Data types and standard visualizations

Anvar Kurmukov, 2022

Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) is an approach for data analysis without statistical model or formulated prior hypothesis

- Maximize insight into a data set
- Uncover underlying structure
- Detect missing data
- Detect outliers and anomalies
- Rank important factors
- Perform sanity check

Descriptive statistics: computing simple summary statistics such as mean, median, standard deviation, plotting box plots, histograms

Visualization: plotting the raw data - data traces, scatter plots, frequency plots, probability plots, multivariate plots

John Tukey, Exploratory Data Analysis, 1977

Data types

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False

- TABLE ROWS = instances, examples, data points, observations, samples
- TABLE COLUMNS = attributes, features, variables

- Categorical data (= labels, nominal, ordinal [ordered], binary)
- Quantitative data (= numbers, discrete [integer], continues [real])

EDA tools Methods and approaches

Summary statistics

- min, max (range)
- mean, median (location)
- variance, standard deviation (dispersion)
- skewness (asymmetry)
- kurtosis (peakedness)

Visualization

- Bar plots
- Scatter plots
- Histograms
- Box plots
- Index chart

Mean

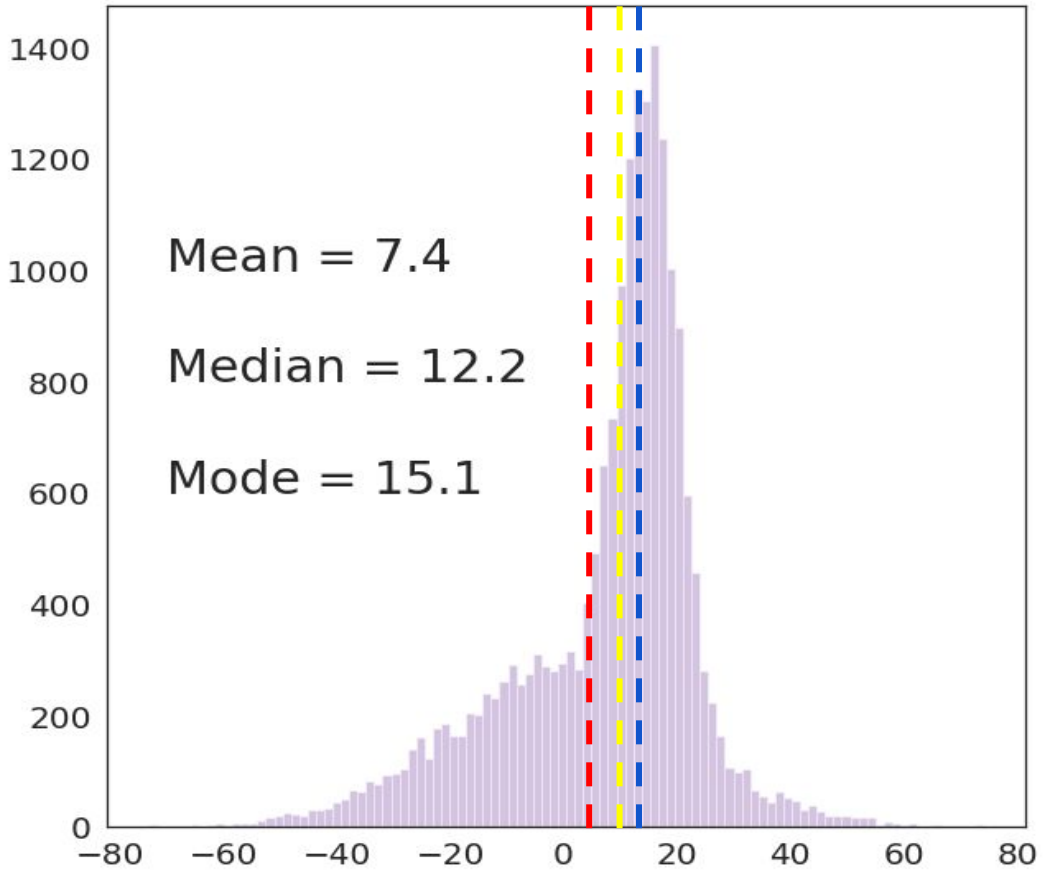
arithmetic mean

Median

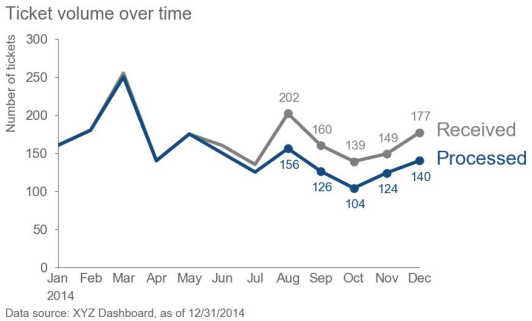
50-th percentile

Mode

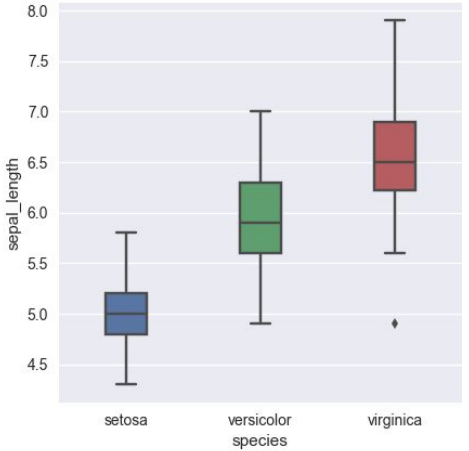
the most frequent value

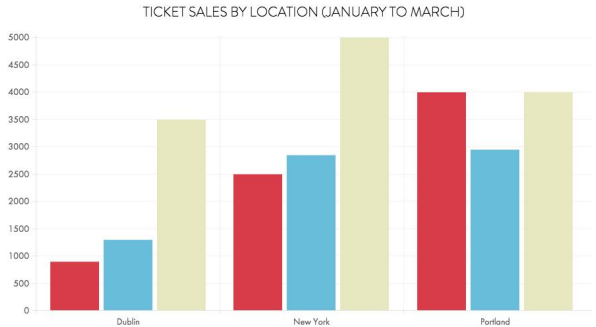


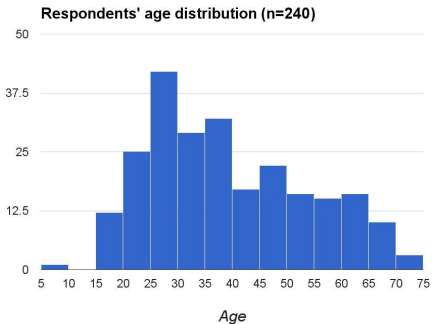
