BASIC DATA ANALYTICS TECHNIQUES

LEARNING OBJECTIVES

- Understand the origin and function of business intelligence.
- Understand and be able to apply a selection of fundamental data analysis concepts and techniques.
- Exposure to a simple but comprehensive data analysis pipeline process from data collection to data presentation, in preparation for carrying out a similar process in the upcoming lab.

OUTLINE

1. Background and Process
2. Insight *via* Number Crunching: Some Core Concepts
3. Insight *via* Number Crunching: Some Core Techniques
4. Exercise

INTRODUCTION AND REVIEW

THE DATA ANALYSIS PIPELINE

- Data modeling and conceptual analysis
- Data collection
- Data transformation
- Data storage
- Data exploration
- Data presentation

THE DATA ANALYSIS PIPELINE

- Data modeling and conceptual analysis
- Data collection
- Data transformation
- Data storage
- Data exploration
- Data presentation

Today:
Putting it all together in the context of business intelligence and data analysis for business intelligence.

BACKGROUND AND PROCESS

BUSINESS INTELLIGENCE – BUSINESS ANALYTICS

Use data (and information) about internal operations and the state of the market to support **informed decision making** about **business operations** and **business strategy**.

No firmly agreed upon definition of these terms – is one a subset of the other?

Goals: increased situational awareness + improved foresight

HISTORY OF BUSINESS INTELLIGENCE

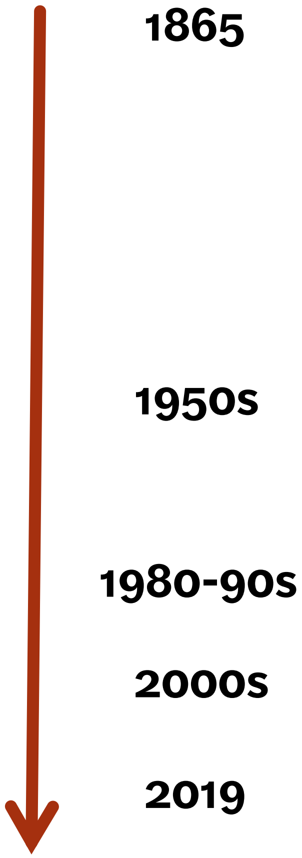
Late 1800s: people started to recognize that they could use data to gain a competitive advantage

1950s: advent of the first business database for decision support

1980s -1990s: computers and data becoming increasingly available - data warehouses, data mining – still very technical and specialized

2000s: trying to take business analytics out of the hands of data miners and other specialists and more into the hands of domain experts

Now: big data and specialized techniques have arrived on the scene, but so has data visualization, dashboards, software as a service



BUSINESS INTELLIGENCE AND DATA SCIENCE

Historically, one of the streams contributing to modern day Data Science

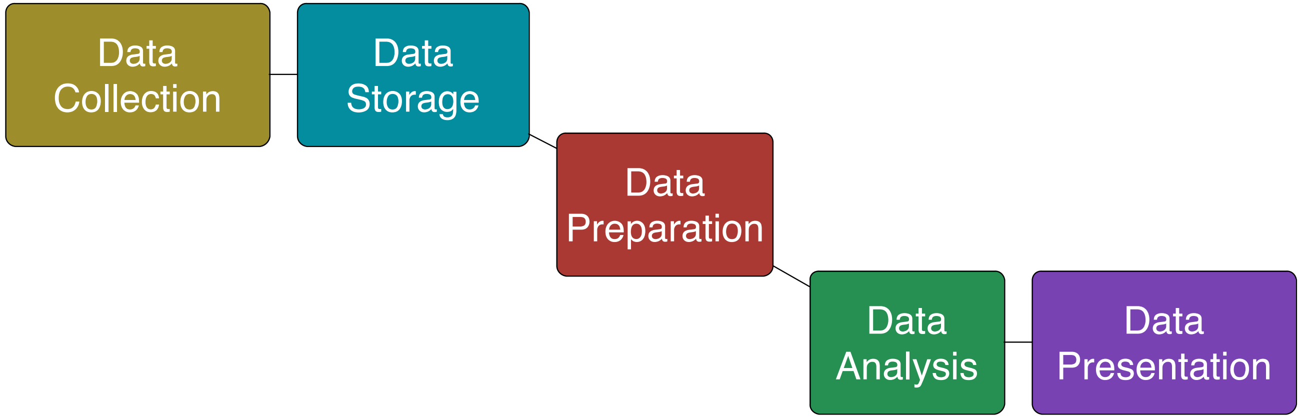
- **System of interest**: the commercial realm - the market in which you are involved
- **Sources of data**: transaction data, financial data, sales data, organizational data
- **Goals**: provide awareness of competitors, consumers and internal activity and use this to support decision making
- **Culture and preferred techniques**: datamarts, key performance indicators, consumer behaviour, slicing and dicing, business 'facts’

The ultimate goal is still the same: **insight into your system of interest**.

