

Data Visualization

Practices and examples

Luis Moneda, Data Scientist at Nubank

About me

Academic

- MSc in Computer Science student (IME-USP)
- Bachelor in Computer Engineering (Poli-USP)
- Bachelor in Economics (FEA-USP)

Work & activities

- Data Scientist at Nubank (2017 - Current)
- Teaching Machine Learning for MBA courses at FIA
- Udacity mentor and project reviewer for data related courses
- Nubank Machine Learning meetup organizer
- Kaggle (competitions and datasets)
- Twitter and Blog: @lgmoneda and lgmoneda.github.io

Outline

1. Goal
2. Problems & Practices
3. Types
4. Data Science
 - a. EDA
 - b. PCA
 - c. t-SNE
5. Tools
6. Example
7. Resources

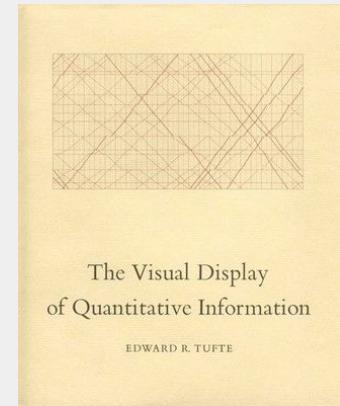
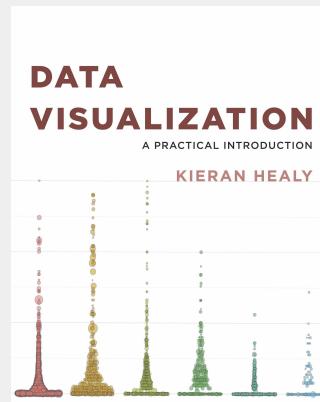
Resources and References

Books

- The visual display of quantitative information. Edward R. Tufte.
- Visual Explanations: Images and Quantities, Evidence and Narrative. Edward R. Tufte.
- Data Visualization, A practical Introduction. Kieran Healey (<http://socviz.co/>)

Links

- Pandas for plotting:
https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
- Visual Vocabulary (types of plots and what they are good for):
<https://journalismcourses.org/courses/DE0618/Visual-vocabulary.pdf>
- Your Friendly Guide to Colors in Data Visualisation:
<https://blog.datawrapper.de/colorguide/>
- The Python Graph Gallery: <https://python-graph-gallery.com/>
- Fundamentals of Data Visualization:
<https://serialmentor.com/dataviz/introduction.html>

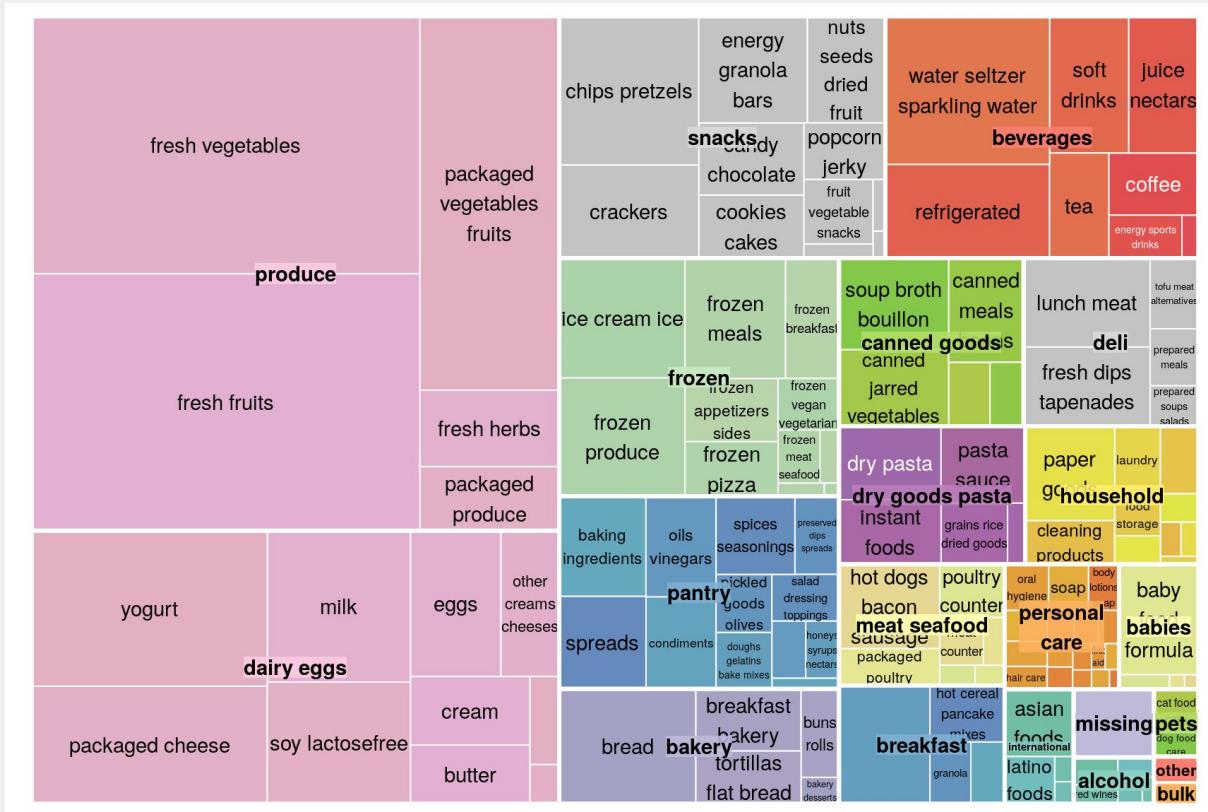


Goal: efficient communication

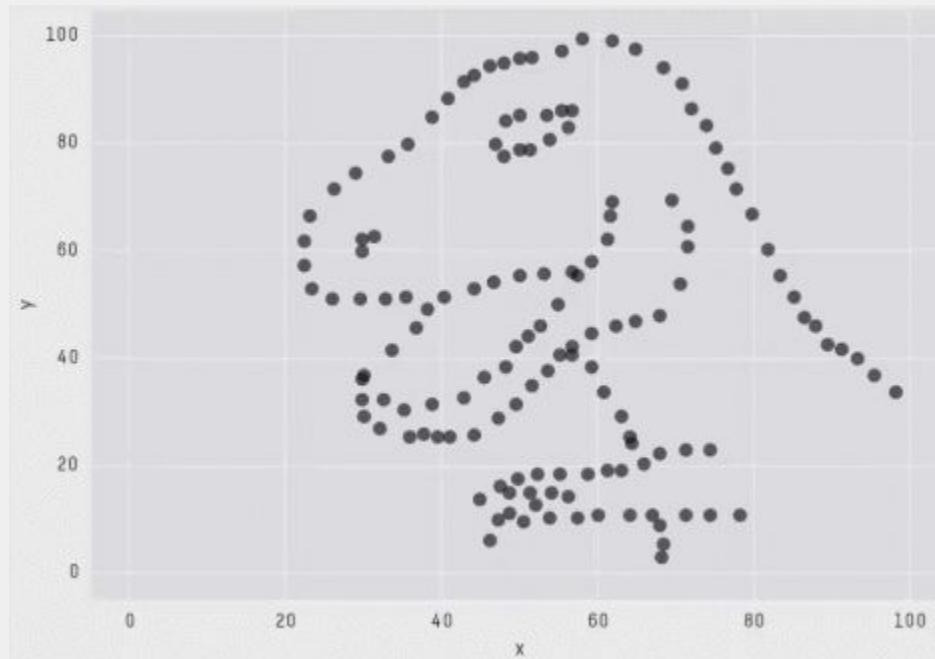
Goal

product_id	proportion_reordered	n	product_name	aisle_id	department_id
1729	0.9347826	92	2% Lactose Free Milk	84	16
20940	0.9130435	368	Organic Low Fat Milk	84	16
12193	0.8983051	59	100% Florida Orange Juice	98	7
21038	0.8888889	81	Organic Spelt Tortillas	128	3
31764	0.8888889	45	Original Sparkling Seltzer Water Cans	115	7
24852	0.8841717	18726	Banana	24	4
117	0.8833333	120	Petit Suisse Fruit	2	16
39180	0.8819876	483	Organic Lowfat 1% Milk	84	16
12384	0.8810409	269	Organic Lactose Free 1% Lowfat Milk	91	16
24024	0.8785249	461	1% Lowfat Milk	84	16

Goal



Goal



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

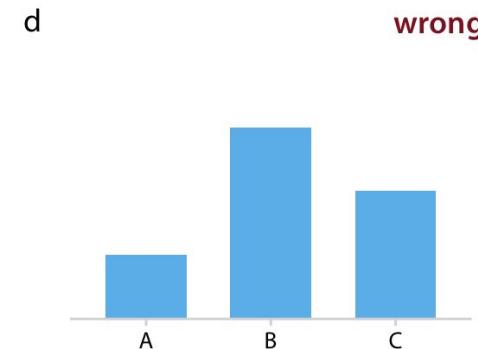
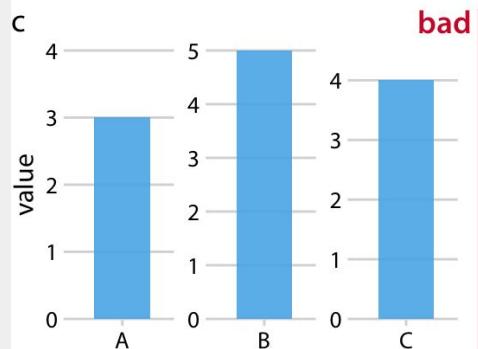
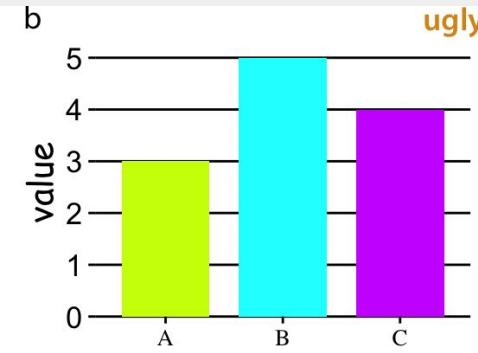
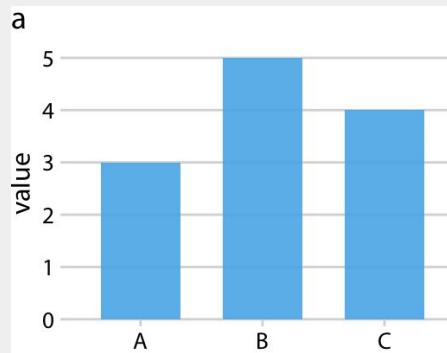
Problems & Practices

Problems & Practices

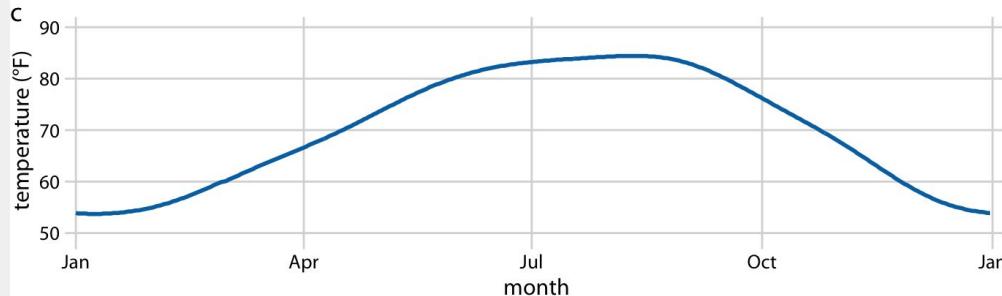
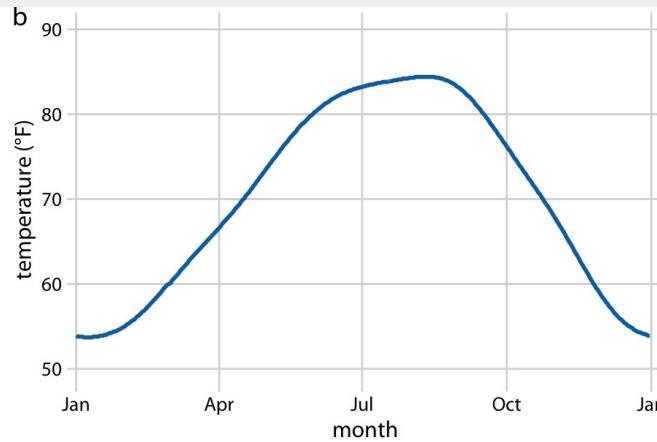
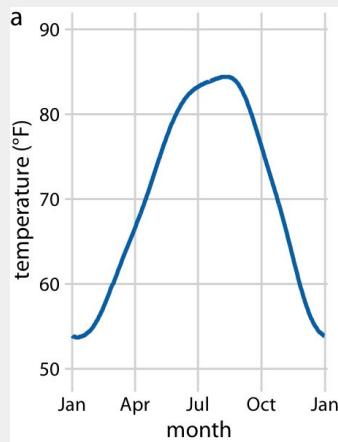
Problems	Practices
<ul style="list-style-type: none">- Bad taste- Bad data- Bad perception	<ul style="list-style-type: none">- Labeling- Plot Design- Context- Honest- Self sufficiency (in terms of data!)- Right plot type

The more complex is the idea you want to communicate, more successful you would need to be in "clarity, precision, and efficiency".

