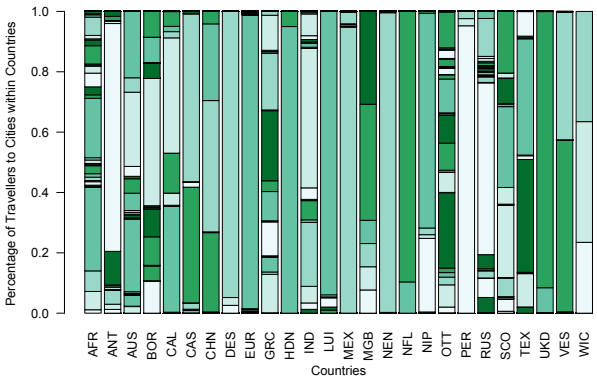


Figure 18. % travellers to cities within foreign countries.

Comparison Between Airfields: a Closer Look

The different patterns of destinations gives a sense of the differences between airfields (from which we could infer potential airfield roles and travel habits).

However, examining travel patterns to specific countries for (international flights) or to specific cities (for domestic flights), as shown in Figure 28, can provide a more detailed understanding of these patterns.

		Destination			
		AUC	CWL	QUE	SAF
Origin	AUC	-	212,794	413,534	87,287
	CWL	211,986	-	90,137	55,714
	QUE	412,709	89,812	-	436,387
	SAF	87,450	54,286	63,931	-

Table 10. Origin-destination traveller count. Notice the discrepancy for travel between SAF and QUE.

We can see from these heatmaps, for instance, that for both domestic and foreign travel, the majority of travellers go to a small number of popular destinations, with the majority of locations being travelled to infrequently.

Exploration of Destination Network

We have explored the popularity of specific destinations. It can also be useful to get an idea of the overall destination patterns. Unfortunately, we only have access to departure information for 4 class-A airfields and so network information (origin-destination traveller pairs) is by necessity incomplete.

Table 10 provides a summary of the number of travellers departing from and arriving at one of the four airfields being studied. We would expect this table to be more or less symmetrical about the diagonal: after all, travellers usually return to their home base.

With one exception, this is mostly what is happening with the entries: discrepancies could be explained by some travellers who are moving to another city, or who are returning using a connecting flight to a different city. But it is difficult to explain away the difference in the SAF → QUE and the QUE → SAF numbers. What might be going on there? Without additional contextual information (which is not directly available from the data), the answer to that question is not obvious.

5.2.4 Miscellaneous Charts

Various charts and tables exploring other aspects of the dataset are provided on pp. 96-96.

5.2.5 Outstanding Questions and Issues

Based on our data cleaning and data exploration results, there is a number of issues that require clarification from the client before we can begin data analysis in earnest:

1. there are 330 instances where the Year value was 1899 or 1900 – is there an explanation for why these values occur?;[9]
2. there are 10,638 CWL rows that are complete duplicates of another observation (including order) – are these data entry errors?

[9] Time stamp error, test entries, etc.

		Destination		
		Domestic	International	Total
Origin	AUC	714,880	331,723	1,046,603
	CWL	2,486,609	1,027,254	3,513,863
	QUE	3,366,370	1,411,328	4,777,698
	SAF	206,209	314,240	520,449
	Total	6,774,068	3,084,545	9,858,613

		Destination		
		Domestic	International	Total
Origin	AUC	7.3%	3.4%	10.6%
	CWL	25.2%	10.4%	35.6%
	QUE	34.1%	14.3%	48.5%
	SAF	2.1%	3.2%	5.3%
	Total	68.7%	31.3%	100.0%

Figure 19. Domestic and international screened traveller count, by airfield (left); relative frequencies (right).