Standard graphs

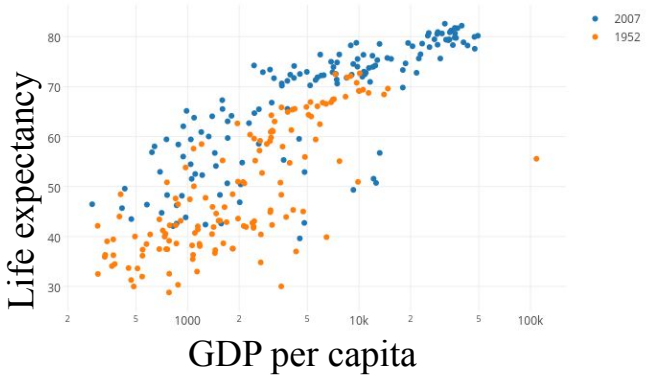


1. Index chart
2. Bar chart (pie chart)
3. Box plot
4. Scatter plot
5. Histogram



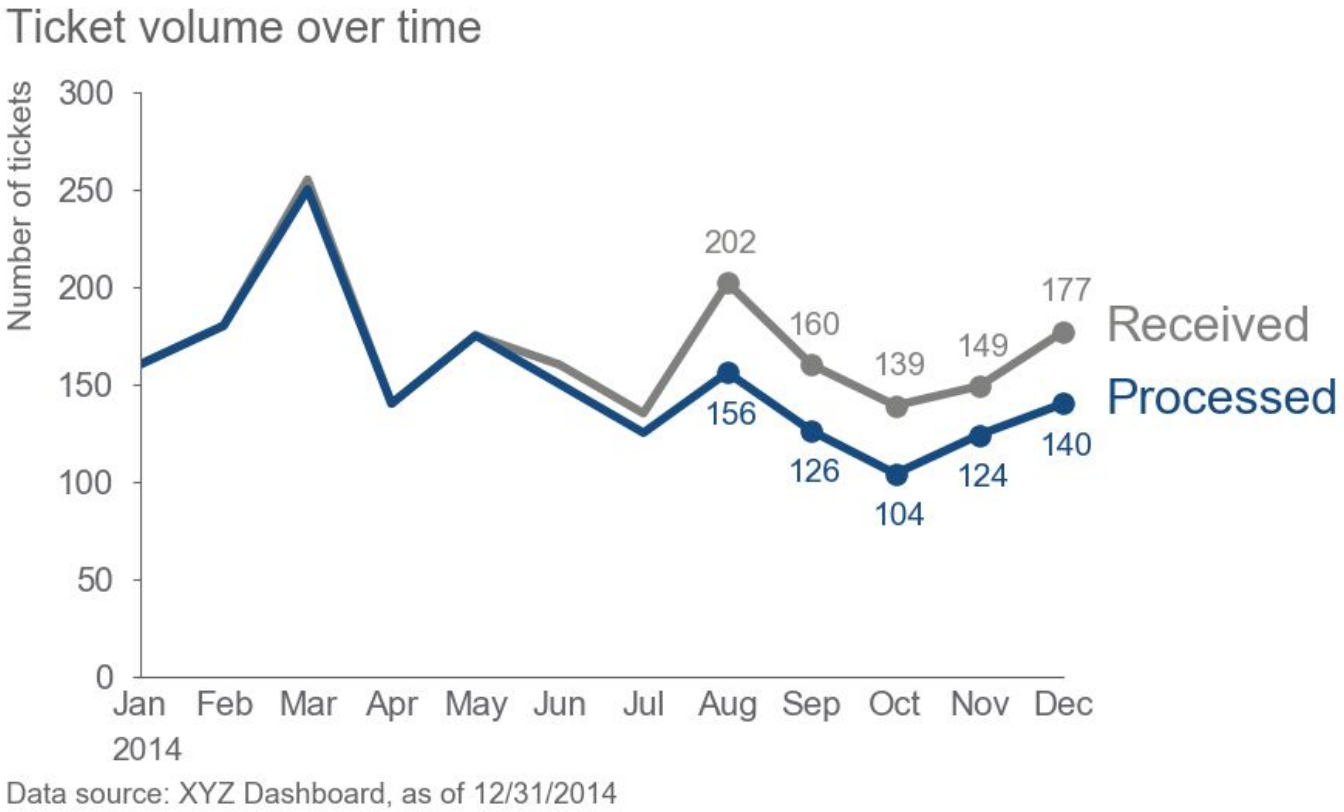




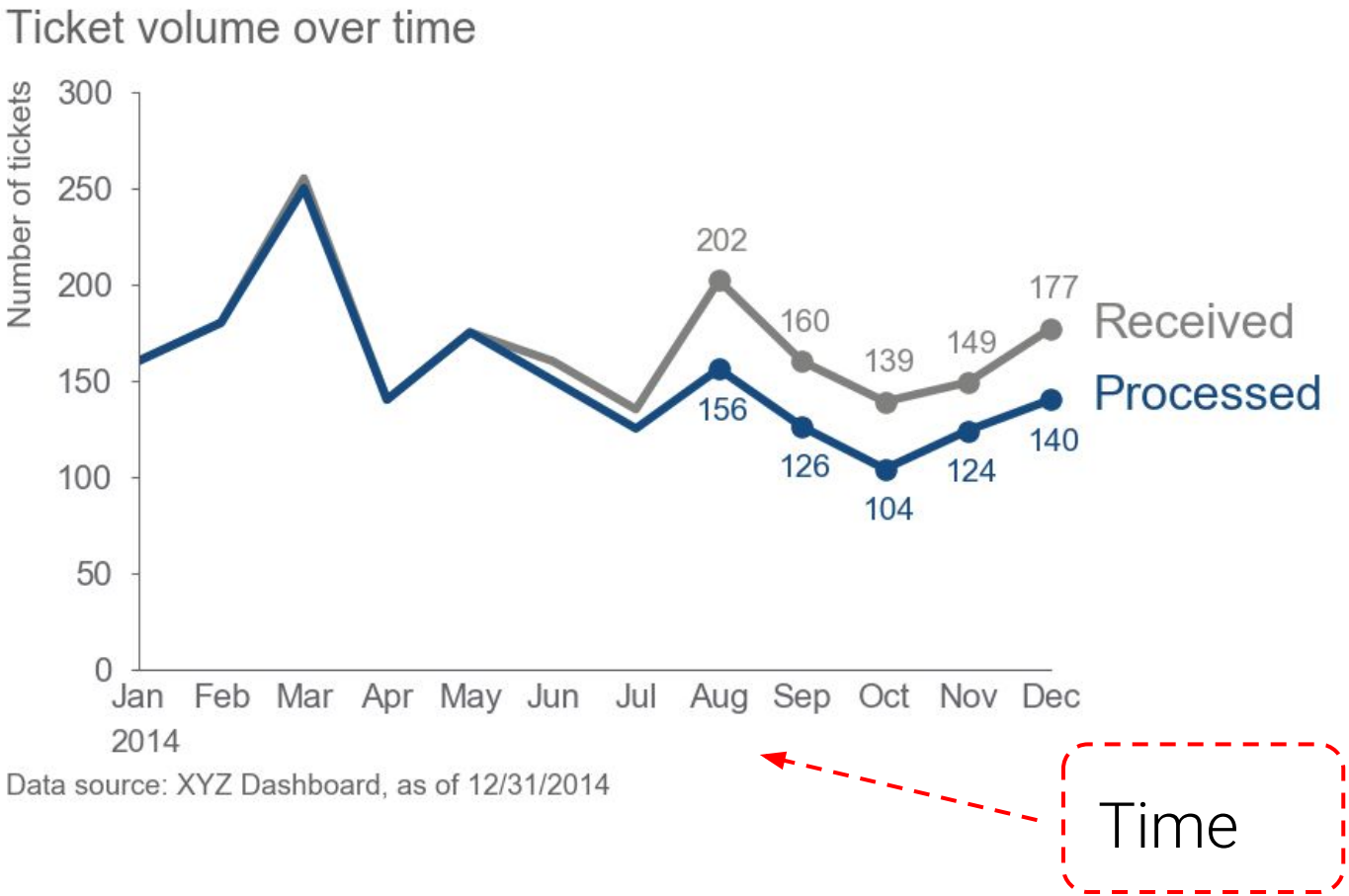


Index chart

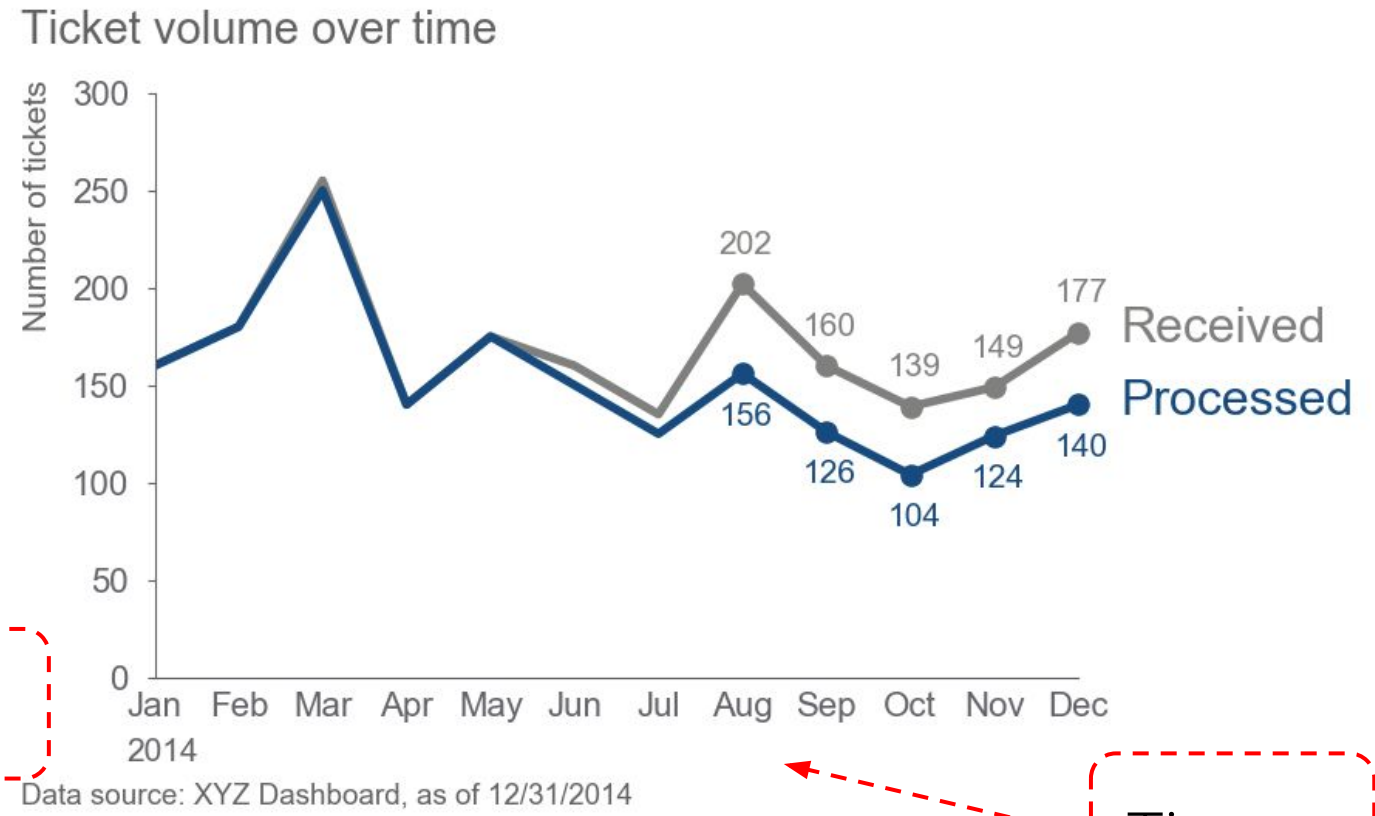
Index chart


Ticket volume over time
Number of tickets
300
250
200
150
100
50
0
202
177
160
139
149
156
126
104
124
140
Received
Processed
Jan
2014
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Data source: XYZ Dashboard, as of 12/31/2014

Index chart


Ticket volume over time
Number of tickets
300
250
200
150
100
50
0
Jan
2014
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
202
160
139
149
177
156
126
104
124
140
Received
Processed
Data source: XYZ Dashboard, as of 12/31/2014
Time

