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
- Kaggler (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

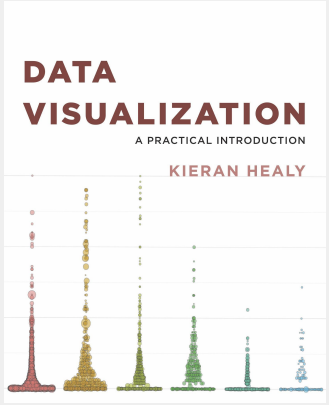
Resources and References

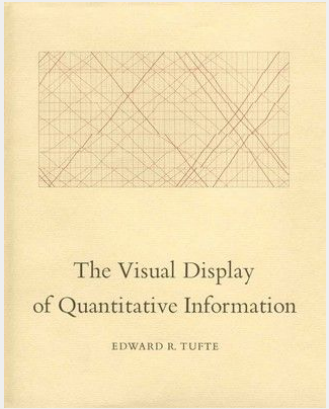
Books

- 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



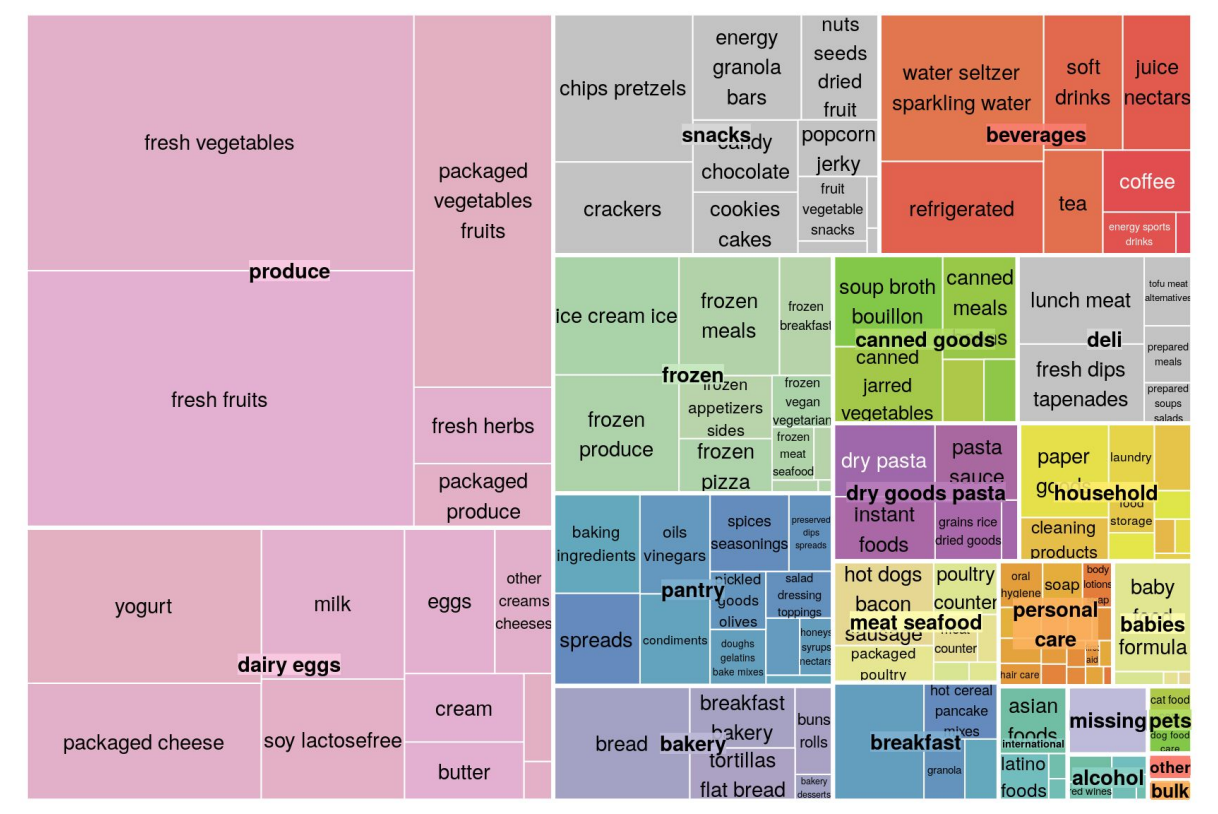


Goal: efficient communication

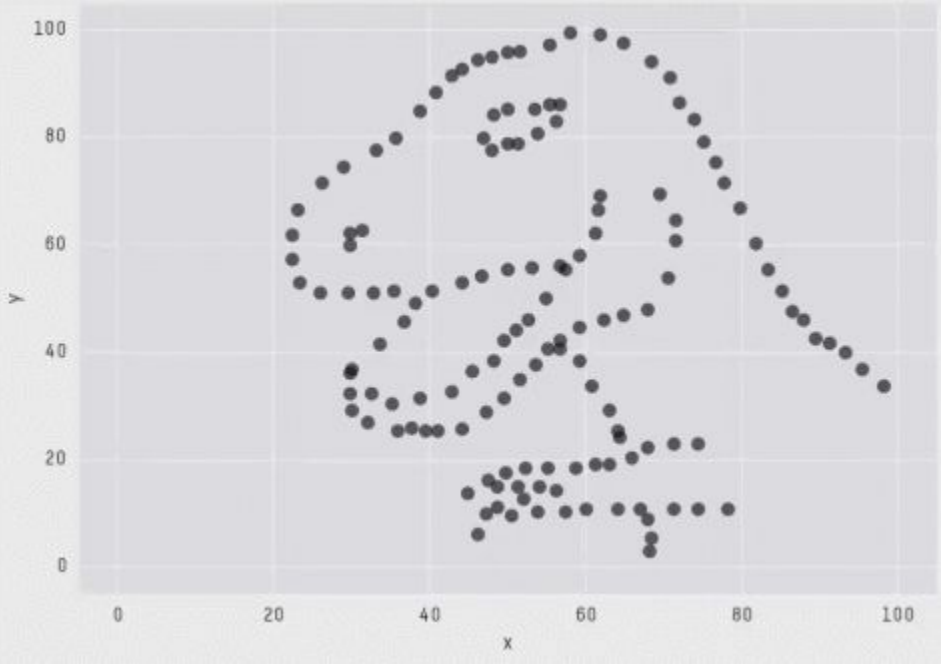
Goal

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


fresh vegetables
produce
fresh fruits
packaged vegetables fruits
fresh herbs
packaged produce
chips pretzels
energy granola bars
nuts seeds dried fruit
snacks
candy
popcorn jerky
chocolate
crackers
cookies cakes
fruit vegetable snacks
water seltzer sparkling water
soft drinks
juice nectars
beverages
refrigerated
tea
coffee
energy sports drinks
ice cream ice
frozen meals
frozen breakfast
frozen
frozen appetizers sides
frozen vegan vegetarian
frozen produce
frozen pizza
frozen meat seafood
soup broth bouillon
canned meals
canned goods
canned jarred vegetables
lunch meat
tofu meat alternatives
deli
prepared meals
fresh dips tapenades
prepared soups salads
dry pasta
pasta sauce
dry goods pasta
instant foods
grains rice dried goods
paper
laundry
household
cleaning products
storage
baking ingredients
oils vinegars
spices seasonings
preserved dips spreads
pantry
pickled goods olives
salad dressing toppings
spreads
condiments
doughs gelatins bake mixes
honeys syrups nectars
hot dogs bacon
poultry counter
meat seafood
sausage packaged poultry
counter
oral hygiene
soap
body lotions
personal care
hair care
baby
babies formula
yogurt
milk
dairy eggs
eggs
other creams cheeses
packaged cheese
soy lactosefree
cream
butter
bread
breakfast bakery
buns rolls
bakery
tortillas flat bread
bakery desserts
breakfast
hot cereal pancake mixes
granola
asian foods
international
missing
cat food
pets
dog food care
latino foods
alcohol
wines
other
bulk

Goal



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

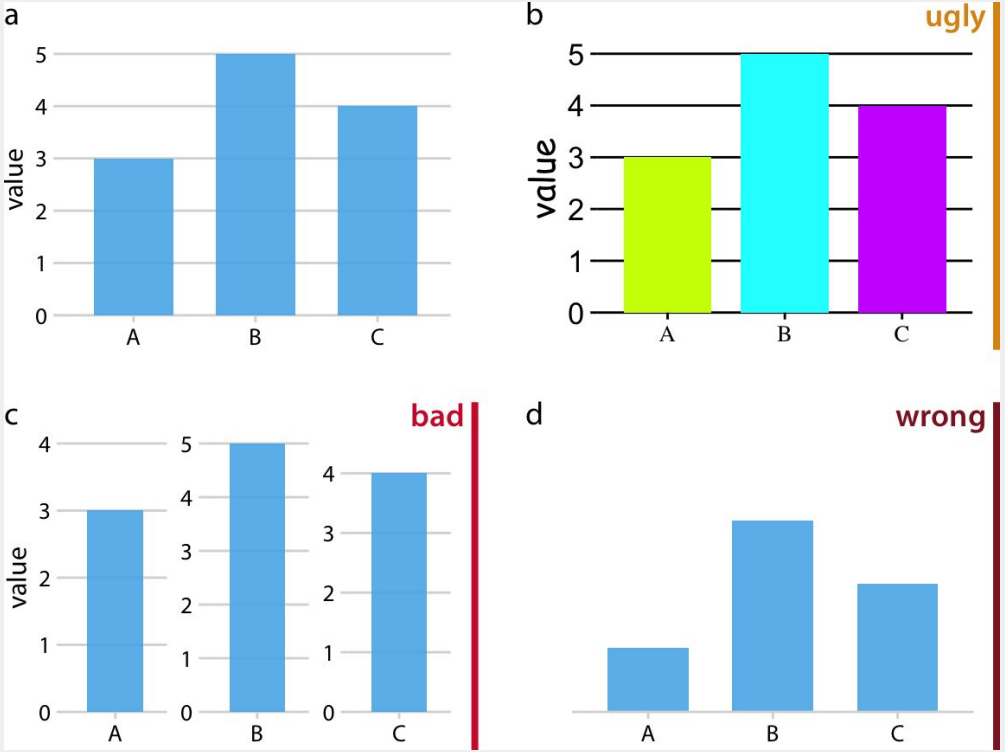
Problems & Practices

Problems & Practices

Problems	Practices
- Bad taste - Bad data - Bad perception	- Labeling - Plot Design - Context - Honest - Self sufficiency (in terms of data!) - Right plot type

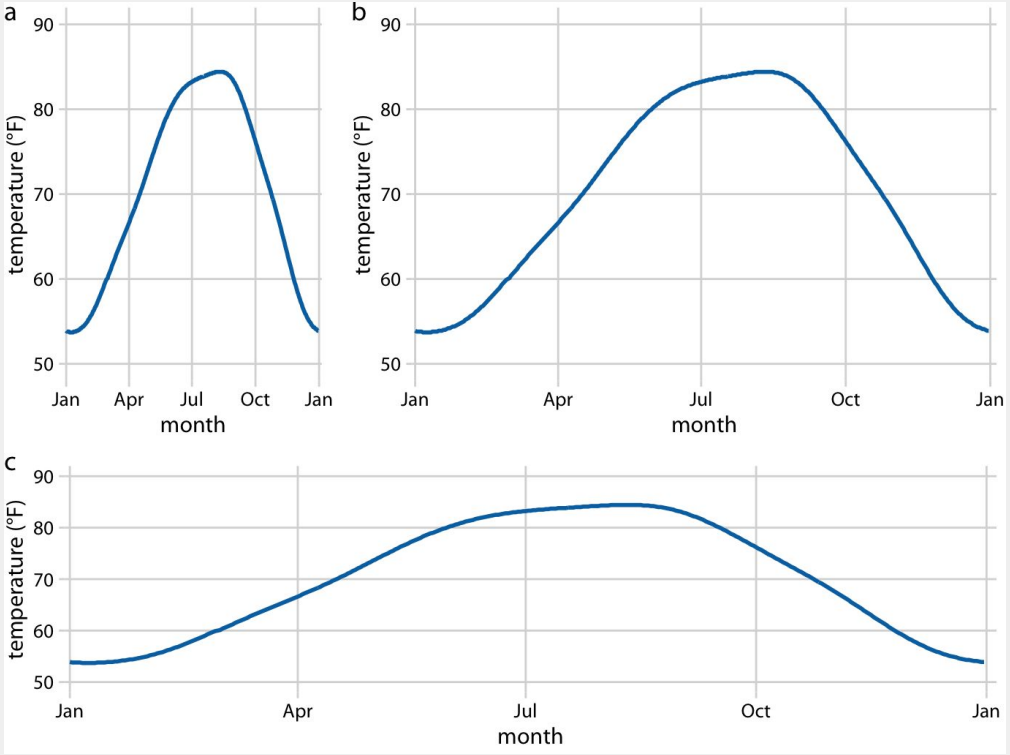
The more complex is the idea you want to communicate, more successful you would need to be in "clarity, precision, and efficiency".