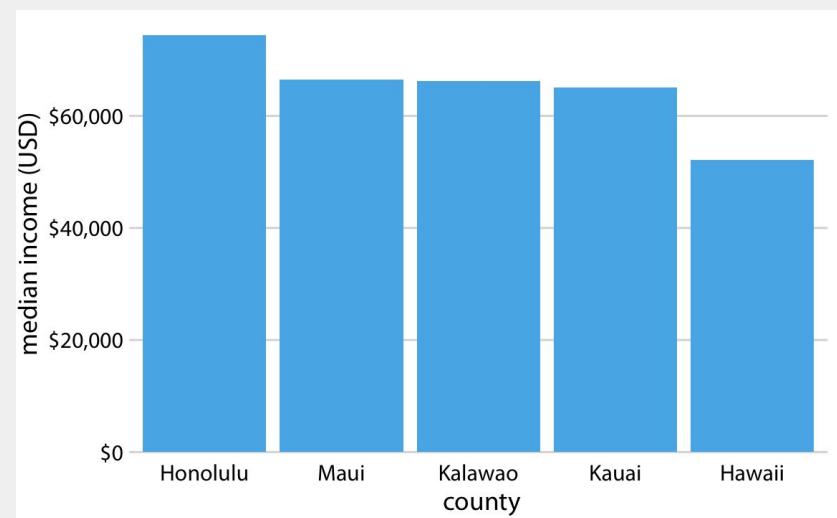
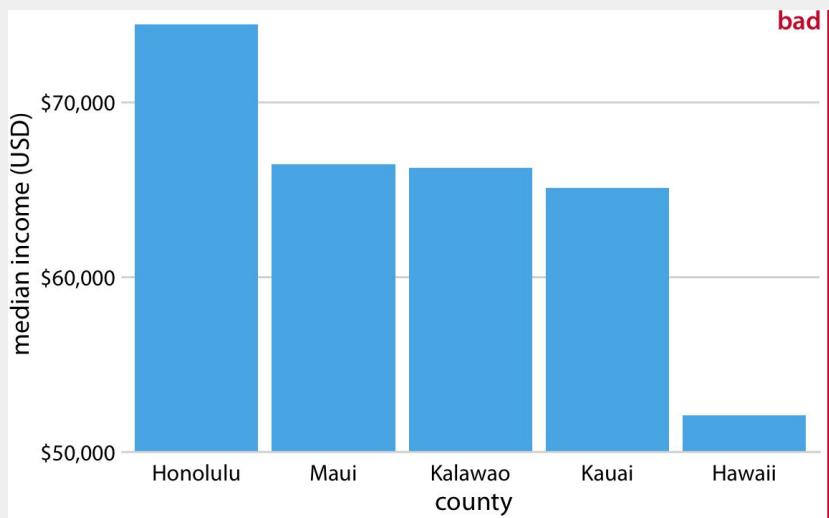
Problems & Practices



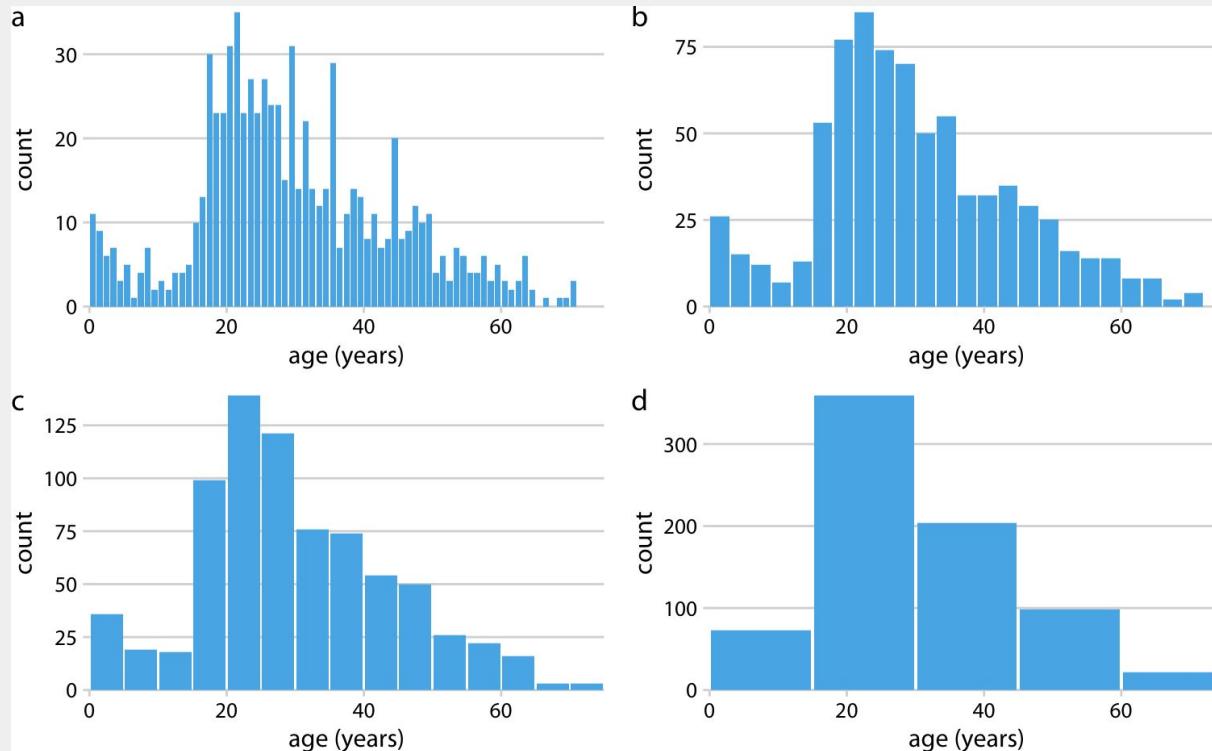
Problems - Scale



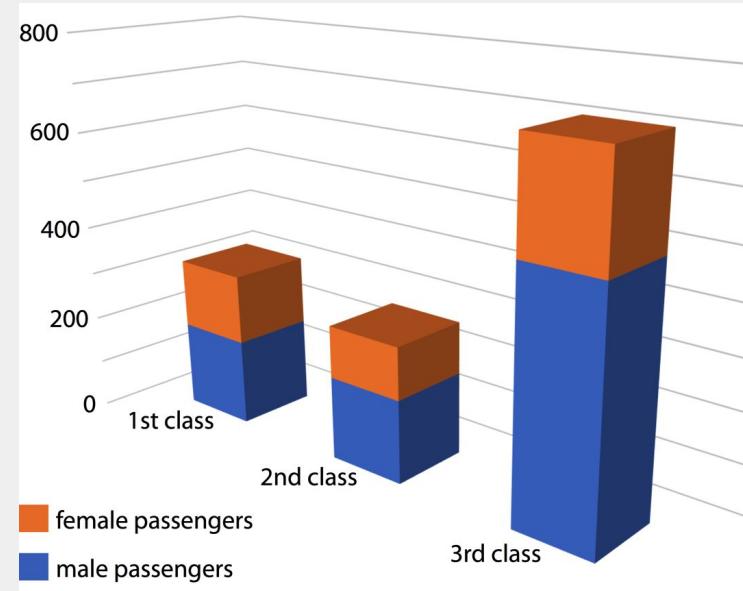
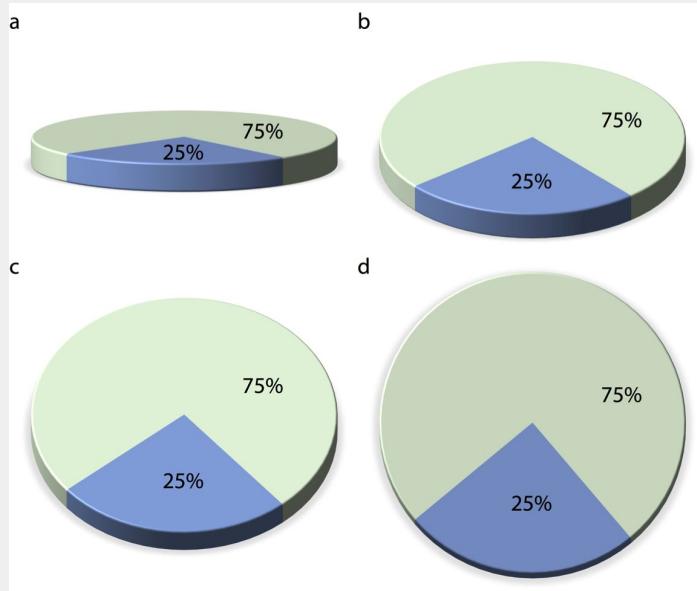
Problems - Scale



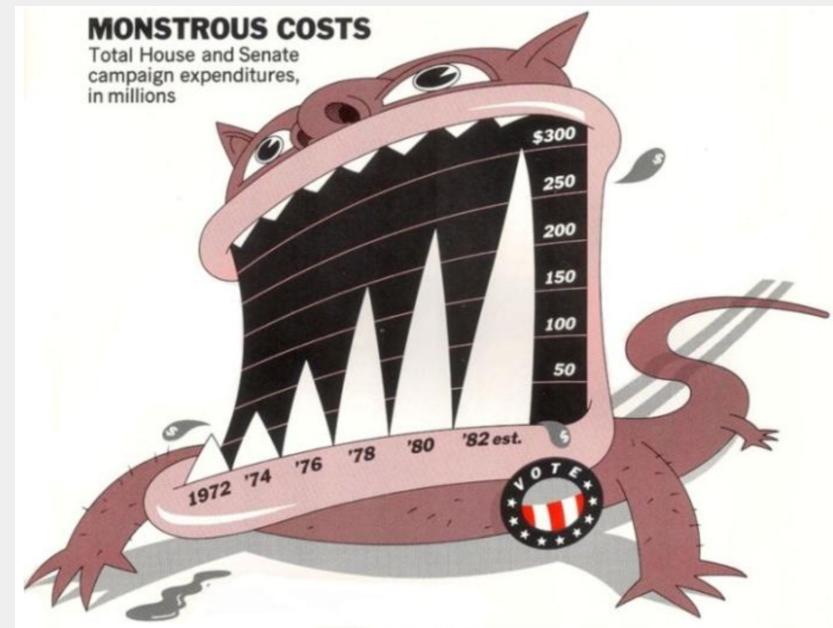
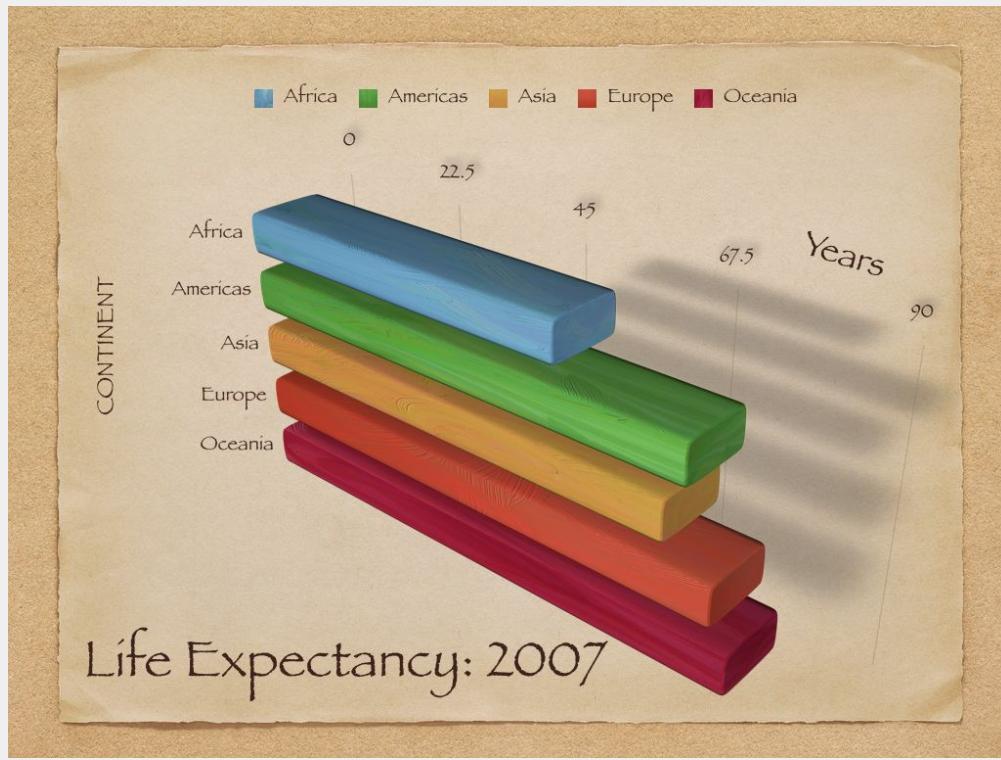
Problems - Distribution



Problems - 3D



Problems - Combining



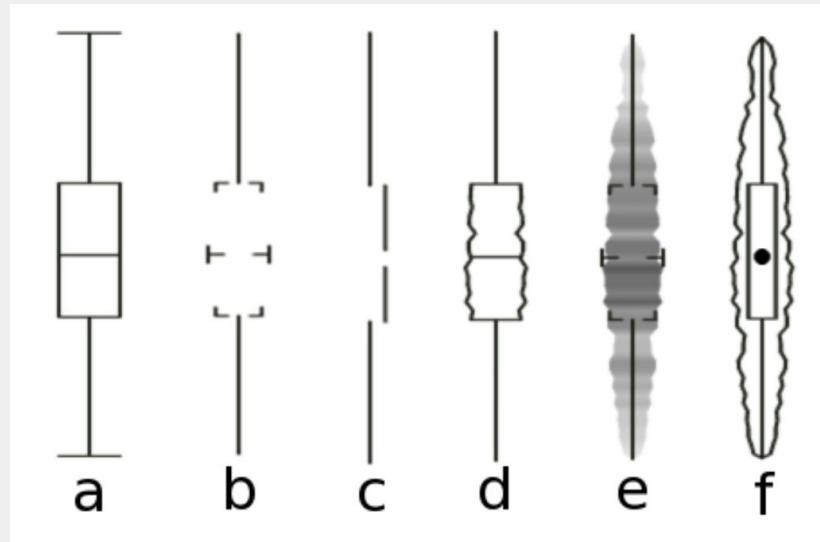
ideias

Arrangements of points and lines on a page can encourage us—sometimes quite unconsciously—to make inferences about similarities, clustering, distinctions, and causal relationships that might or might not be there in the numbers. Sometimes these perceptual tendencies can be honestly harnessed to make our graphics more effective. At other times, they will tend to lead us astray, and must take care not to lean on them too much.