Index chart



Amount

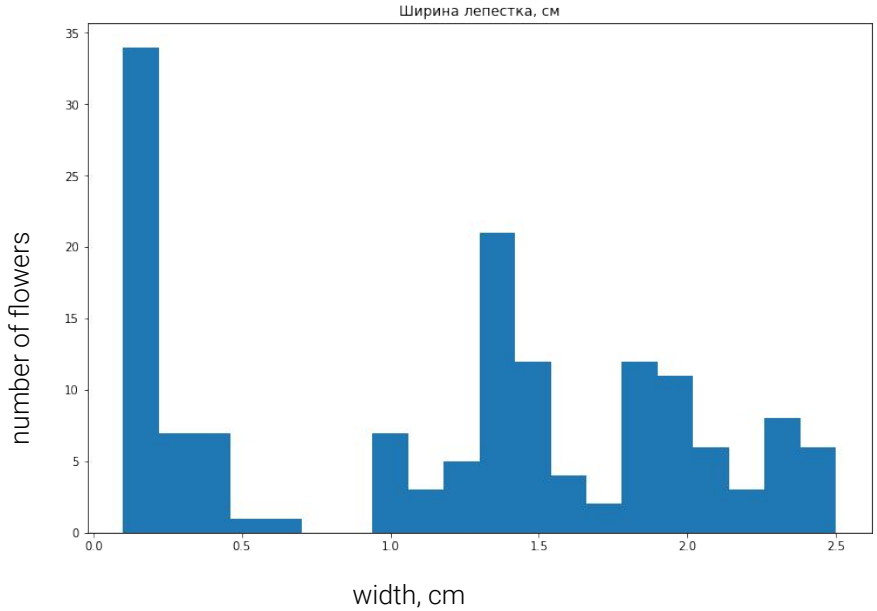
Time

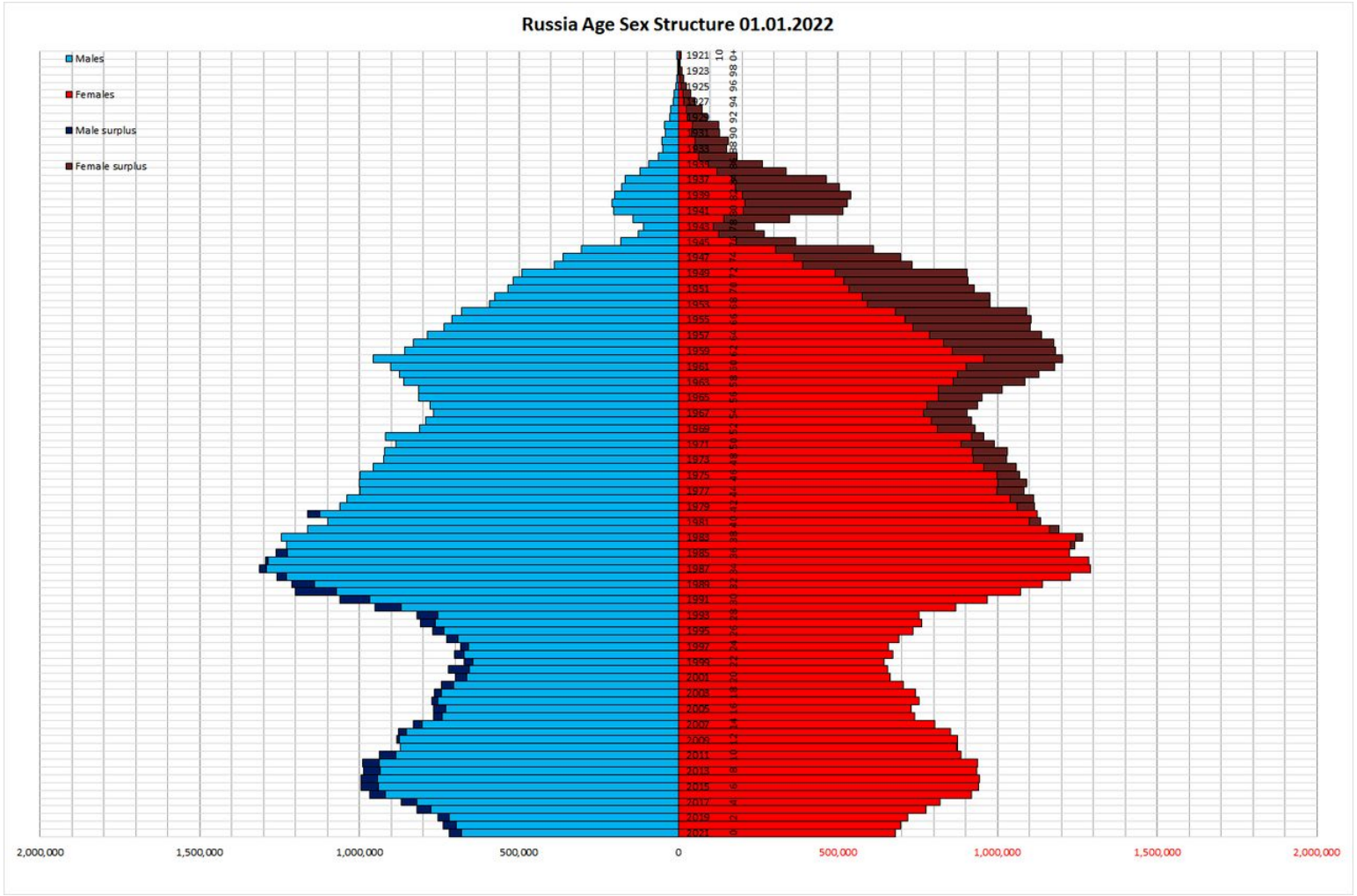
Histogram

Ирис щетинистый
(*Iris setosa*)

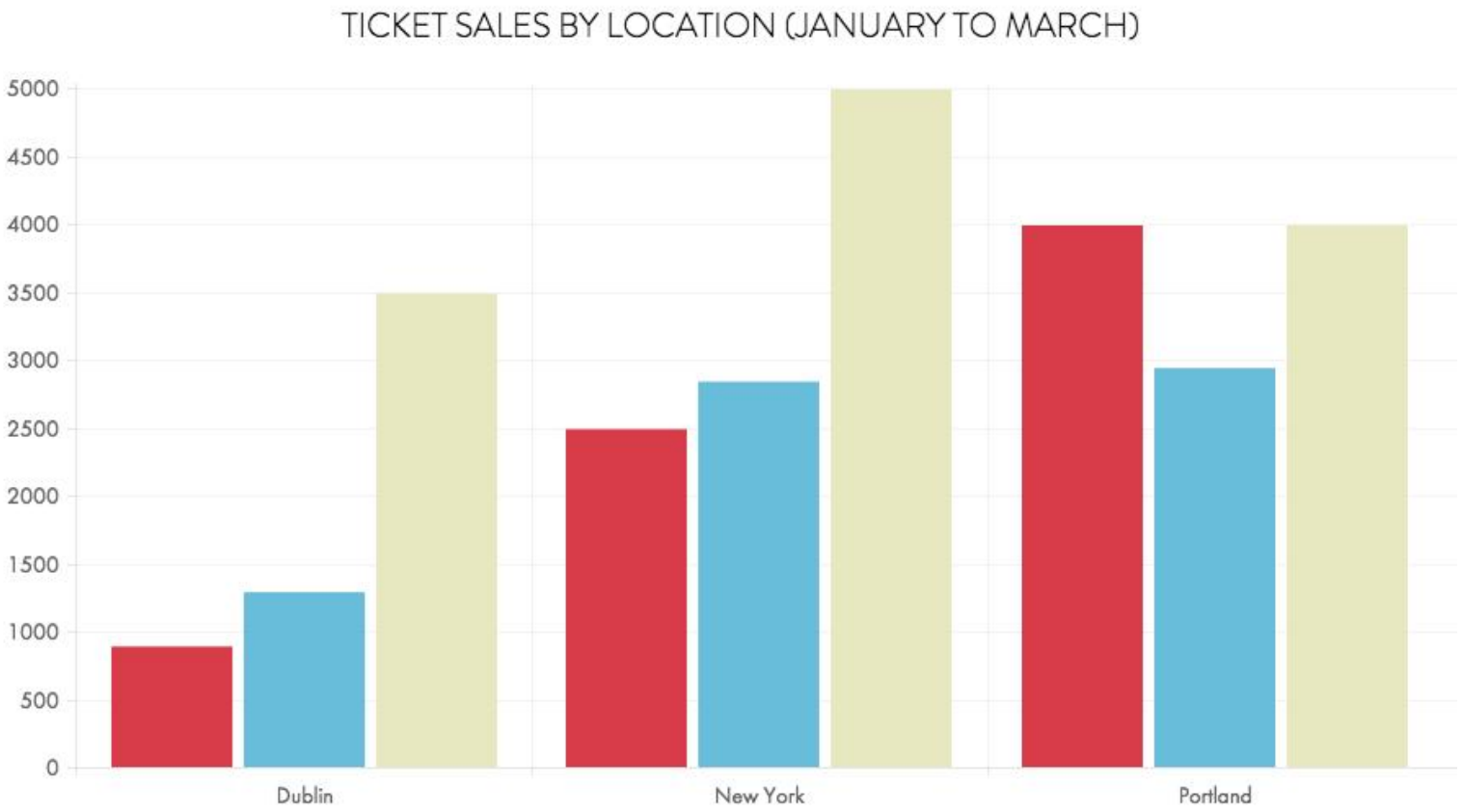
Ирис разноцветный
(*Iris versicolor*)