BUSINESS INTELLIGENCE AND THE DATA PIPELINE

Our general data pipeline model also works for business intelligence.

What are some aspects of the business intelligence pipeline that might distinguish it from a more generic analysis pipeline?



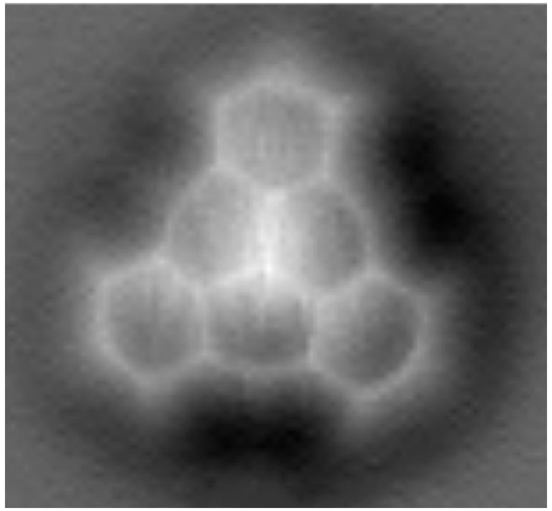
INSIGHT VIA # CRUNCHING: CORE CONCEPTS

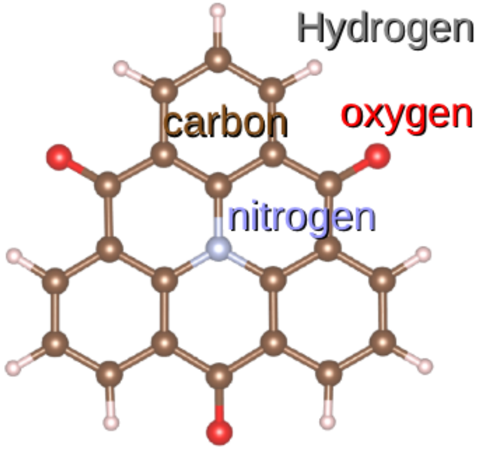
FINDING PATTERNS, GENERALIZATIONS AND STRUCTURE

- **Pattern**: A predictable, repeating regularity
- **Structure**: An organization of elements in a system
- **Generalization**: Creation of more general or abstract concepts from more specific concepts or instances

Underlying goal during analysis - find patterns or structure in our data, draw conclusions via these patterns or structures.

Finding patterns and structure is not bad or wrong, per se, it's how you use these discoveries – the conclusions that you draw - that is important.





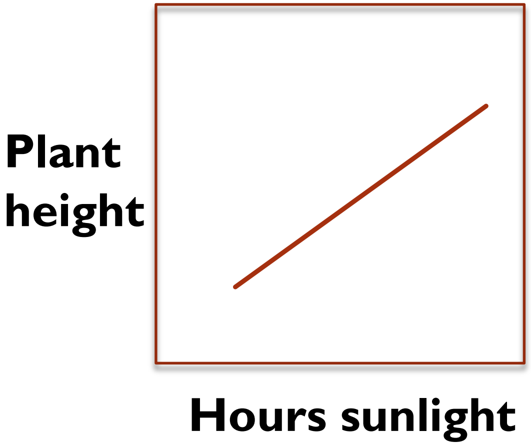
https://upload.wikimedia.org/wikipedia/commons/d/d0/TOAT_AFM.png

INDEPENDENT VS DEPENDENT VARIABLES

In an experimental setting:

- **Control/Extraneous Variables**: We do our best to keep these controlled and unchanging while other variables are changed
- **Independent**: We control the values of the variable. We suspect that they influence the dependent variables.
- **Dependent**: We don't control the values - they are generated in some way during the experiment, and presumably are dependent on everything

How do these translate over to other datasets?



TYPES OF DATA

Numerical Data: integers or continuous numbers

- 1, 7, 34.654, 0.000004

Text Data: strings of text – may be restricted to a certain number of characters

- “Welcome to the park”, “AAAAA”, “345”, “45.678”

Categorical Data: a fixed number of values, may be numeric or represented by strings. **There is no specific or inherent ordering**

- (‘red’,’blue’,’green’),(‘1’,’2’,’3’)

Ordinal Data: Categorical data with an inherent ordering. Unlike integer data, the spacing between values is **not** defined

- (very cold, cold, tepid, warm, super hot)

TURNING CATEGORICAL DATA INTO NUMERICAL (COUNT) DATA

We can useful turn categorical data into numeric data by generating frequency counts of the different values of the categorical variable.

This in turn allows us to apply numerical analysis techniques.

House Colour	Frequency
red	40
blue	13
green	2

THE SPECIAL ROLE OF CATEGORICAL DATA

Categorical data play a special role:

- In *data science*, we talk about a categorical variable with a pre-defined set of values
- In *experimental science*, a factor is an independent variable with the levels of the variable defined - it may also be viewed as a category of treatment
- In *business analytics,* people talk about dimensions (with members) vs measures

However we label these types of variables, we can use these them to **subset** our data, or **roll up/summarize** our data.