In short, good visualization methods offer extremely valuable tools that we should use in the process of exploring, understanding, and explaining data. But they are not some sort of magical means of seeing the world as it really is. They will not stop you from trying to fool other people if that is what you want to do; and they may not stop you from fooling yourself, either.

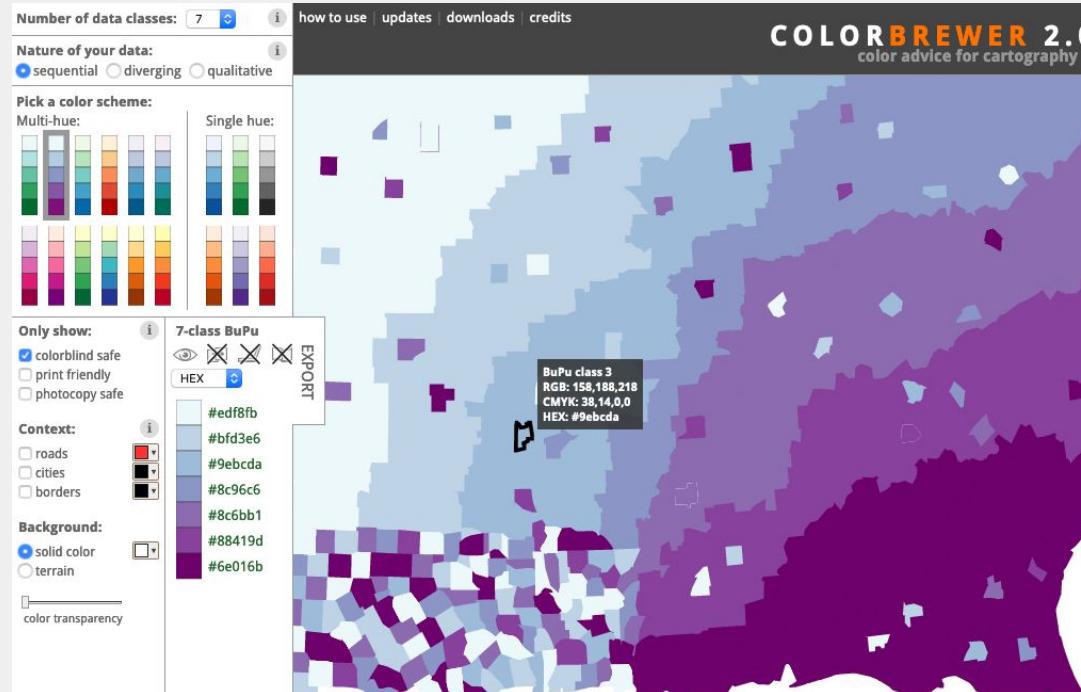
We will not automatically get the right answer to our questions just by looking.

Principle: Maximize data-to-ink ratio



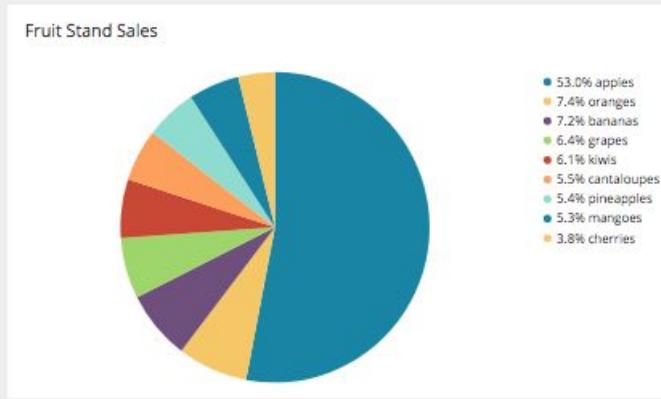
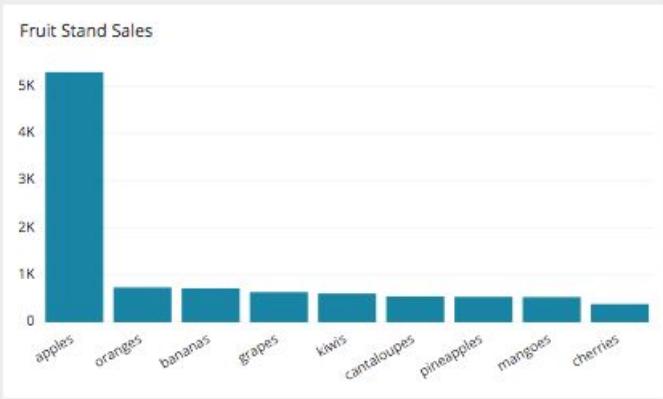
Six Kinds of summary boxplots.
Type (c) if rom Tufte

Principle: Colors



Examples

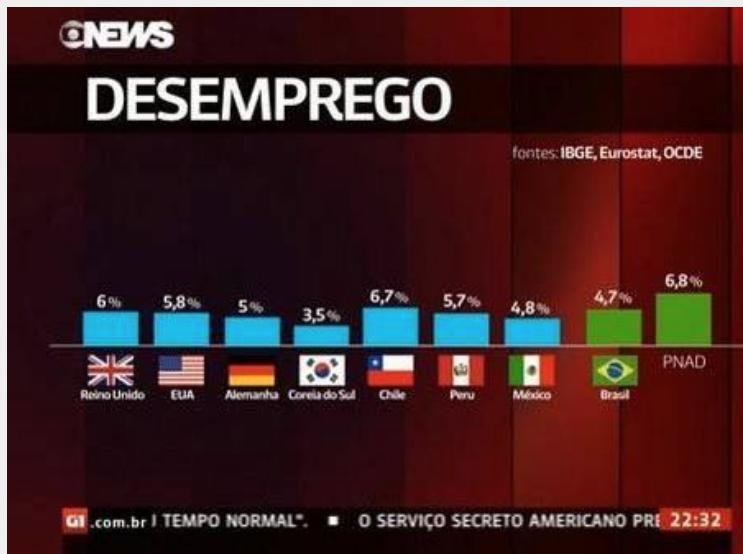
Bad examples



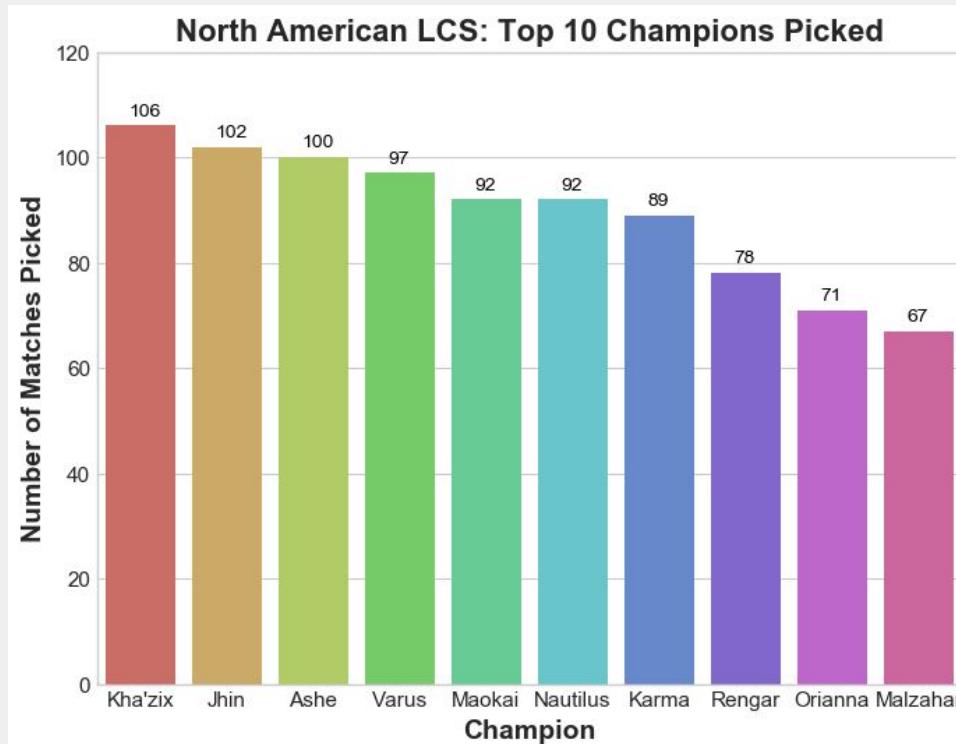
Bad examples



Bad examples

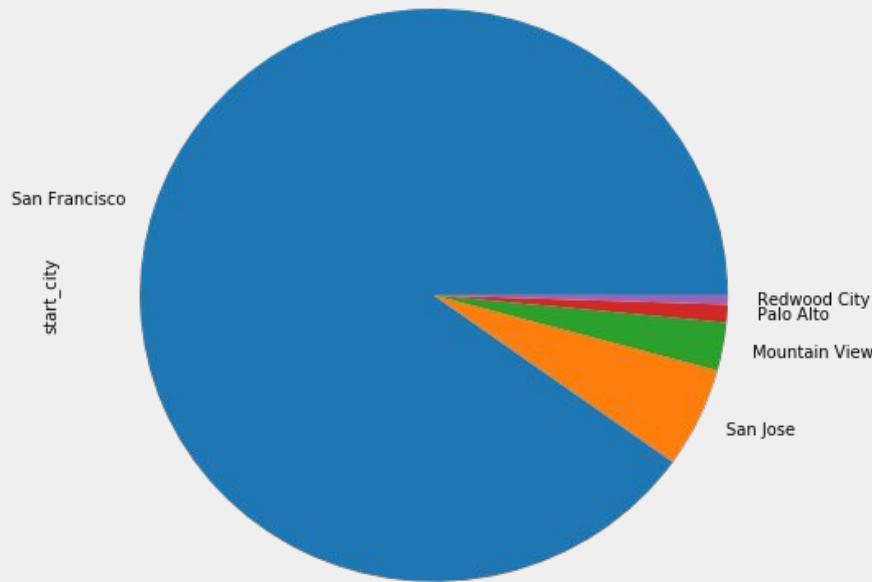


Bad examples



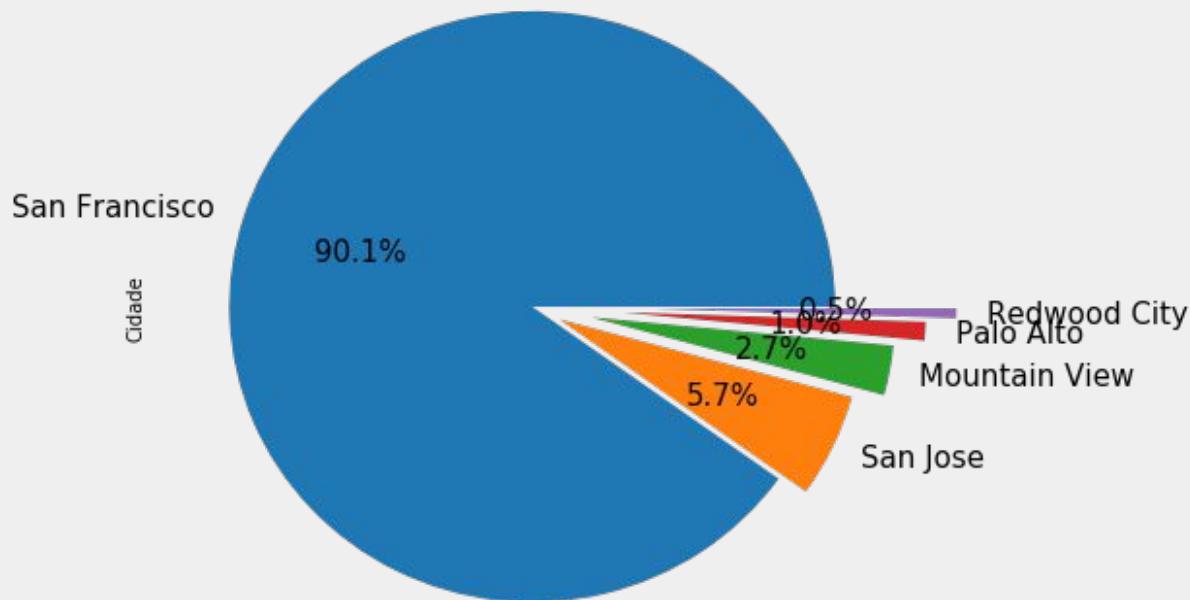
Does it really happen?