3. there are many instances where `Pass_ID` is repeated: there are 304 occurrences of `Pass_ID` = 2, 98 instances of `Pass_ID` = 7,073,179, and so on – do these values represent certain types of passengers or event?;[10]

4. does `Pass_ID` uniquely identify travellers?[11]

5. the combination `Airfield` × `Pass_ID` × `S_2` does not uniquely define a row as some of them have different order values – what might explain this discrepancy?;

6. are there systematic differences between cases where `BFO_Dest_City`=" " and "."? A quick look at these cases reveal no particular pattern, other than that cases with an empty city code have only 6,639 instances out of 125,571 that had some information regarding actual departure time, while 497 out of 6,639 cases were found to have the same information when the city code is given as a dot;

7. why are so many passengers missing their flight out of `CWL`? Is it related to the weather, or to the temperament of Chebugonians? Are flights indeed being missed in such large quantities? Is there some recording/reporting error with the scheduled and actual departure times? With the PBS scan times?;

8. is there a server vacation policy[12] in place for the 4 class-A airfields? If not, how do security services decide when to make a new server available?

9. what is happening to passengers travelling to `SAF` from `QUE`? Is there an exodus from `QUE`? Are they returning to `QUE` using different modes of transportation? Or from another airfield?;

10. we have assumed that the destination city corresponds to the ultimate passenger destination rather than the next destination (otherwise, there would seem to be too many flights leaving an airfield at the exact same time) – is that indeed the case or is there some other explanation?

[10] Security inspectors, airfield employees, frequent flyers, etc.

[11] If the same person travels goes through PBS for different trips, are they given a different `Pass_ID`?

[12] The policy used to decide when to open or close PBS servers.

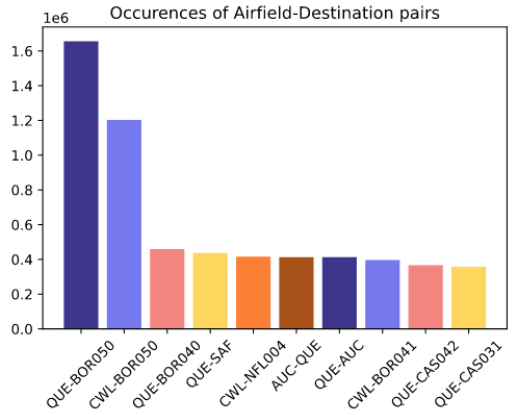


Figure 20. Top 10 airfield-city destinations.

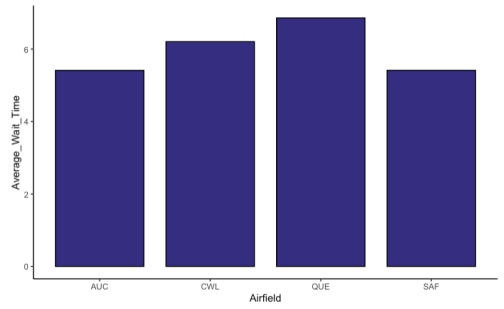


Figure 21. Average wait time per airfield.

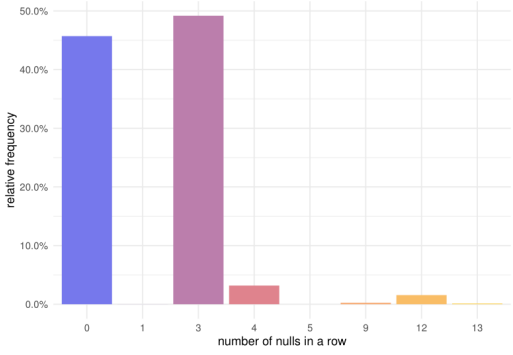


Figure 22. Number of missing values per row.

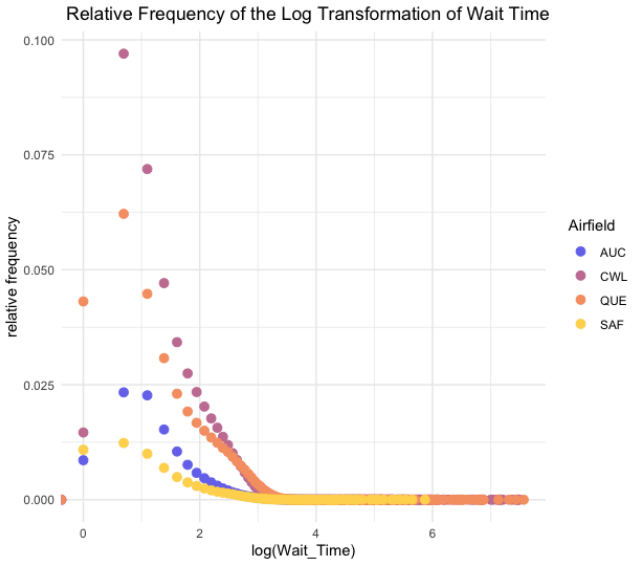


Figure 23. Relative frequencies of the natural logarithm of wait time per airfield.