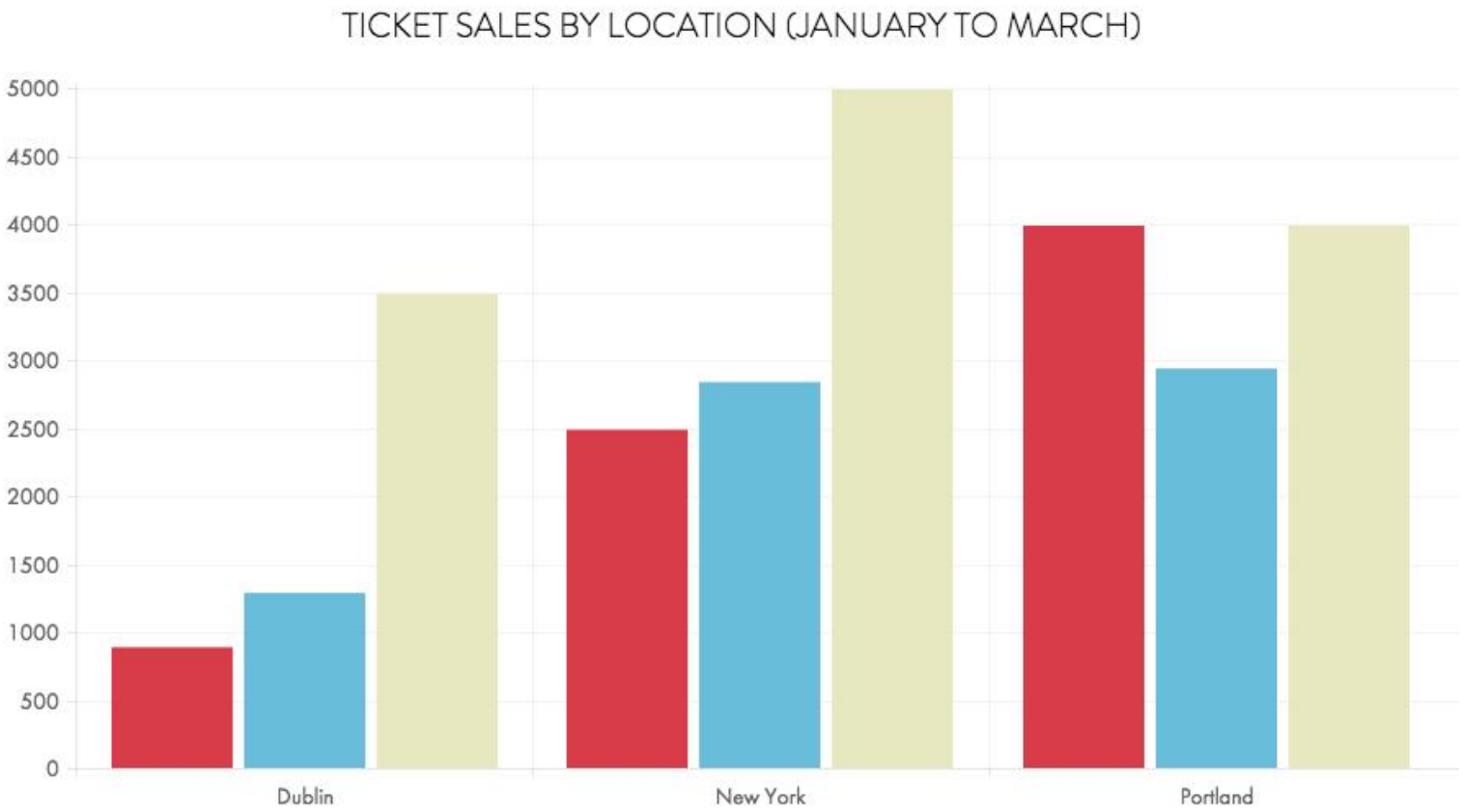
Ирис виргинский
(*Iris virginica*)



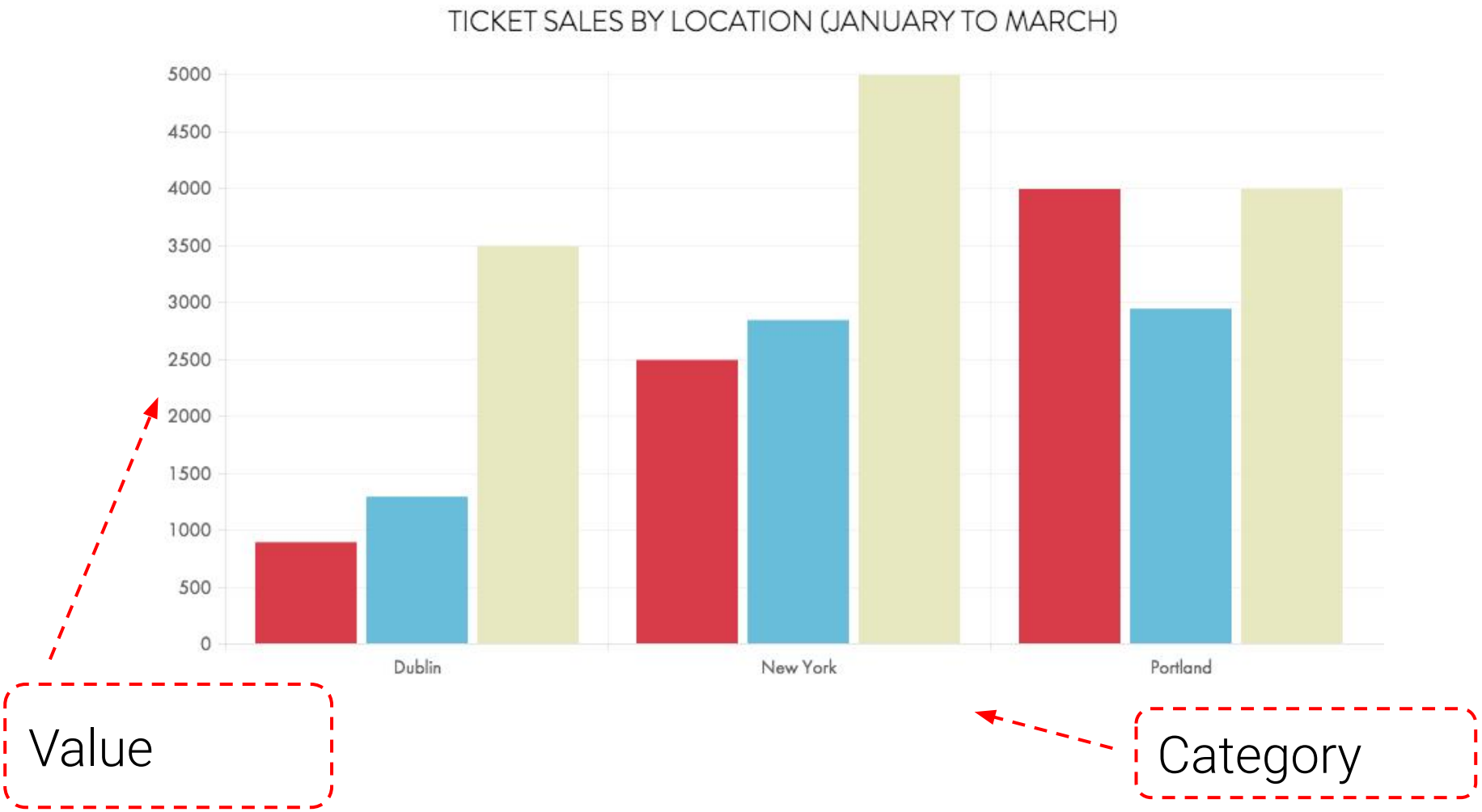

Russia Age Sex Structure 01.01.2022
Males
Females
Male surplus
Female surplus
2,000,000
1,500,000
1,000,000
500,000
0
500,000
1,000,000
1,500,000
2,000,000

Bar chart

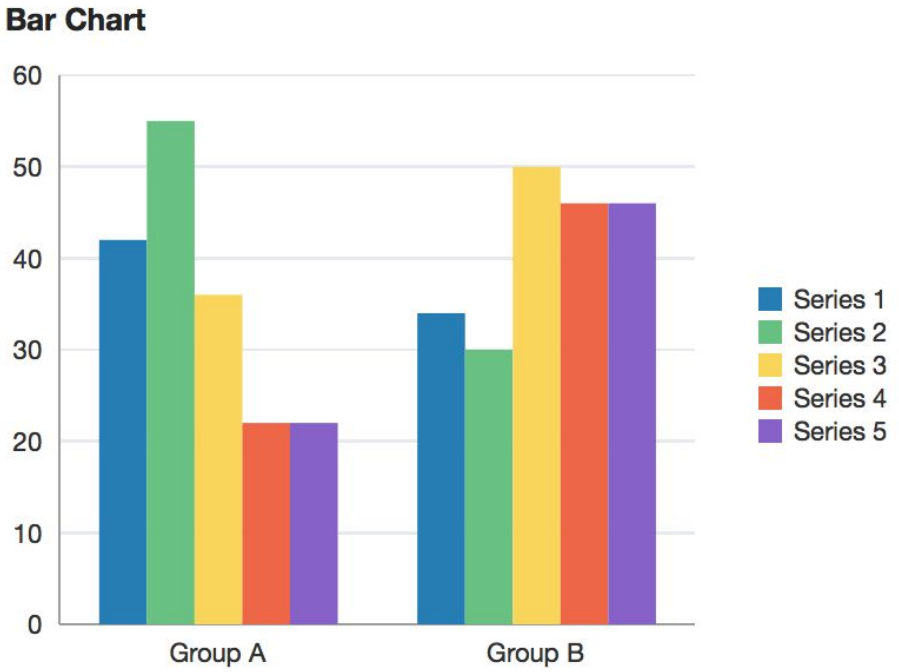

TICKET SALES BY LOCATION (JANUARY TO MARCH)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Dublin
New York
Portland