Examples

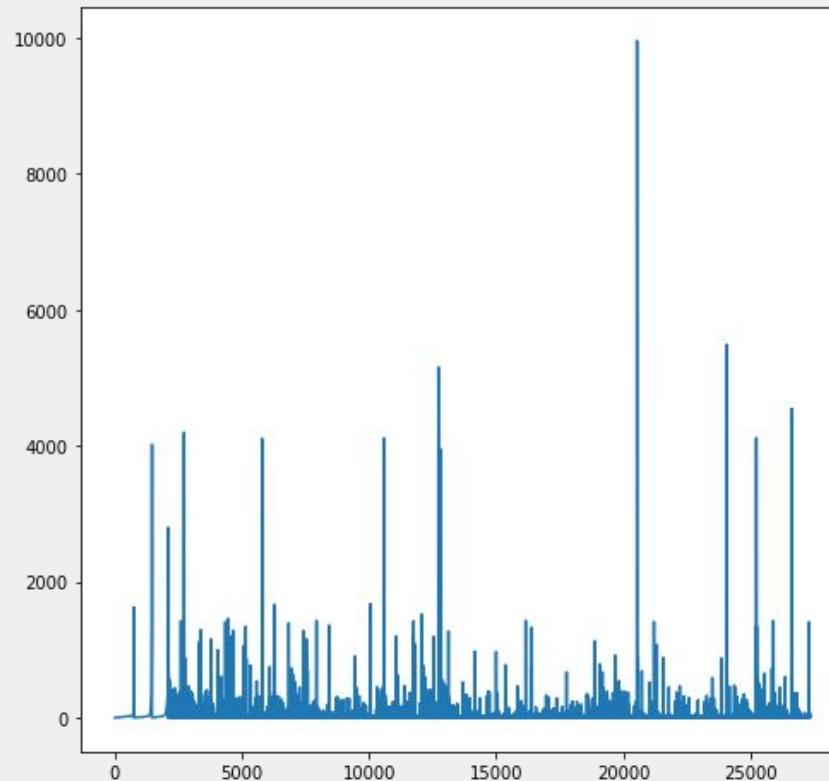


Examples

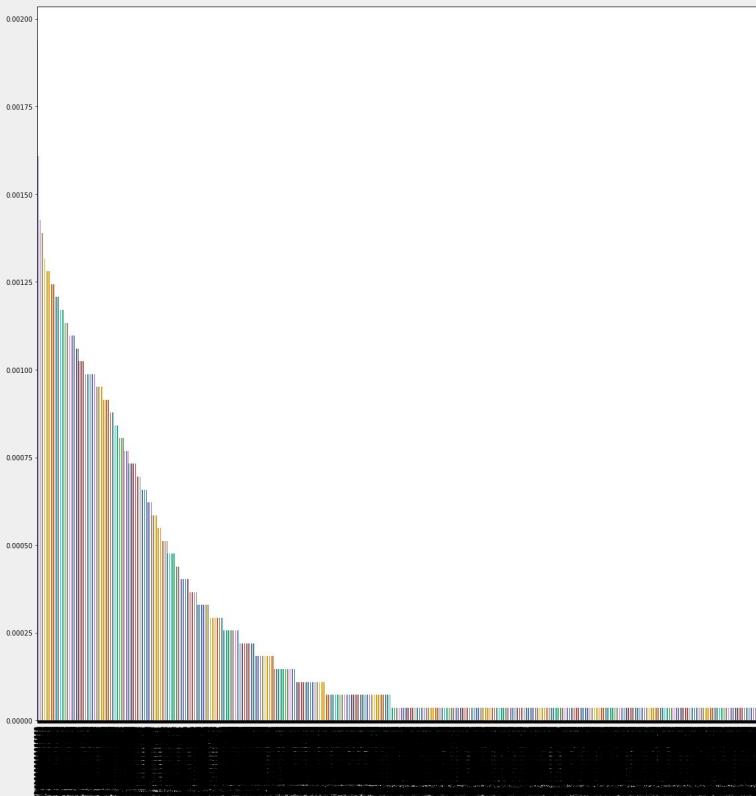
Estação inicial Mais Utilizada por Cidade (%)