HIERARCHICAL / NESTED / MULTILEVEL DATA / MODELS

If a categorical variable has multiple levels of abstraction, we can create levels out of this variable.

We can view these levels as new categorical variables, in a sense.

The ‘new’ categorical variable has a pre-defined relationship with the more detailed level.

This is common with time and space variables – we can ‘zoom’ in or out.

This lets us talk about the **granularity** of the data – what is the ‘maximum zoom’?

Year	Quarter	Count
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

INSIGHTS VIA # CRUNCHING: CORE TECHNIQUES

DATA SUMMARIZING

Min: Smallest value of variable

Max: Largest value of variable

Median: Middle value of variable

Mode: Most frequent value

Unique Values: List of unique values

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

ROLLING UP YOUR DATA

We can perform an operation over a set (or subset) of the data - typically over a column of the data.

When we do this, we can think of this as compressing or 'rolling up' the many data values into a single representative value.

Typical roll up functions are 'mean', 'sum' and 'count'.

If we apply the same roll up function to many different columns we can think of this as **mapping** (a list of) columns to functions.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

CONTINGENCY TABLES / PIVOT TABLES

Contingency Table: A table used to examine the relationship between two categorical variables - specifically the frequency of one variable relative to a second variable (cross tabulation).

Pivot Table: A table generated in a software application by applying operations (e.g. sum, count, mean) to variables, possibly based on another (categorical) variable. Can be used to create a contingency table.

	Large	Medium	Small
Blue	1	32	31
Orange	14	11	0
Green	5	5	5

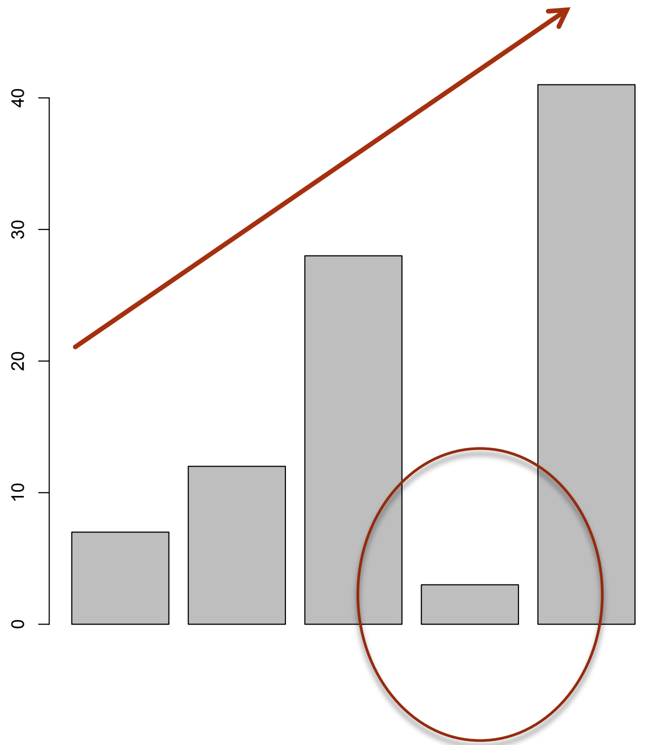
ANALYSIS THROUGH VISUALIZATION

Analysis broadly defined:

- identifying patterns or structure
- adding meaning to these patterns or structure by **interpreting** them in the context of your system.

Option 1: use analysis techniques to do this.

Option 2: visualize the data and use the analytic power of our (perceptual) brain to come to meaningful conclusions about these patterns.

