Modulo 5

Data Analytics

Lesson 5: Big Data Analytics & Visualization

Alejandro Maté
Juan Carlos Trujillo

Table of contents

BIG DATA INTEGRATION

- Typically, when working with Big Data, we gather information from multiple sources and **put it all together** in a **repository**
  - This is known as ***Data Lake***
- Data lakes **do not** perform any **transformation** over the data sources

BIG DATA INTEGRATION

- However, **isolated** data is **rarely useful**
- For example:
  - The **number of calls** to the Fire Department have **increased** over the years
    - Have there been **more fires**?
    - How much has the **population increased**?
    - Did the Fire Department **reduce** the amount of **fire deaths**?
- We need to **integrate** the information in order to draw conclusions

BIG DATA INTEGRATION

- Variability of formats in 2017 for Open Data

	format	|resources|	%	|portals|
1	**HTML**	491,891	25	74
2	PDF	182,026	9.2	83
3	**CSV**	179,892	9.1	108
4	XLS(X)	120,703	6.1	89
5	**XML**	90,074	4.6	79
6	**ZIP**	50,116	2.5	74
	...			
11	**JSON**	28,923	1.5	77
16	**RDF**	10,445	0.5	28

BIG DATA INTEGRATION

- Nevertheless, even if the format is **homogeneous,** we still have a **number of challenges** to face:
  - Structure and metadata
  - Accessibility and timeliness
  - Trust and data provenance
  - Multiculturalism and semantics

BIG DATA INTEGRATION

- Structure and metadata:
  - There are a **high number of different sources** and datasets available in Big and Open Data. In many cases, the order and number of columns do not match **even across datasets** with the **same file extension**
    - Sometimes data is even published in **non-machine readable** or proprietary formats, making it inaccessible for processing
  - **Metadata** refers to the names of the columns and other **complementary information** to the data provided
    - **Without** metadata, **choosing the right interpretation** for a value or a dataset is practically **impossible**
    - Rarely well-formed

BIG DATA INTEGRATION

- Accessibility and timeliness:
  - Information gathered from external sources may **stop being accessible** at any point
    - However, we may not have enough **storage capacity** at our disposal to store everything
  - Information may **stop being updated** by third parties
    - May **still be useful** depending on our requirements and the context
    - !!! CAREFUL **DATES SHOULD MATCH** ON DIFFERENT DATASETS !!!
  - Certain data may be **missing**
    - Due to being collected incorrectly or due to errors in the system
  - For Open Data it is **unrealistic** to think that information will be provided at the **same level of quality** as commercial sources

BIG DATA INTEGRATION

- Trust and data provenance:
  - What to do if data **may** not be exact?
    - Wikipedia?
  - How much **confidence** can we have? How much do we **need**?
  - What if the data is provided by a **facilitator or portal**, how has it been **altered**?

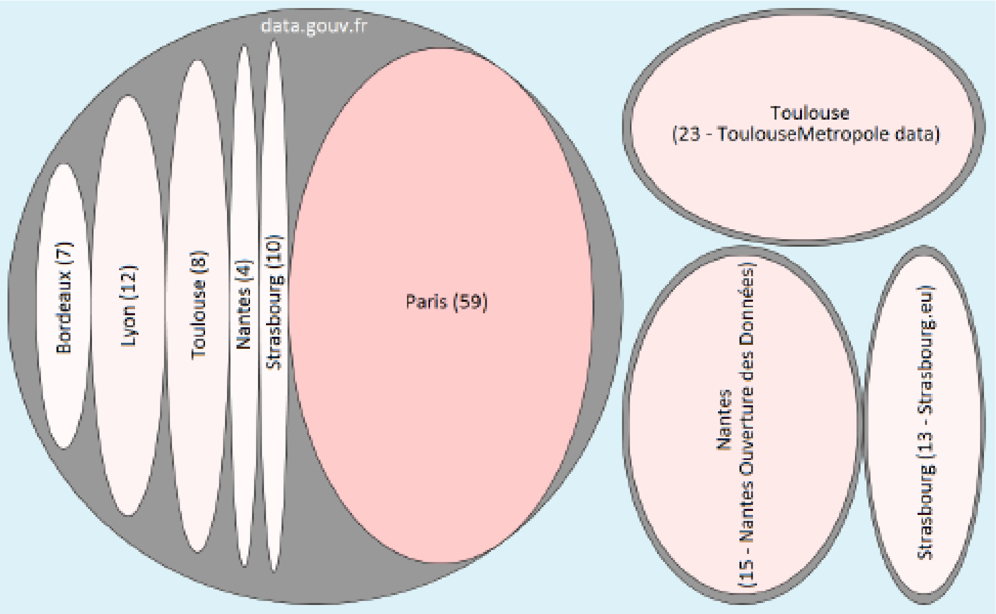
- **Data provenance** registers the transformations applied to the data **until** it has reached its **current version**
  - In absence of data provenance we must depend on:
    - **Redundancy**: How often a certain piece of information appears exactly the same?
    - **Provider trust**: How much do we trust whoever is providing us the data?

BIG DATA INTEGRATION

- Multilingualism and semantics:
  - Assume data is available, its structure is correct and the metadata is well-formed
  - The data may still be registered in **different languages**
    - Canada has two official languages – French and English
    - EU has 28 (27?) states with 24 languages
  - The data may refer to the **same concept** with **different semantics**
    - “Address” in France may not refer to the same concept as “Address” in Germany
    - National ID only exists in Europe
    - National ID and Fiscal ID are separate in some countries
    - Date is registered differently in English 11/30/2018 than in Spanish 30/11/2018

DATA ANALYTICS PROCESS: DATA VISUALIZATION

- Many people consider that data visualizations are the end of the data analytics process – **Communication** of results
- However, data visualization can be **effective** for aiding **during the analysis** and **integration of data**



Source:
Carvalho, Paulo & Hitzelberger, Patrik & Otjacques, Benoit & Bouali, Fatma & Venturini, Gilles. (2014). Open Data Integration -Visualization as an Asset. 10.13140/2.1.1788.5440.

DATA ANALYTICS PROCESS: DATA VISUALIZATION

- Visualizations can help us to **understand and identify**:
  - How the **data behaves**, especially across time
  - Existing **patterns** within the data
  - Potential attributes for **classification** and machine learning
  - Evaluate and initially **discard** some **hypotheses**
  - What **other datasets** could **complement** our current analysis

BIG DATA TOOLS FOR DATA ANALYTICS

- In order to perform analytics over Big Data sources we need **tools** that facilitate our task:
  - We require **flexibility** to read multiple and varying data formats
  - We require **visualization capabilities** to analyze and present the data
  - We require **efficiency** in processing data in order to scale

DATA VISUALIZATION

- Visualizations in Jupyter:
  - Jupyter (and Python) have a wide arrange of visualization libraries to visualize data
  - For the purpose of this module we will use the **Seaborn library**

seaborn

DATA VISUALIZATION

- Visualizations in Jupyter:
  - To create a visualization we first need to calculate the data values

In [168]:
```
# Calcular el número de distritos operados por cada unidad
SFFD.groupby(['Unit ID'],as_index=False).agg({'City': pd.Series.nunique})
```

Out[168]:

	Unit ID	City
0	27	1
1	30	1
2	45	1
3	46	1
4	47	4
...	...	...
913	VAN5	1
914	VAN6	1
915	VAN7	1
916	VAN8	2
917	VAN9	1

918 rows × 2 columns

DATA VISUALIZATION

- Visualizations in Jupyter:
  - Once we have the correct data, we pass the dataframe to the seaborn plot we wish to create
  - Depending on the plot we will need to specify multiple axis

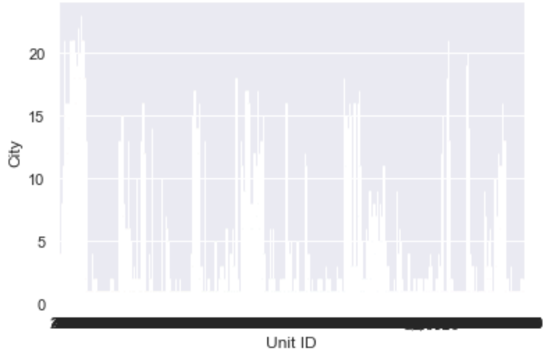
In [169]:

```python
# Para visualizar los datos, hacemos el calculo anterior y lo cargamos en seaborn
# As_index = Fales nos devuelve un dataframe plano para poder acceder a los datos desde seaborn
import seaborn as sns

values=SFFD.groupby(['Unit ID'],as_index=False).agg({'City': pd.Series.nunique})
sns.set_theme()
barplot = sns.barplot(data=values,x='Unit ID',y='City')
barplot
```

Out[169]:

```
<AxesSubplot:xlabel='Unit ID', ylabel='City'>
```



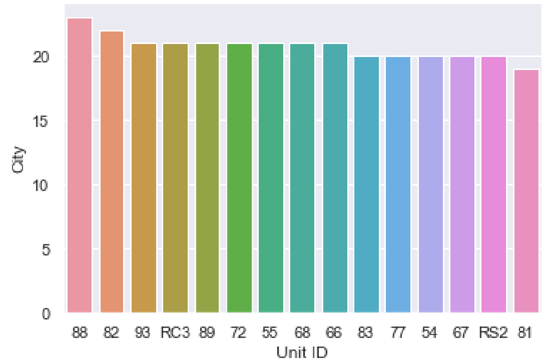
DATA VISUALIZATION

- Visualizations in Jupyter:
  - If there are too many items, the visualization will not show any meaningful information
  - In these cases we can filter the data to reduce the amount of items

In [177]:
```
# Como el número de unidades a representar es excesivo, tendremos que filtrar.
# Una manera es ordenar de mayor a menor número de distritos las unidades y quedarnos con aquellas que menos distrit
# Otras alternativas serían centrar el análisis únicamente en aquellas de una cierta zona, cierto tipo,
# Las que más llamadas reciben, etc

values=SFFD.groupby(['Unit ID'],as_index=False).agg({'City': pd.Series.nunique})\
.sort_values(by=['City'],ascending=False).head(15)
barplot = sns.barplot(data=values,x='Unit ID',y='City')
barplot
```

Out[177]: <AxesSubplot:xlabel='Unit ID', ylabel='City'>

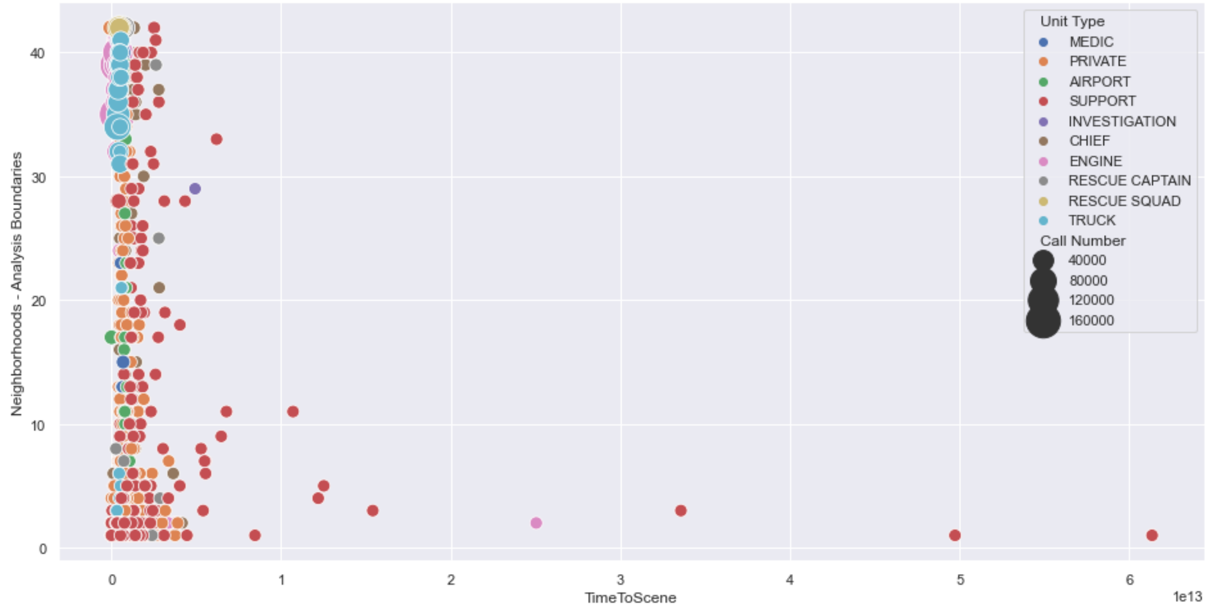


DATA VISUALIZATION

- Visualizations in Jupyter:
  - In some cases, there will be multiple dimensions to represent at once
    - For example, the time to scene and unit type for each unit considering the number of calls of each unit
  - These cases require more advanced visualizations

In [202]:

```
sns.set(rc={"figure.figsize":(16, 8)})
bubbleplot = sns.scatterplot(data=values,x='TimeToScene',y='Neighborhooods - Analysis Boundaries', \
                              size='Call Number',hue='Unit Type', sizes=(100,800))
```

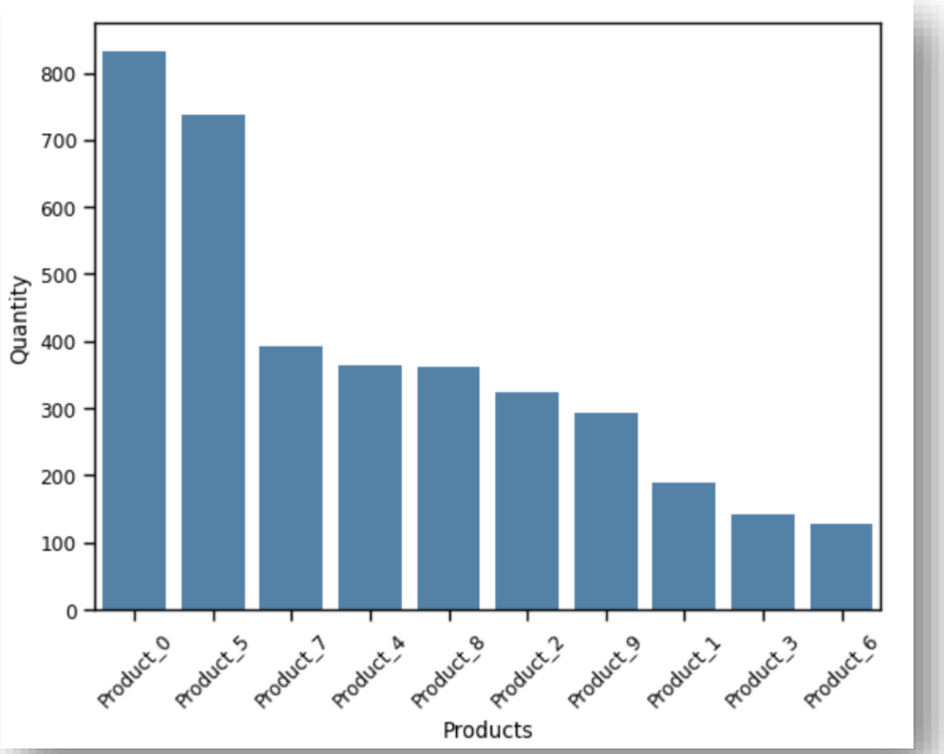


DATA VISUALIZATION

- Let’s dig deeper into how to select the **most adequate visualizations**
  - A combination of **objectives pursued, intended use, receiver** and **dataset constraints**
- Visualizations depend heavily on the **analysis/communication goals**. Some of the common ones are:
  - Ranking
  - Comparatives
  - Correlations
  - Part to Whole
  - Distribution
  - Evolution across time
  - Flow