Problems & Practices



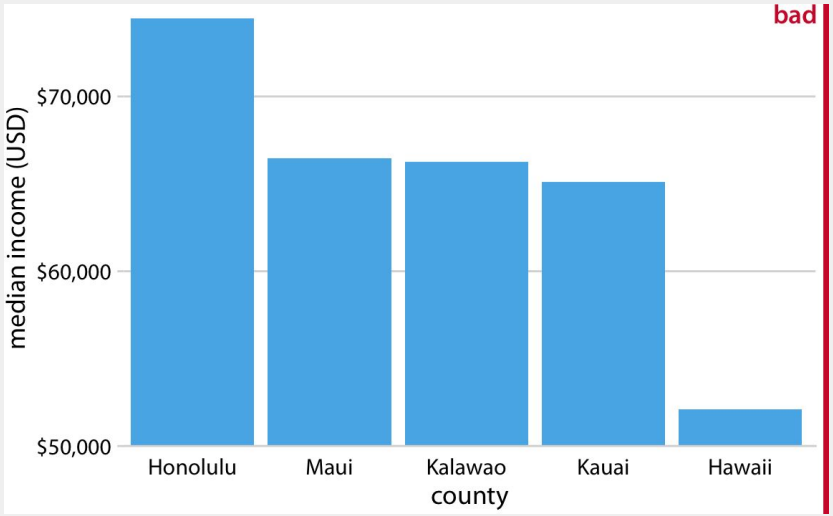
Reference: https://serialmentor.com/dataviz/introduction.html

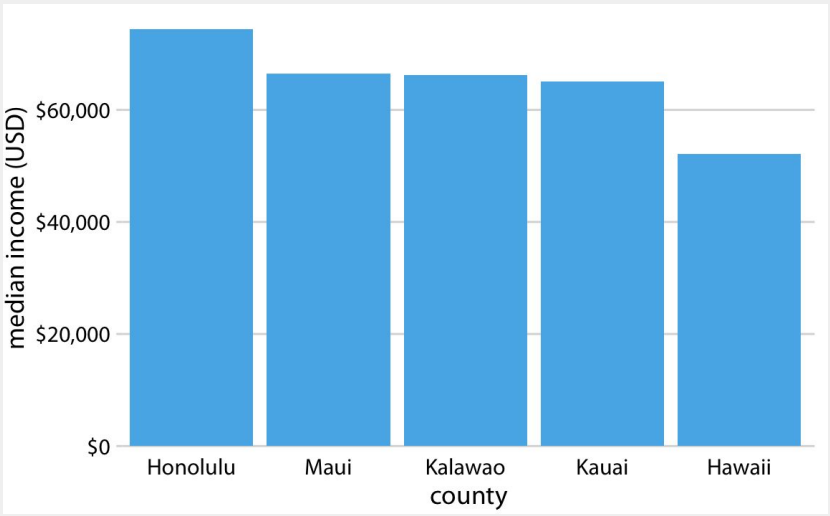
Problems - Scale



Reference: https://serialmentor.com/dataviz/coordinate-systems-axes.html

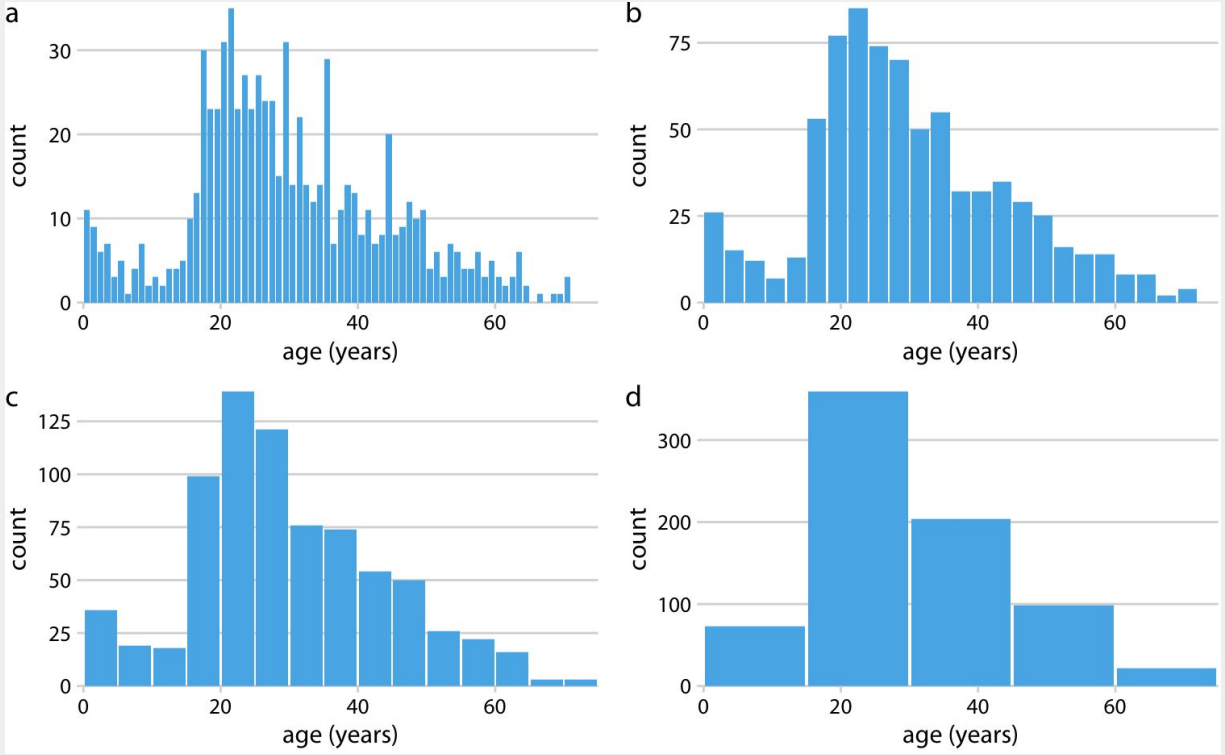
Problems - Scale





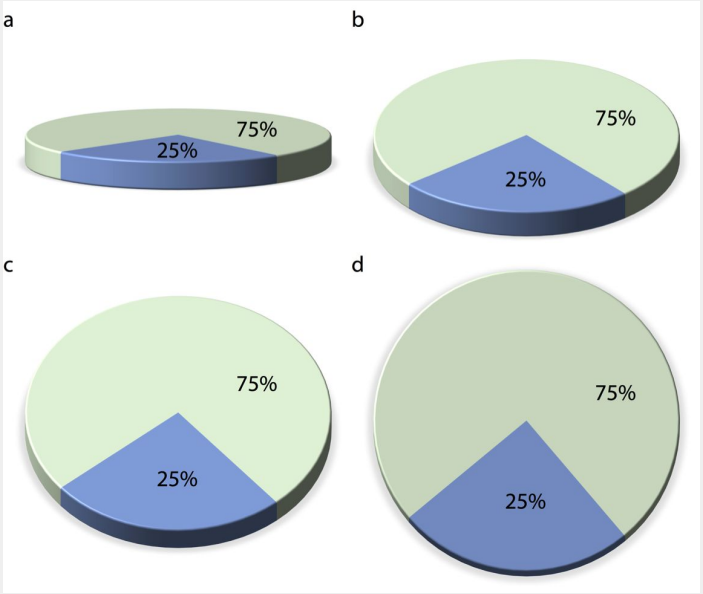
Reference: https://serialmentor.com/dataviz/proportional-ink.html

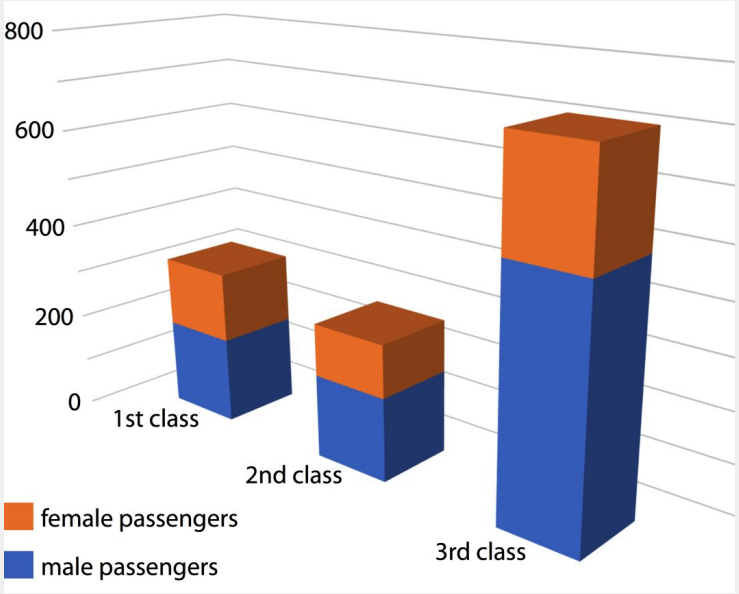
Problems - Distribution



Reference: https://serialmentor.com/dataviz/histograms-density-plots.html

Problems - 3D





Reference: https://serialmentor.com/dataviz/no-3d.html

Problems - Combining

Reference: https://serialmentor.com/dataviz/no-3d.html

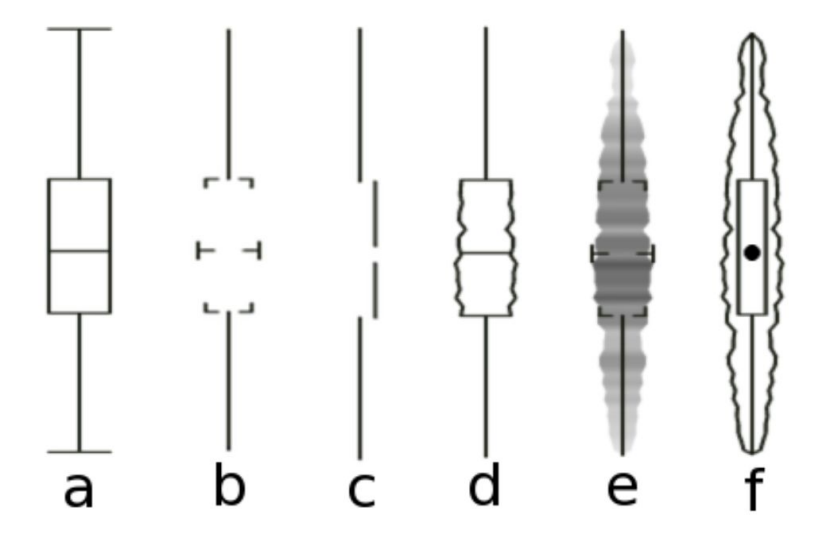
ideias

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.

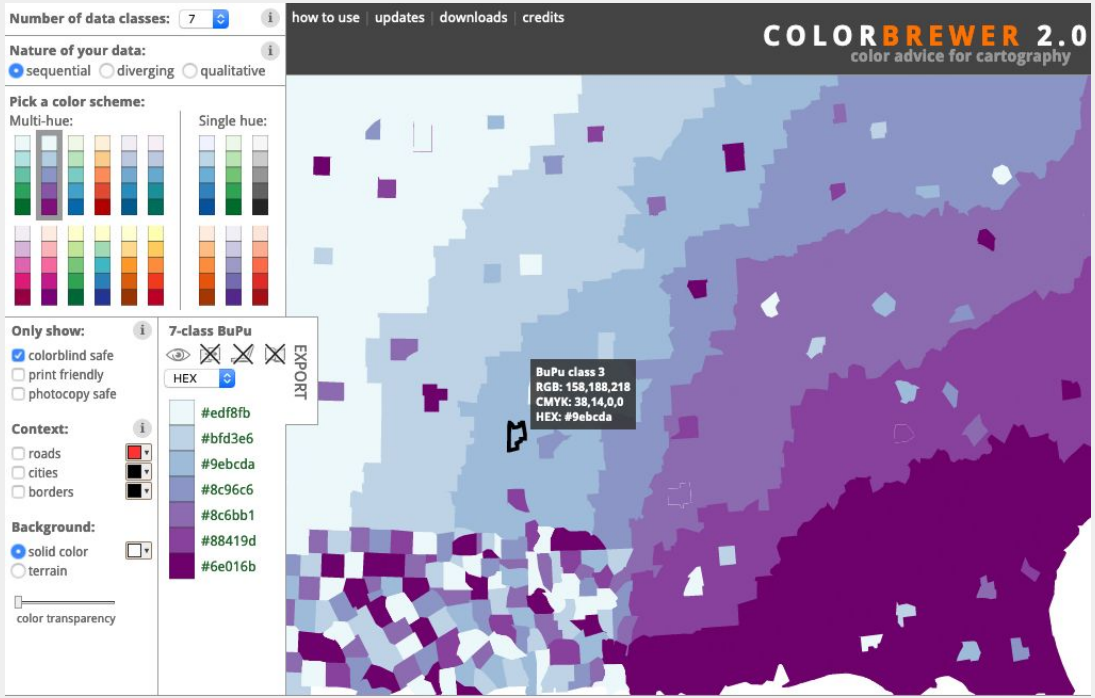
Principle: Maximize data-to-ink ratio



Six Kinds of summary boxplots.
Type (c) is from Tufte

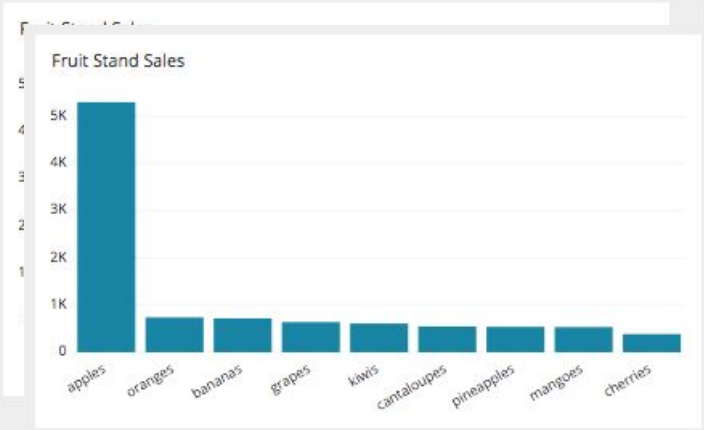
Reference: http://socviz.co/lookatdata.html#why-look-at-data