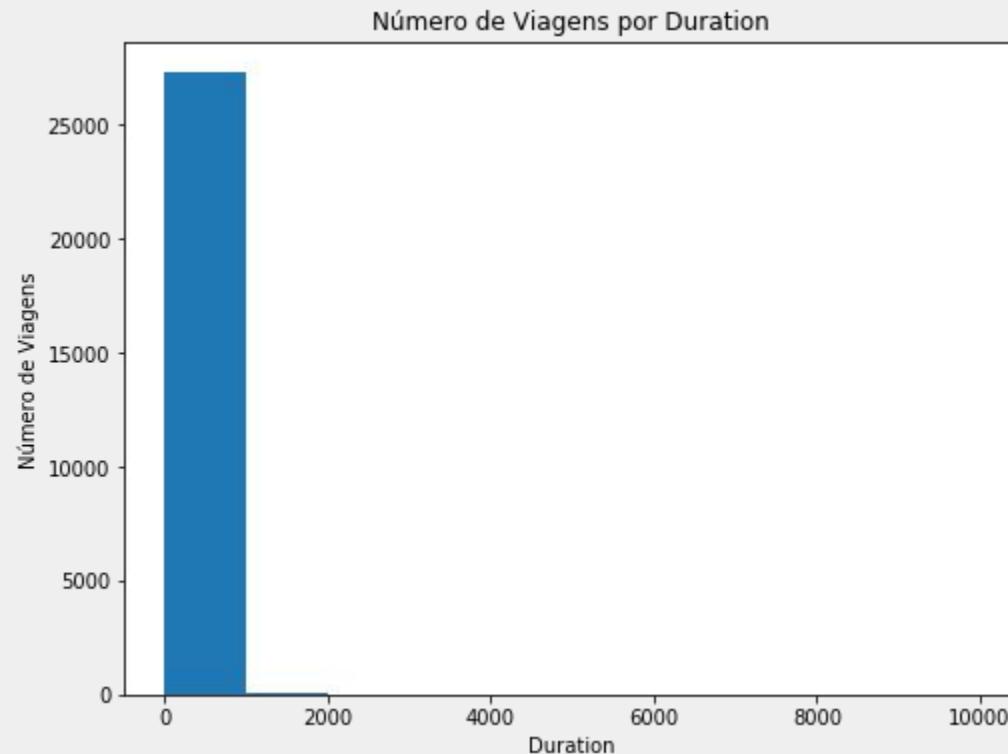
Examples



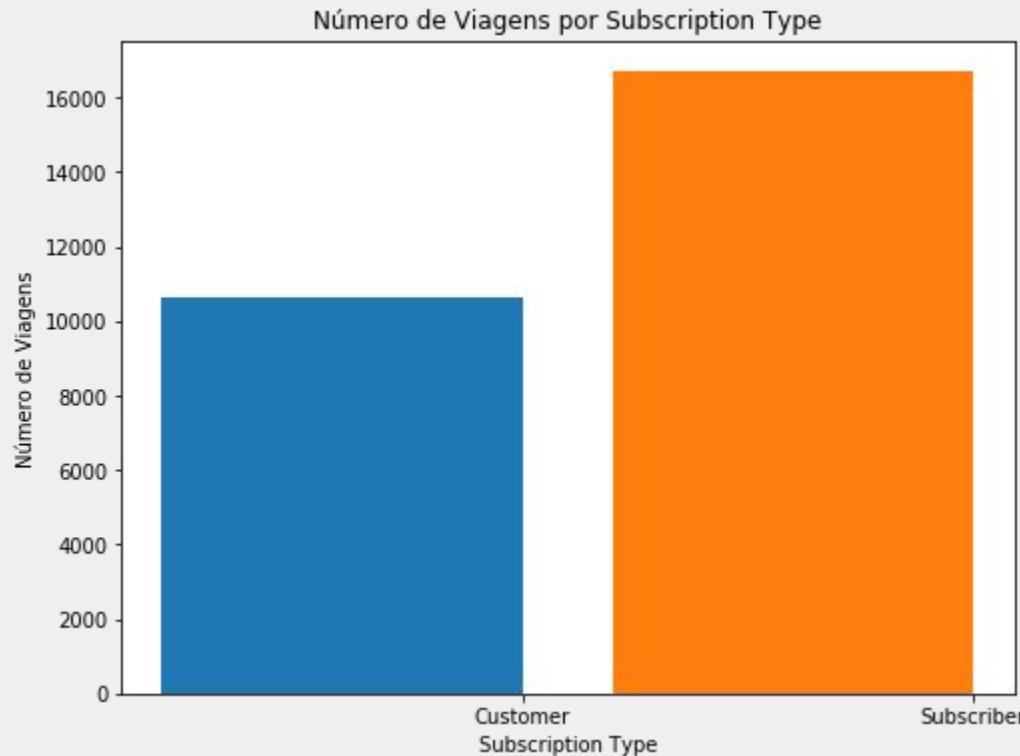
Examples



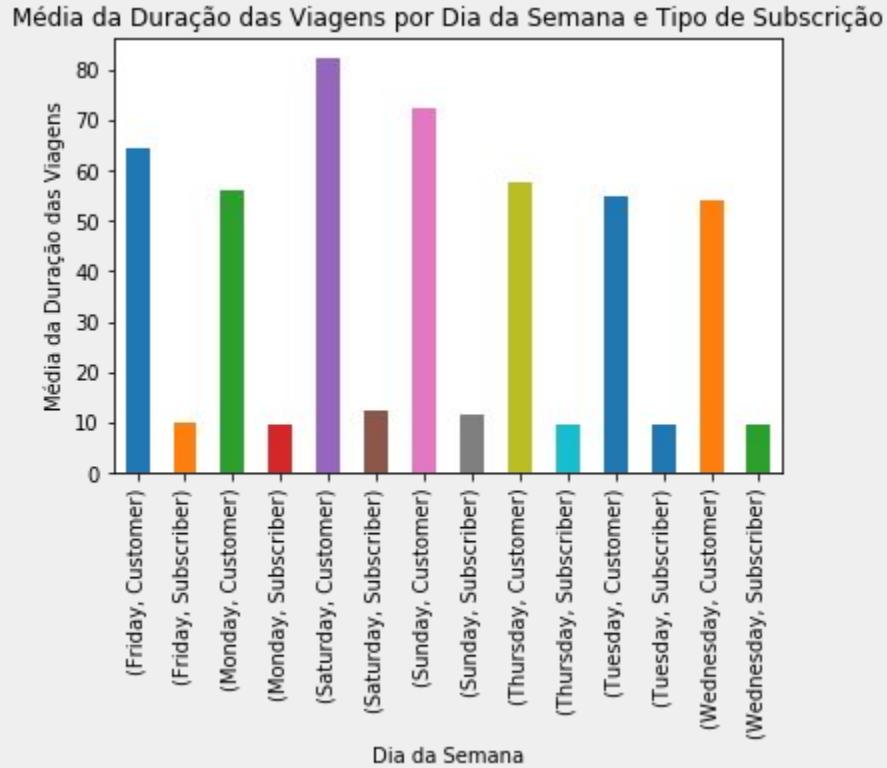
Examples



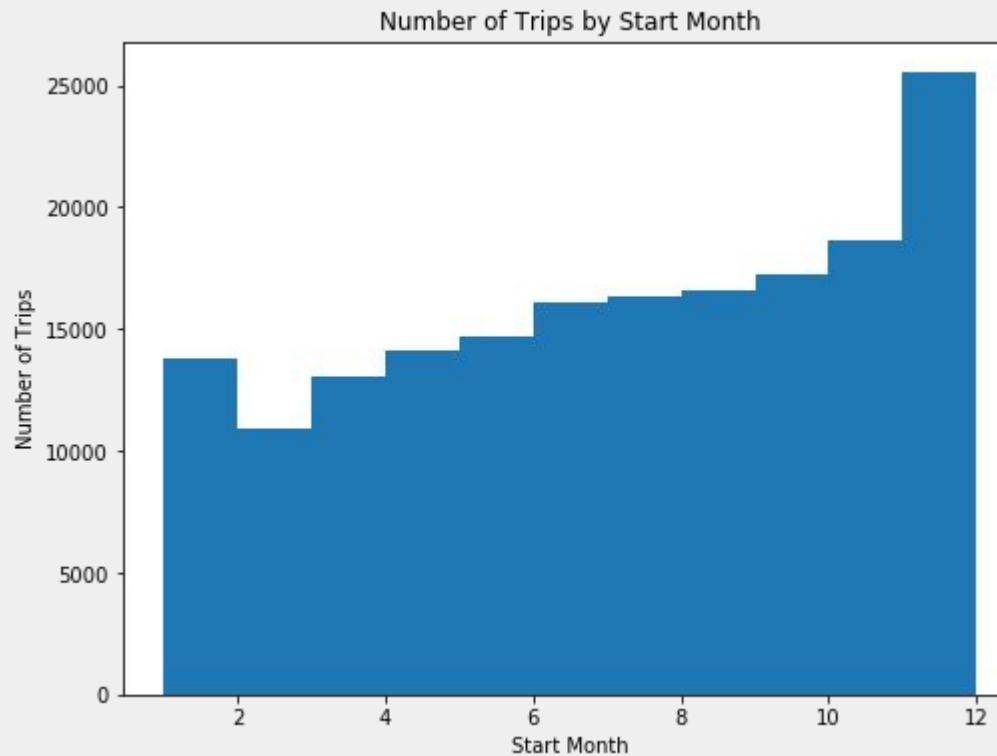
Examples



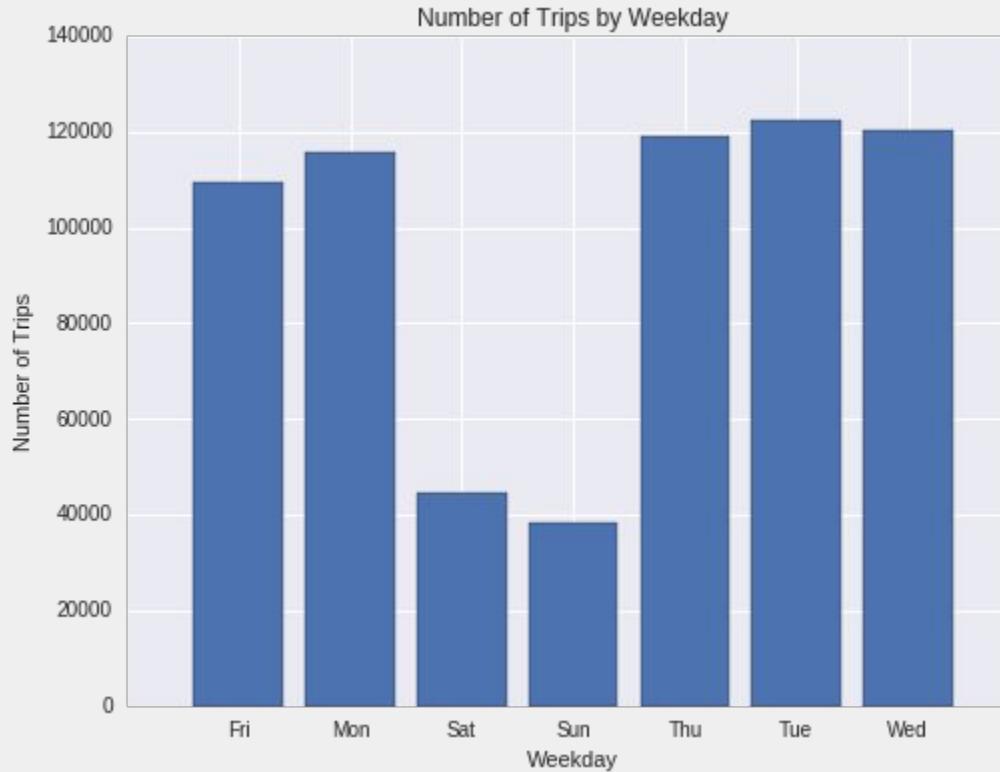
Examples



Examples



Examples



Examples



Good examples!

Examples

Figure 3: Daily and weekly habits of annual subscribers

Figure 3.A: Number of trips by hour

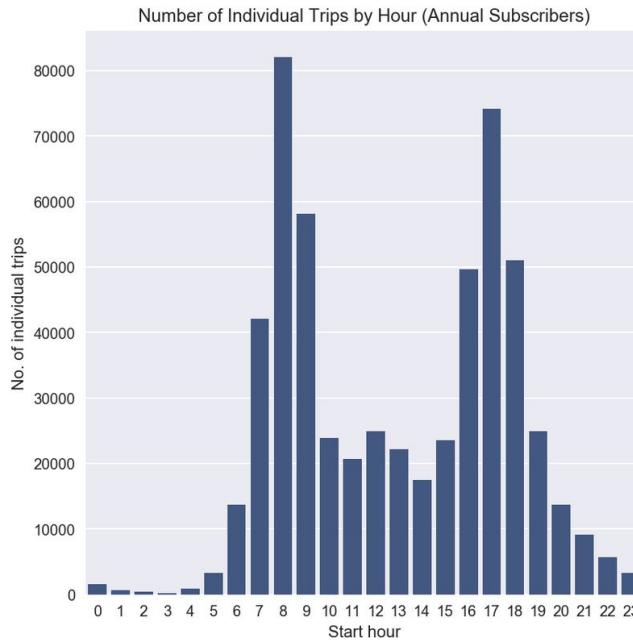
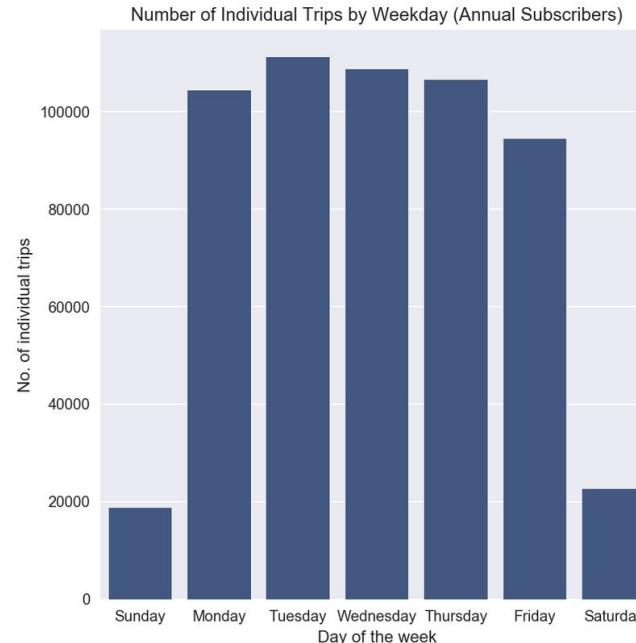


Figure 3.B: Number of trips by weekday



Examples

Figure 6: Distribution of overtime trips duration by subscription type

Figure 6.A: Overtime trips by annual subscribers

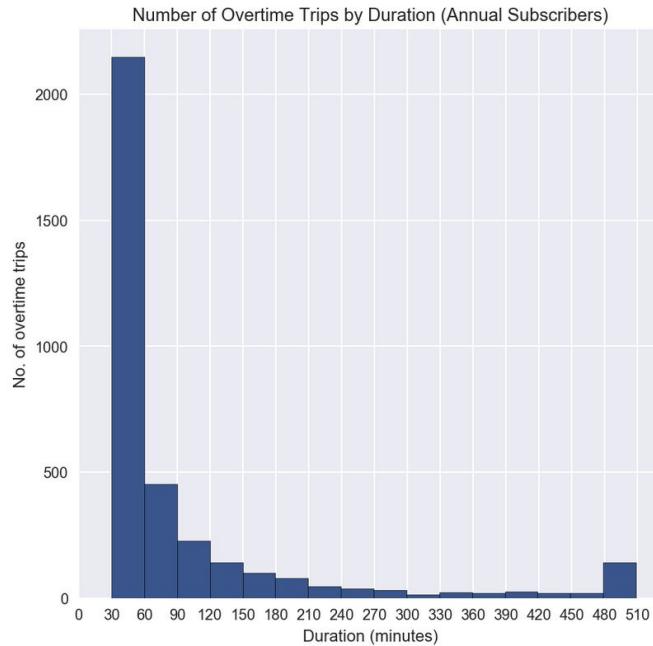
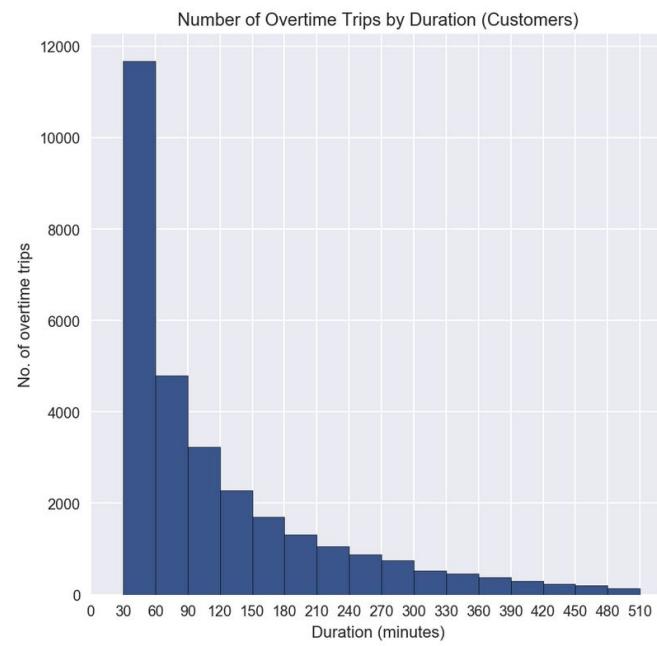
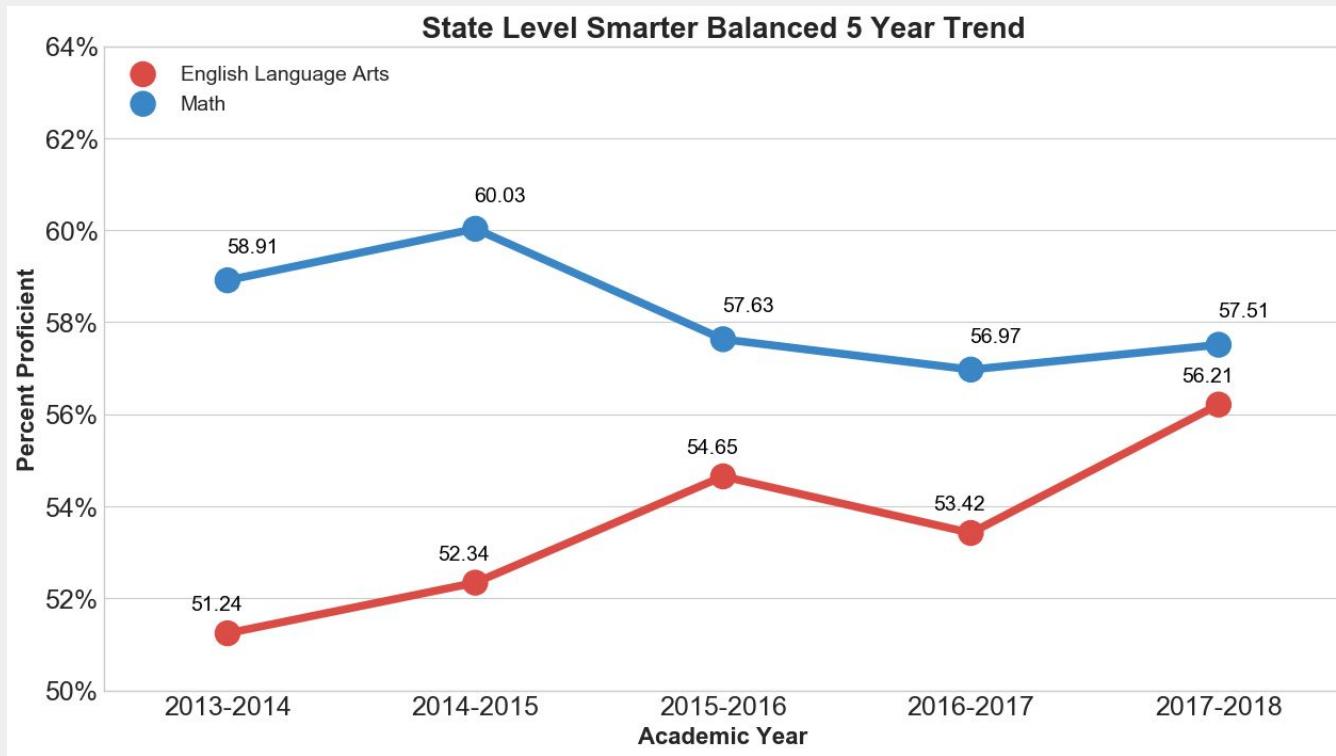