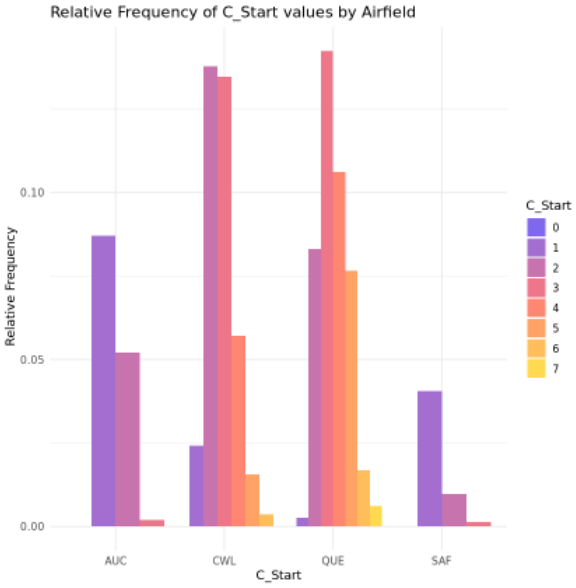


Figure 25. Relative frequency of the number of active servers when passengers arrive at S_1, per airfield.

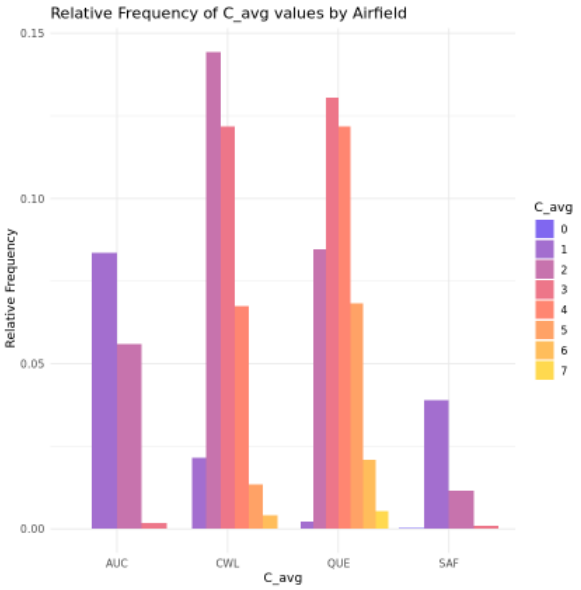


Figure 24. Relative frequency of the average number of active server while passengers are waiting to be screened, per airfield.

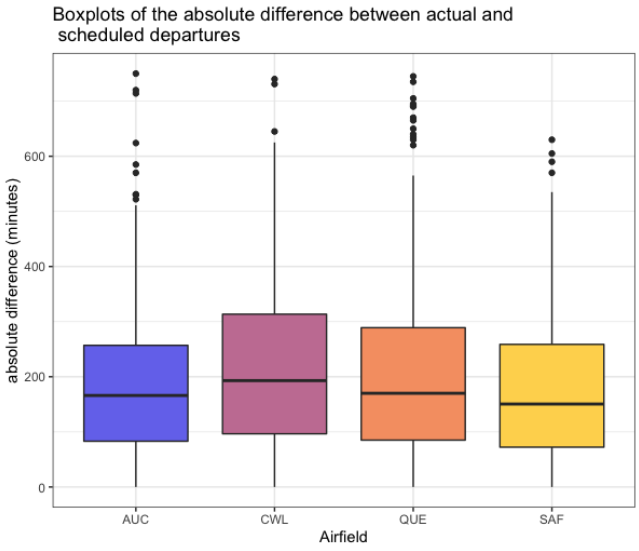


Figure 26. The distribution of the absolute difference between actual and scheduled departure, per airfield.