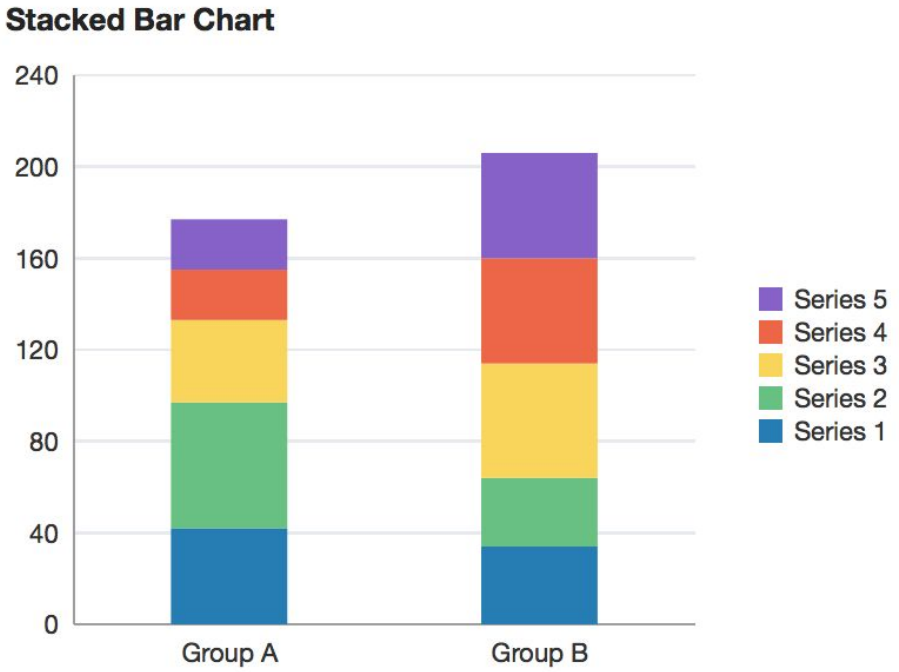

TICKET SALES BY LOCATION (JANUARY TO MARCH)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Dublin
New York
Portland

Category

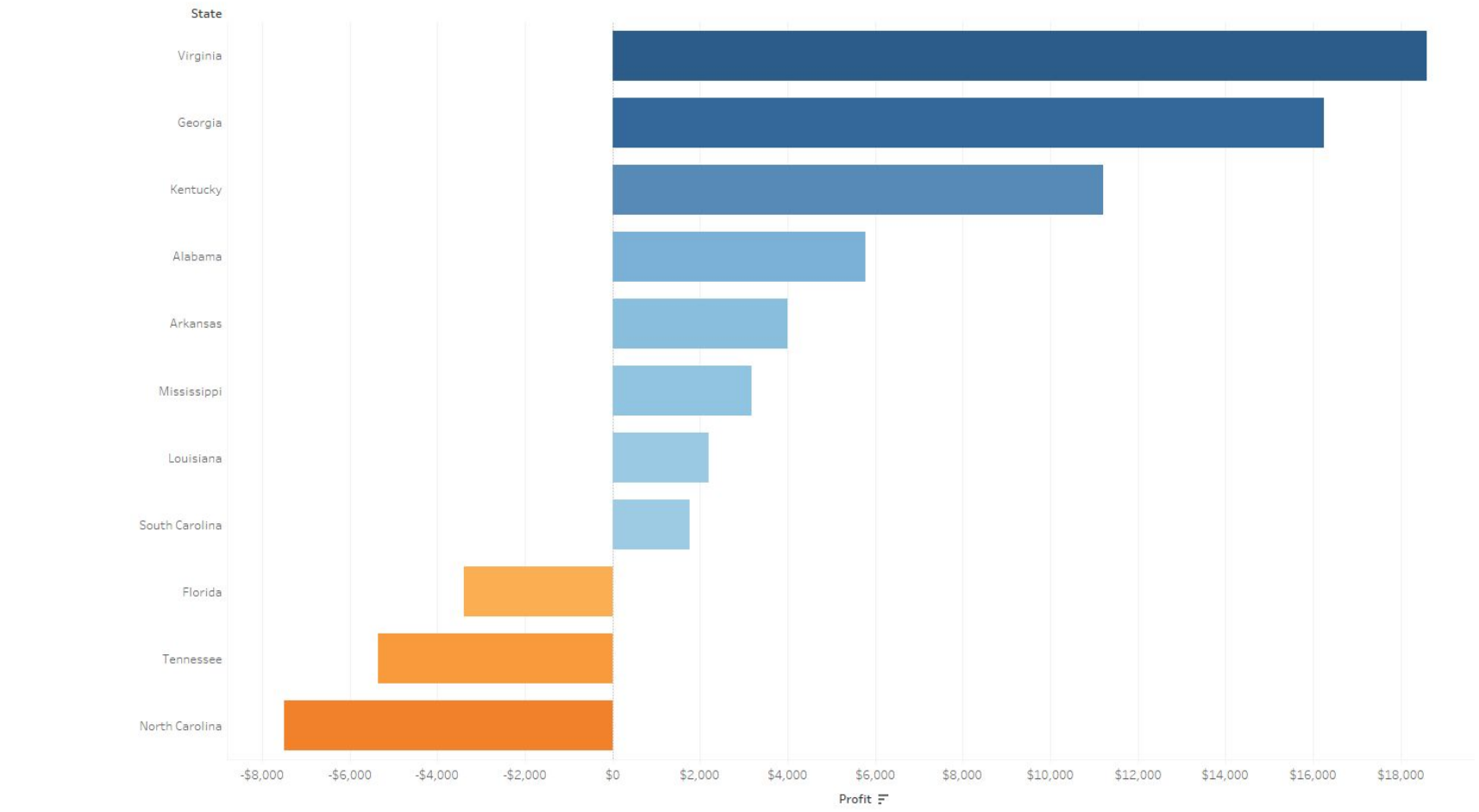

TICKET SALES BY LOCATION (JANUARY TO MARCH)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Dublin
New York
Portland
Value
Category