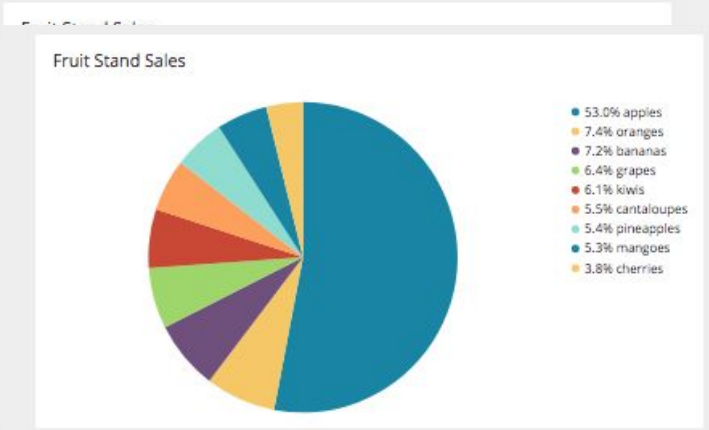
Principle: Colors



Examples

Bad examples

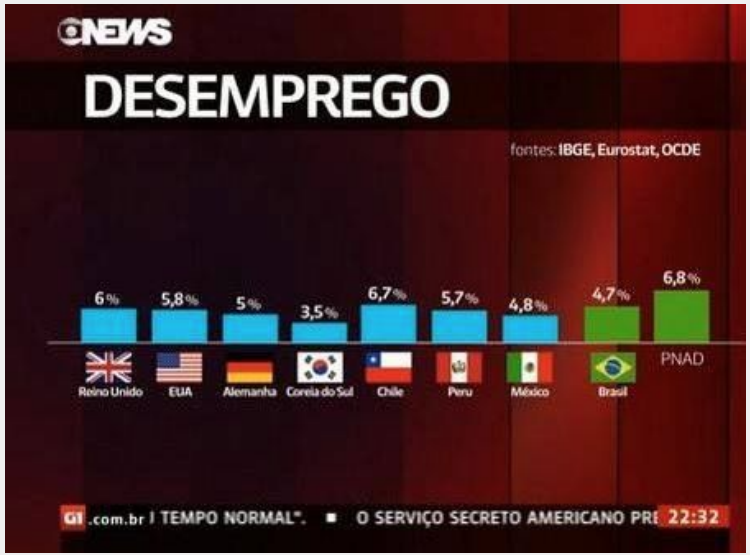




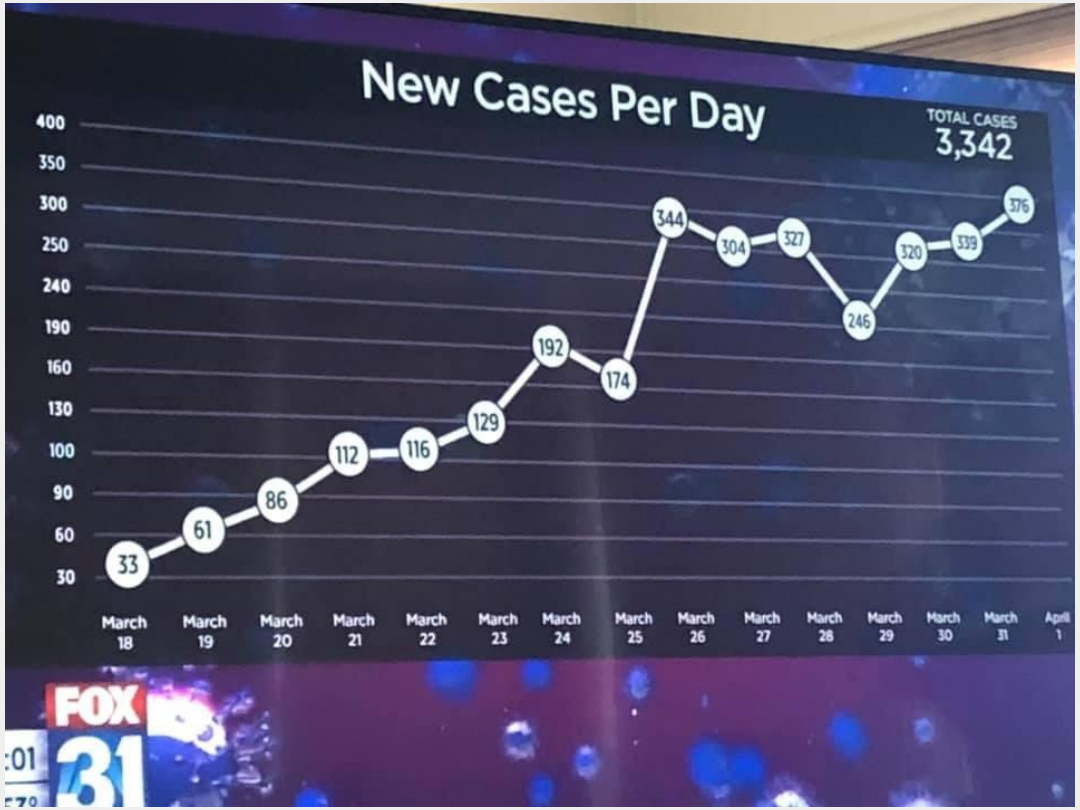
Reference: https://chartio.com/learn/dashboards-and-charts/what-is-the-difference-between-a-pie-chart-and-a-bar-chart/

Bad examples

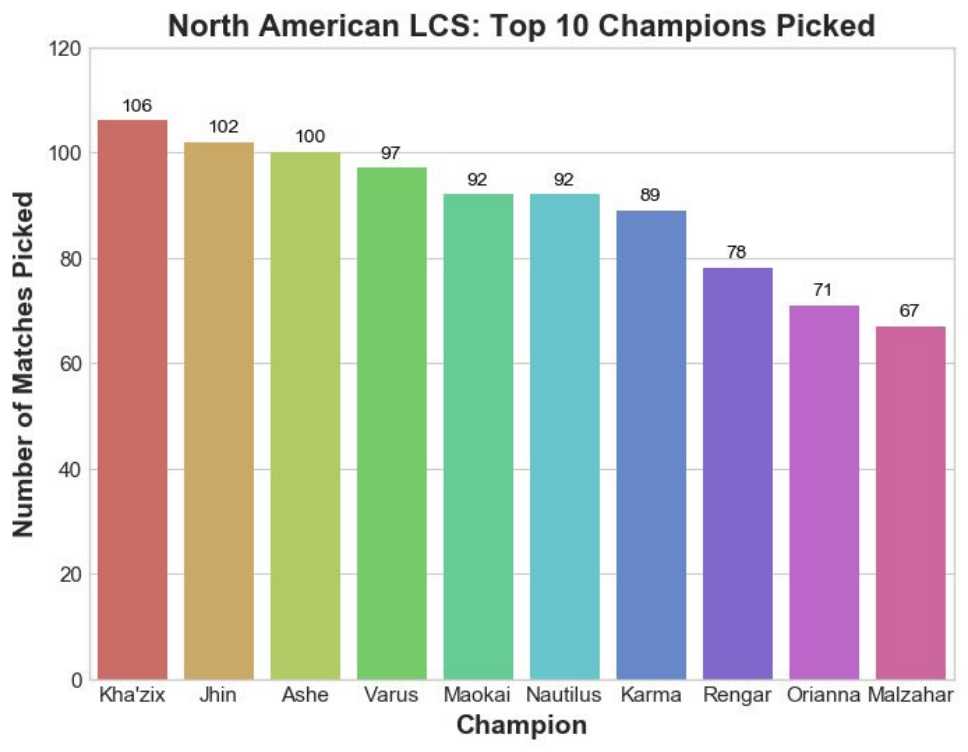
Bad examples



Bad examples



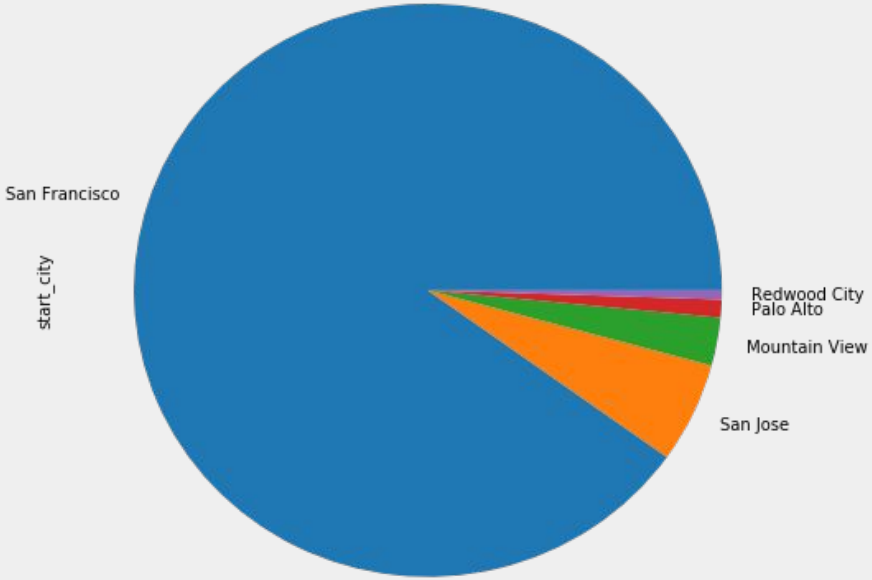
Bad examples



Reference: https://omegadescent.wordpress.com/2018/06/11/seaborn1/

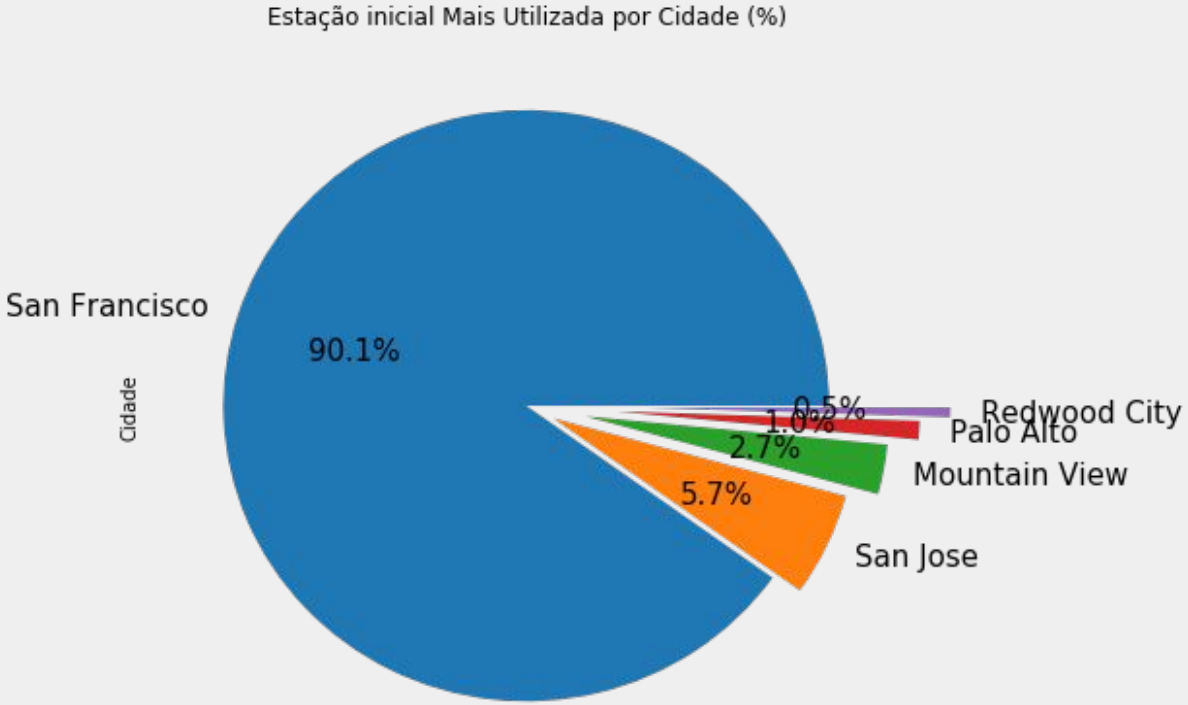
Does it really happen?

Examples

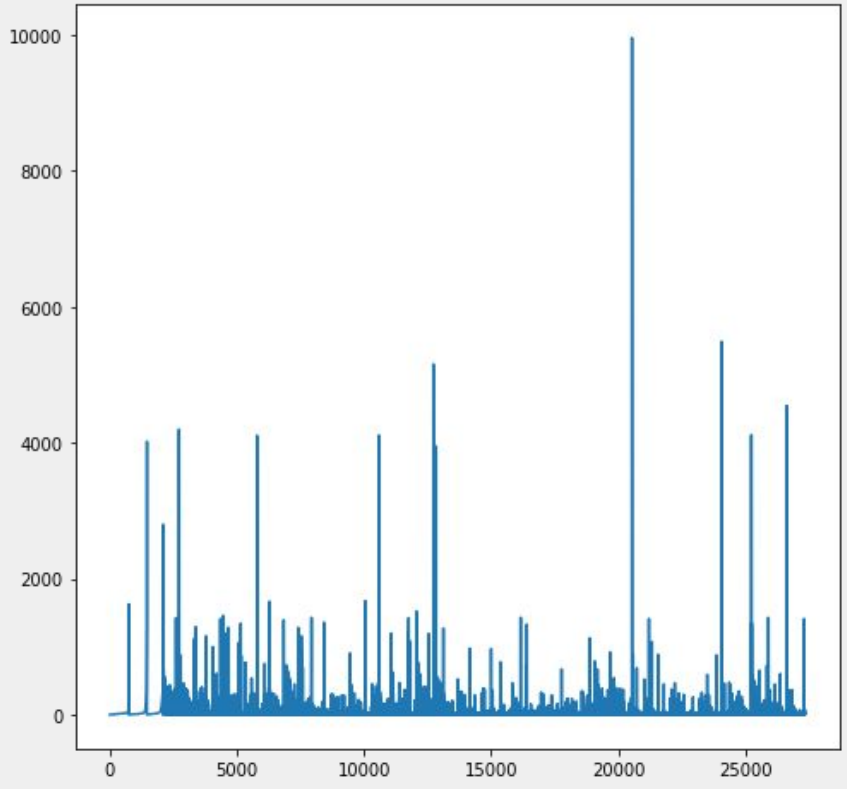


Reference: all the following examples are from open Udacity projects in Github.