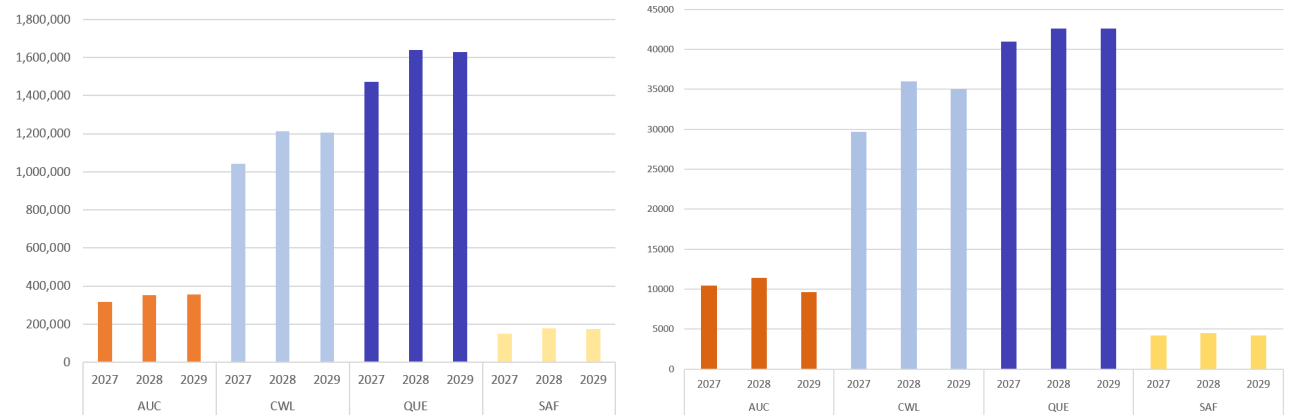


Figure 27. Count of passengers through PBS per airfield per year (left); number of flights per airfield per year (right).

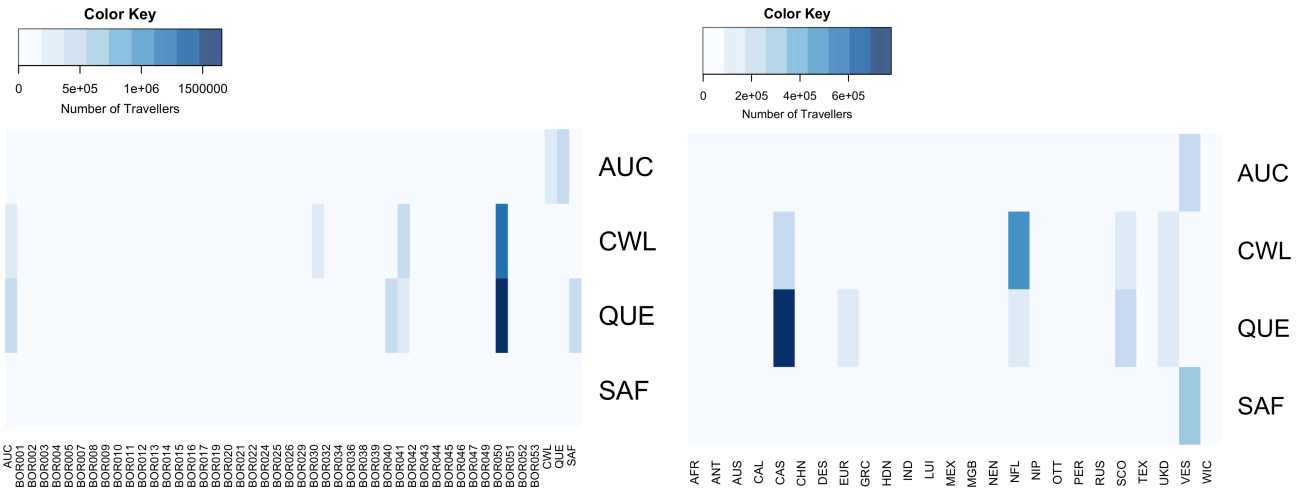


Figure 28. Number of domestic travellers, by city (left); number of travellers to international destinations, by country (right), for the 4 Borealian class-A airfields.

6. Conclusion

Exploration of the dataset such as the one conducted in this chapter provides a better understanding of the underlying situation, by identifying data quality issues and follow-up questions to bring to the client. Data visualizations, especially, can help consultants and analysts spot problem areas.