Stacked bar chart

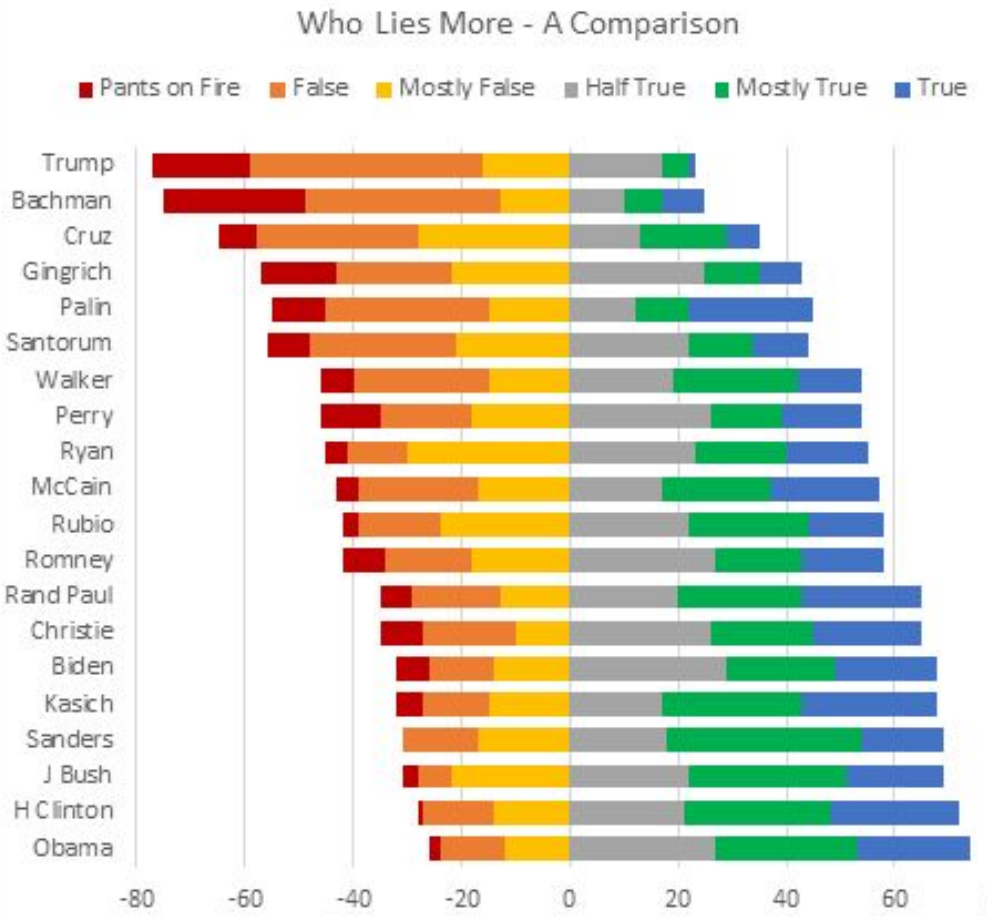




Vertical bar chart



Vertical stacked bar chart



Box-and-whiskers diagram (box plot)

Ящик с усами

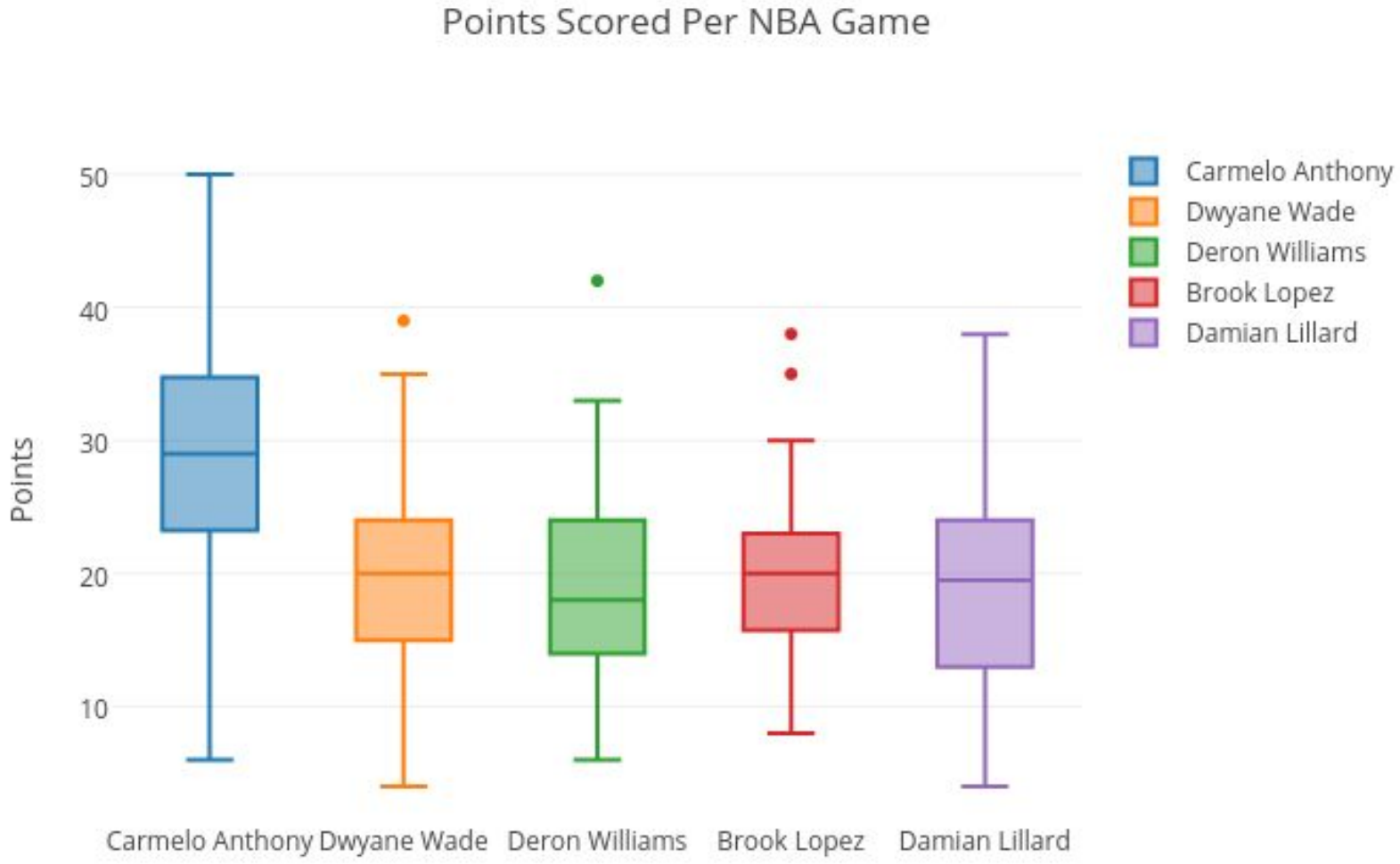
Материал из Википедии — свободной энциклопедии

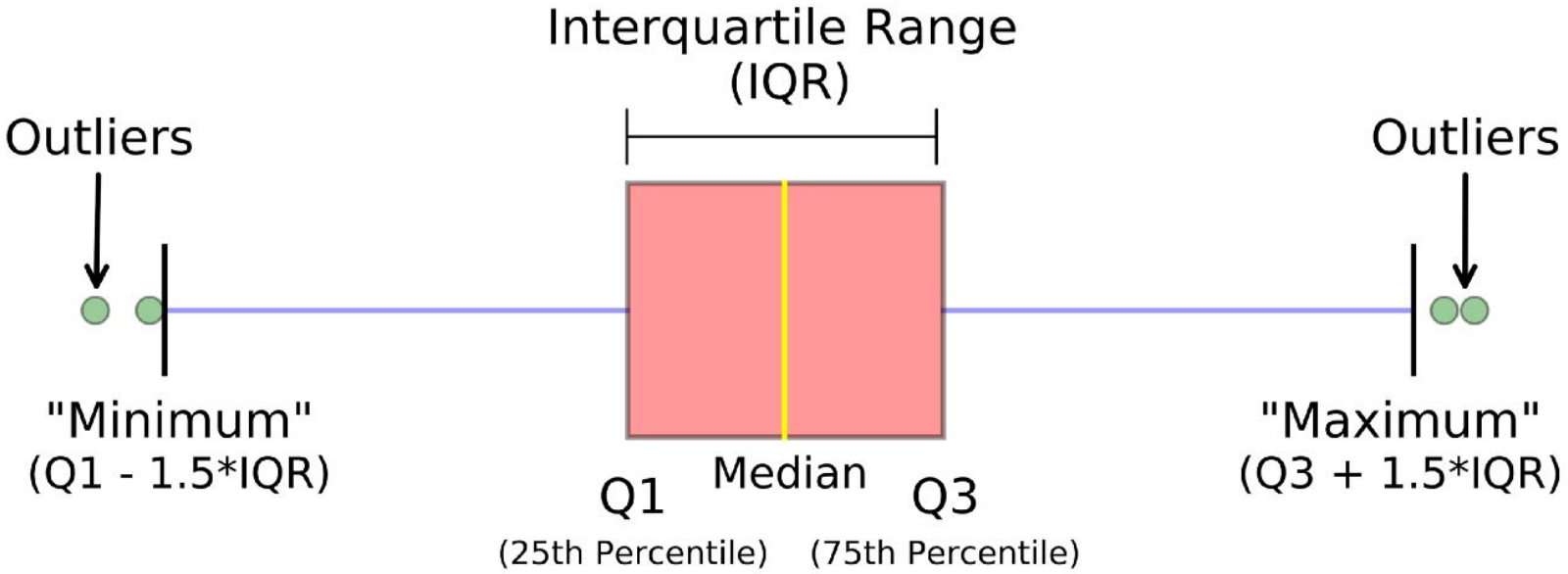
Не следует путать с японскими свечами.