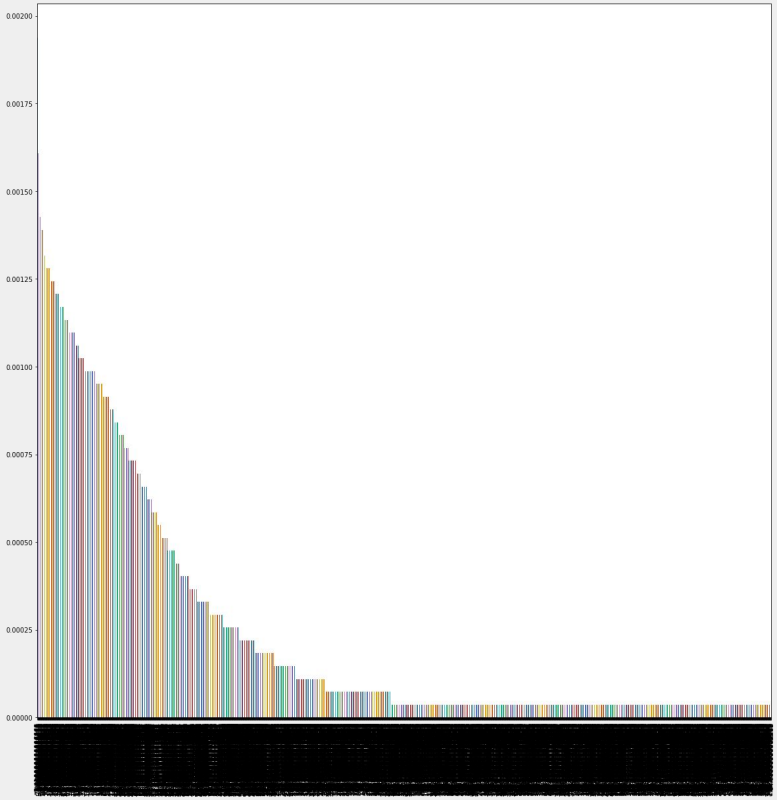
Examples


Estação inicial Mais Utilizada por Cidade (%)
Cidade
San Francisco
90.1%
5.7%
2.7%
1.0%
0.5%
San Jose
Mountain View
Palo Alto
Redwood City

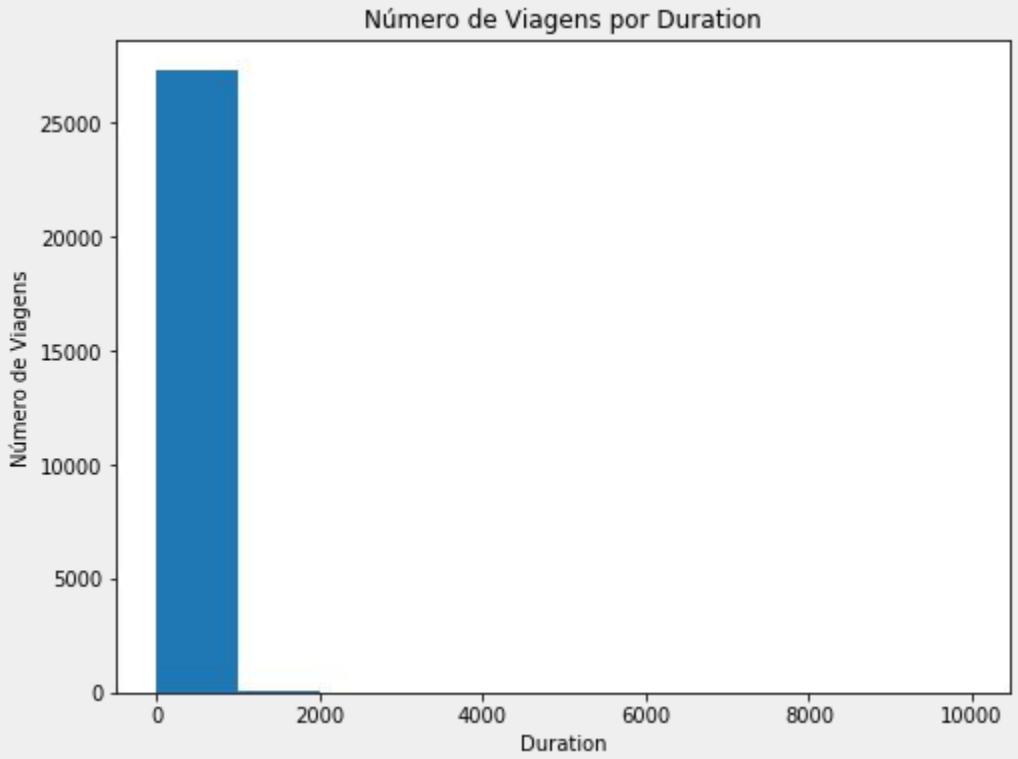
Examples



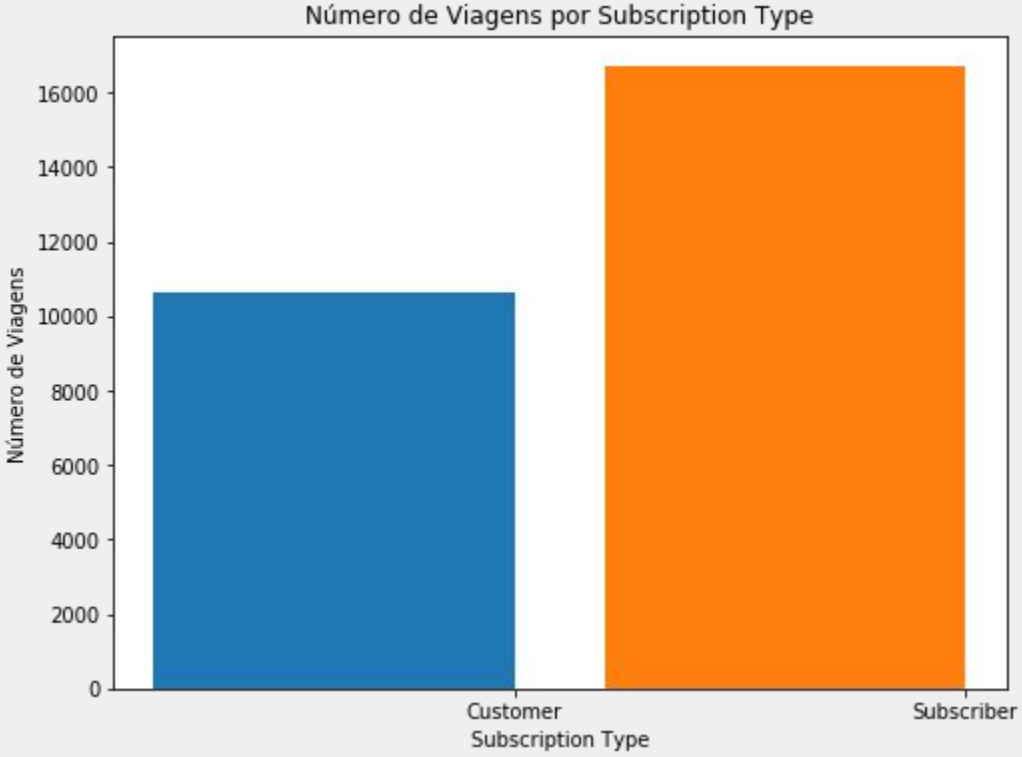
Examples


0.00200
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000

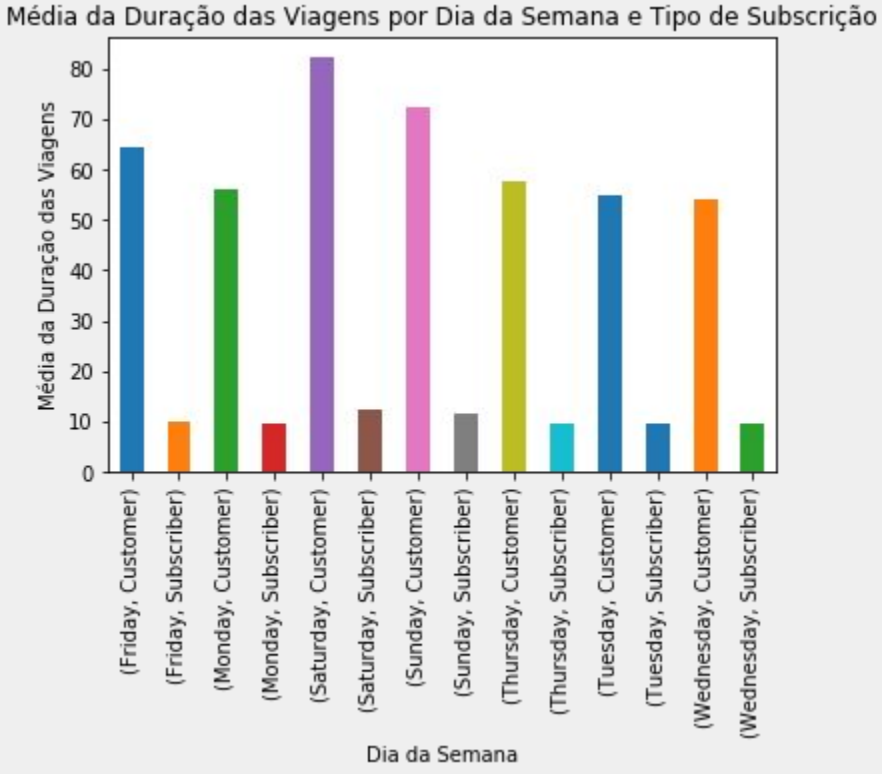
Examples


Número de Viagens por Duration
Número de Viagens
25000
20000
15000
10000
5000
0
0
2000
4000
6000
8000
10000
Duration

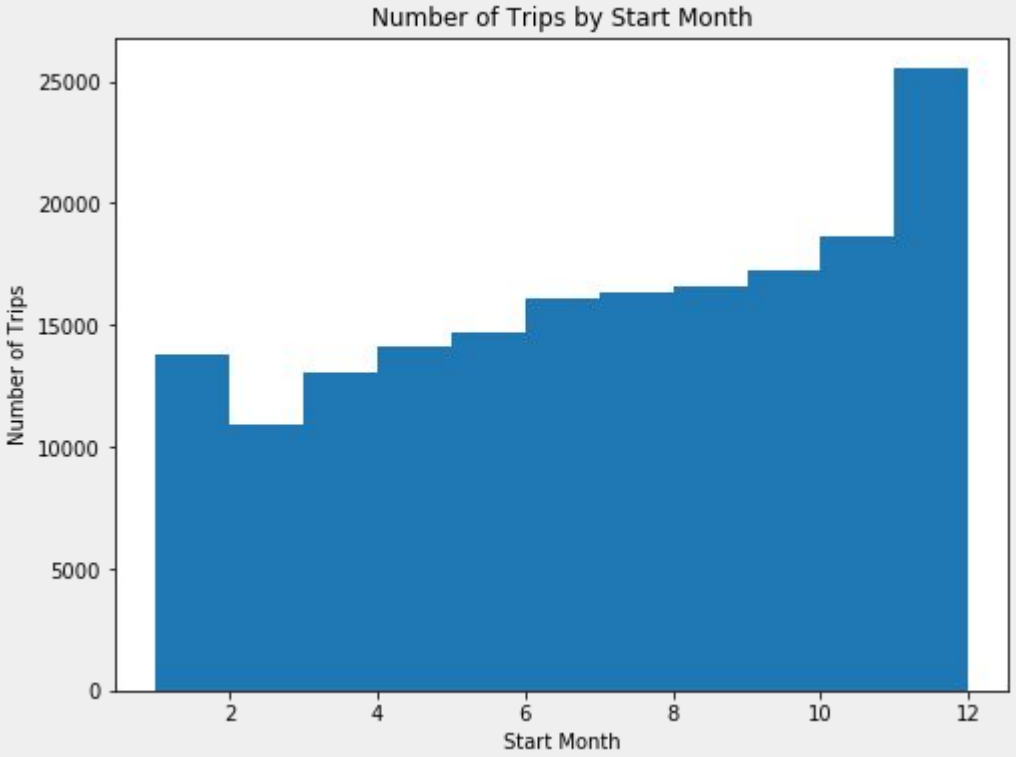
Examples


Número de Viagens por Subscription Type
Número de Viagens
16000
14000
12000
10000
8000
6000
4000
2000
0
Customer
Subscriber
Subscription Type

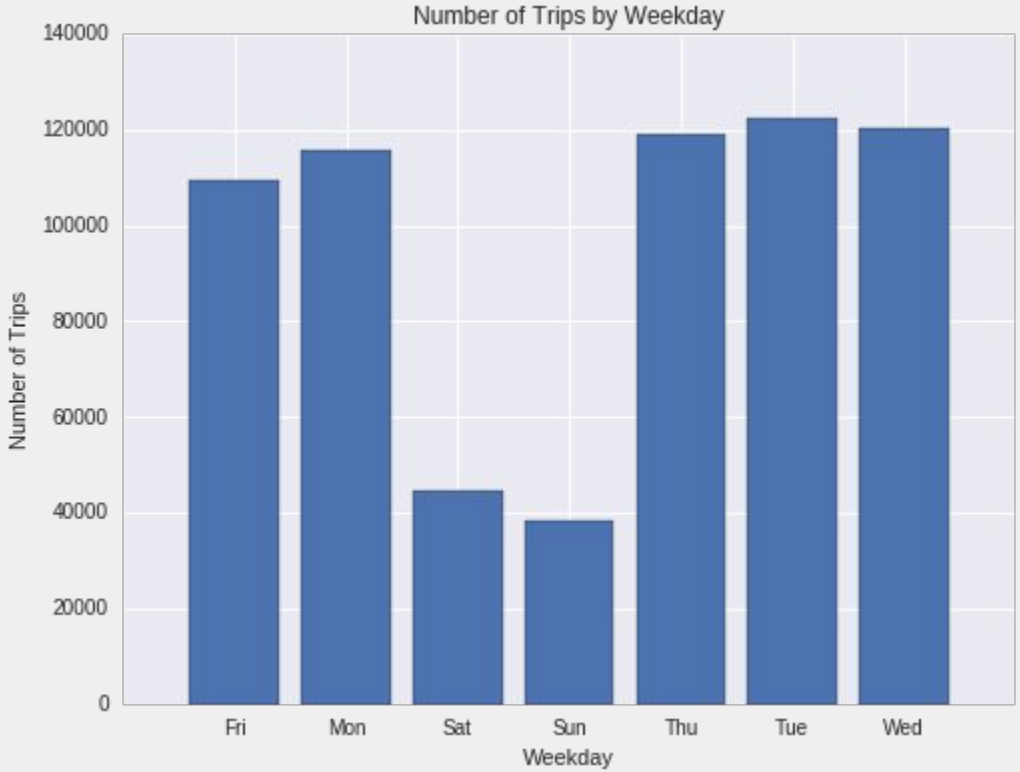
Examples


Média da Duração das Viagens por Dia da Semana e Tipo de Subscrição
Média da Duração das Viagens
0
10
20
30
40
50
60
70
80
(Friday, Customer)
(Friday, Subscriber)
(Monday, Customer)
(Monday, Subscriber)
(Saturday, Customer)
(Saturday, Subscriber)
(Sunday, Customer)
(Sunday, Subscriber)
(Thursday, Customer)
(Thursday, Subscriber)
(Tuesday, Customer)
(Tuesday, Subscriber)
(Wednesday, Customer)
(Wednesday, Subscriber)
Dia da Semana