DATA VISUALIZATION

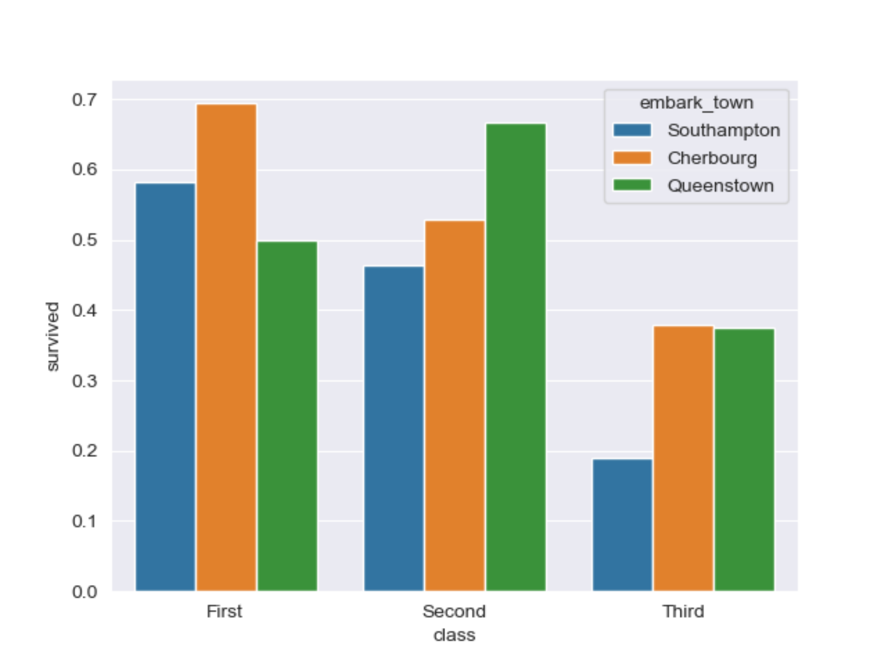
- **Ranking:** Aims to show the **order relationship** in a dataset



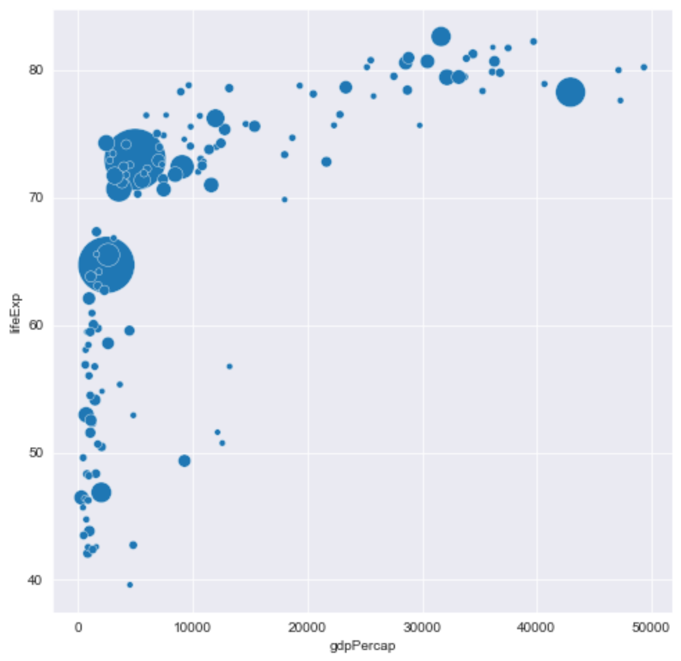
Bar Chart

DATA VISUALIZATION

- **Comparative:** Compares **numeric values** associated to different elements. Order is irrelevant but numeric differences are important



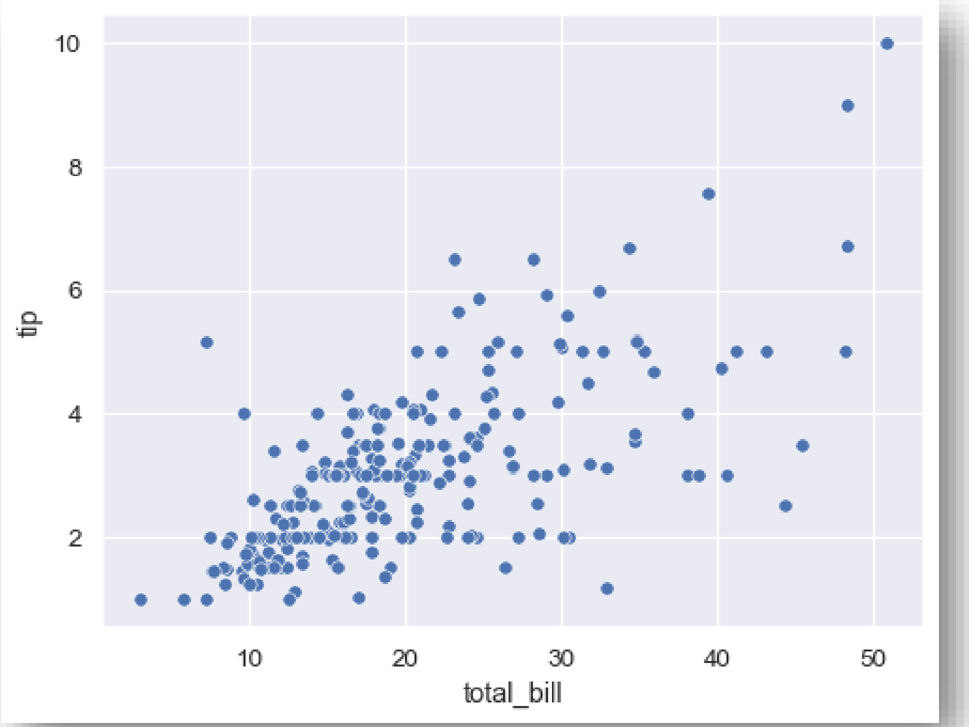
Grouped Bar Chart



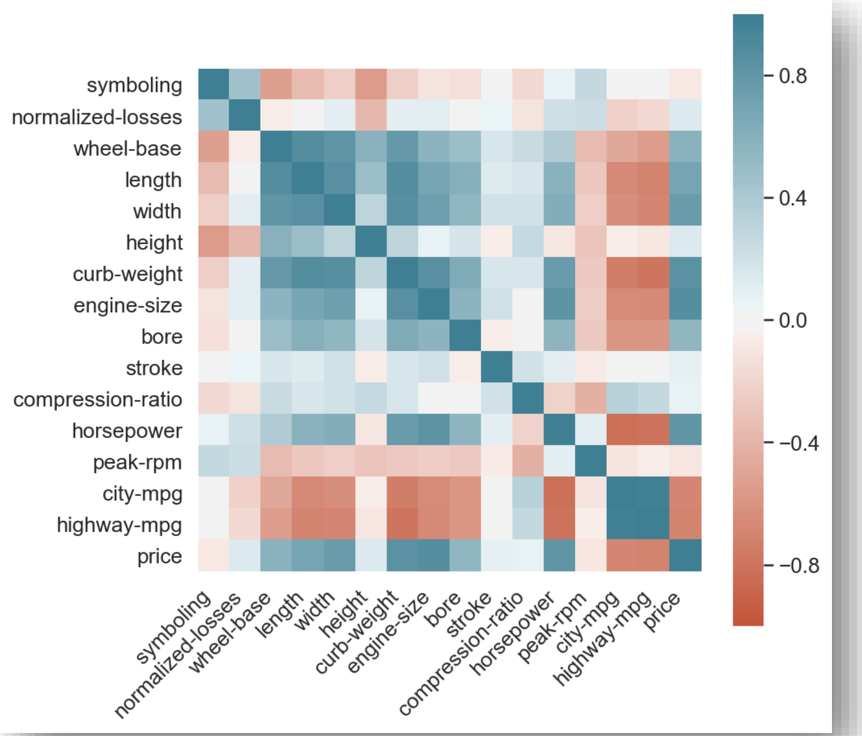
Bubble Chart

DATA VISUALIZATION

- **Correlation:** Analyzes how one variable **changes according to another variable**



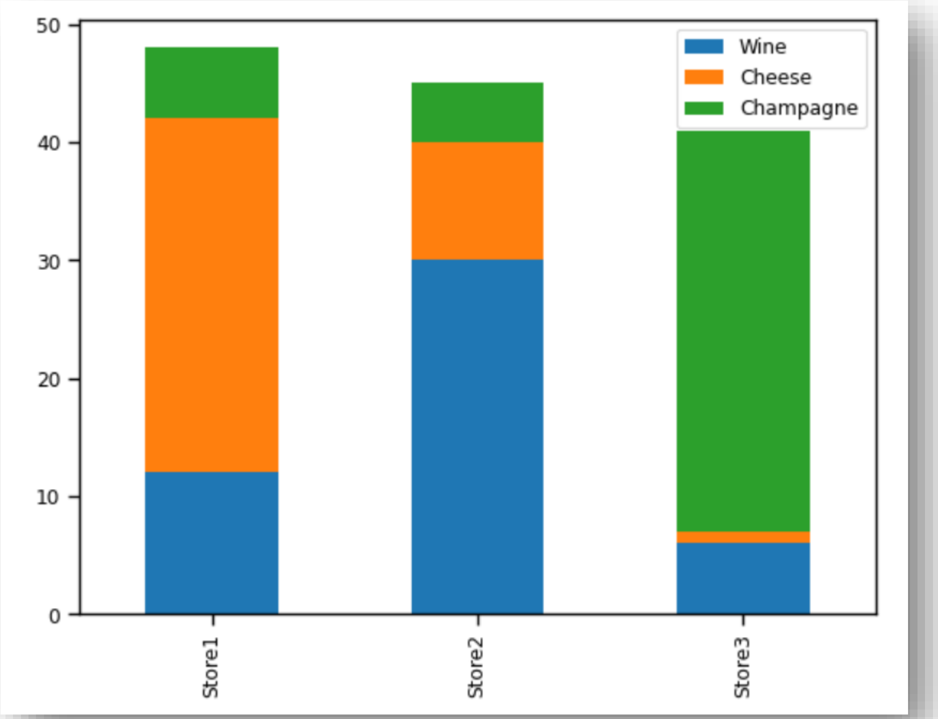
Scatterplot



Heatmap

DATA VISUALIZATION

- **Part to Whole:** Divides an element into its **components**

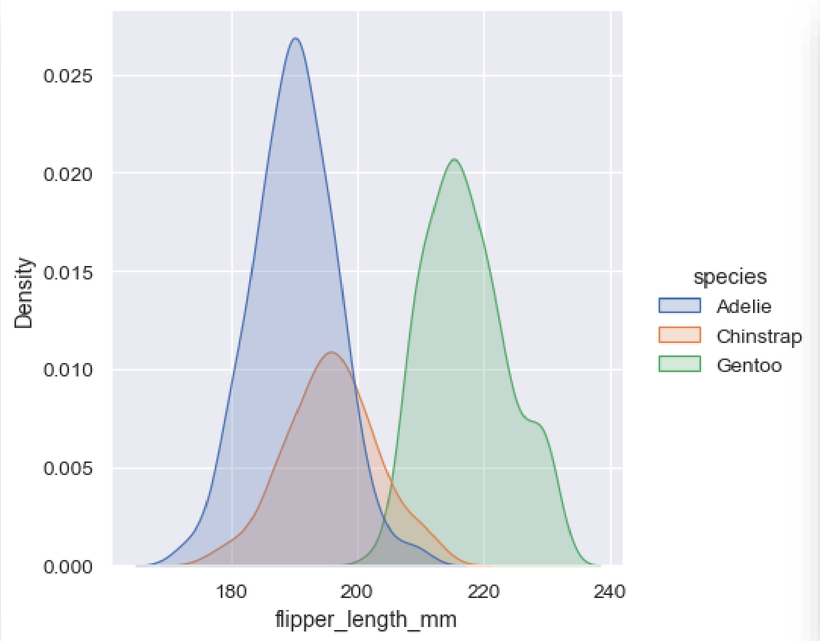


Stacked bar chart

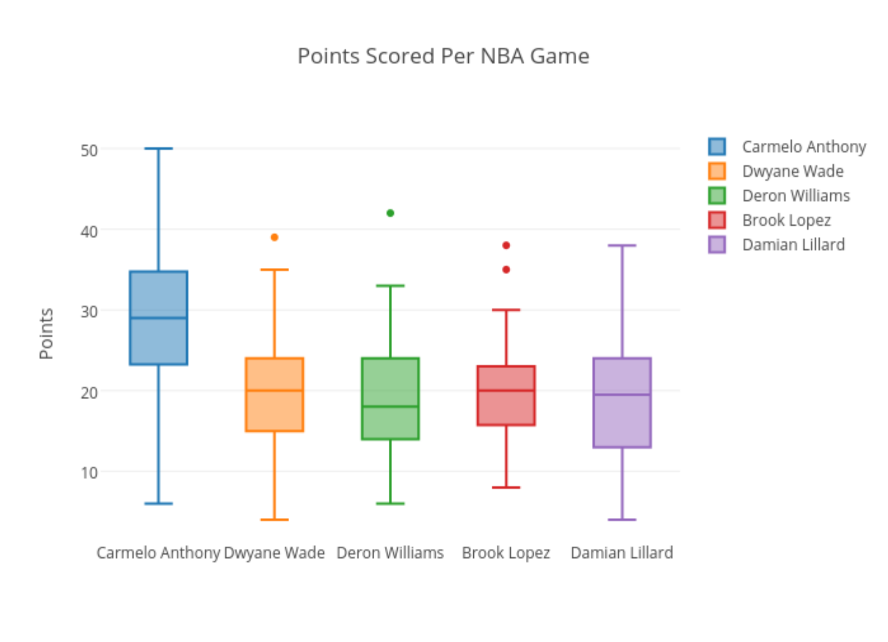
Tree Map

DATA VISUALIZATION

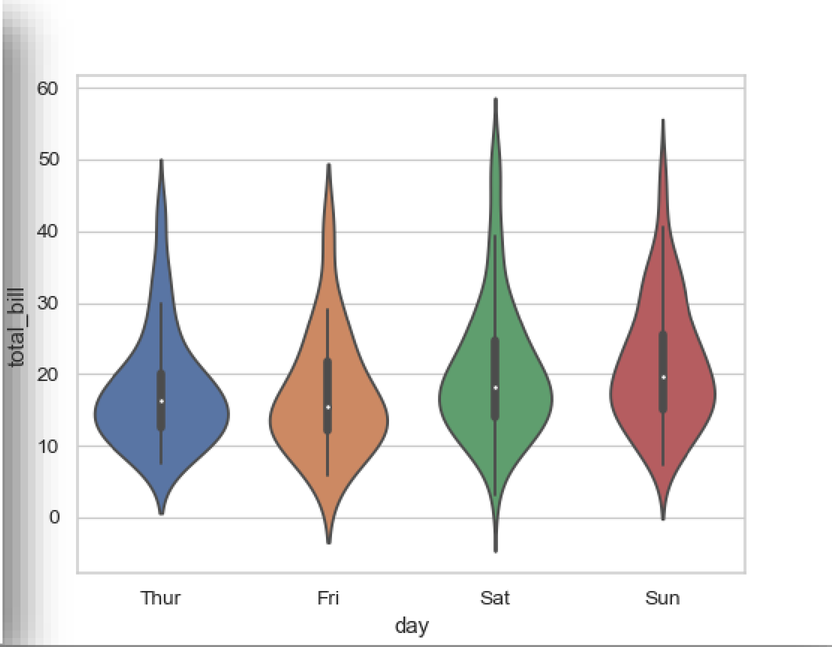
- **Distribution**: Describes the **frequency** of elements according to their value