Figure 6.B: Overtime trips by customers

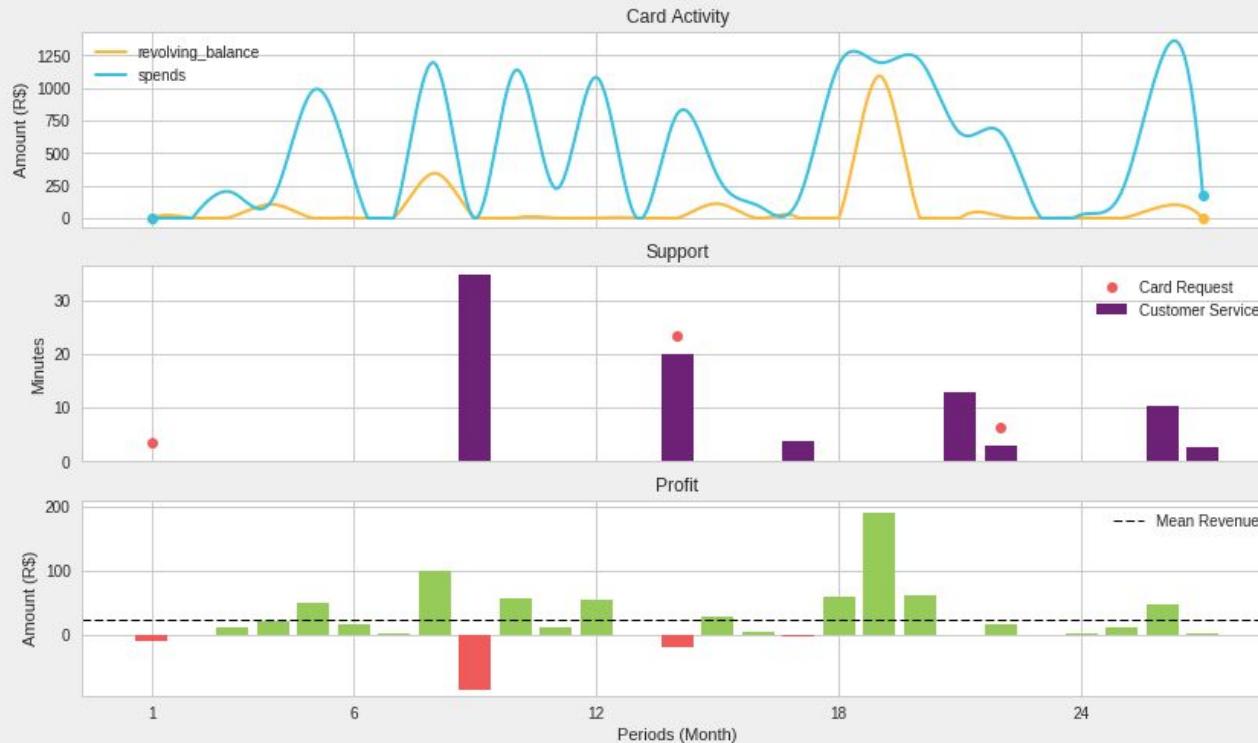


Good examples

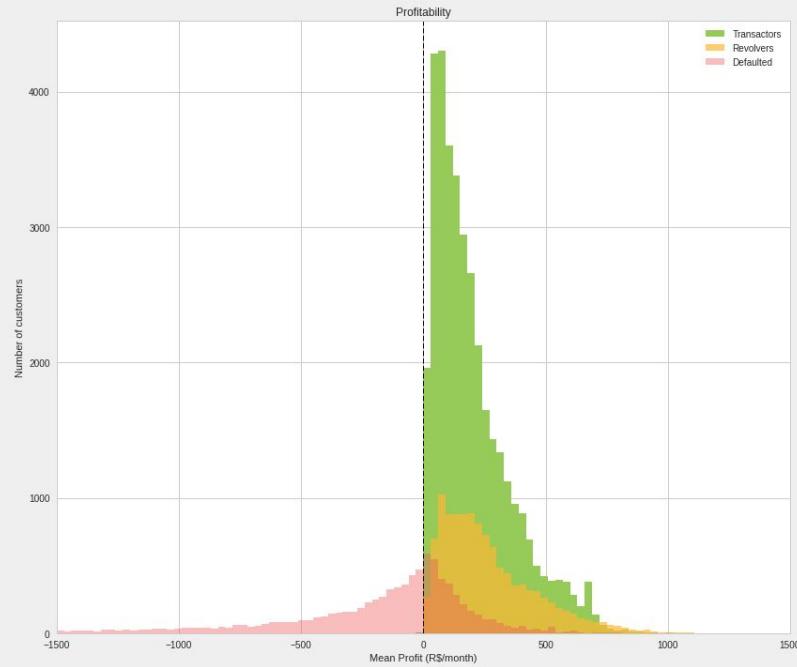


Examples

Behavior and Profit for "810e3277-6..."



Examples

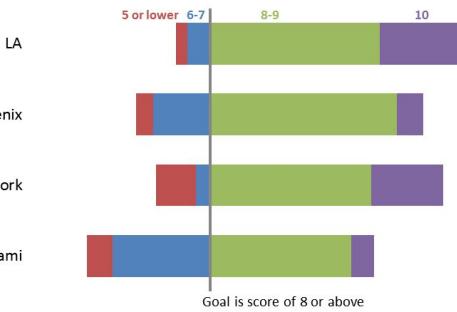


Types

Types - Deviation

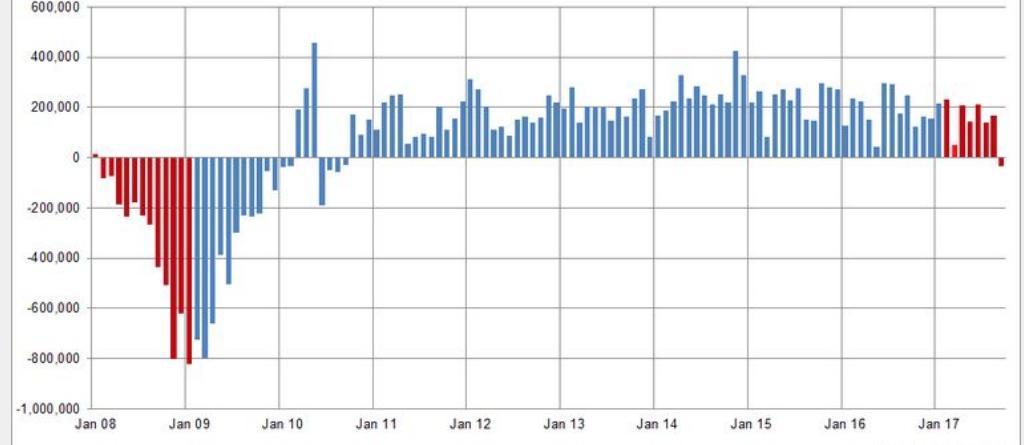
Diverging stacked bar

The LA call center has the best ratings from customers

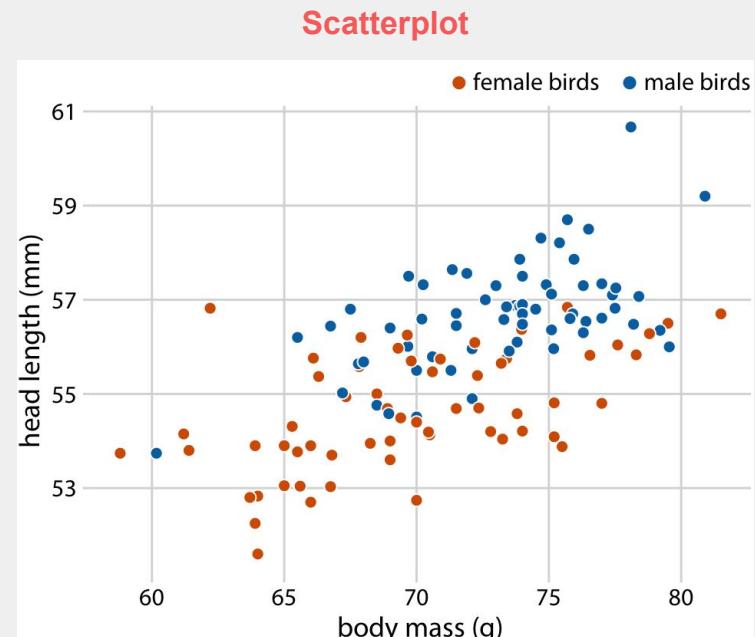
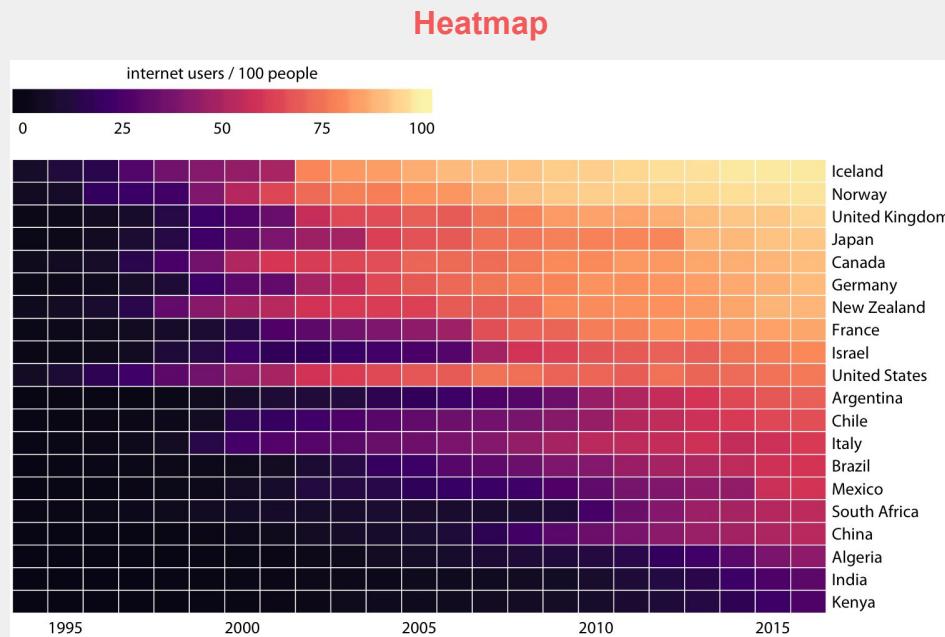


Surplus/deficit filled line

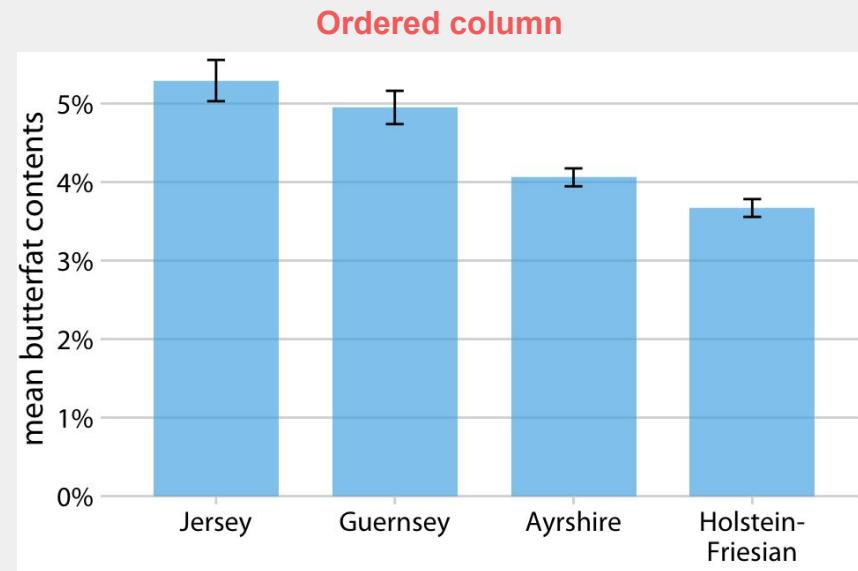
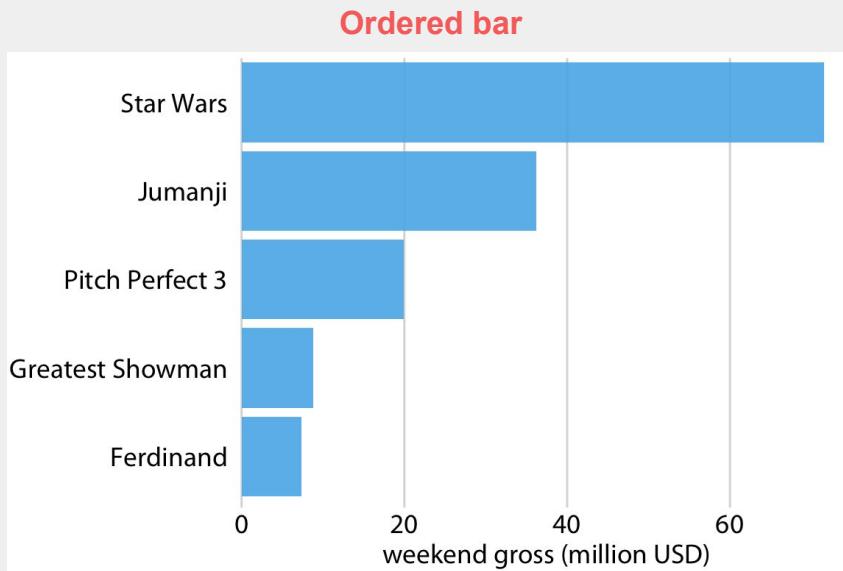
Monthly U.S. Job Totals, 2008-2017