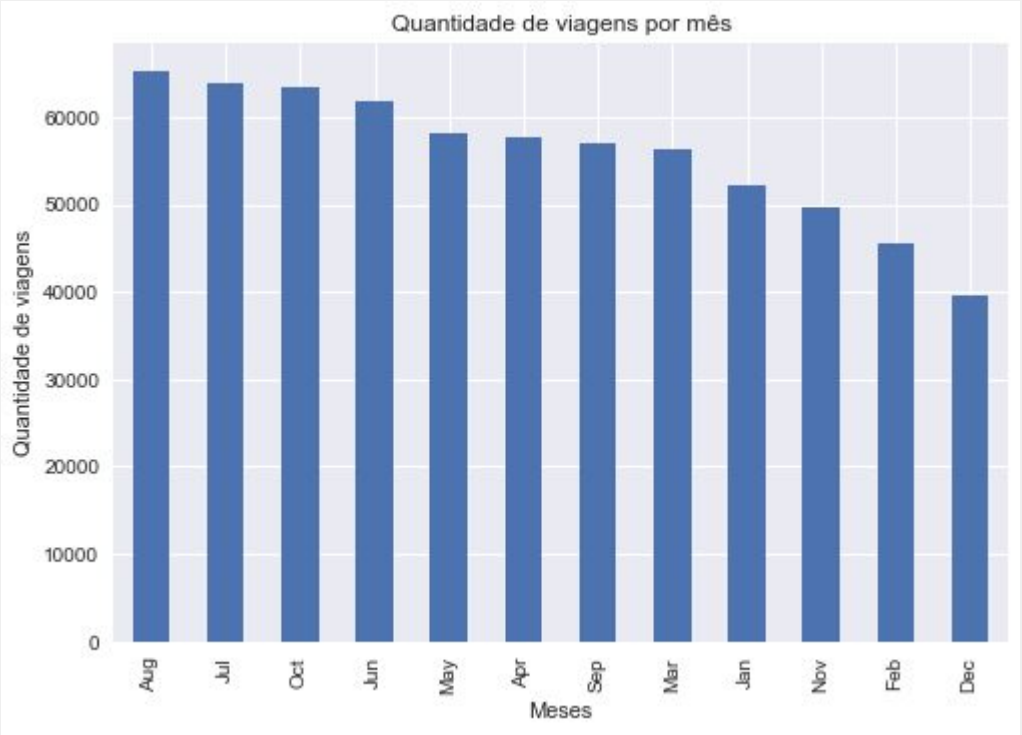
Examples


Number of Trips by Start Month
Number of Trips
25000
20000
15000
10000
5000
0
2
4
6
8
10
12
Start Month

Examples


Number of Trips by Weekday
140000
120000
100000
80000
60000
40000
20000
0
Number of Trips
Fri
Mon
Sat
Sun
Thu
Tue
Wed
Weekday

Examples


Quantidade de viagens por mês
Quantidade de viagens
60000
50000
40000
30000
20000
10000
0
Aug
Jul
Oct
Jun
May
Apr
Sep
Mar
Jan
Nov
Feb
Dec
Meses

Good examples!

Examples

Figure 3: Daily and weekly habits of annual subscribers

Figure 3.A: Number of trips by hour

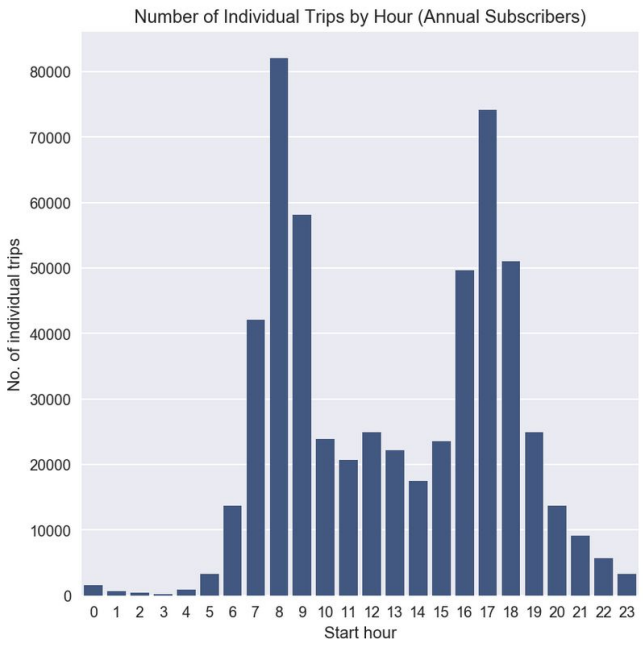
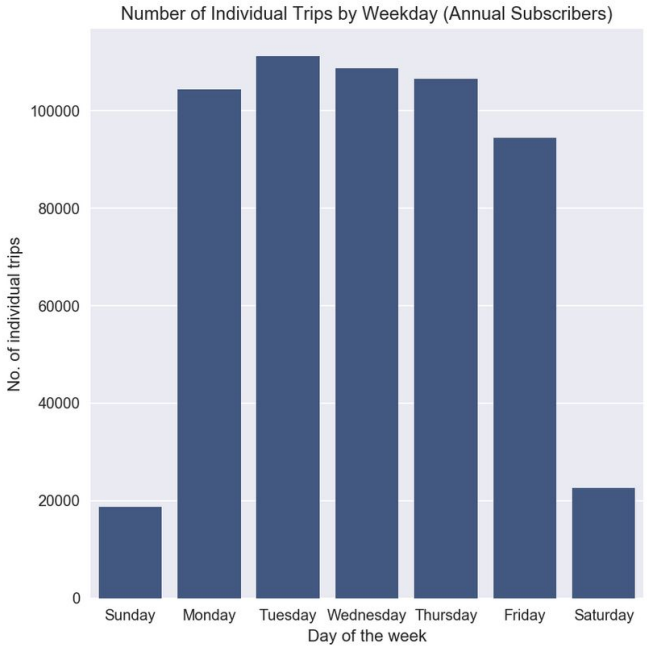


Figure 3.B: Number of trips by weekday



Examples

Figure 6: Distribution of overtime trips duration by subscription type

Figure 6.A: Overtime trips by annual subscribers

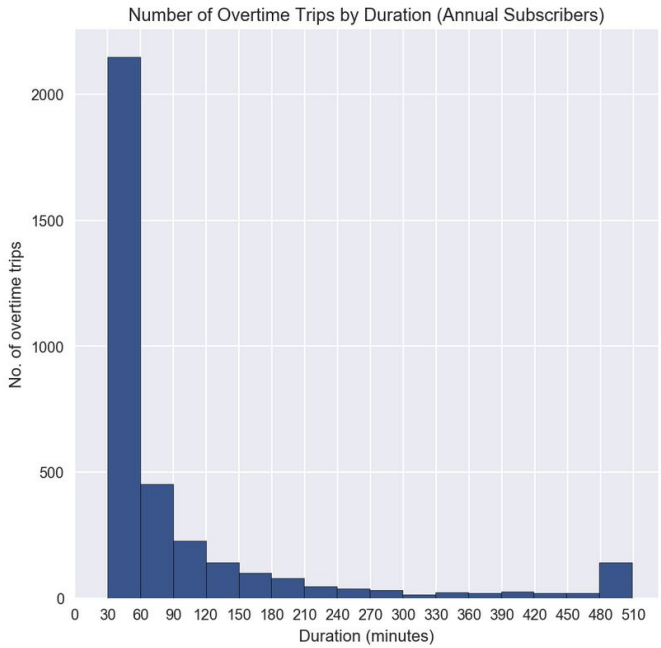
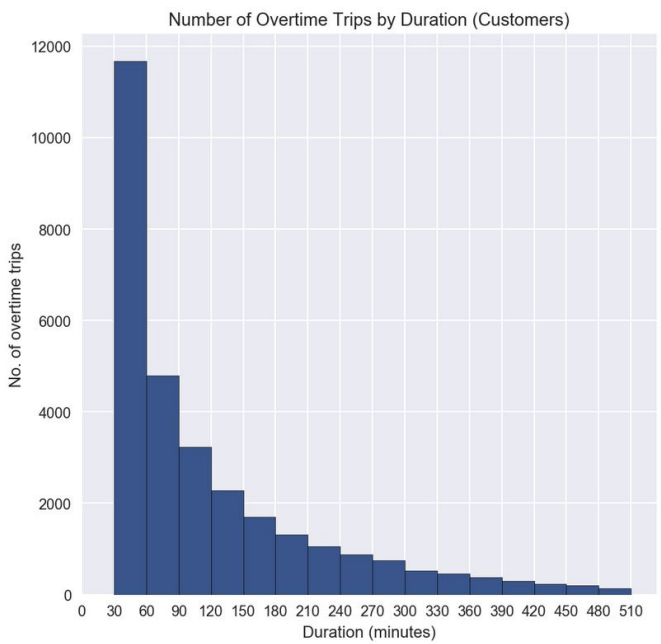
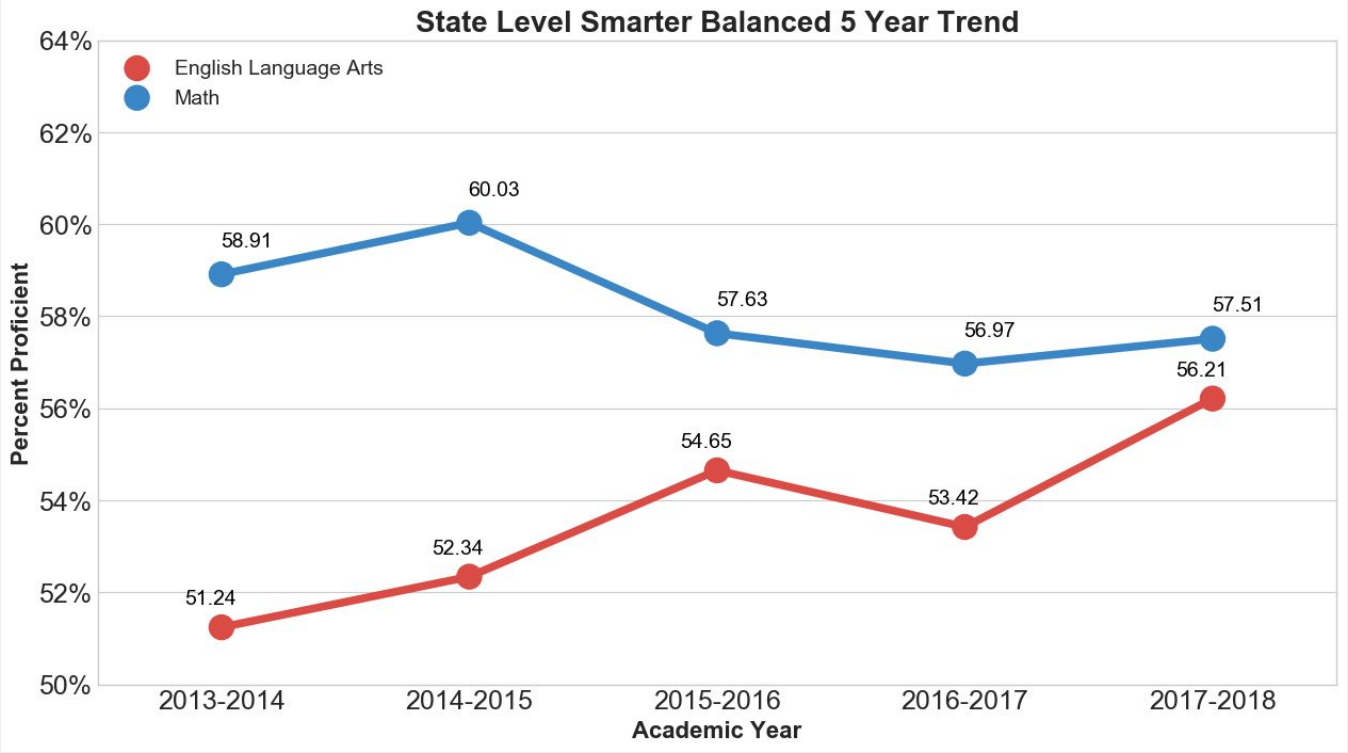