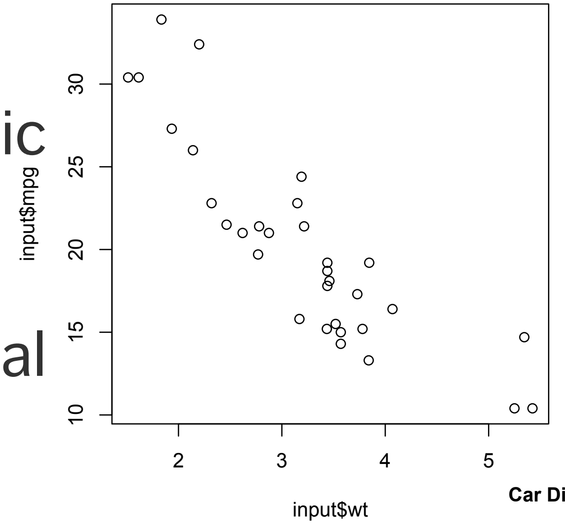


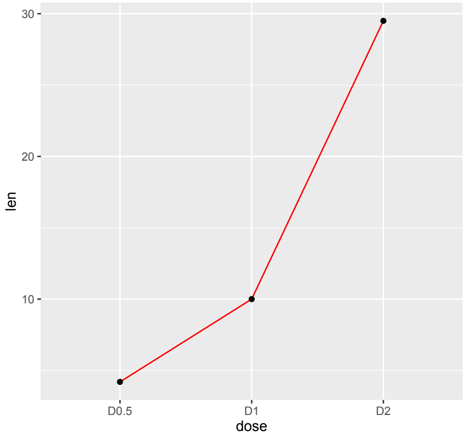
SOME SIMPLE VISUALIZATIONS TO REVEAL PATTERNS

Scatter plot: best suited for two numeric variables

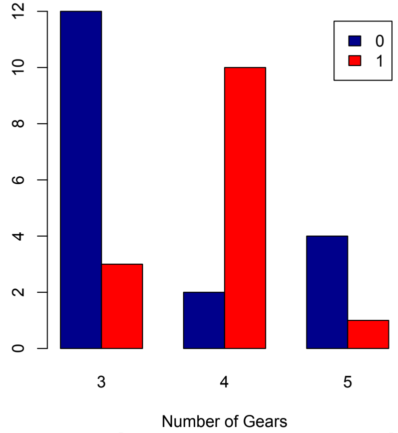
Line chart: numeric variable and categorical variable

Bar chart: best suited for one categorical and one numeric - or multiple categorical/nested categorical data and





Car Distribution by Gears and VS



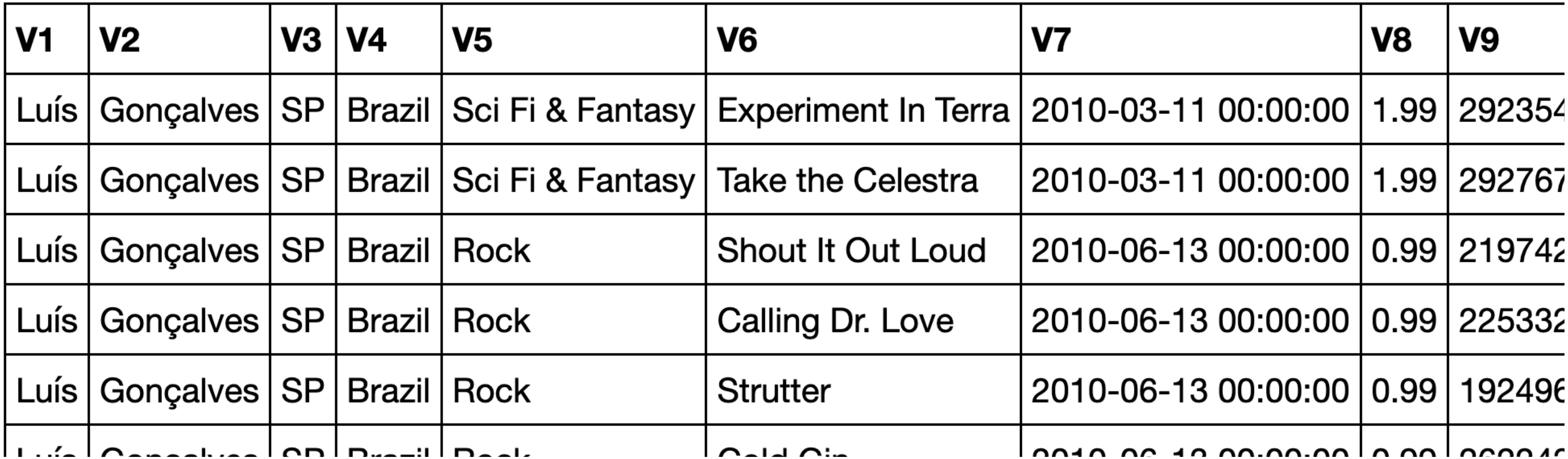
EXAMPLE (ARTIFICIAL)

LOADING THE DATA

Use your software/analysis tool’s **functionality** (you will need to learn its subtleties) to **load the data**.

How can you confirm that the data has been loaded successfully?

LOADING THE DATA



V1	V2	V3	V4	V5	V6	V7	V8	V9
Luís	Gonçalves	SP	Brazil	Sci Fi & Fantasy	Experiment In Terra	2010-03-11 00:00:00	1.99	292354
Luís	Gonçalves	SP	Brazil	Sci Fi & Fantasy	Take the Celestra	2010-03-11 00:00:00	1.99	292767
Luís	Gonçalves	SP	Brazil	Rock	Shout It Out Loud	2010-06-13 00:00:00	0.99	219742
Luís	Gonçalves	SP	Brazil	Rock	Calling Dr. Love	2010-06-13 00:00:00	0.99	225332
Luís	Gonçalves	SP	Brazil	Rock	Strutter	2010-06-13 00:00:00	0.99	192496
Luís	Gonçalves	SP	Brazil	Rock	Cold Gin	2010-06-13 00:00:00	0.99	262340