Histogram



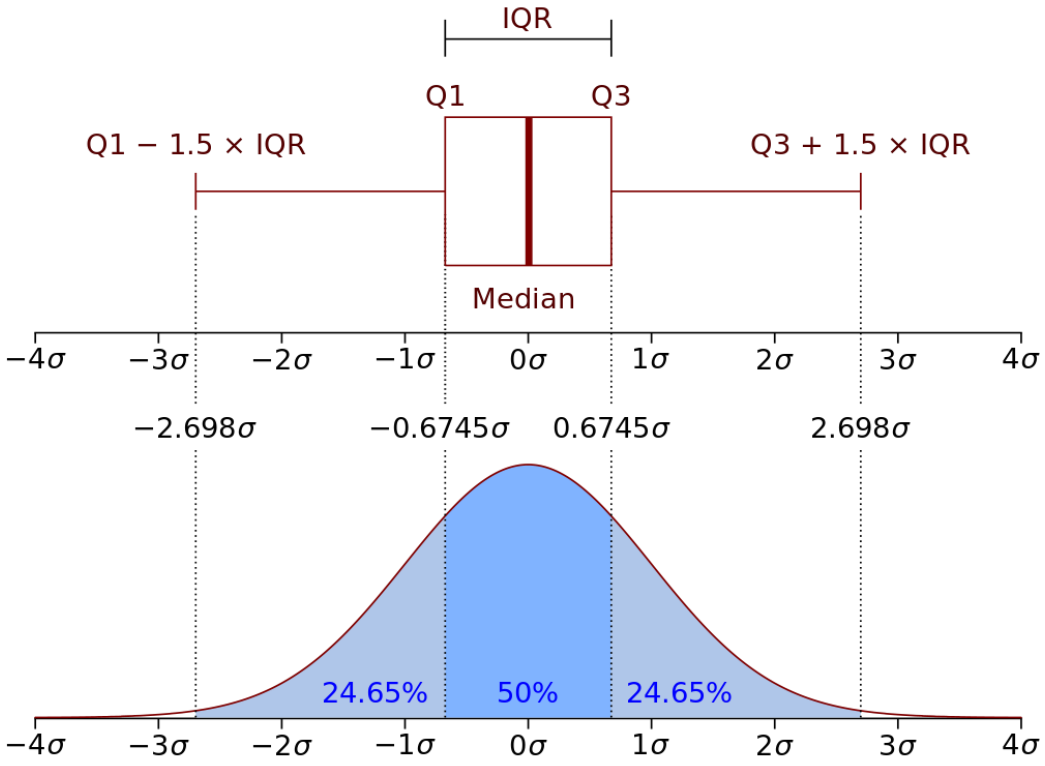
Boxes



Violin

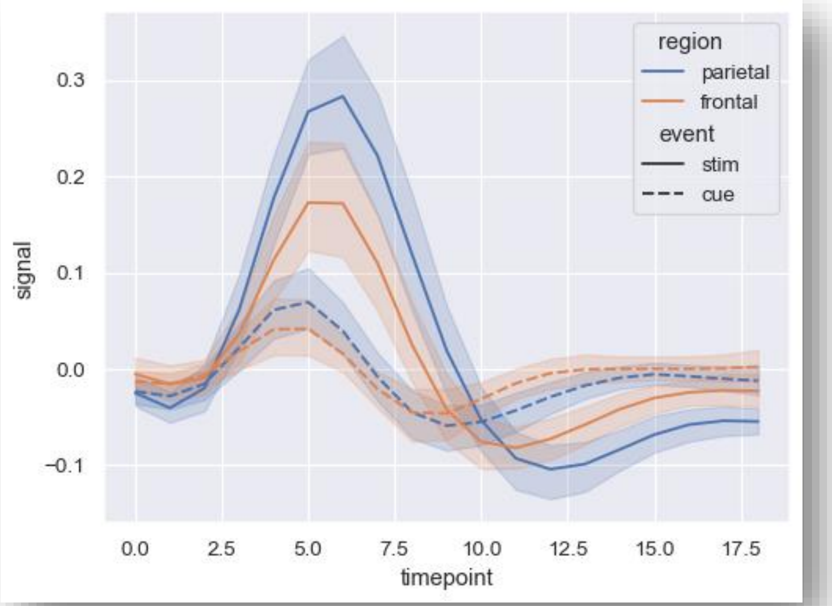
DATA VISUALIZATION

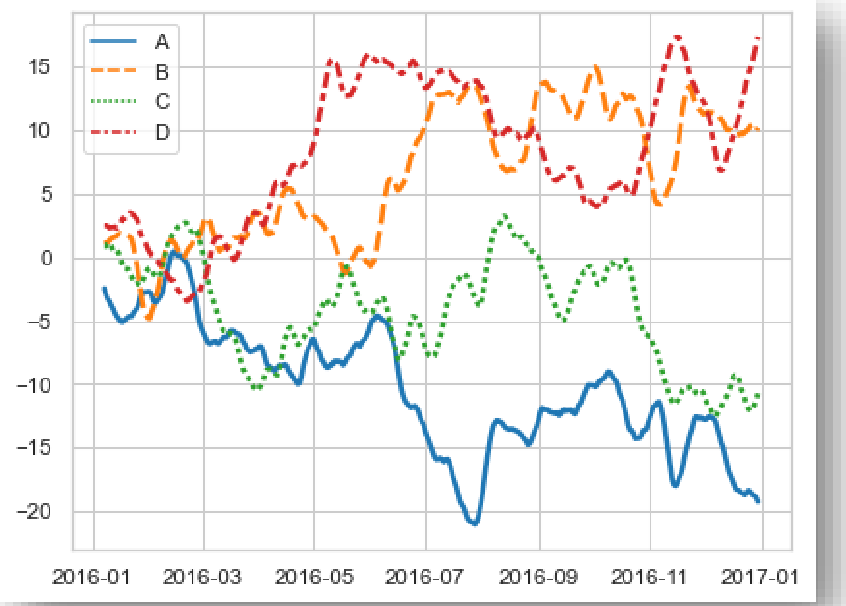
- **Distribution**: To understand distribution visualizations we must understand how a distribution behaves



DATA VISUALIZATION

- **Evolution across time:** Reflects how the value of a variable **evolves as time passes**

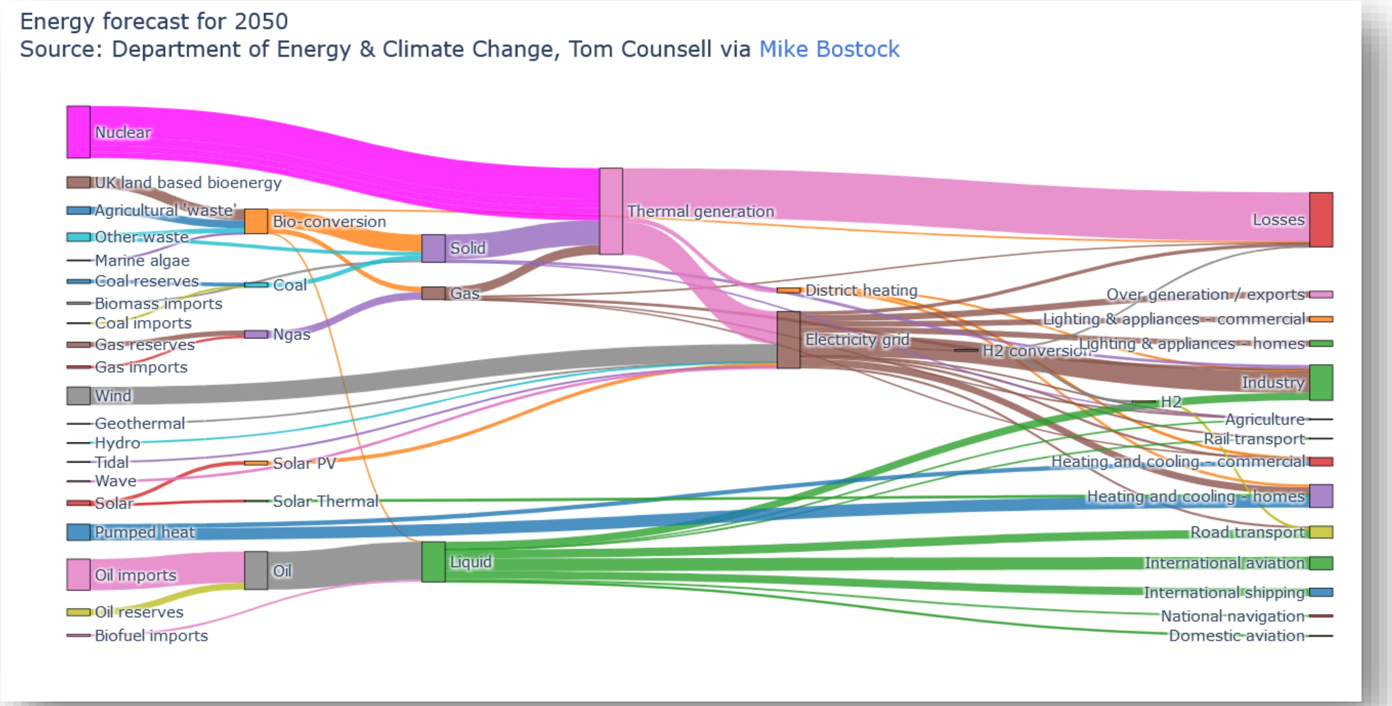




Line Chart

DATA VISUALIZATION

- **Flow:** Connects **sources and destinations** for an element. It can describe **trips** or **transformations** for materials and money



Sankey

DATA VISUALIZATION

- Selecting the most adequate visualization:
  - **Goal**:
    - Evolution across time? Distribution? Value comparison?
  - **Receiver**:
    - Visualization expert? Domain expert? Non-informed user?
  - **Use**:
    - Explore / Communicate
  - **Dataset characteristics**:
    - What kind of data will be used? How many data are we crunching? How many different values exist? How many dimensions should be considered?

DATA VISUALIZATION

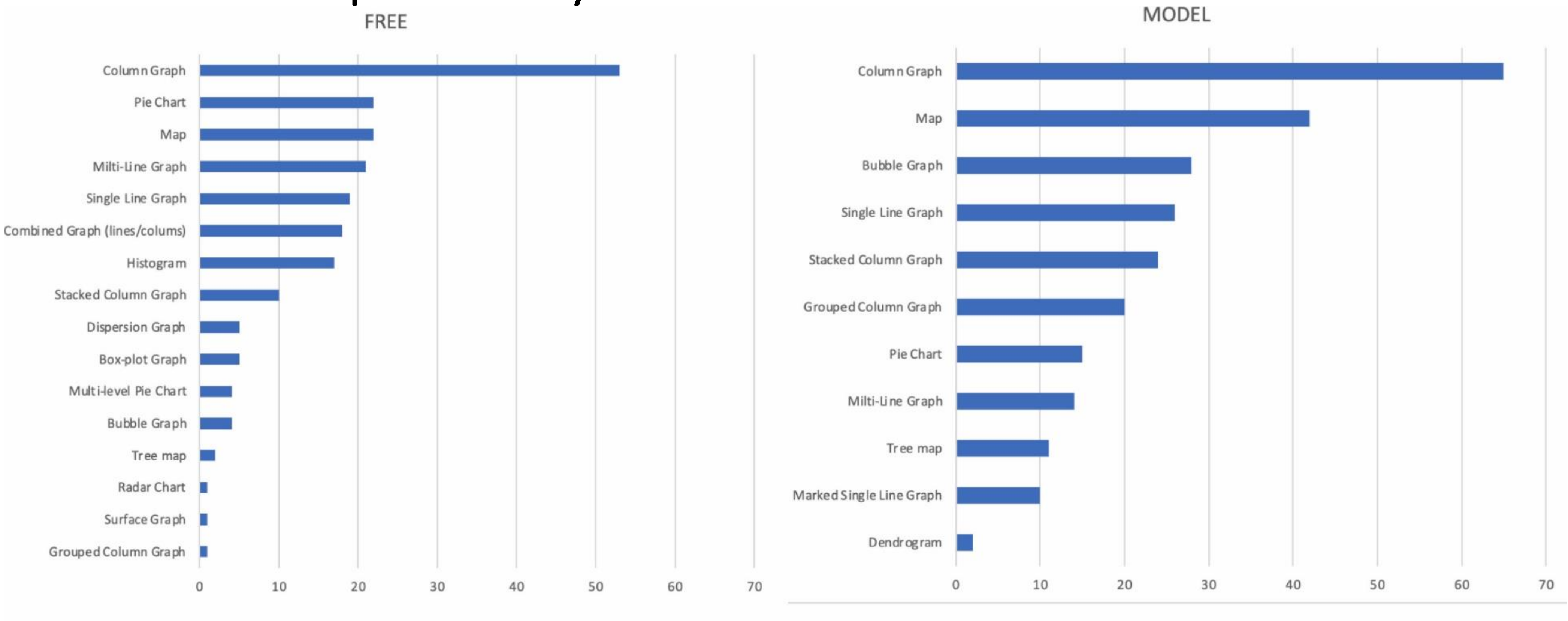
- There are proposals to select the best visualization according to the analysis context

	VISUALIZATION CONTEXT	Stacked Column Chart	Bubble Chart	Pie Chart
Goal:	Composition	fit	unfit	fit
	Comparison	fit	fit	unfit
Interaction:	Overview	acceptable	acceptable	fit
User:	Lay	fit	acceptable	fit
Dimensionality:	2-dimensional	unfit	unfit	fit
	n-dimensional	fit	fit	unfit
Cardinality:	Low	fit	acceptable	fit
Independent Type:	Nominal	fit	unfit	fit
Dependent Type:	Ratio	fit	fit	fit

Fuente: Lavalle, A., Maté, A., Trujillo, J., & Rizzi, S. (2019, September). Visualization requirements for business intelligence analytics: a goal-based, iterative framework. In *2019 IEEE 27th International Requirements Engineering Conference (RE)* (pp. 109-119). IEEE

DATA VISUALIZATION

- Selecting charts without prior analysis leads to **less-than-ideal selection**



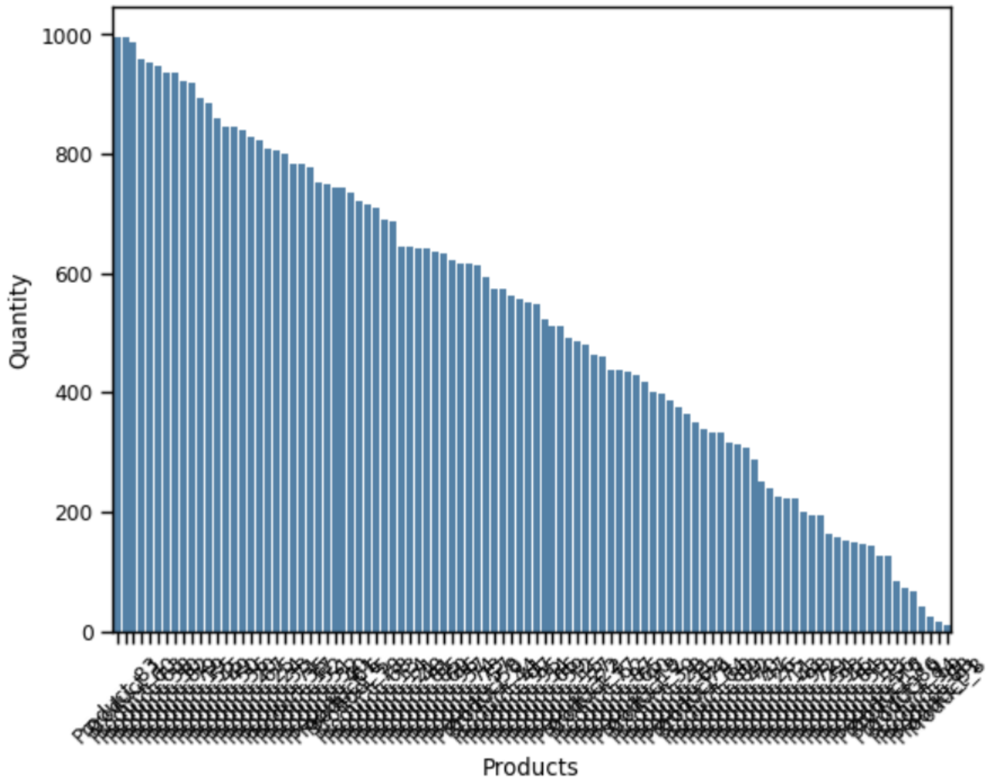
(a) Without using our methodology

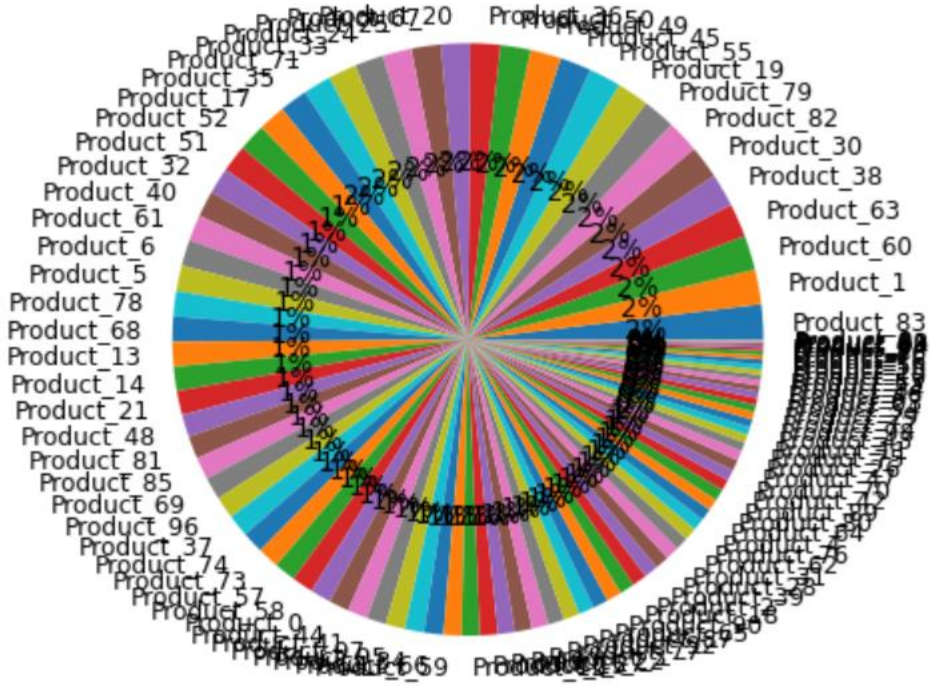
(b) Using our methodology

Fuente: Lavalle, A., Maté, A., Trujillo, J., Teruel, M. A., & Rizzi, S. (2021). A methodology to automatically translate user requirements into visualizations: Experimental validation. *Information and Software Technology*, *136*, 106592.