Types - Correlation

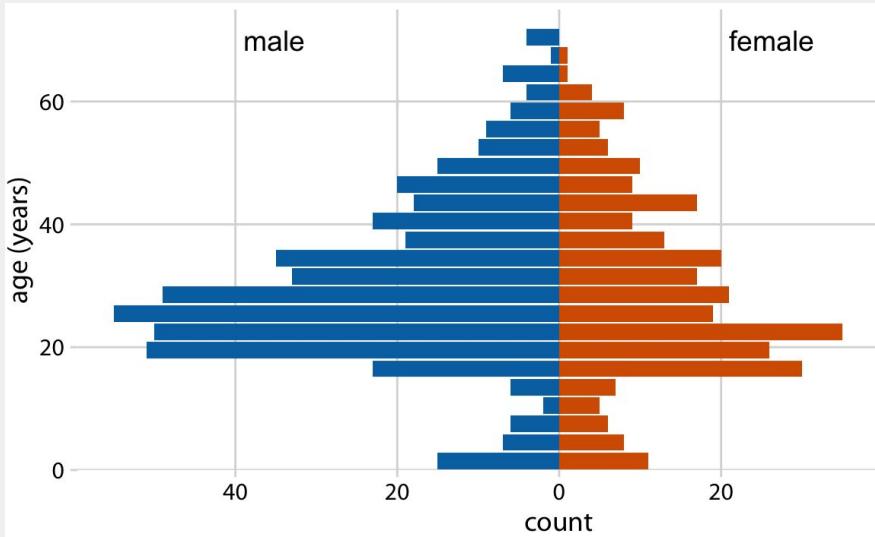


Types - Ranking

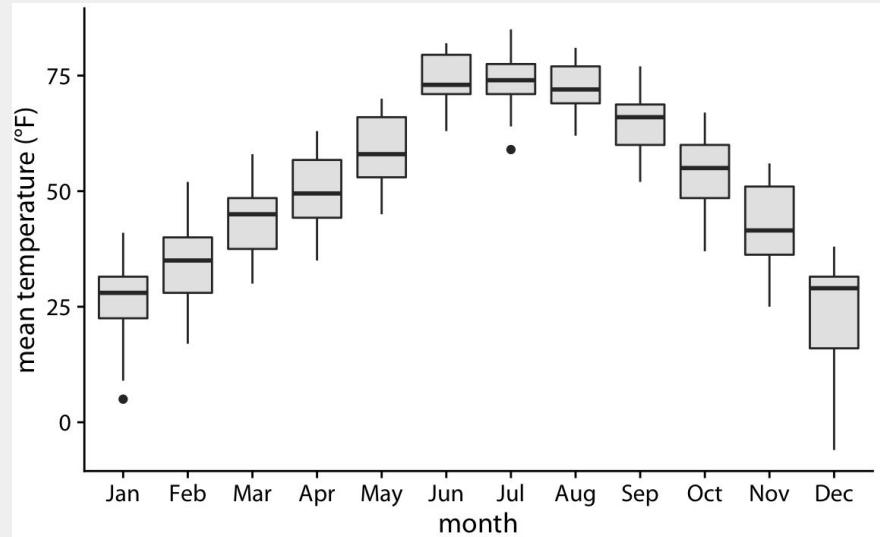


Types - Distribution

Population pyramid

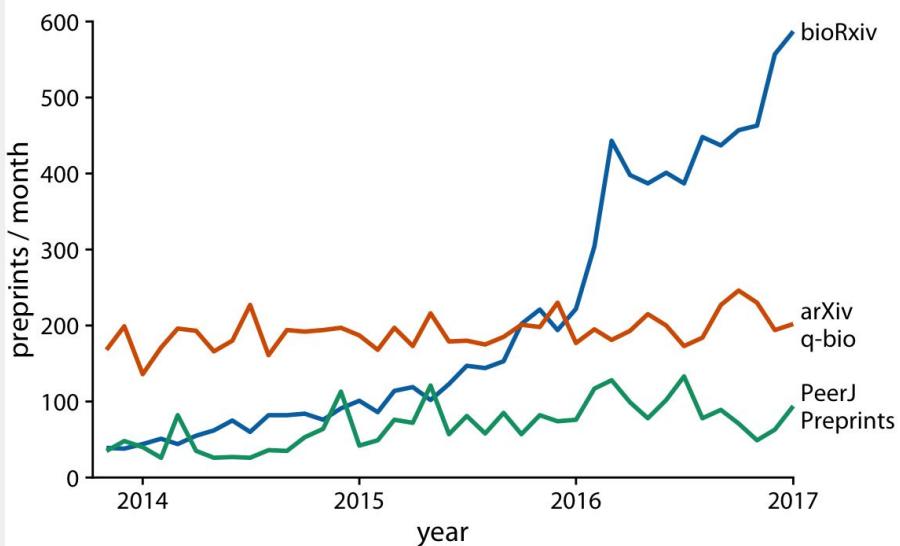


Box plot

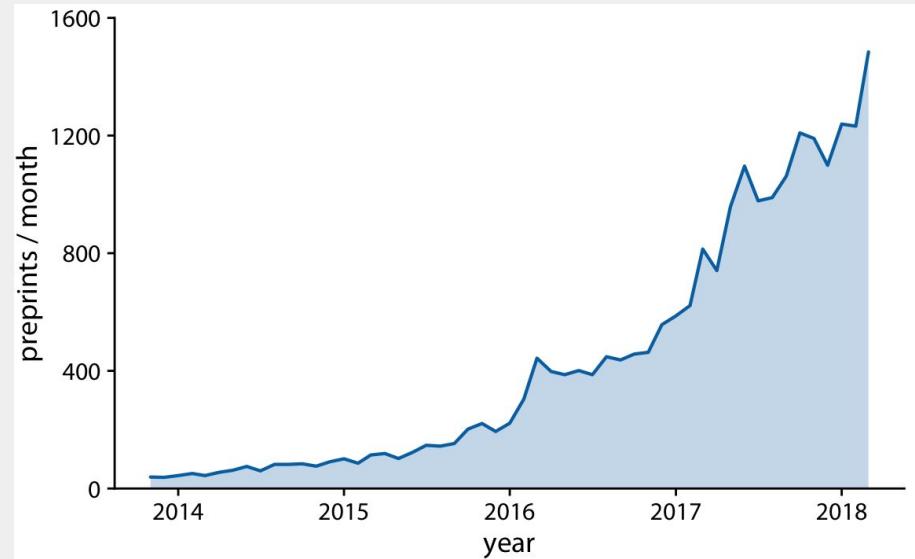


Types - Change overtime

Line plot

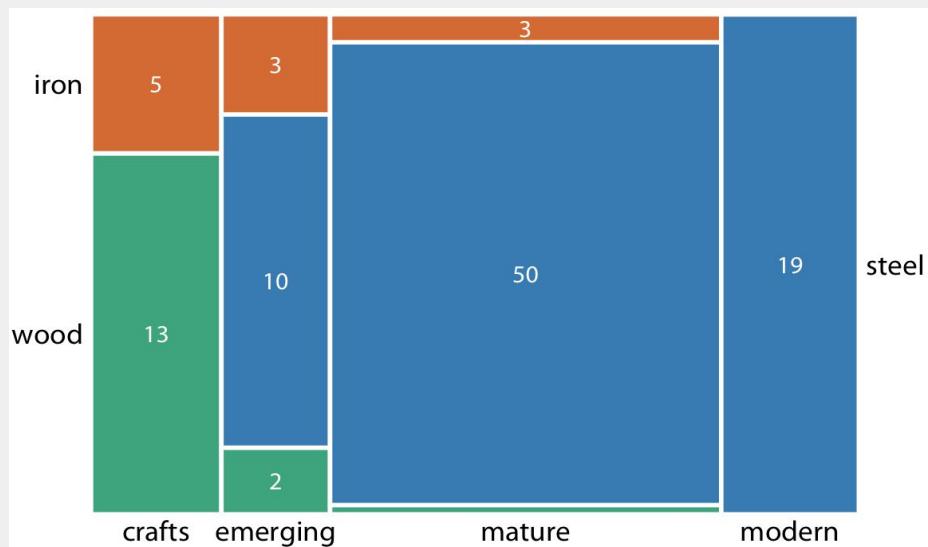


Area chart

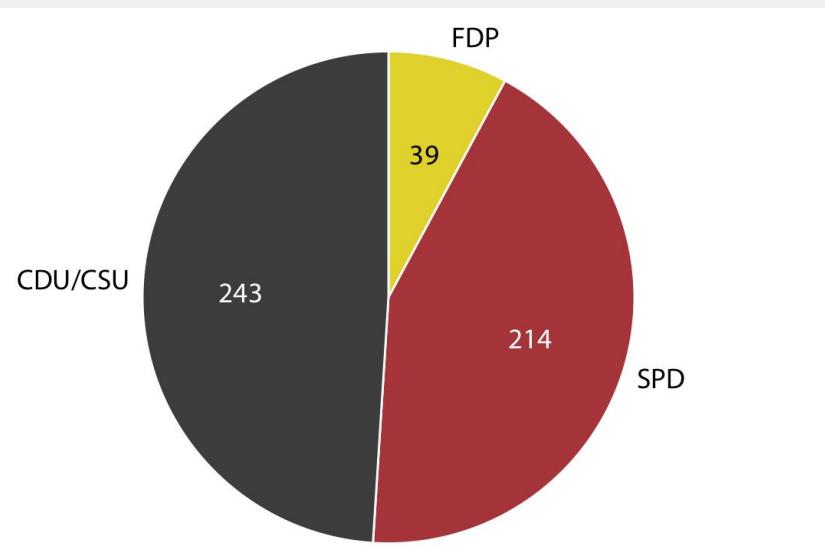


Types - Part-to-whole

Tree map

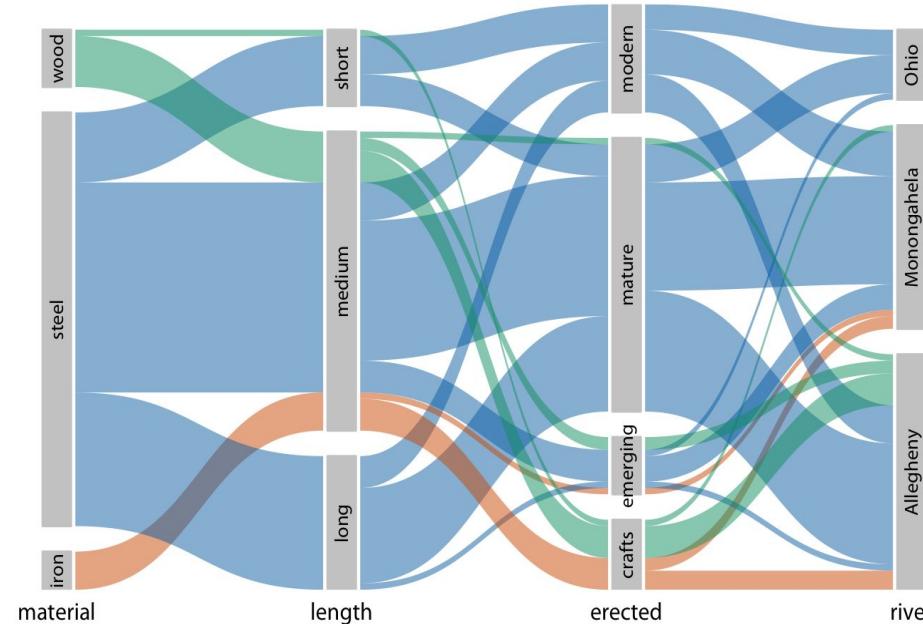


Pie chart



Types - Flow

Sankey



Data Science

Data Science

Exploratory Data Analysis

A series of hypothesis and procedures to gather evidence to them or suggest new ones.

If the problem is well defined (you know the target, it's just a matter of modeling), the hypothesis should be mostly **about the validity of your data in order to feed a model**.

If you're exploring a bunch of data to look for an opportunity, you need to plot them business oriented and understand **how the data relates to the business**.

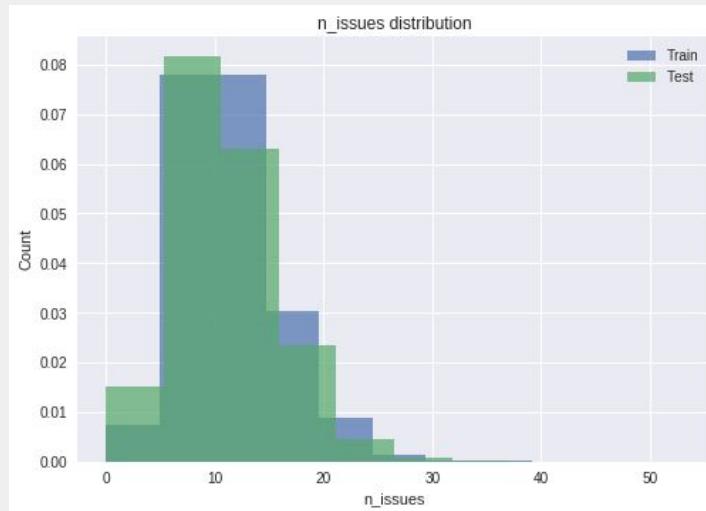
Results and presentation

Communicating a project result - Storytelling.

EDA

Hypothesis: a specific features distribution follows what I expect considering my business knowledge.