DATA PREPARATION

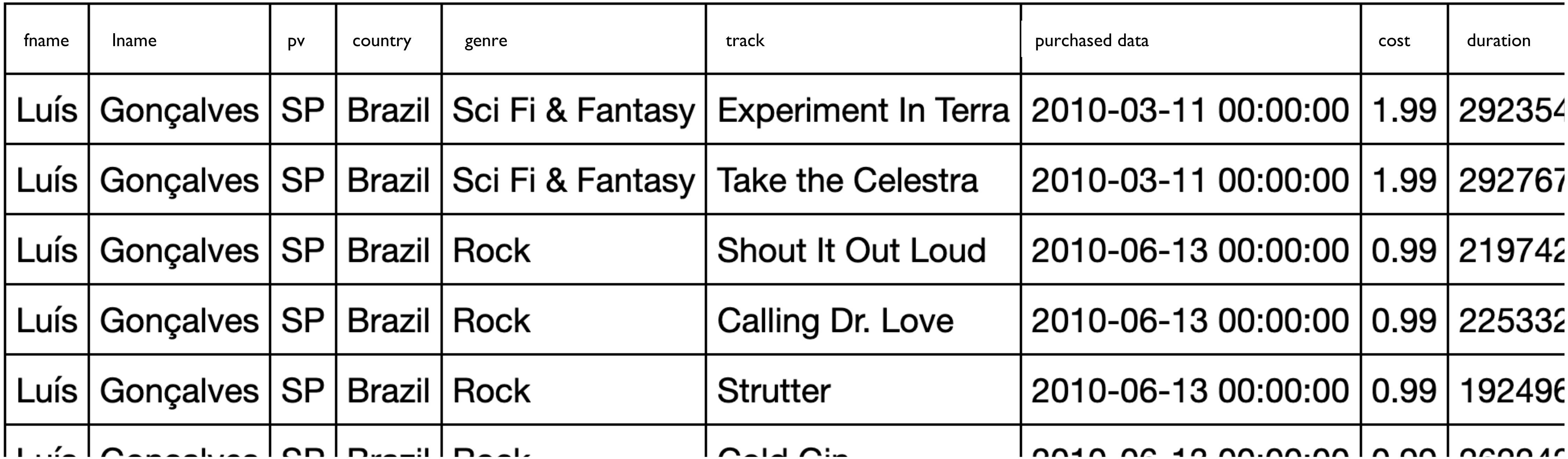
Prepare a **back-up** version of the original dataset.

Make back-ups **regularly** as you transform the original dataset.

Tips:

- Label the data for ease of use.
- Assigned the correct data types (strings, numbers, factors, dates, etc.) to features
- Create new variables
- Subset the data for ease of visualization/analysis

DATA PREPARATION



fname	lname	pv	country	genre	track	purchased data	cost	duration
Luís	Gonçalves	SP	Brazil	Sci Fi & Fantasy	Experiment In Terra	2010-03-11 00:00:00	1.99	292354
Luís	Gonçalves	SP	Brazil	Sci Fi & Fantasy	Take the Celestra	2010-03-11 00:00:00	1.99	292767
Luís	Gonçalves	SP	Brazil	Rock	Shout It Out Loud	2010-06-13 00:00:00	0.99	219742
Luís	Gonçalves	SP	Brazil	Rock	Calling Dr. Love	2010-06-13 00:00:00	0.99	225332
Luís	Gonçalves	SP	Brazil	Rock	Strutter	2010-06-13 00:00:00	0.99	192496
Luís	Gonçalves	SP	Brazil	Rock	Cold Gin	2010-06-13 00:00:00	0.99	[illegible]