DATA VISUALIZATION

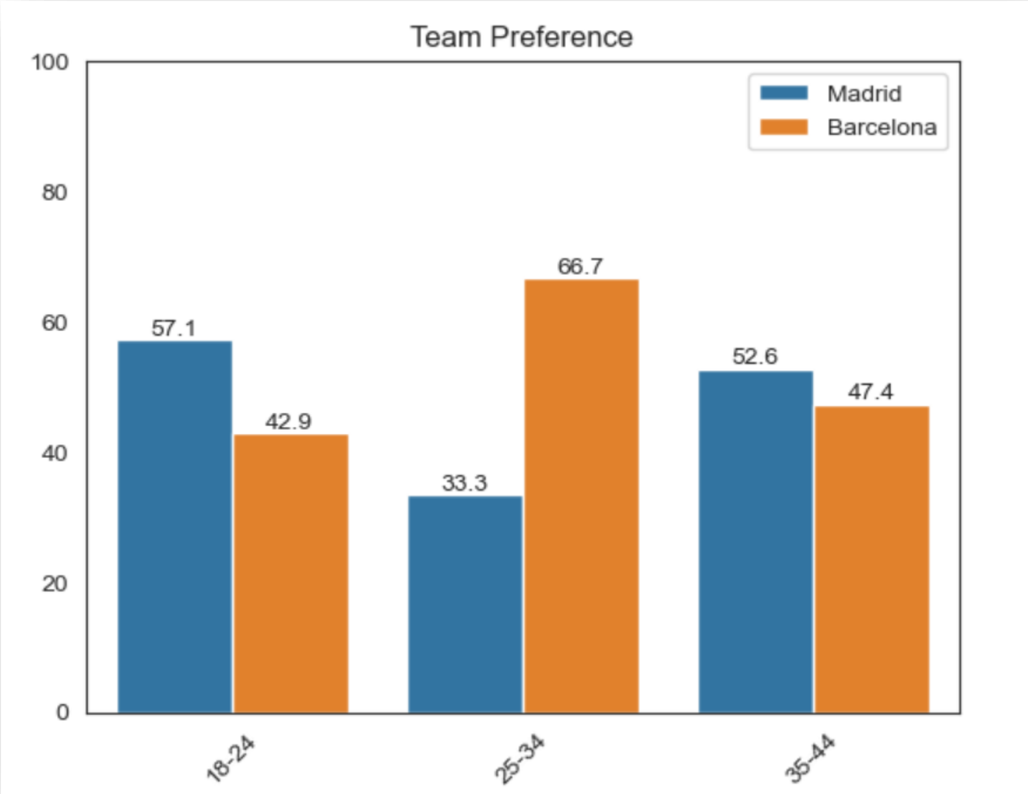
- **Dataset characteristics:** Data **volume** problems

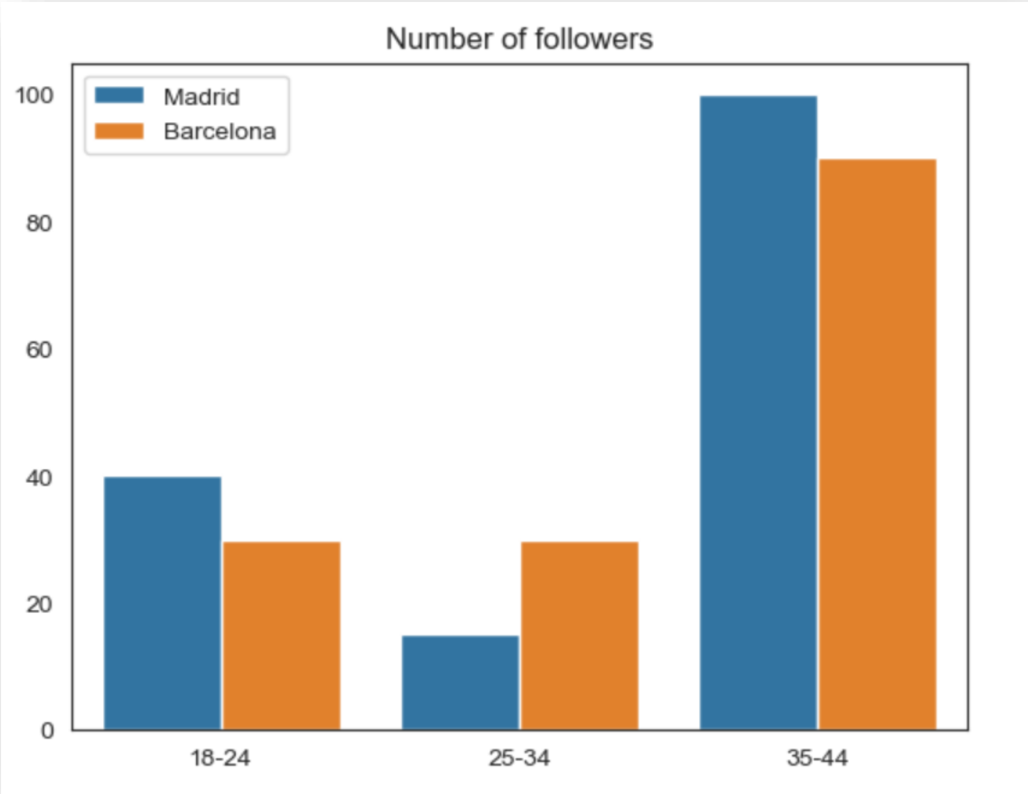




DATA VISUALIZATION

- **Dataset characteristics: Dimensionality** and **extreme values**



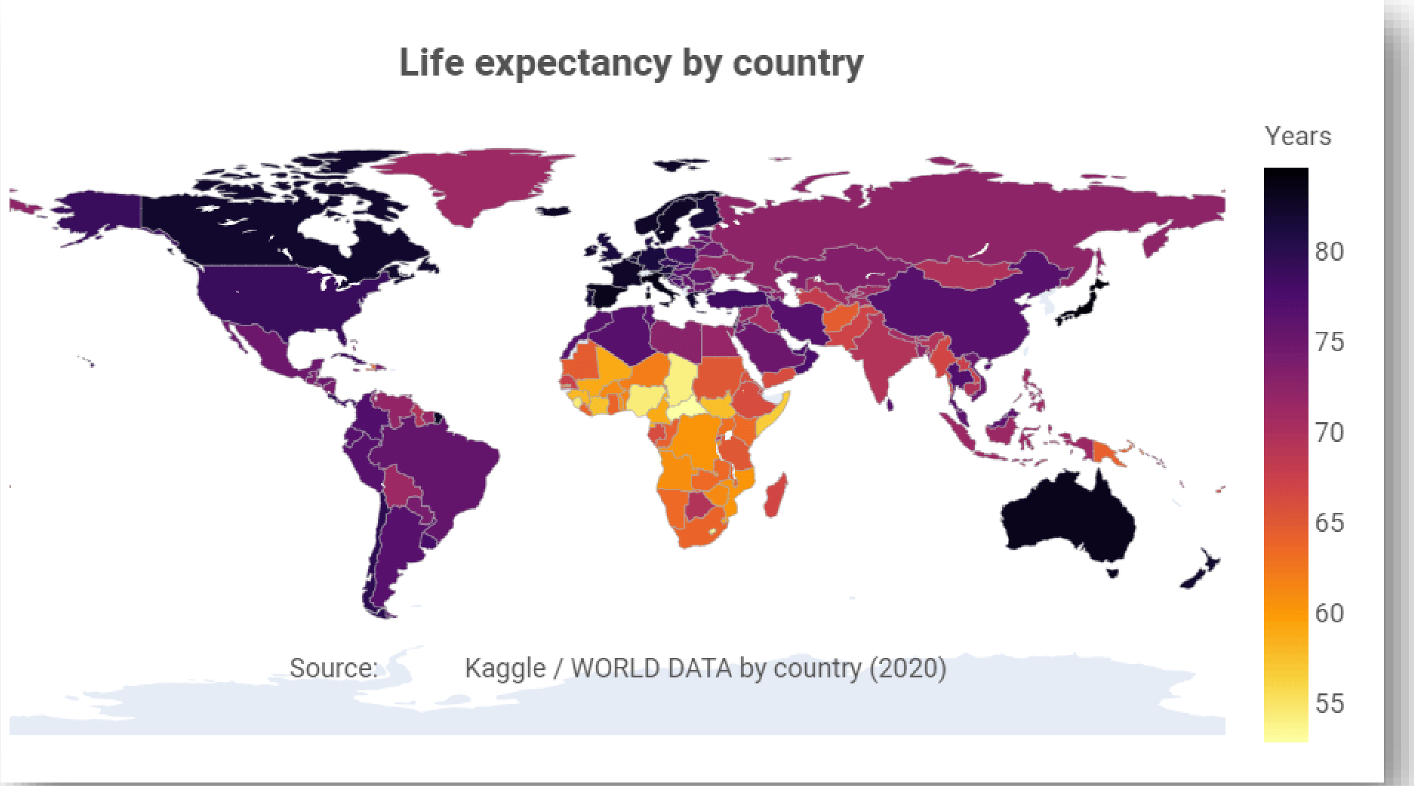


DATA VISUALIZATION

- Problems related to **data types**
  - Numeric vs String
  - Continuous vs Discrete
  - With or without order relationship

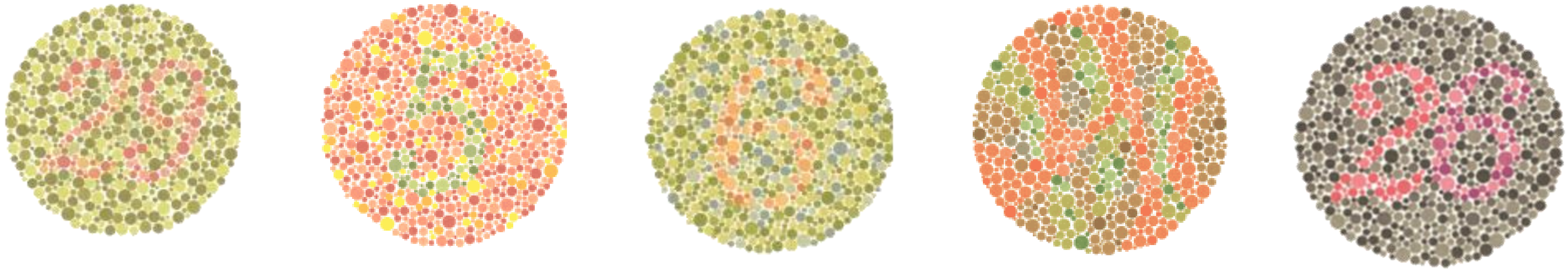
DATA VISUALIZATION

- **Color:** Additional **dimension** to represent information



DATA VISUALIZATION

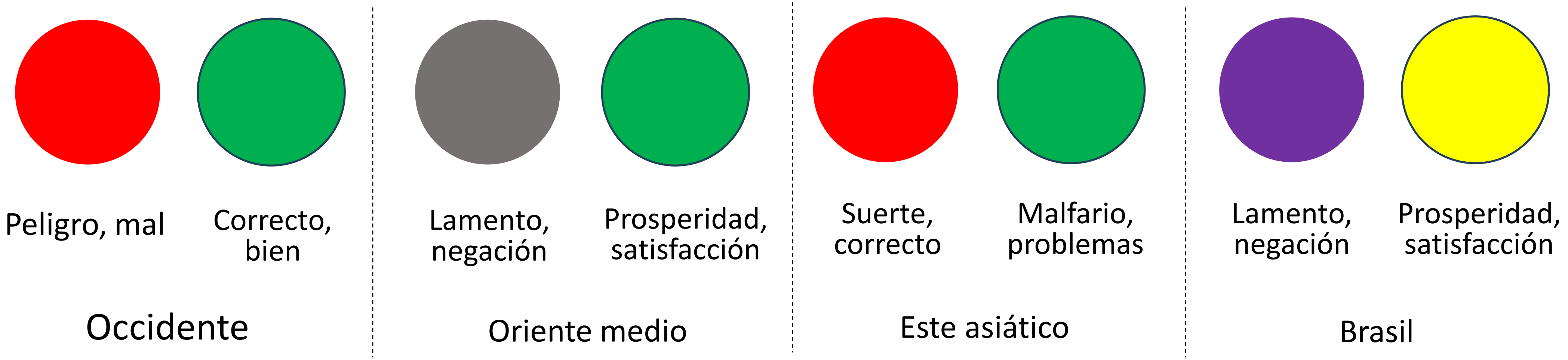
- **Color:** There are certain **limitations** to its use, for example **daltonism**



Source: Grupo Visioon. Test de daltonismo
https://visioon.es/blog/test-daltonismo-oftalmologia/

DATA VISUALIZATION

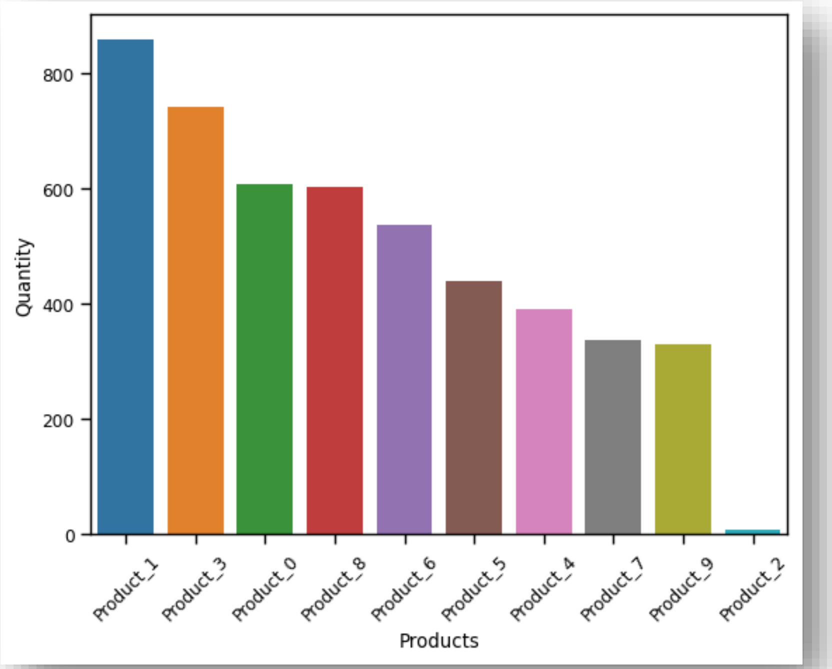
- **Color:** Also inadverted ones, such as **cultural clashes**



Fuente: Ayed, C. B., Halili, S., Tan, Y., & Grubb, A. M. (2023). Toward Internationalization and Accessibility of Color-based Goal Model Interpretation.