Figure 6.B: Overtime trips by customers

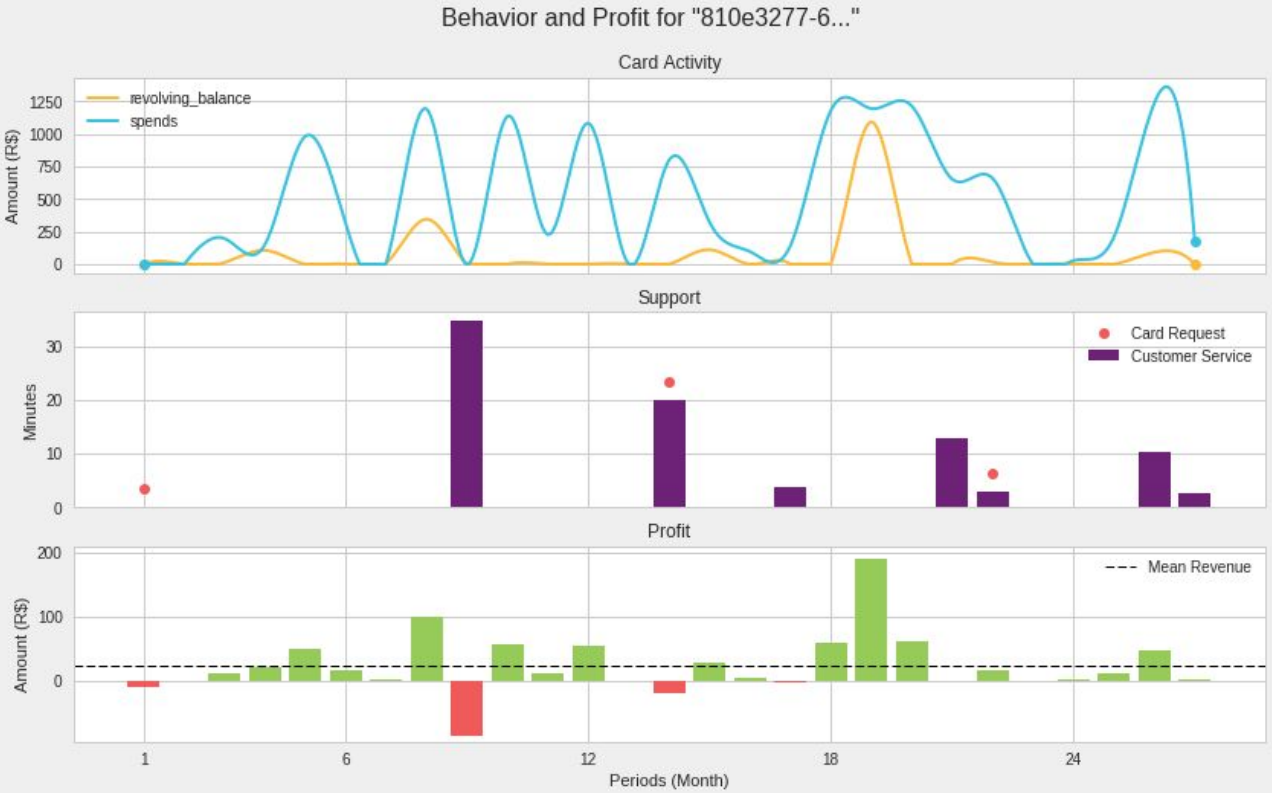


Good examples

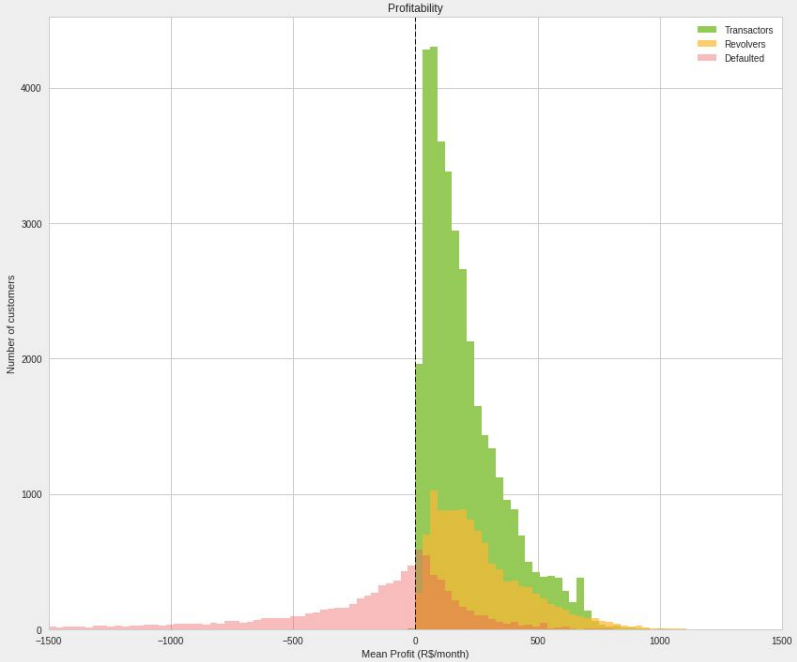


Reference: https://omegadescent.wordpress.com/2018/06/11/seaborn1/

Examples


Behavior and Profit for "810e3277-6..."
Card Activity
revolving_balance
spends
Amount (R$)
1250
1000
750
500
250
0
Support
Card Request
Customer Service
Minutes
30
20
10
0
Profit
Mean Revenue
Amount (R$)
200
100
0
1
6
12
18
24
Periods (Month)

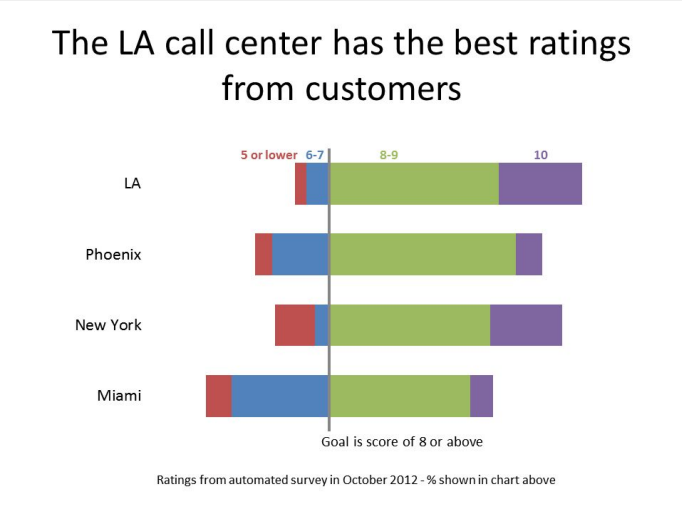
Examples


Profitability
Transactors
Revolvers
Defaulted
Number of customers
4000
3000
2000
1000
0
−1500
−1000
−500
0
500
1000
1500
Mean Profit (R$/month)

Types

Types - Deviation

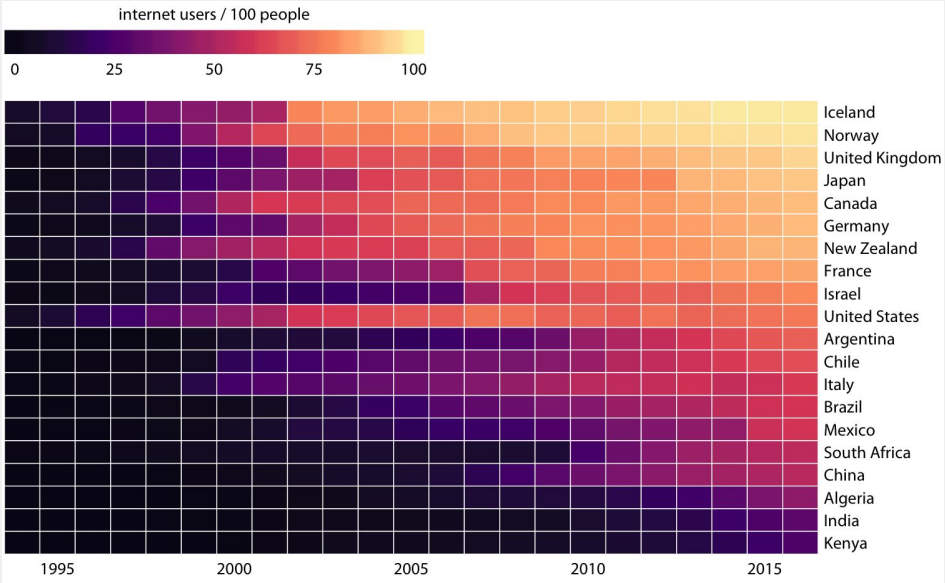
Diverging stacked bar



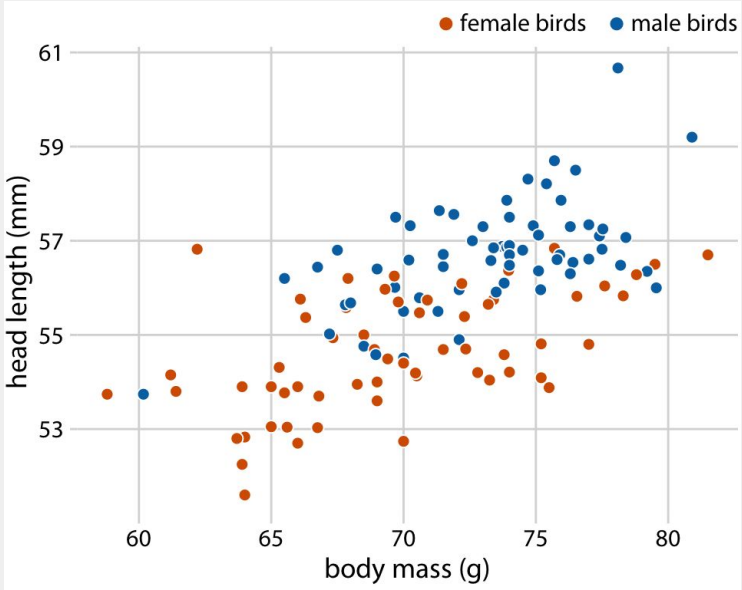
Surplus/deficit filled line

Types - Correlation

Heatmap



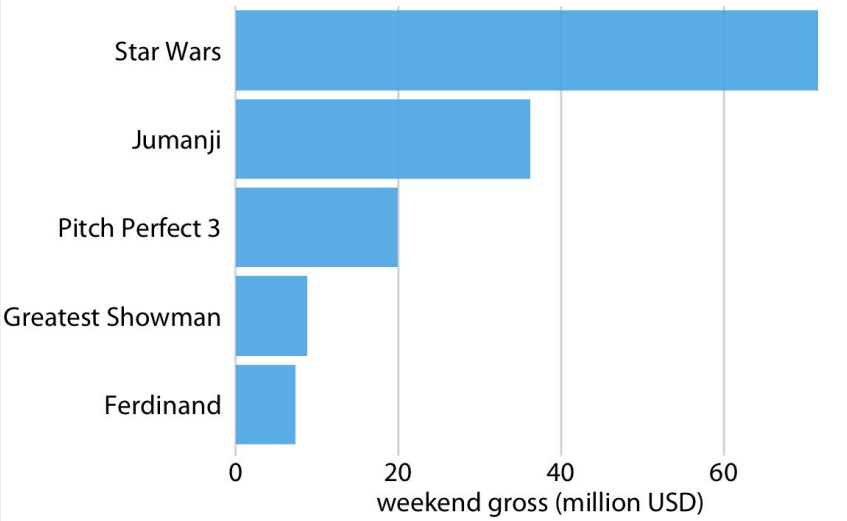
Scatterplot



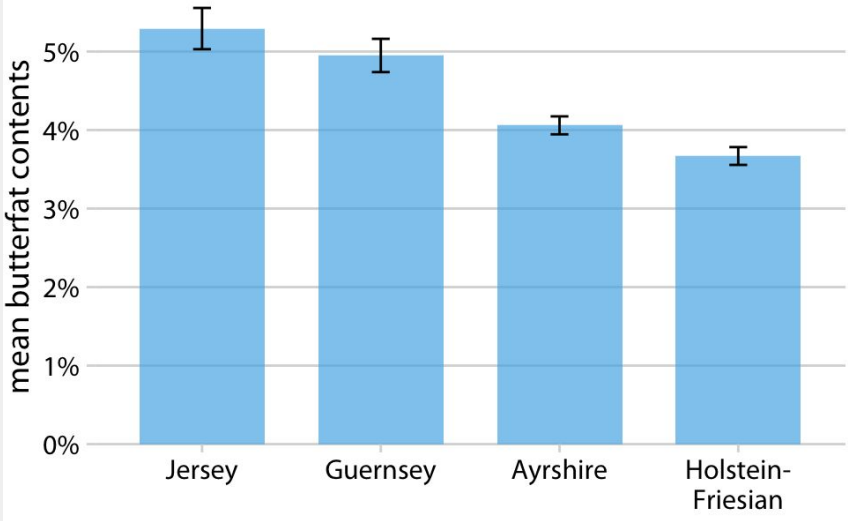
Reference: (left) https://serialmentor.com/dataviz/visualizing-amounts.html (right) https://serialmentor.com/dataviz/visualizing-associations.html