NUMBER CRUNCHING AND ANALYSIS

Generate 1-way/n-way **contingency tables** on the whole dataset or on subsets

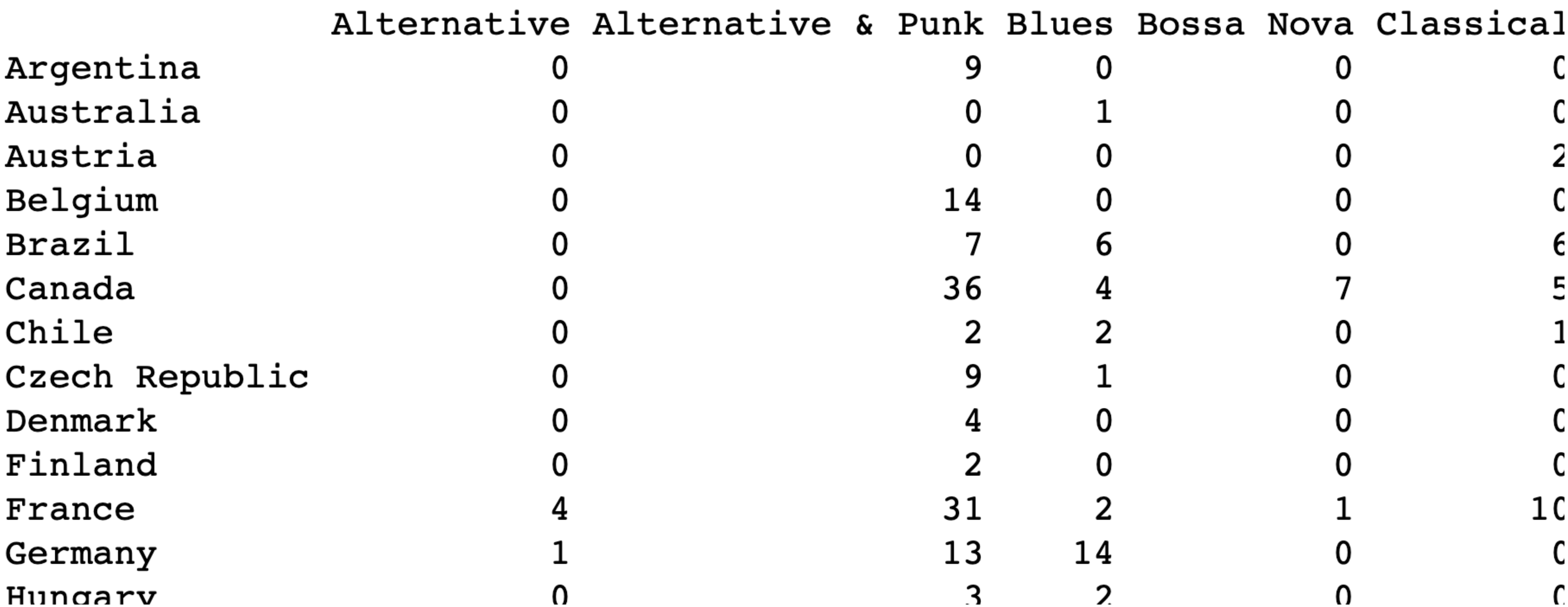
Use variations to get more information

Create **new data frames** from contingency tables (using counts and other functions)

NUMBER CRUNCHING AND ANALYSIS

```
     Argentina      Australia        Austria        Belgium         Brazil 
            38             38             38             38            190 
        Canada          Chile Czech Republic        Denmark        Finland 
           304             38             76             38             38 
        France        Germany        Hungary          India        Ireland 
           190            152             38             74             38 
         Italy    Netherlands         Norway         Poland       Portugal 
            38             38             38             38             76 
         Spain         Sweden United Kingdom            USA 
            38             38            114            494 
```

NUMBER CRUNCHING AND ANALYSIS



```
                Alternative Alternative & Punk Blues Bossa Nova Classical
Argentina                 0                  9     0          0         0
Australia                 0                  0     1          0         0
Austria                   0                  0     0          0         2
Belgium                   0                 14     0          0         0
Brazil                    0                  7     6          0         6
Canada                    0                 36     4          7         5
Chile                     0                  2     2          0         1
Czech Republic            0                  9     1          0         0
Denmark                   0                  4     0          0         0
Finland                   0                  2     0          0         0
France                    4                 31     2          1        10
Germany                   1                 13    14          0         0
Hungary                   0                  3     2          0         0
```

NUMBER CRUNCHING AND ANALYSIS

2-way province/genre contingency table for **Brazil**

```
   Alternative & Punk Blues Classical Hip Hop/Rap Latin Metal Pop R&B/Soul
DF                  0     6         4           0     8     6   1        2
RJ                  2     0         0           0    16     2   0        0
SP                  5     0         2           2    29     7   2        1

   Reggae Rock Sci Fi & Fantasy Soundtrack World
DF      0   11                0          0     0
RJ      0   16                0          0     2
SP      6   54                2          4     0
```

NUMBER CRUNCHING AND ANALYSIS

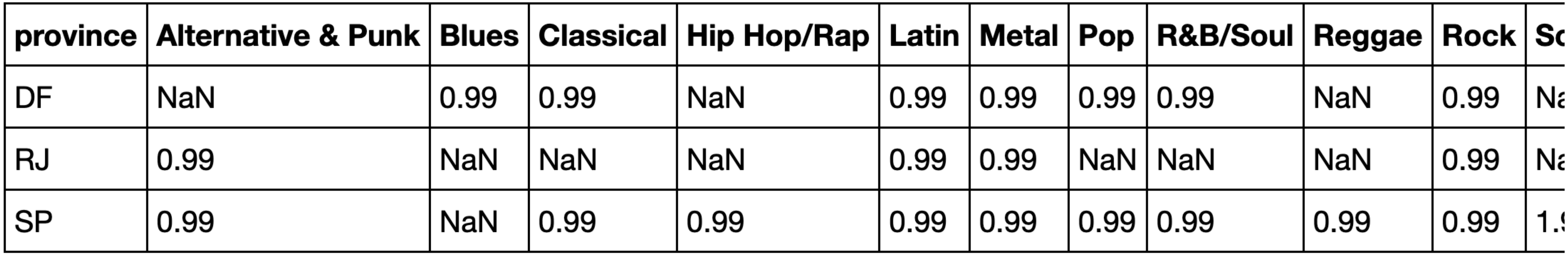
2-way province/genre **proportion** contingency table for Brazil

```
   Alternative & Punk Blues Classical Hip Hop/Rap Latin Metal  Pop R&B/Soul
DF               0.00  0.03      0.02        0.00  0.04  0.03 0.01     0.01
RJ               0.01  0.00      0.00        0.00  0.08  0.01 0.00     0.00
SP               0.03  0.00      0.01        0.01  0.15  0.04 0.01     0.01

   Reggae Rock Sci Fi & Fantasy Soundtrack World
DF   0.00 0.06             0.00       0.00  0.00
RJ   0.00 0.08             0.00       0.00  0.01
SP   0.03 0.28             0.01       0.02  0.00
```