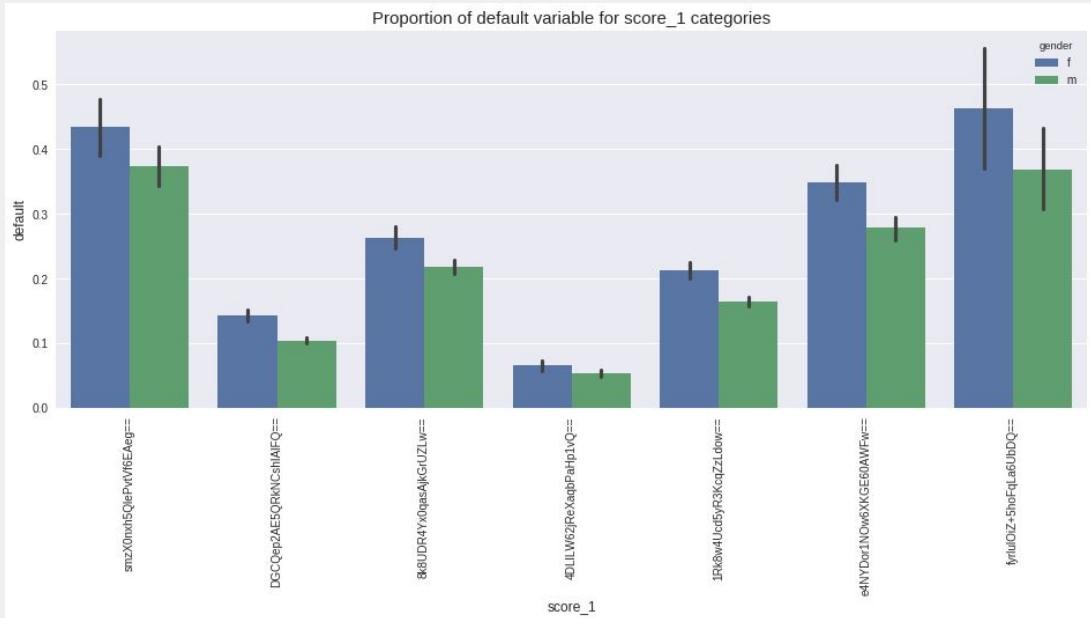
Conclusion: yes, it follows.



EDA

Hypothesis: different categories from a variable have different target proportion and they differ by gender.

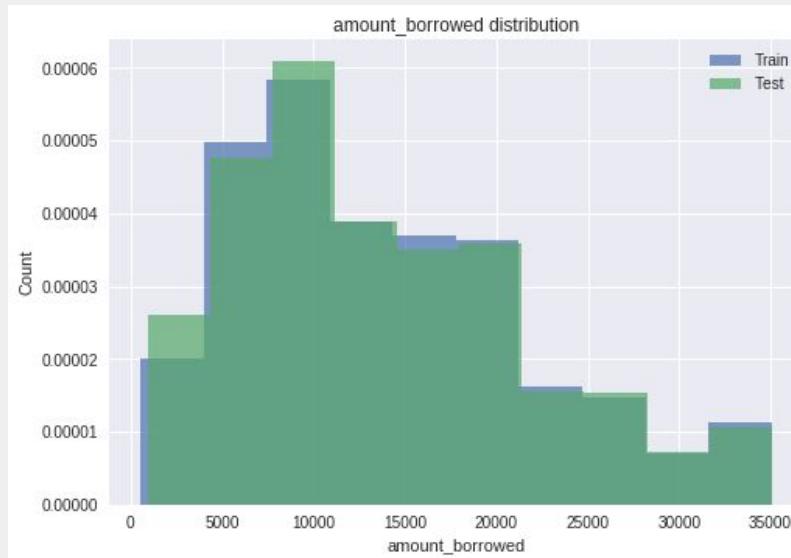
Conclusion: yes, different classes have different averages and for some it's gender sensitive.



EDA

Hypothesis: test set distribution follows the train set one.

Conclusion: yes, it follows.



EDA Examples

EDA

[Kaggle](#) is a great place to check EDA examples!

Warnings:

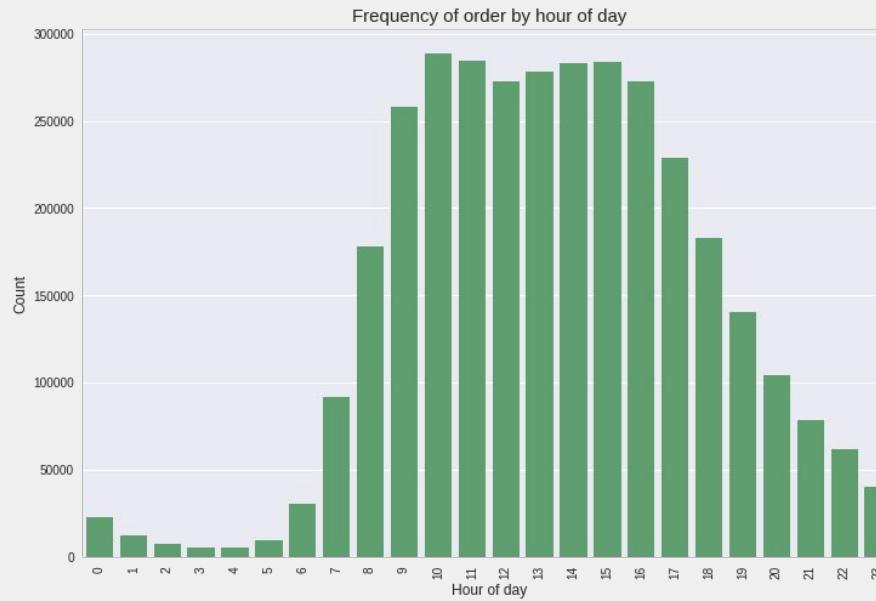
- People tend to do some useless plots there to build long kernels
- Not everyone follow the best practices for plots
- They try to do fancy plots to impress the community and get upvotes

Instacart Market Basket Analysis competition.

The screenshot shows the top six notebooks from a Kaggle competition. Each entry includes the upvote count, author profile, notebook title, submission time, and tags. The notebooks are:

- 463 upvotes: Simple Exploration Notebook - Instacart (2y ago, eda, data visualization)
- 156 upvotes: Instacart XGBoost Starter - LB 0.3791 (2y ago, @ 0.3793224)
- 142 upvotes: Customer Segments with PCA (2y ago, marketing analytics)
- 138 upvotes: light GBM benchmark 0.3692 (2y ago, @ 0.3759265)
- 132 upvotes: F1-Score Expectation Maximization in O(n²) (2y ago, optimization)
- 112 upvotes: Instacart Simple Data Exploration (2y ago)

EDA



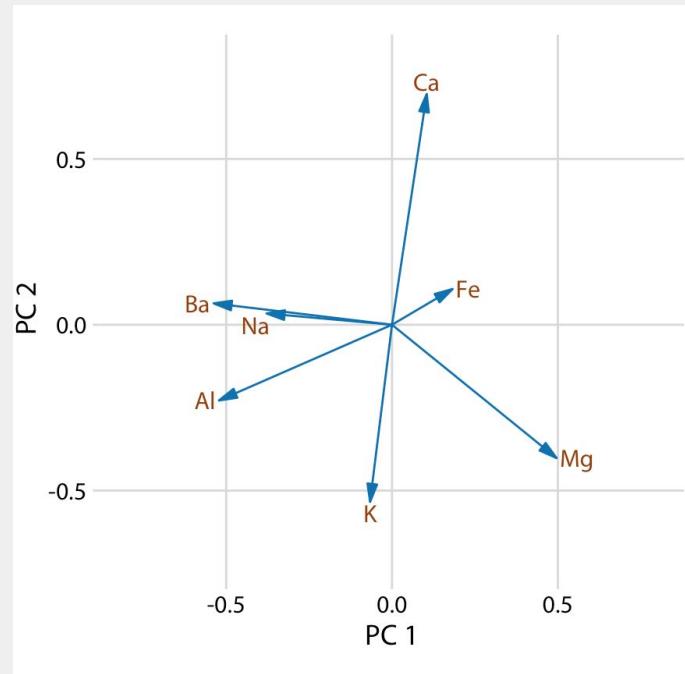
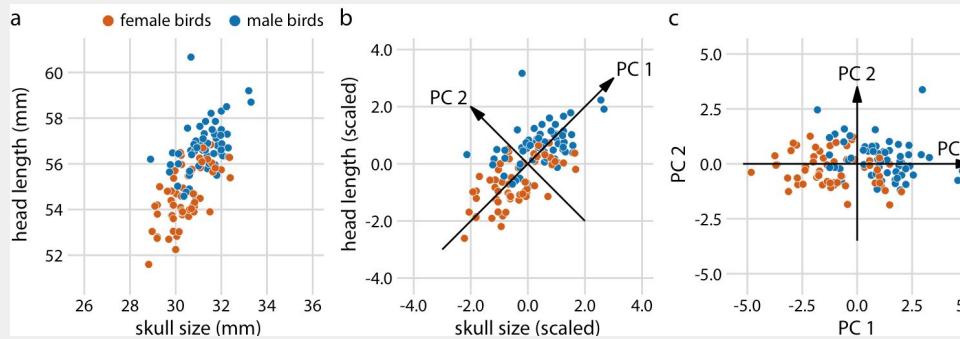
EDA



PCA and t-SNE

Data Science - PCA

Principal Component Analysis



Data Science - t-SNE

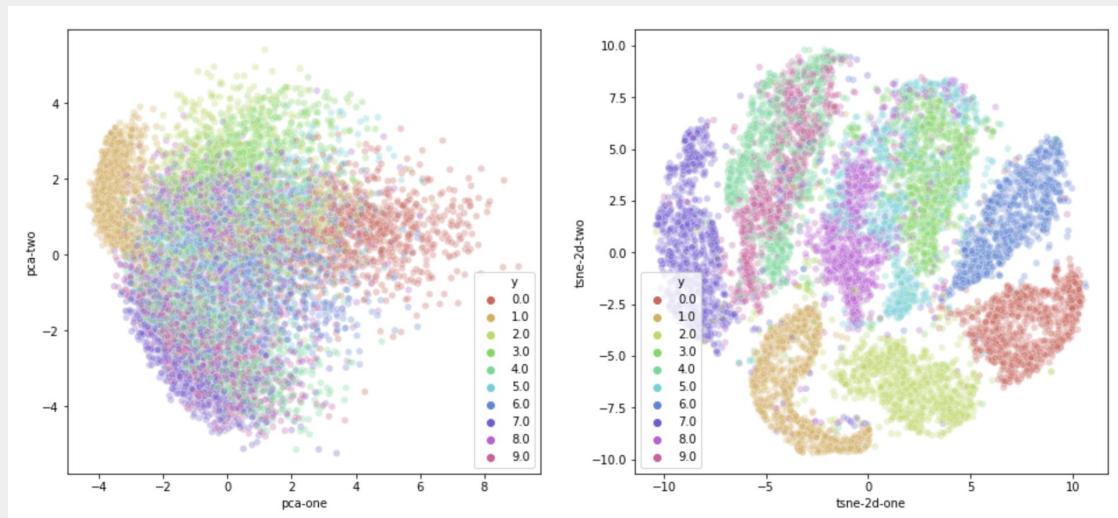
t-Distributed Stochastic Neighbor Embedding

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)},$$

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}.$$

MNIST

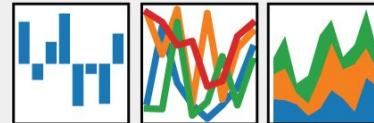


Tools

Tools

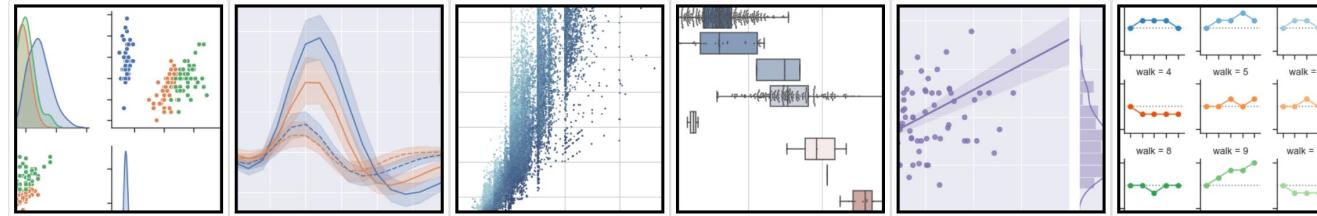
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



matplotlib

seaborn: statistical data visualization



Notebook examples & Hands-on

Takeaways

It's hard to define a recipe for data visualization, but keep in mind the general idea of **clarity, precision and efficiency**:

- Plot for a **reason**, take conclusions from every plot or discard them
- **State a plot reason before design it**, the type plays an important role
- Use it to fastly frame **important business reflection on the data**
- Don't use **pie charts!**
- The audience should be able **to understand your plot without further info!**
- **Don't bother about tool syntax**, plots are probably the most googled part of a data scientist job.



Questions?

Twitter: @lgmoneda

E-mail: lgmoneda@gmail.com

Blog: <http://lgmoneda.github.io/>