DATA VISUALIZATION

- **Color:** It is important to consider **object similarity** and **avoid abusing color** to make the chart "pretty"



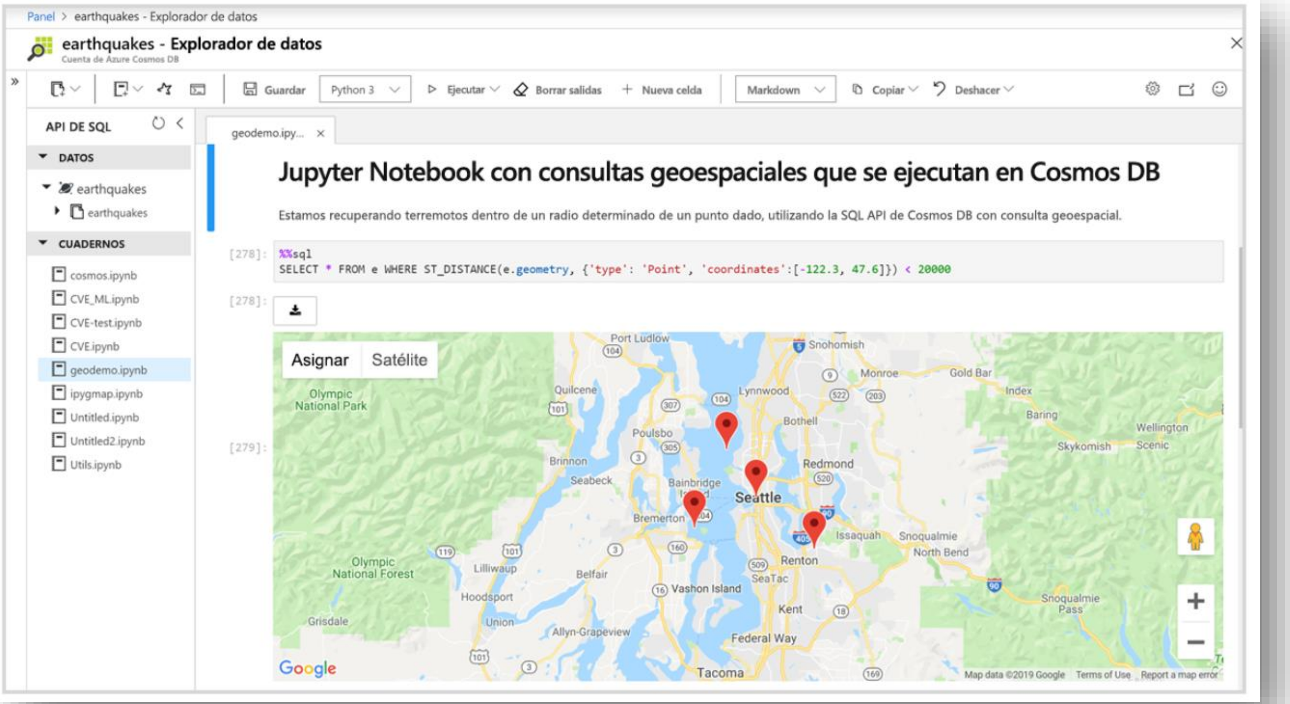
DATA VISUALIZATION

- Typically, it is **insufficient with a single visualization** to communicate. It is necessary to **combine multiple ones**.
- We can classify the traditional combinations as:
  - Notebooks
  - Dashboards (or CMOs)
  - Scorecards (or CMIs)
  - Infographics

DATA VISUALIZATION

Notebooks:

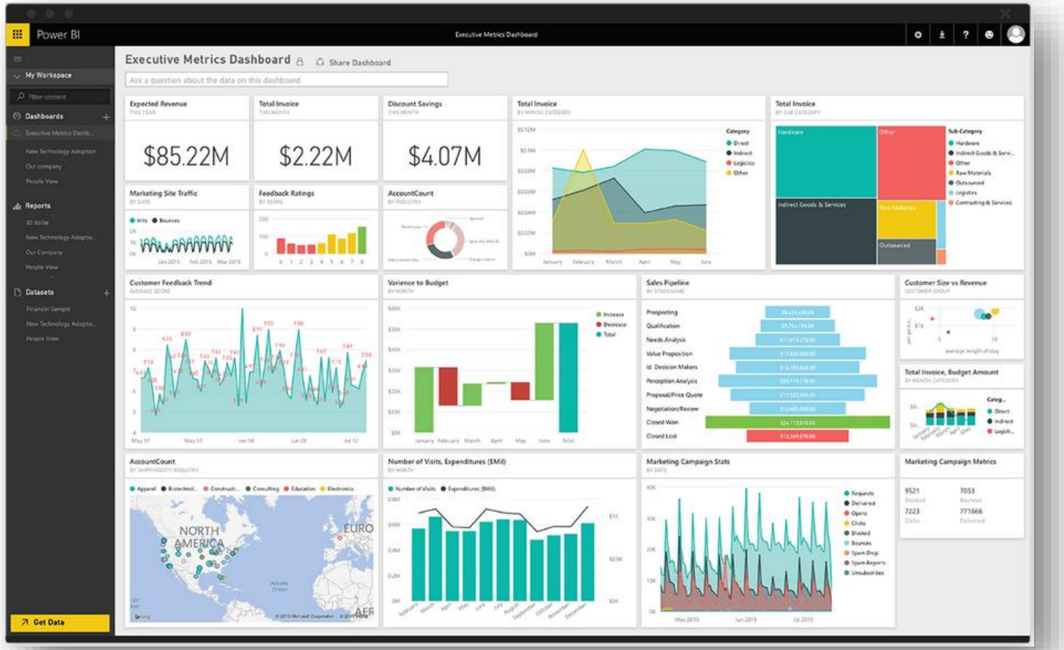
- Organize visualizations in a **narrative or explicative succession**



DATA VISUALIZATION

Dashboards or CMOs:

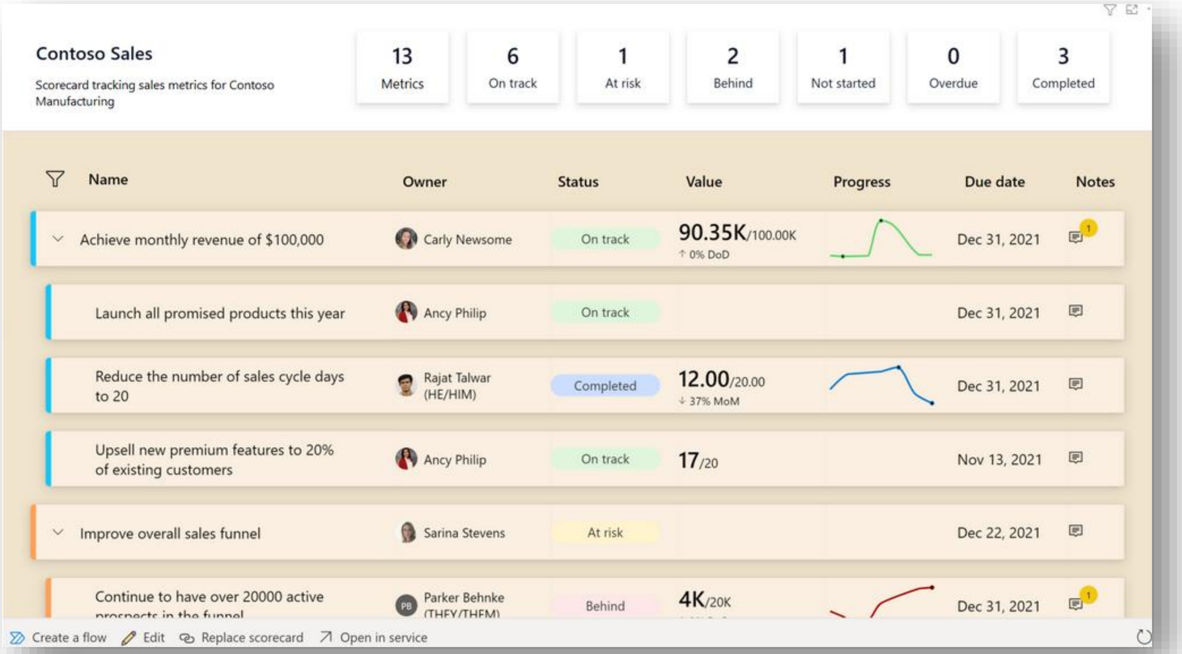
- Organize visualizations from an **analytical perspective** focused on an **analysis context**. They are typically related to business processes



DATA VISUALIZATION

Scorecards or CMIs:

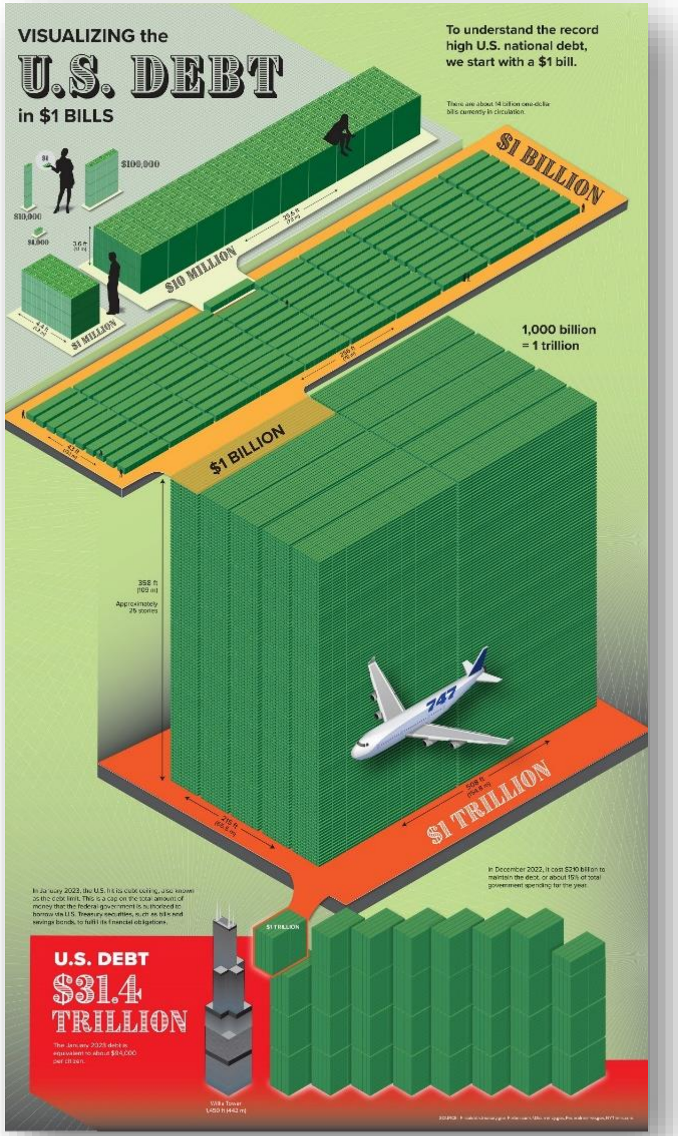
- Organize visualizations to **summarize** the current **status of the organization**



DATA VISUALIZATION

Infographics:

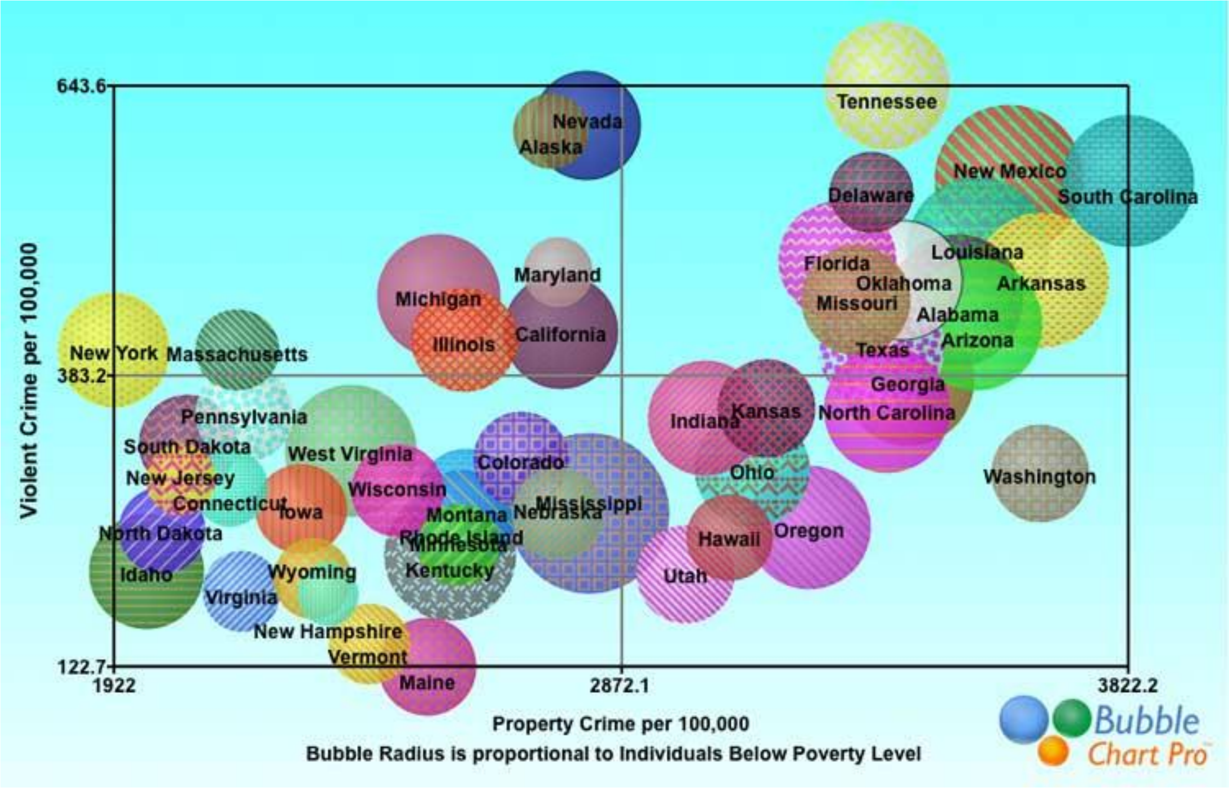
- Aim to **communicate and increase awareness** about a certain topic or situation