Types - Ranking

Ordered bar



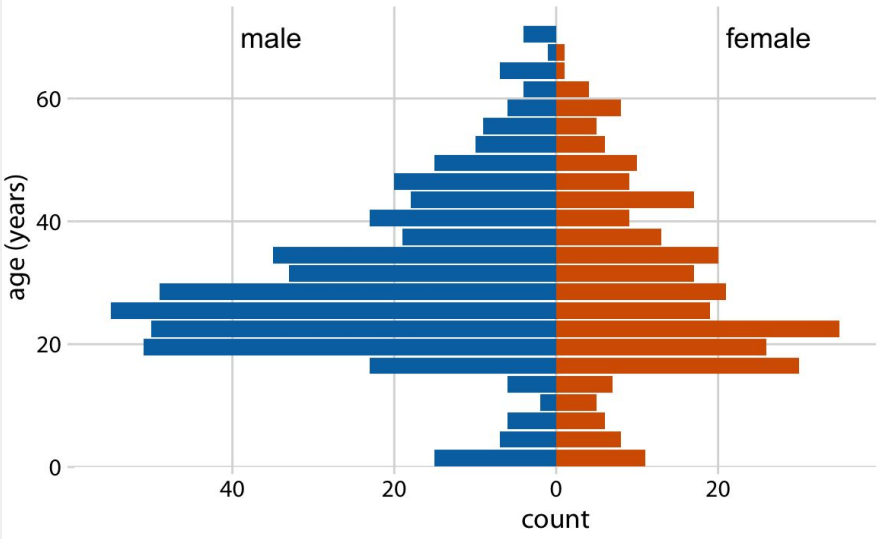
Ordered column



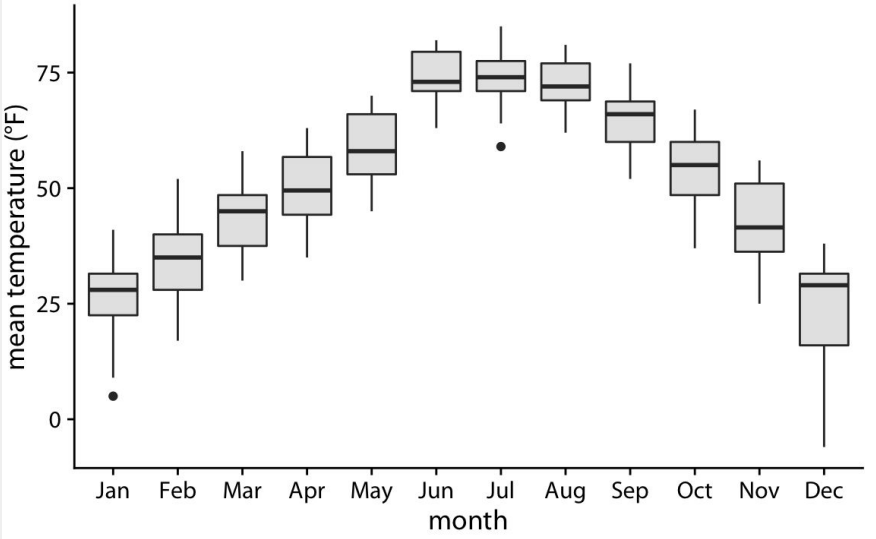
Reference: (left) https://serialmentor.com/dataviz/visualizing-amounts.html (right) https://serialmentor.com/dataviz/visualizing-uncertainty.html

Types - Distribution

Population pyramid



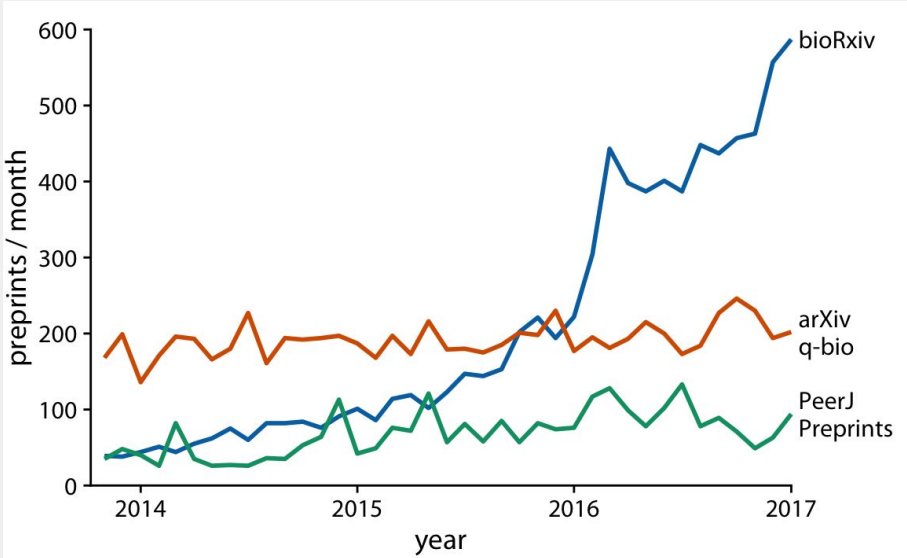
Box plot



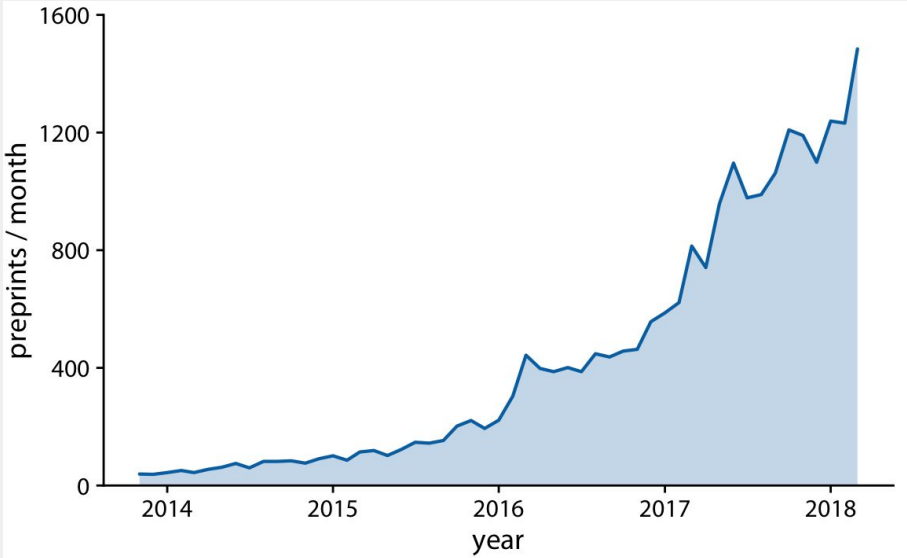
Reference: (left) https://serialmentor.com/dataviz/histograms-density-plots.html (right) https://serialmentor.com/dataviz/boxplots-violins.html

Types - Change overtime

Line plot



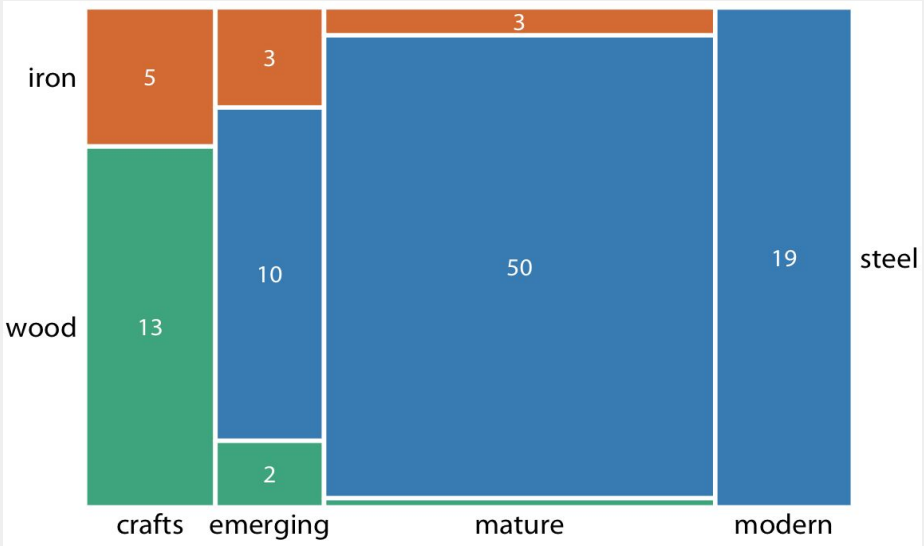
Area chart

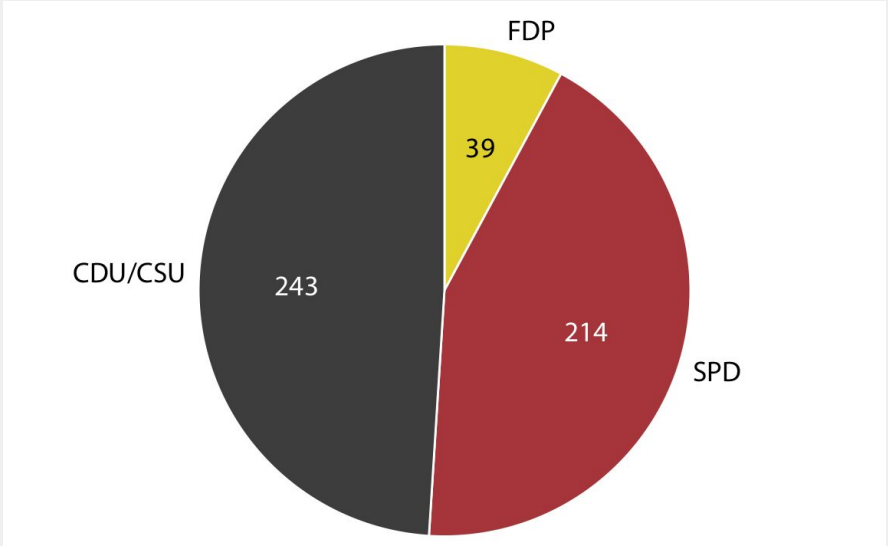


Reference: https://serialmentor.com/dataviz/time-series.html

Types - Part-to-whole

Tree map

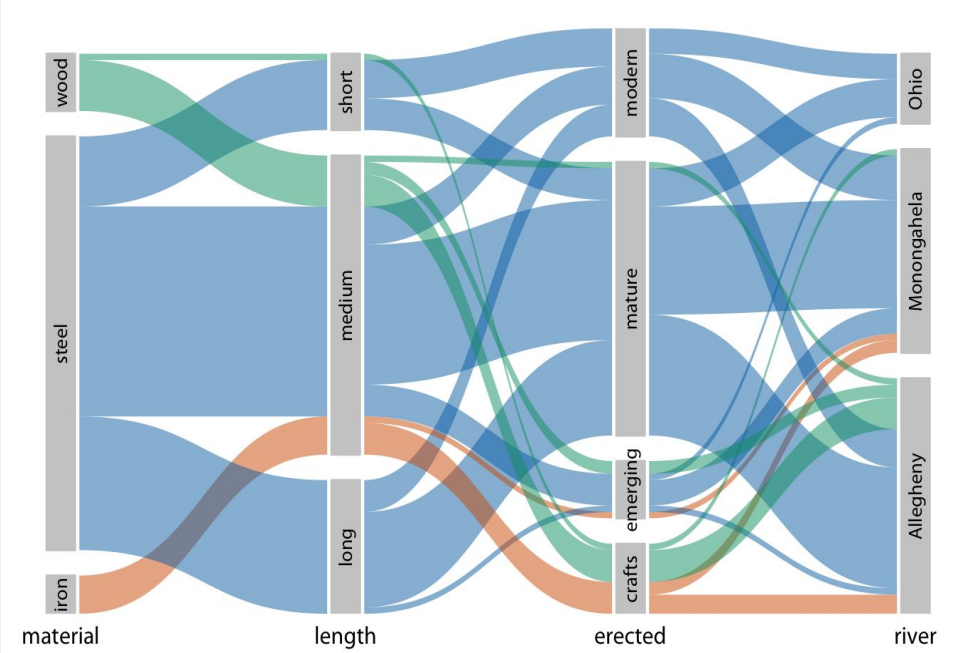




Reference: (left) https://serialmentor.com/dataviz/nested-proportions.html (right) https://serialmentor.com/dataviz/visualizing-proportions.html

Types - Flow

Sankey



Reference: https://serialmentor.com/dataviz/nested-proportions.html

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**.

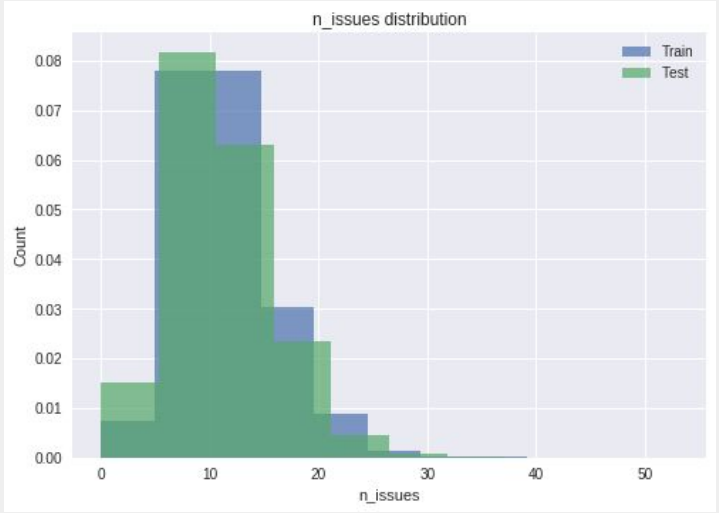
Results and presentation

Communicating a project result - Storytelling.

EDA

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

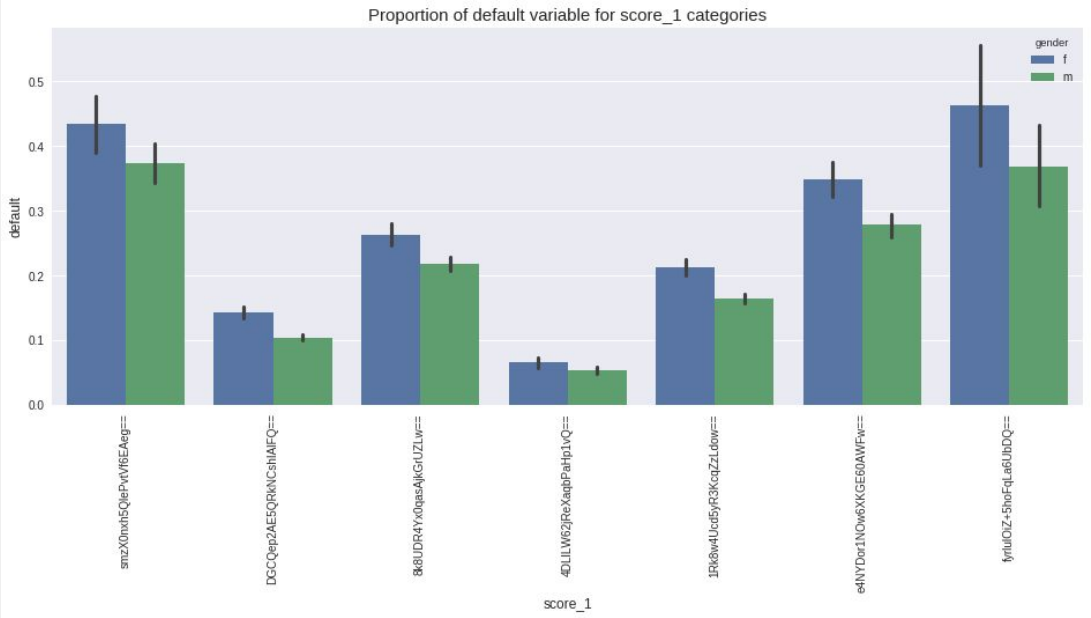
Conclusion: yes, it follows.