NUMBER CRUNCHING AND ANALYSIS

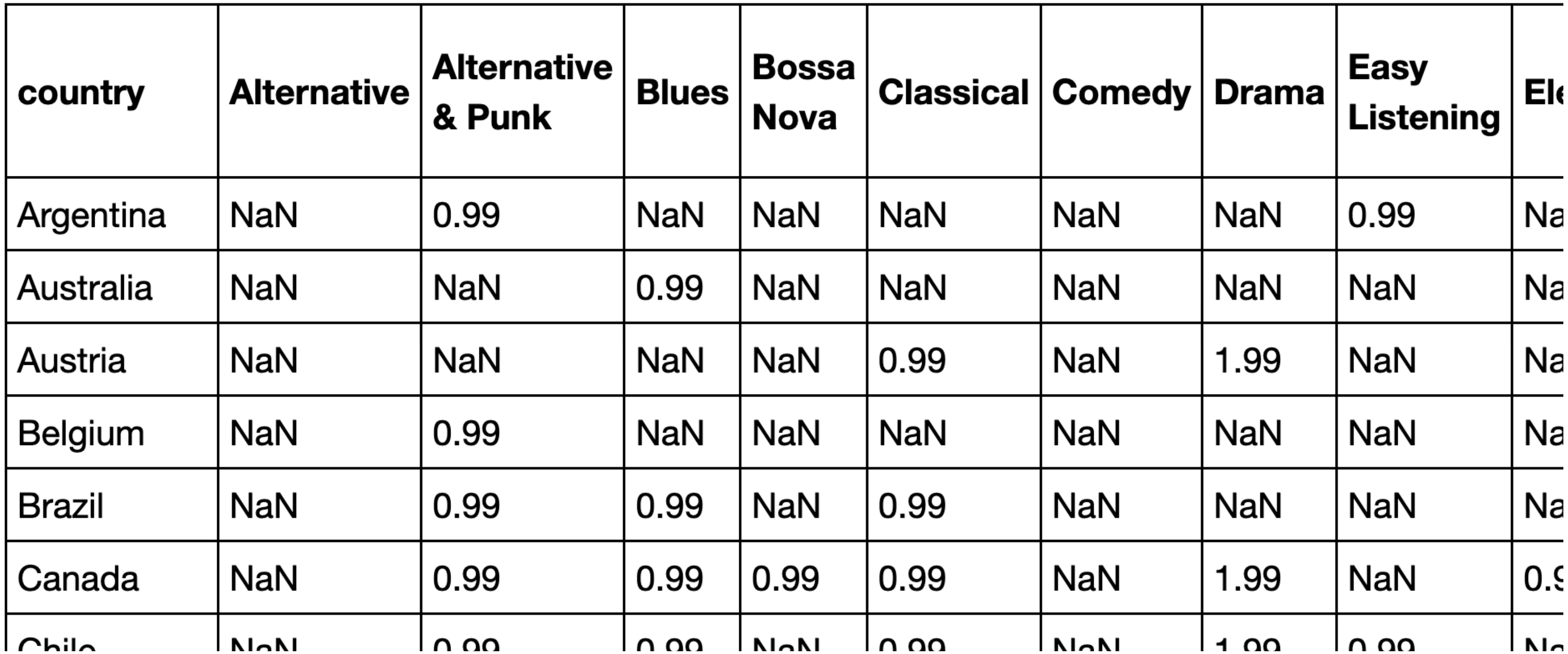
2-way province/genre **purchase cost information** for Brazil



province	Alternative & Punk	Blues	Classical	Hip Hop/Rap	Latin	Metal	Pop	R&B/Soul	Reggae	Rock	Sc
DF	NaN	0.99	0.99	NaN	0.99	0.99	0.99	0.99	NaN	0.99	Na
RJ	0.99	NaN	NaN	NaN	0.99	0.99	NaN	NaN	NaN	0.99	Na
SP	0.99	NaN	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.