The tower is 442m tall

DATA VISUALIZATION

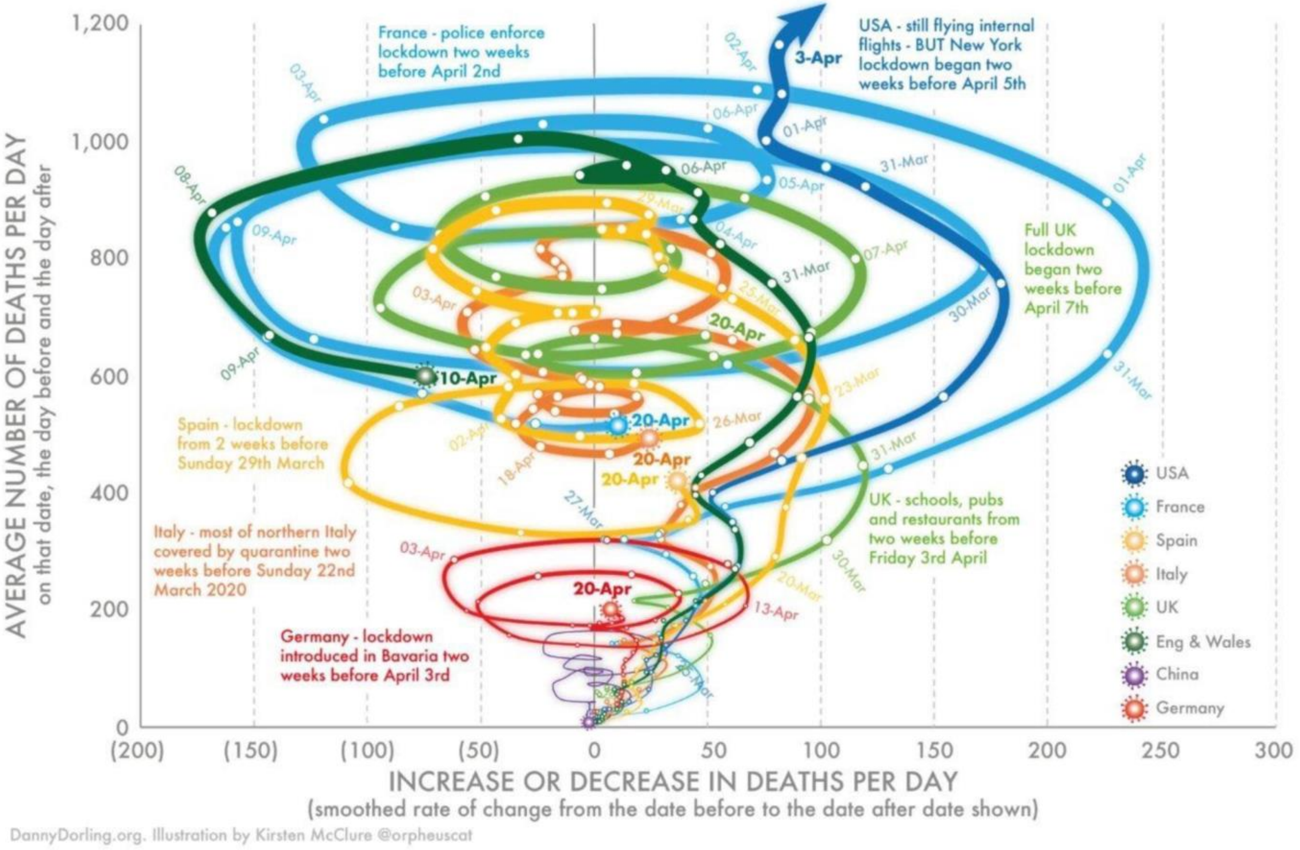
- Mistakes (maybe) to avoid: **Overloading** the chart with information



Source: Polymer. 10 Good and Bad Examples of Data Visualization (2023). https://www.polymersearch.com/blog/10-good-and-bad-examples-of-data-visualization

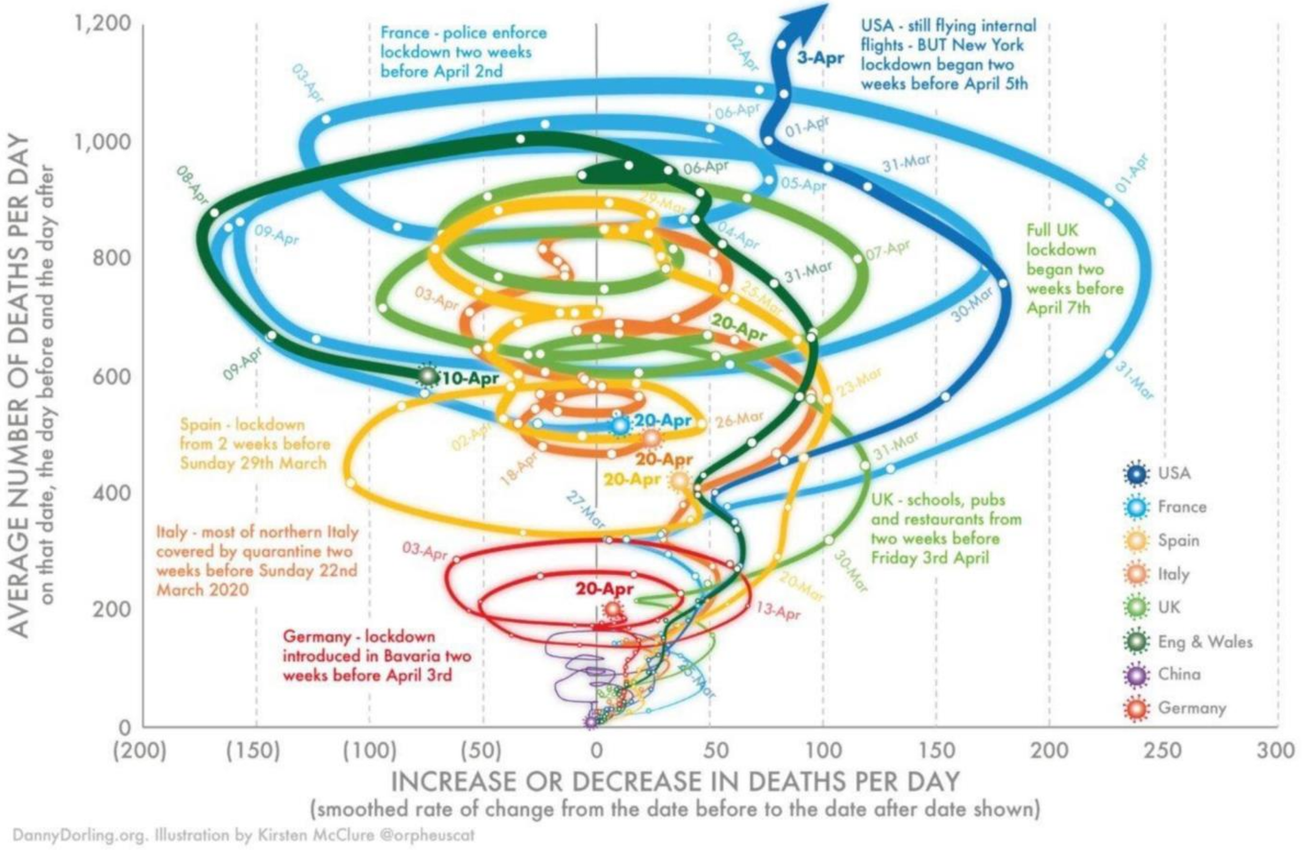
DATA VISUALIZATION

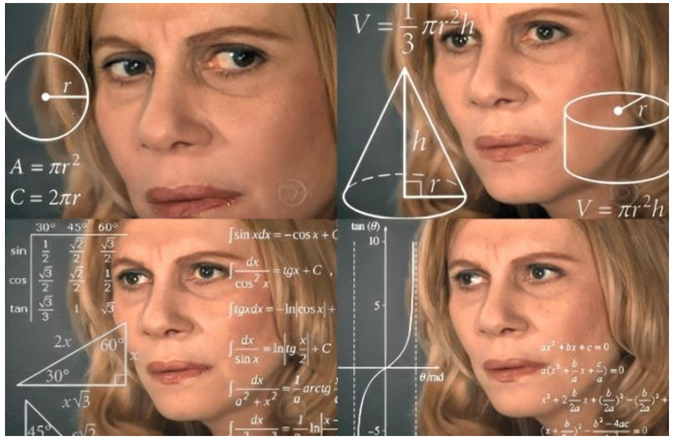
- Mistakes (maybe) to avoid: **Overloading** the chart with information



DATA VISUALIZATION

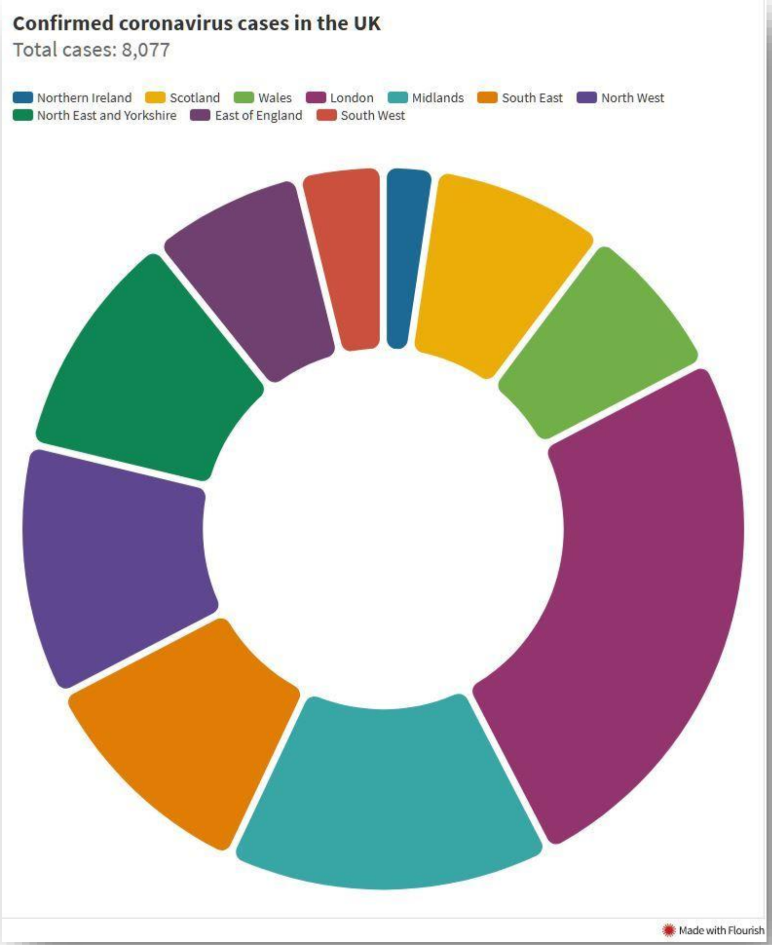
- Mistakes (maybe) to avoid: **Overloading** the chart with information





DATA VISUALIZATION

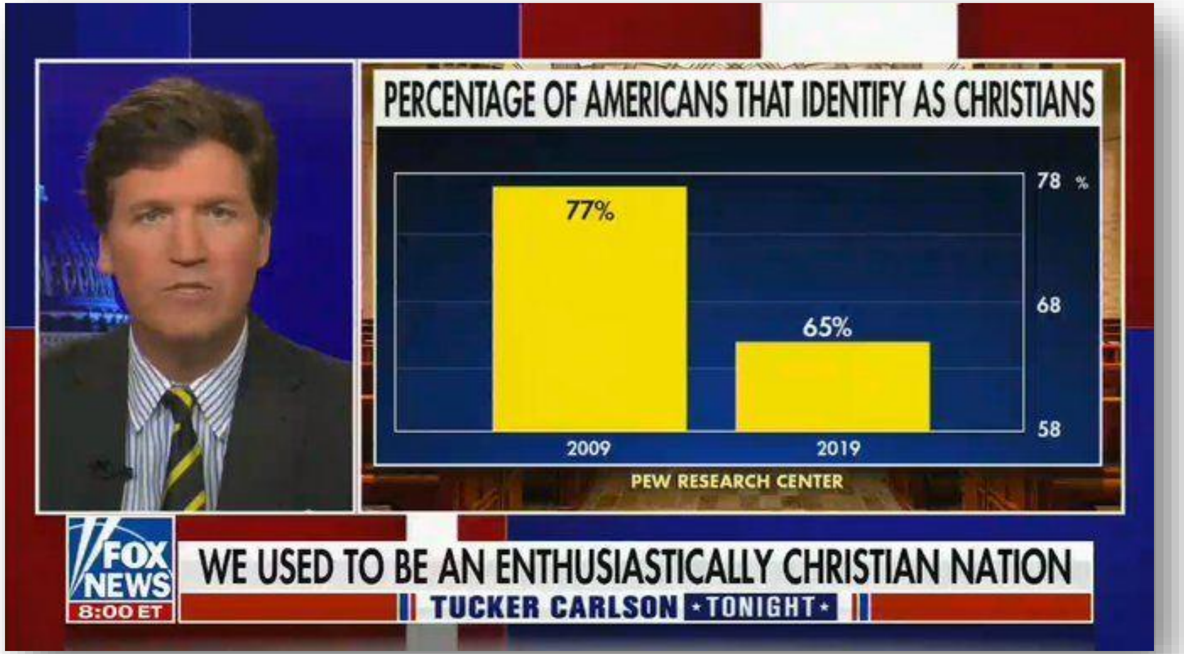
- Mistakes (maybe) to avoid: Presenting a chart with **insufficient** information

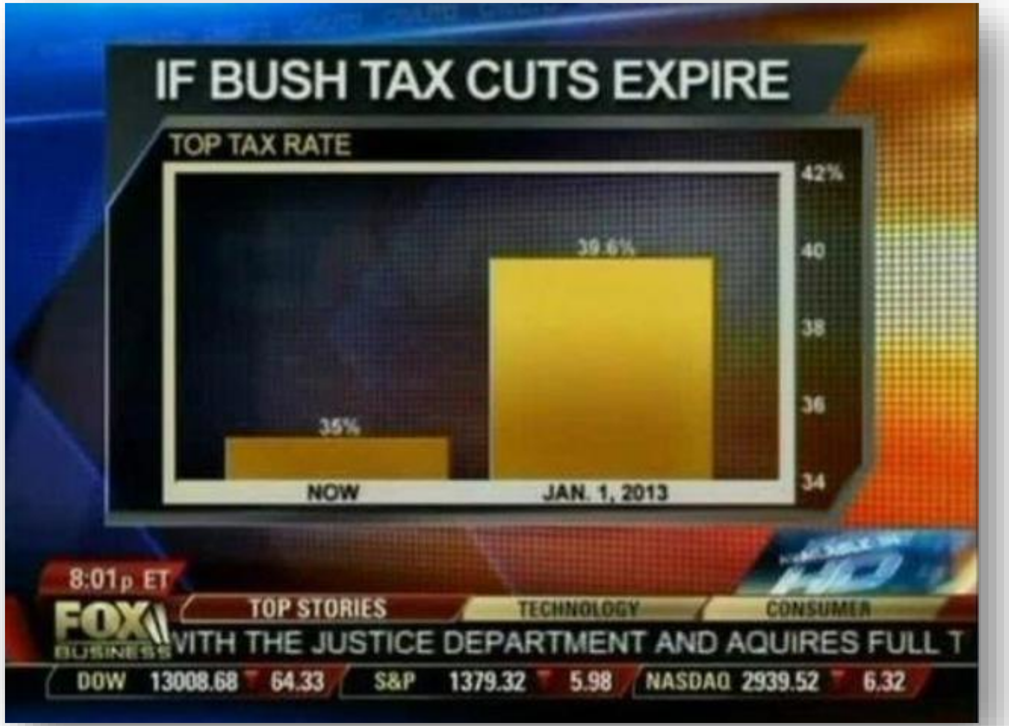


Source: Venggage. Bad infographics. https://venngage.com/blog/bad-infographics/

DATA VISUALIZATION

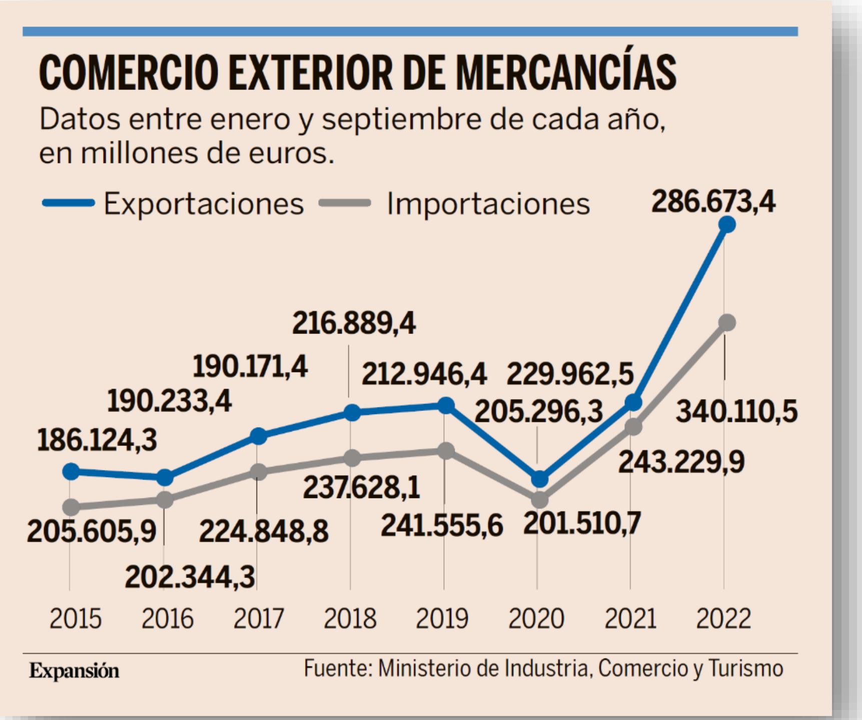
- Mistakes (maybe) to avoid: **Distortion** of axis and proportions (common)