EDA

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

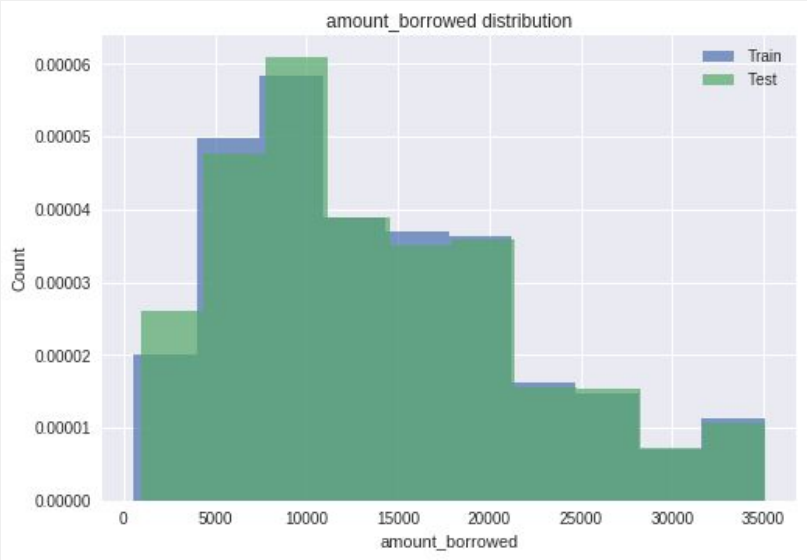
Conclusion: yes, different classes have different averages and for some it's gender sensitive.



EDA

Hypothesis: test set distribution follows the train set one.

Conclusion: yes, it follows.



EDA Examples

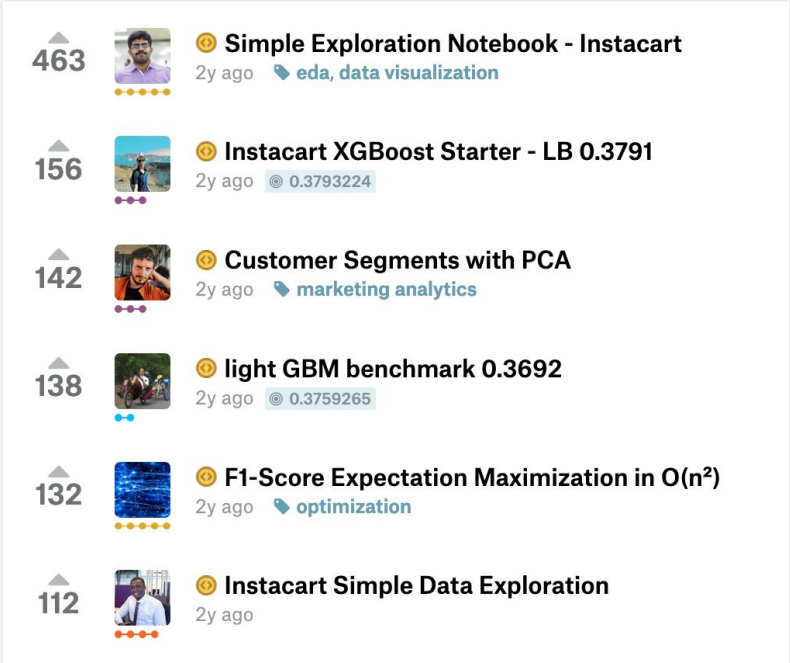
EDA

Kaggle is a great place to check EDA examples!

Warnings:

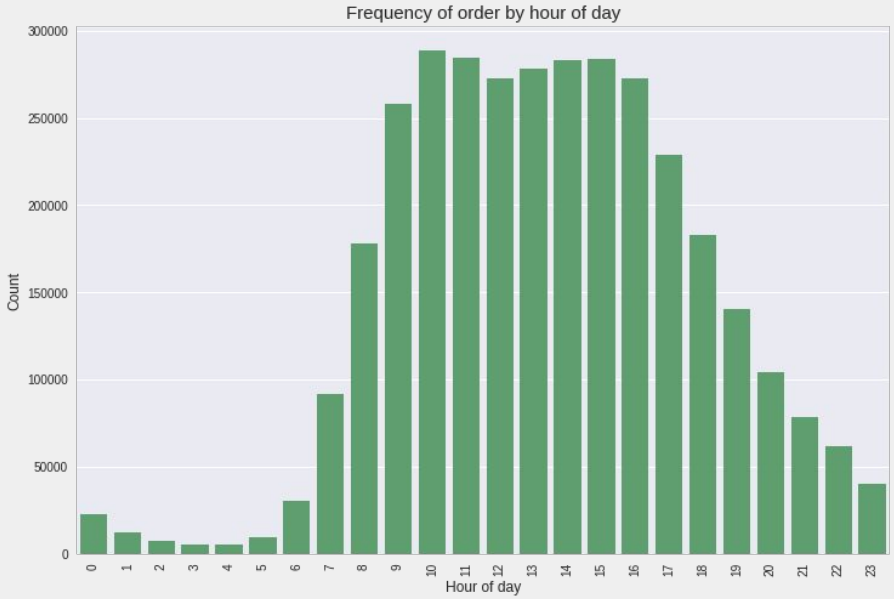
- People tend to do some useless plots there to build long kernels
- Not everyone follow the best practices for plots
- They try to do to fancy plots to impress the community and get upvotes

Instacart Market Basket Analysis competition.

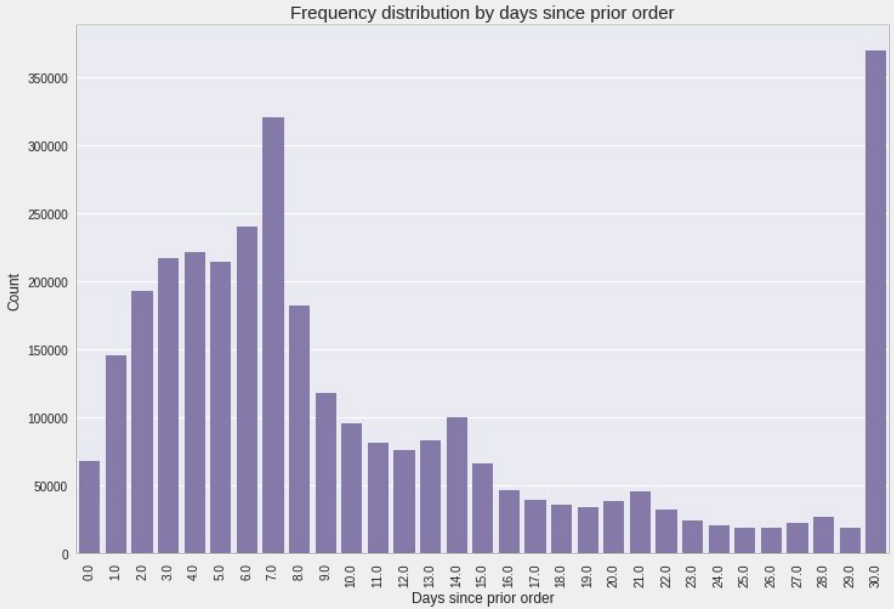


Reference: https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart

EDA


Frequency of order by hour of day
Count
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of day

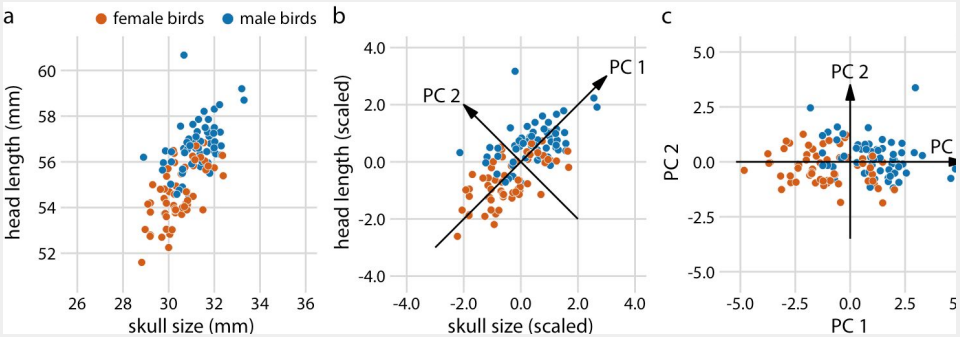
EDA

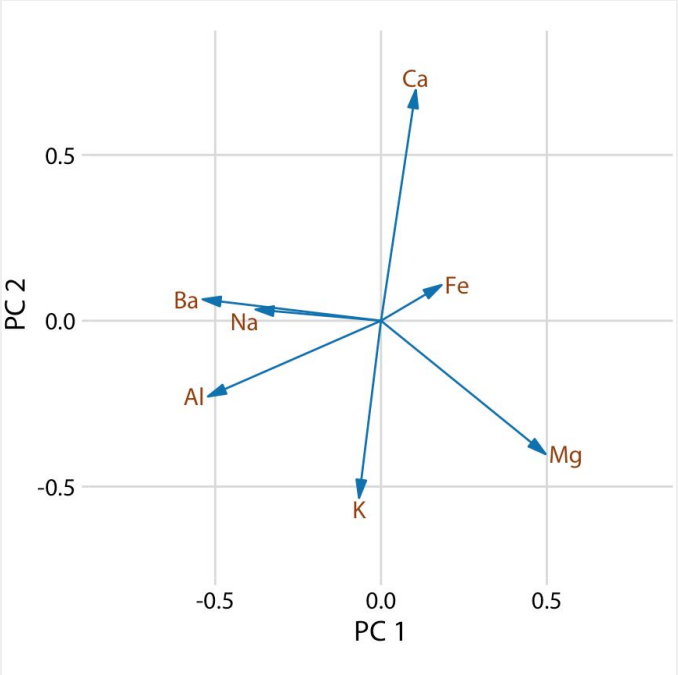

Frequency distribution by days since prior order
Count
0
50000
100000
150000
200000
250000
300000
350000
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0
Days since prior order

PCA and t-SNE

Data Science - PCA

Principal Component Analysis





Reference: https://serialmentor.com/dataviz/visualizing-associations.html

Data Science - t-SNE

t-Distributed Stochastic Neighbor Embedding

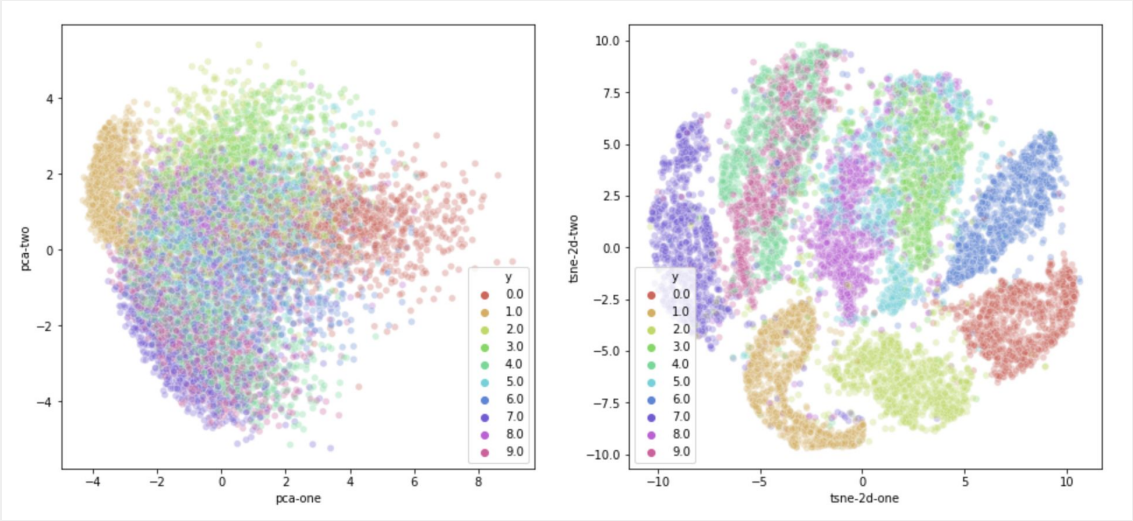
$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

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

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

Reference: http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

MNIST



Reference: https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b

Tools

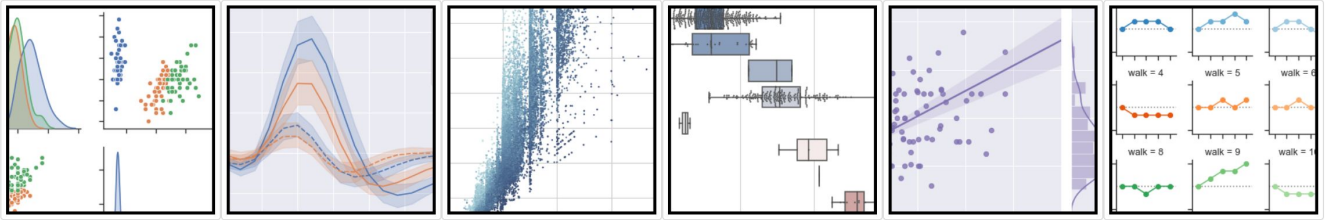
Tools

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

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.

Twitter: @lgmoneda
E-mail: lgmoneda@gmail.com
Blog: http://lgmoneda.github.io/

Questions?