The questions that could be asked are not limited to those presented in Sections 3 and 5.2.5 – virtually any attempt to understand the scenario and explore the dataset is guaranteed to give rise to new problems and issues, which lead to new analysis approaches, and so forth.

What other insights lurk in the BASA system? Are there emerging patterns for the flights (instead of passengers)? Can the entire network be simulated to discover how to counteract disruptions? Can the factors that influence passenger waiting time and missed flights be identified? Can PBS server policy be derived from the data?

Consulting Post-Mortem There is no other way to say it: an 11-person project is difficult to manage. There are multiple pieces to juggles, including various personalities, roles to fill, competing deadlines, different expectations, and so much more.

The tendency for most of us is to jump right into the analysis, and to try to form an understanding of the system through bits and pieces that fall into place over the project's duration. It is crucial to take the time to build an explicit conceptual model (with schematic diagrams and data dictionaries), to set-up the data infrastructure, and to clean and explore the data before tackling the analytical tasks; simply said, quantitative analysis requires patience.

References

[1] Abou-Nasr, M., Lessman, S., Stahlbock, R., Weiss, G.W. (eds.) [2013], *Real World Data Mining Applications*, Annals of Information Systems 17, Springer.

[2] Bruce, P., Bruce, A. [2017], *Practical Statistics for Data Scientists: 50 Essential Concepts*, O'Reilly.

[3] Burke, P.J. [1956], "The Output of a Queuing System", *Operations Research* vol **4** (6): 699–704.

[4] MapReduce , retrieved from Wikipedia.org on November 1, 2020.

[5] Classic Data Sets , retrieved from Wikipedia.org on August 26, 2017.

[6] Deng, N., Tian, Y., Zhang, C. [2013], *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*, Chapman & Hall/CRC Press.

[7] Géron, A. [2017], *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly.

[8] Hastie, T., Tibshirani, R., Friedman, J. [2008], *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

[9] Iris Flower Data Set , retrieved from Wikipedia.org on August 26, 2017.

[10] Kuhn, M., Johnson, K. [2013], *Applied Predictive Modeling*, Springer.

[11] Li, T. (ed.) [2016] *Event Mining: Algorithms and Applications*, Chapman & Hall/CRC Press.

[12] Meirelles, I. [2013], *Design for Information: an Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*, Rockport.

[13] Mitsa, T. [2010], *Temporal Data Mining*, Chapman & Hall/CRC Press.

[14] Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. [1996], *Applied Linear Statistical Models*, Irwin.

[15] Newell, G.F. [1971], *Applications of Queuing Theory*, Chapman & Hall.

[16] Ng, A., Soo, K. [2017], *Numsense! Data Science for the Layman (No Math Added)*.

[17] Ross, S.M. [2010], *Introduction to Probability Models*, 10th ed., Academic Press.

[18] Shumway, R.H., Stoffer, D.S. [2010], *Time Series Analysis and its Applications (with R examples)*, 3rd ed., Springer.

[19] Torgo, L. [2017], *Data Mining with R: Learning with Case Studies*, 2nd ed., Chapman & Hall/CRC Press.

[20] Tufte, E.R. [2006], *Beautiful Evidence*, Graphics Press LLC.

[21] Tufte, E.R. [2001], *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press LLC.

[22] VanderPlas, J. [2017], *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly.

[23] Walrand, J. [1983], "A probabilistic look at networks of quasi-reversible queues", *IEEE Transactions on Information Theory*, vol **29** (6): 825–831.

[24] Wickham, H. [2016], *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed., Springer.

[25] Wickham, H., Grolemund, G. [2016], *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly.

[26] Wu, J., Coggeshall, S. [2012], *Foundations of Predictive Analytics*, Chapman & Hall/CRC Press.

[27] Yau, N. [2011], *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, Wiley.

[28] de Jonge, E., van der Loo, M. [2013], "An introduction to data cleaning with R," Discussion Paper, Statistics Netherlands.

[29] Boily, P., Schellinck, J. [2020], *Fundamentals of Data Insights*, Data Science Report Series, Data Action Lab.