NUMBER CRUNCHING AND ANALYSIS

2-way province/genre purchase (mean) cost information for **countries**



country	Alternative	Alternative & Punk	Blues	Bossa Nova	Classical	Comedy	Drama	Easy Listening	Ele
Argentina	NaN	0.99	NaN	NaN	NaN	NaN	NaN	0.99	Na
Australia	NaN	NaN	0.99	NaN	NaN	NaN	NaN	NaN	Na
Austria	NaN	NaN	NaN	NaN	0.99	NaN	1.99	NaN	Na
Belgium	NaN	0.99	NaN	NaN	NaN	NaN	NaN	NaN	Na
Brazil	NaN	0.99	0.99	NaN	0.99	NaN	NaN	NaN	Na
Canada	NaN	0.99	0.99	0.99	0.99	NaN	1.99	NaN	0.9
Chile	NaN	0.99	0.99	NaN	0.99	NaN	1.99	0.99	Na

DATA VISUALIZATION

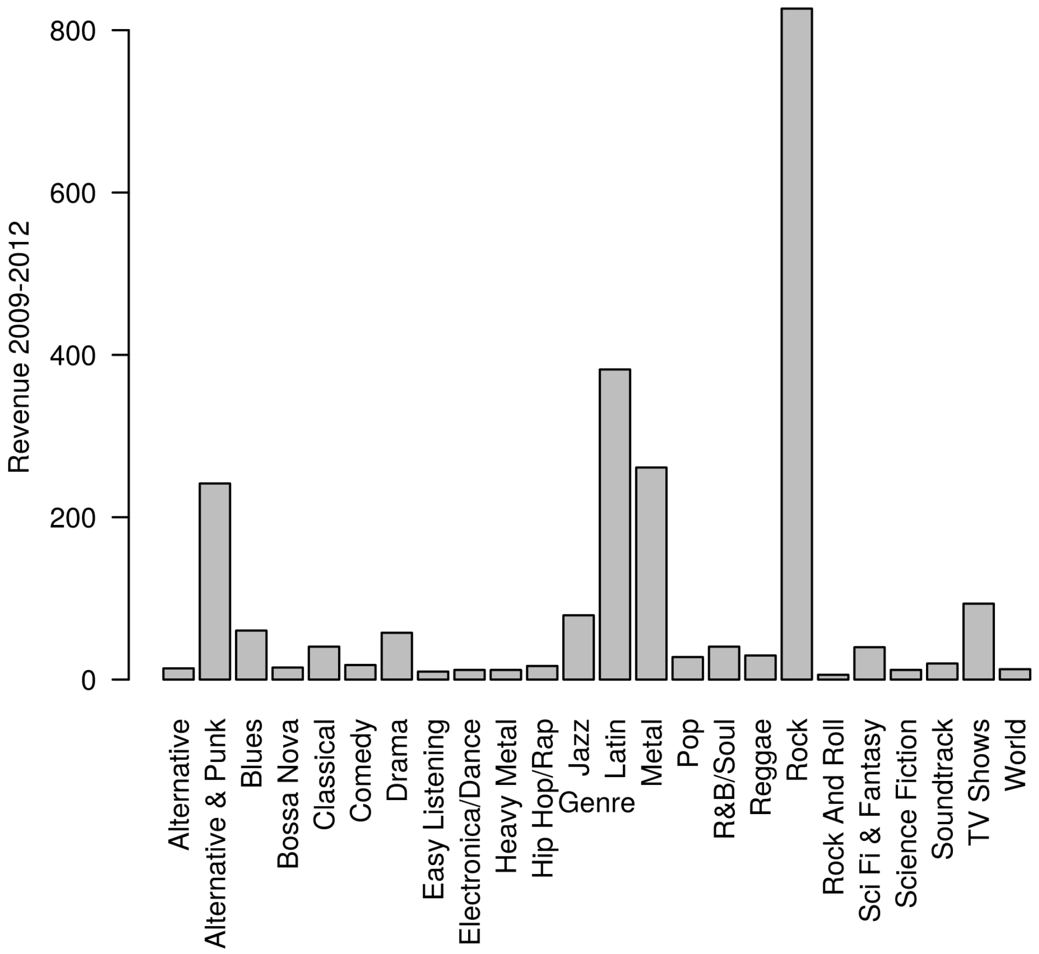
Generate 1D and multivariate charts to analyze/understand the data

Use variations in style/format

Use **multiple visualizations** for the same variables to find optimal methods

- barcharts
- scatterplots
- line graphs
- etc.

NUMBER CRUNCHING AND ANALYSIS



EXERCISE

DATA AND SYSTEM IDENTIFICATION

Identify a system of interest OR identify a dataset that exists and the system that it represents.

DATA RELEVANCE (I)

Determine what data you have about the system of interest OR what data you **could** have, if you were to collect it.

CONCEPTUAL MODEL

Develop a conceptual model of the system.

GOALS OF ANALYSIS

Determine:

- what questions do you want to answer
- what issue do you want to address
- what decision do you want to make
- what relationships do you want to explore

DATA RELEVANCE (II)

Determine: if you have the data you need to answer these questions, explore these relationships, etc.

If not determine how to get that data OR go back and come up with a question you can answer using your data, a relationship you can explore using your data, etc.

DATA STORAGE

Determine how to store your data.

DATE EXPORT

Determine how to export, transform and/or load your data so that it's available for analysis.

DATA VALIDATION

Determine the quality of your data - what it is representing and how well it is representing that which it is representing.

DATA EXPLORATION

Explore your data and discover patterns, relationships, structures.

DATA PRESENTATION

Present your findings.