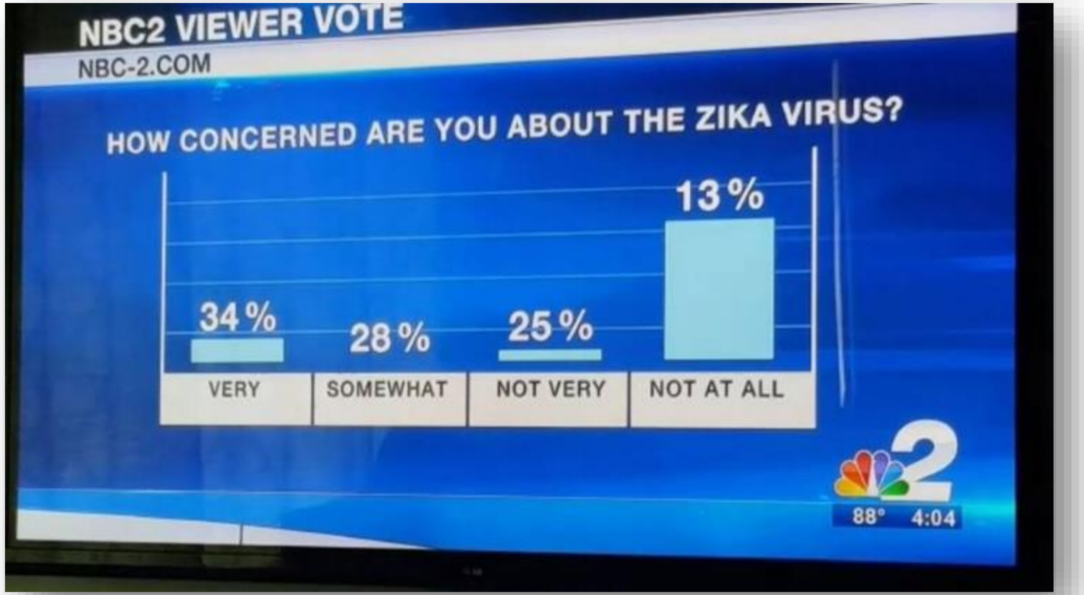




DATA VISUALIZATION

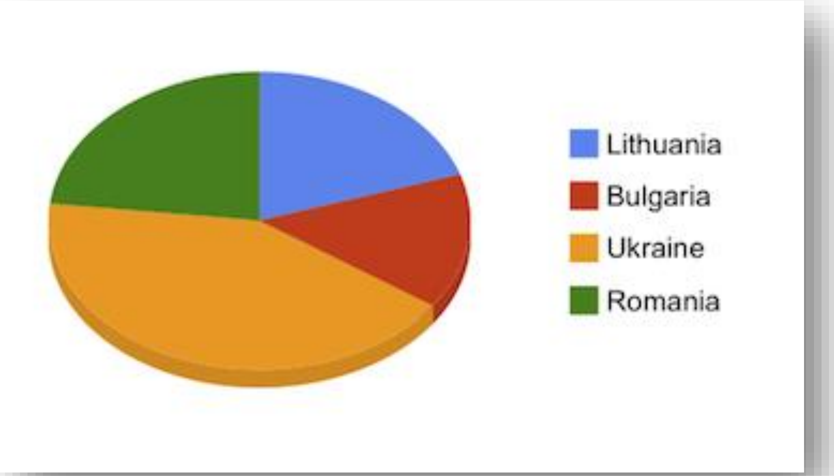
- Mistakes (maybe) to avoid: **Distortion** of relative positions and values



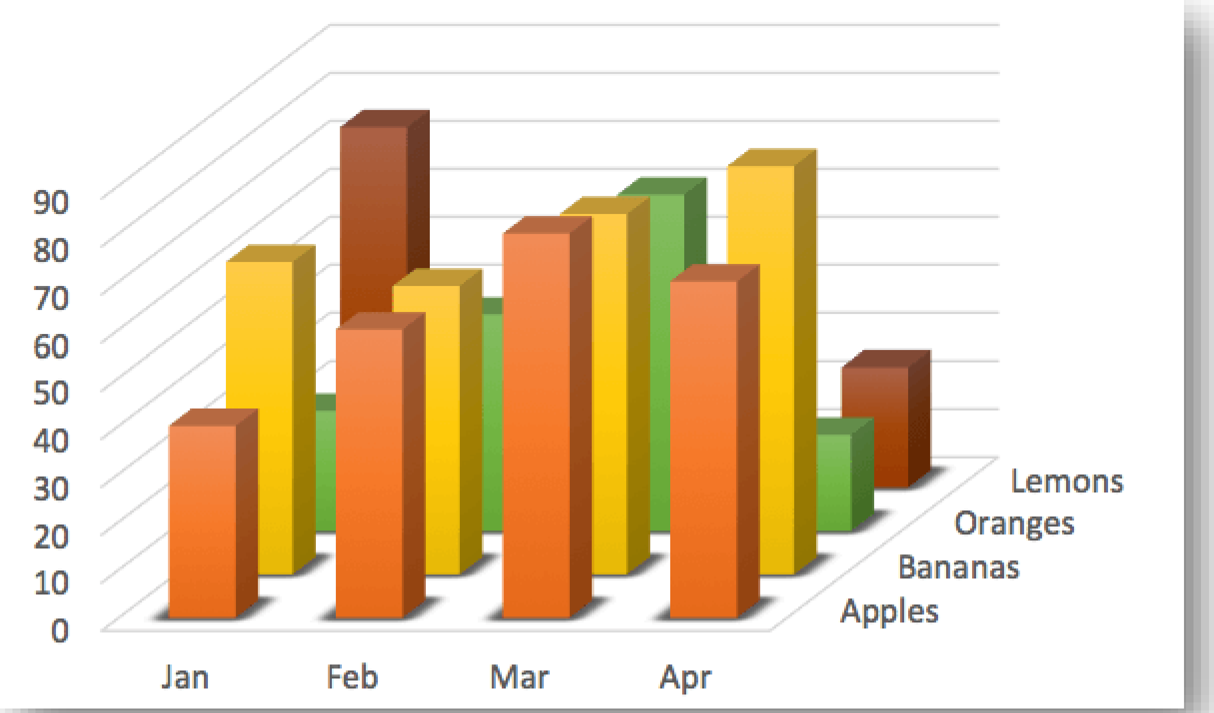


DATA VISUALIZATION

- Mistakes (maybe) to avoid: **Ambiguity** and unclearness

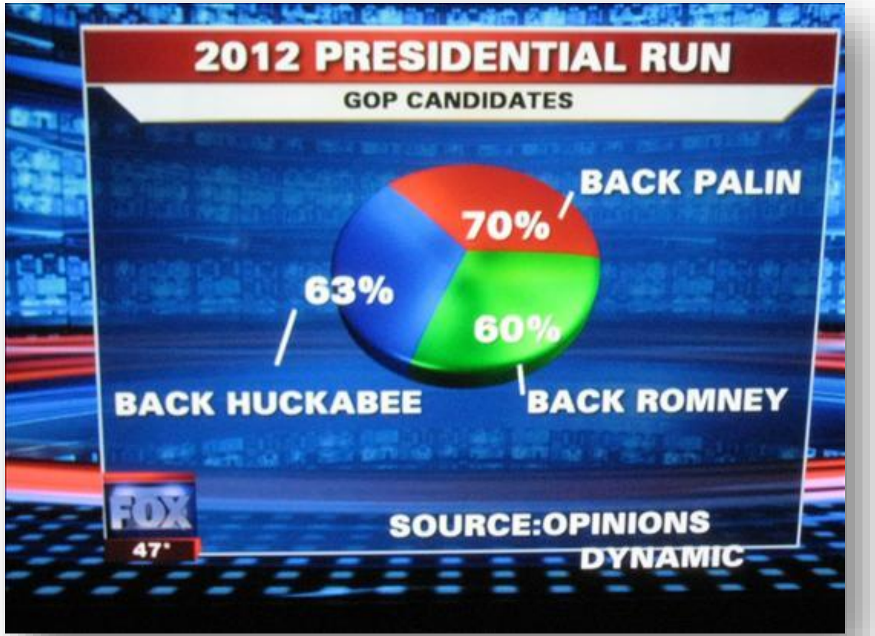


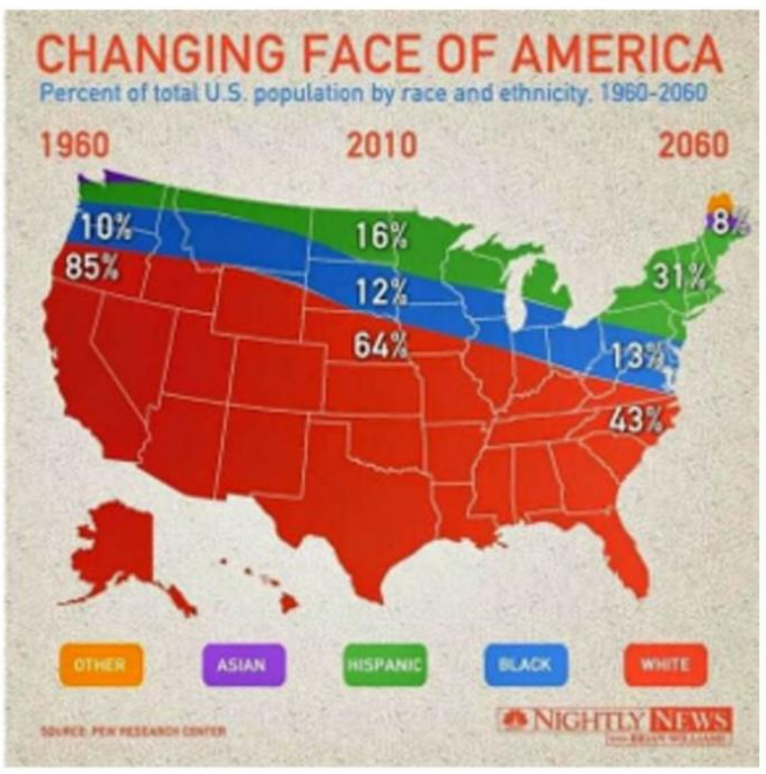
Los datos son (por orden):
398, 294, 840, 462



DATA VISUALIZATION

- Mistakes (maybe) to avoid: When 100% is **simply not enough** (or it is too much)





DATA VISUALIZATION

- Mistakes (maybe) to avoid: When you make all kinds of mistakes **simultaneously**

Via Reddit. https://www.reddit.com/r/chile/comments/fpe6j5/megavisi%C3%B3n_being_very_megavisi%C3%B3n/

DATA VISUALIZATION

- Mistakes (maybe) to avoid: When you make all kinds of mistakes **simultaneously**
  - There is **no Y axis**
  - The X axis is **not proportional**
  - **The Y axis isn't *either***
  - Values **make no sense** after a certain point
  - Worst part: The chart ***doesn't support at all* the thesis suggested**

Via Reddit. https://www.reddit.com/r/chile/comments/fpe6j5/megavisi%C3%B3n_being_very_megavisi%C3%B3n/

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **First question:** Let us compare the average time to arrive to scene
    - Which is the goal of our visualization?

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

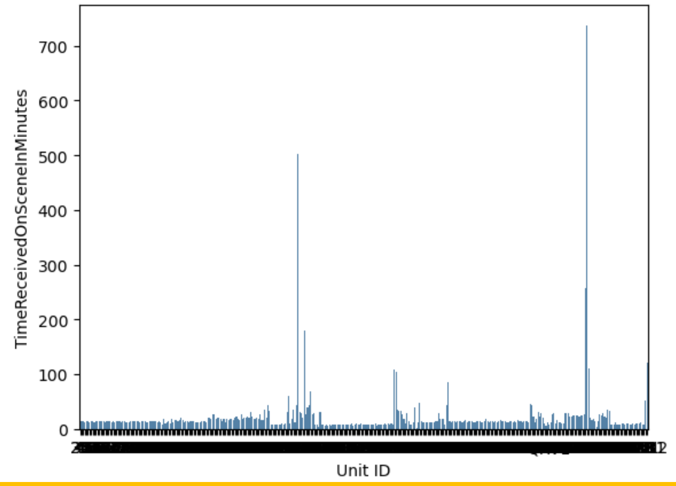
- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **First question:** Let us compare the average time to arrive to scene
    - Which is the goal of our visualization? To compare/rank
    - What charts can we use?

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **First question:** Let us compare the average time to arrive to scene
    - Which is the goal of our visualization? To compare/rank
    - What charts can we use? Bar chart/scatterplot (bubble graph)
    - Since we have only one categorical axis (Unit ID) and a value (Time) we will use the Bar Chart

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **First question:** Let us compare the average time to arrive to scene
    - Which is the goal of our visualization? To compare/rank
    - What charts can we use? Bar chart/scatterplot (bubble graph)
    - Since we have only one categorical axis (Unit ID) and a value (Time) we will use the Bar Chart

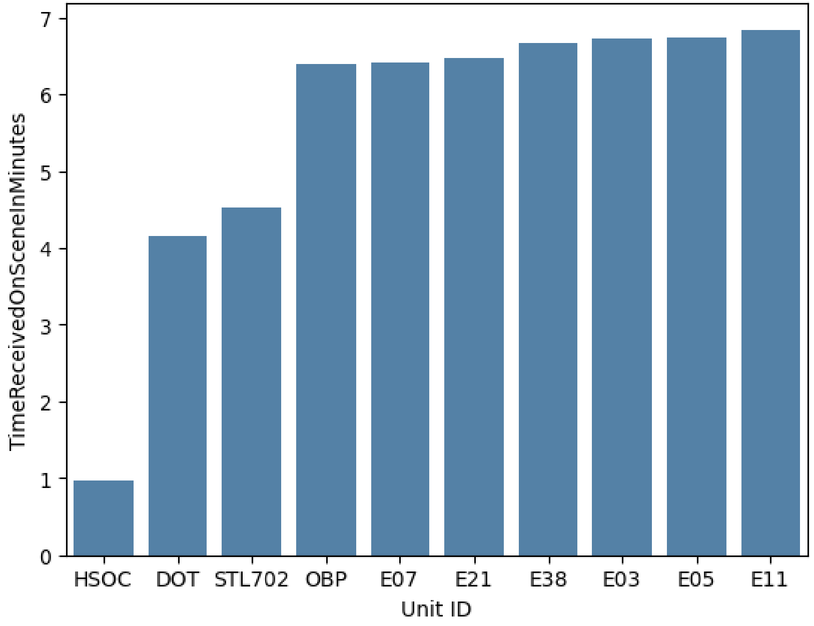


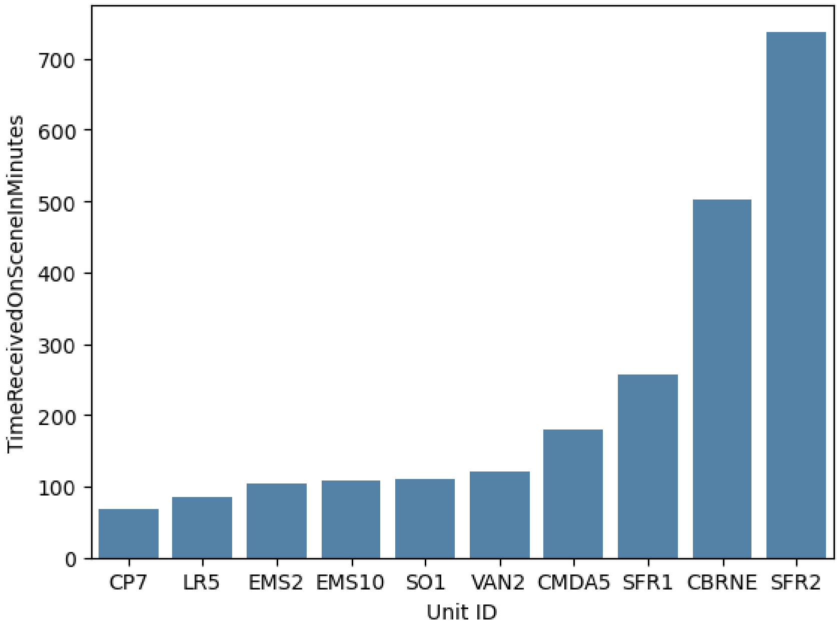
DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Problem! -** Too many values for detail (Volume). Alternatives:
    - Filter (Top/Bottom X or using a threshold value)
    - Group aided by hierarchies and dimensions (By Unit Type, By City, By Neighborhood or by Priority)
    - Take a different perspective on the data (e.g. analyze the distribution of mean times)

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

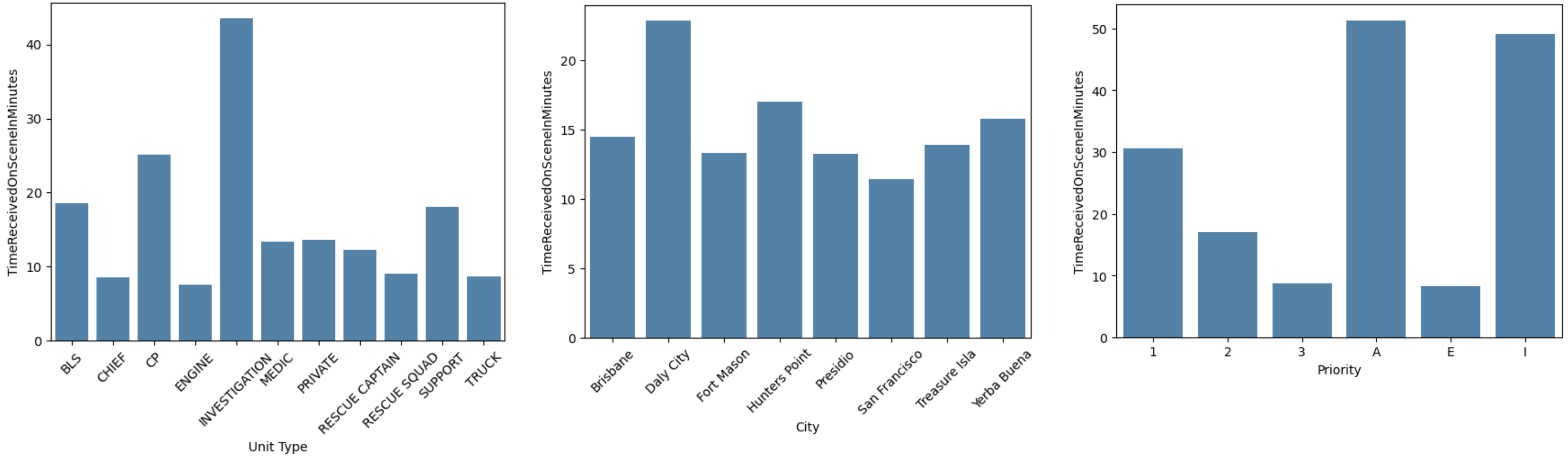
- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Problem! -** Too many values for detail (Volume). Alternatives:
    - Filter (Top/Bottom X or using a threshold value)





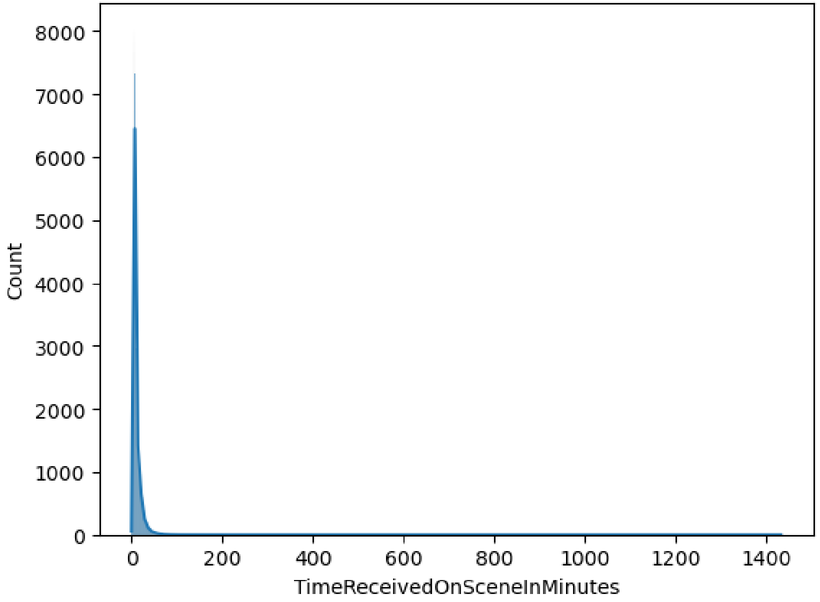
DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Problem! -** Too many values for detail (Volume). Alternatives:
    - Group aided by hierarchies and dimensions (By Unit Type, By City, By Neighborhood or by Priority)



DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Problem! -** Too many values for detail (Volume). Alternatives:
    - Take a different perspective on the data (e.g. analyze the distribution of mean times)



DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **First question:** Let us compare the average time to arrive to scene.
    - Conclusions so far:
      - There are **several outliers and extreme outliers** in the data
      - **Average** time is **around 12-13** minutes for most units, which is **also the most common time**
      - There are **clear differences depending on Unit Types**, with Investigation taking considerably longer
      - Cities present an average around **10 to 20 minutes**, with **Daly City being the slowest**
      - Units react fast to Priority 1 and Emergencies. **Priority A and I are the slowest to react to**

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

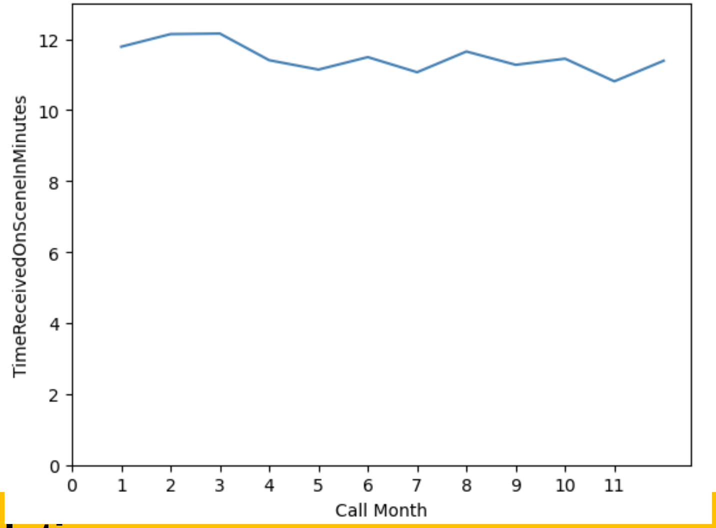
- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Second question:** Does the average time to arrive change across the year?

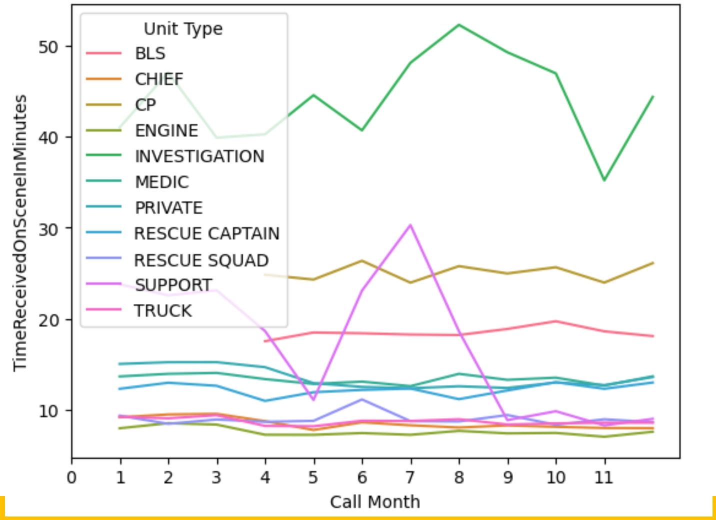
DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Second question:** Does the average time to arrive change across the year?
    - Which is the goal of our visualization? To analyze evolution across time
    - What charts can we use? Line chart
    - Consideration: Overall average time for all units or Average time per unit/unit type?

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Second question:** Does the average time to arrive change across the year?
    - Which is the goal of our visualization? To analyze evolution across time
    - What charts can we use? Line chart
    - Consideration: Overall average time for all units or Average time per unit/unit type?





DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Second question:** Does the average time to arrive change across the year?
    - Conclusions so far:
      - For some reason units **have become quicker** as months have passed
      - This is **mainly related to Support units** as other units remain relatively stable

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

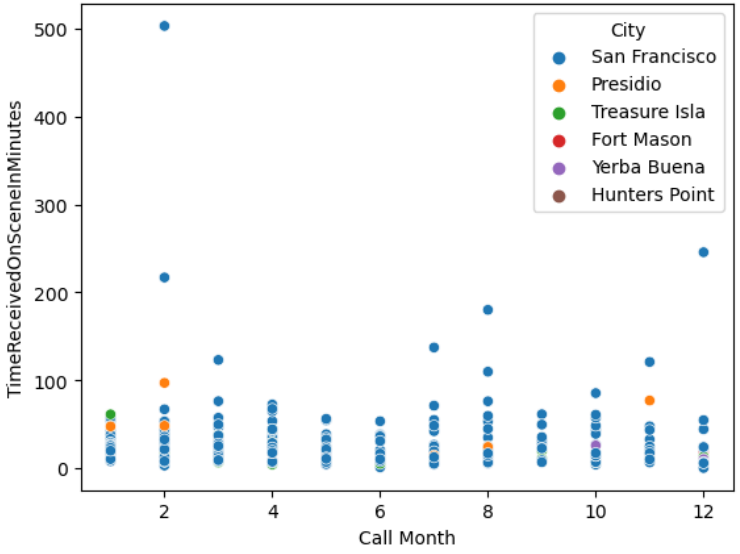
- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Third question:** Have support units become quicker in all cities?

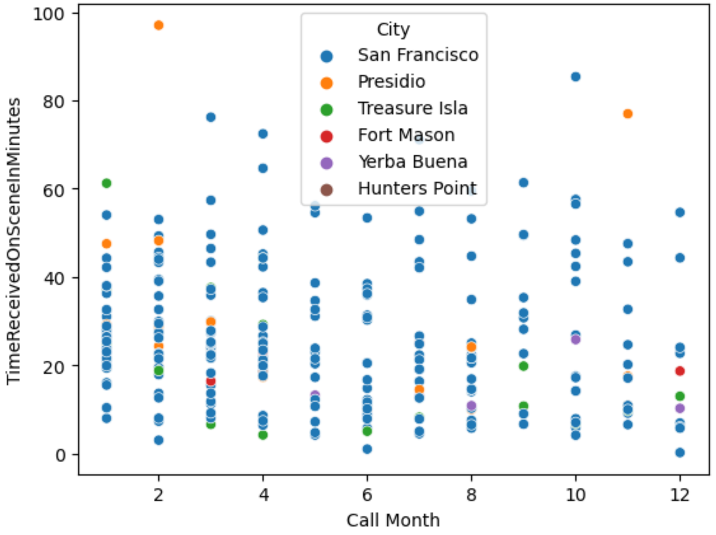
DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Third question:** Have support units become quicker in all cities?
    - Which is the goal of our visualization? To compare unit response time across cities and months
    - What charts can we use? Bar chart/scatterplot (bubble graph)
    - Since we have two categorical axis (Unit ID, City) and several values (response time, month) we can use Bubble graph to support multiple dimensions
    - **Warning!:** This is a heavily-loaded chart, we probably can interpret it because we have been digging into the data. A normal user probably would have a hard time following it

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Third question:** Have support units become quicker in all cities?
    - Configuration:
      - Categorical axis must go into Bubble ID/Bubble Color → Row(Unit ID, rest is covered in other axis)/City
      - Numeric axis can use X/Y axis (first), then bubble size → Months/Time to arrive



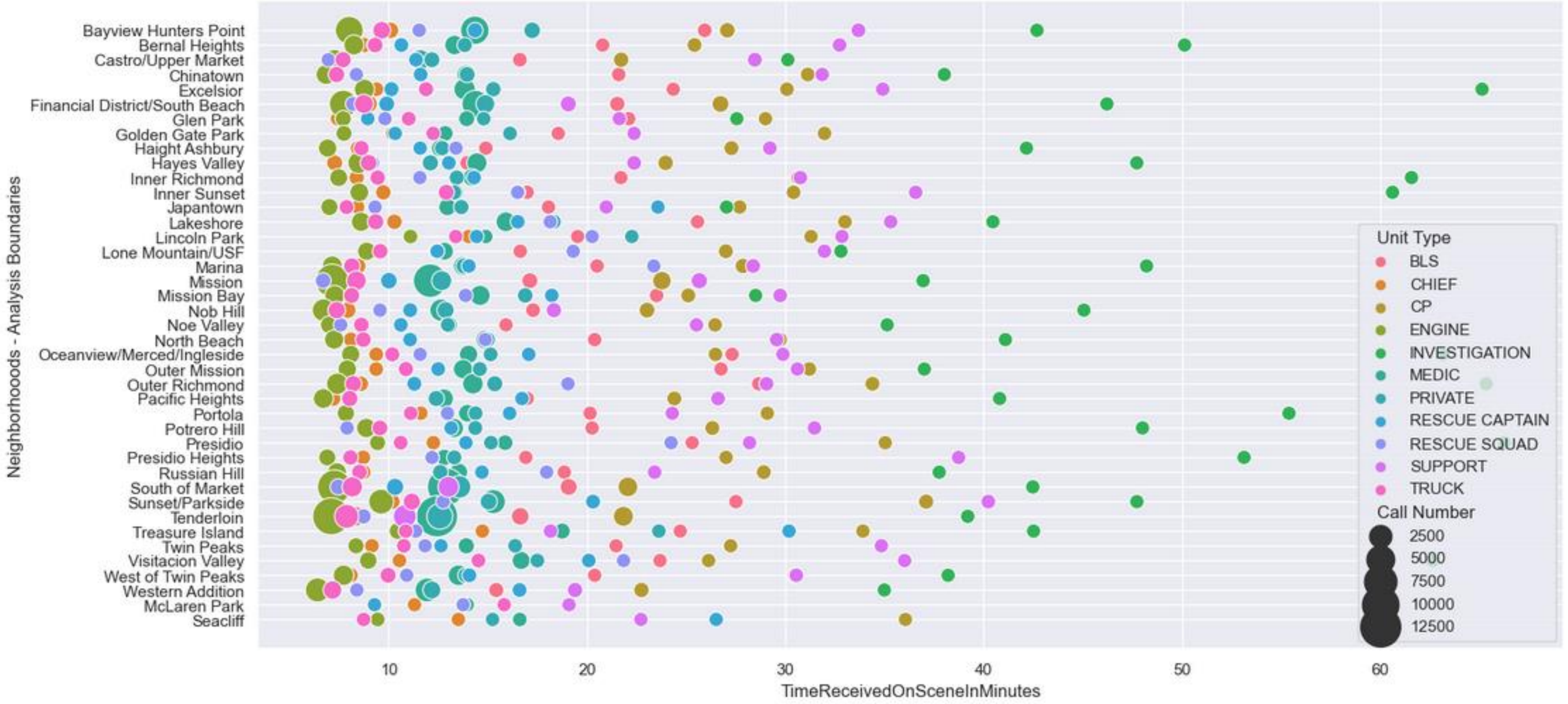


DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- Using Jupyter and Seaborn let us carry out the EDA step from the previous lesson:
  - **Third question:** Have support units become quicker in all cities?
    - Conclusions:
      - Outliers aside all cities seem to **follow more or less the same trend**
      - Daly City **does not appear** → Has no support units

DATA ANALYTICS PROCESS: DATA VISUALIZATION – Classroom exercise

- To finish, bubble graph can also be used to find patterns using bubble size:



Master en
Big Data e Inteligencia Artificial

Big Data Analytics

Alejandro Maté
Juan Carlos Trujillo