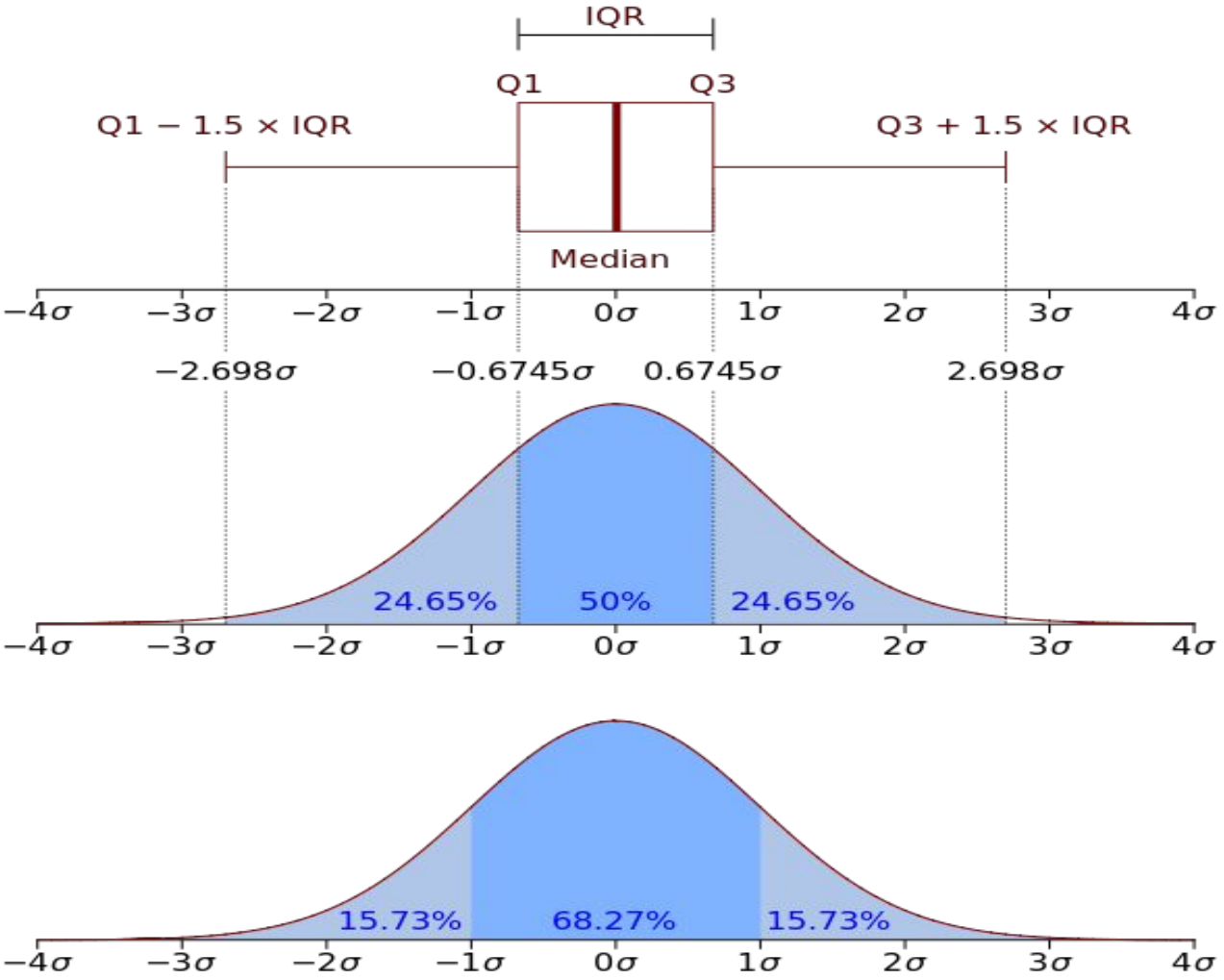
Ящик с усами, **диаграмма размаха** (англ. *box-and-whiskers diagram or plot, box plot*) вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, сред

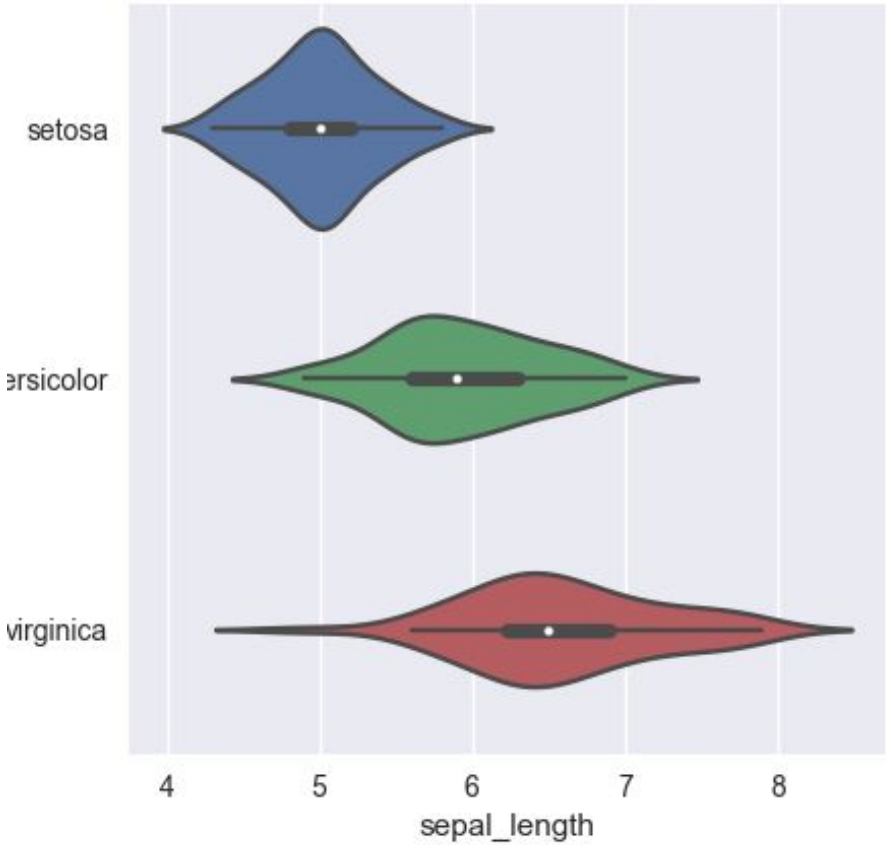
Wikipedia

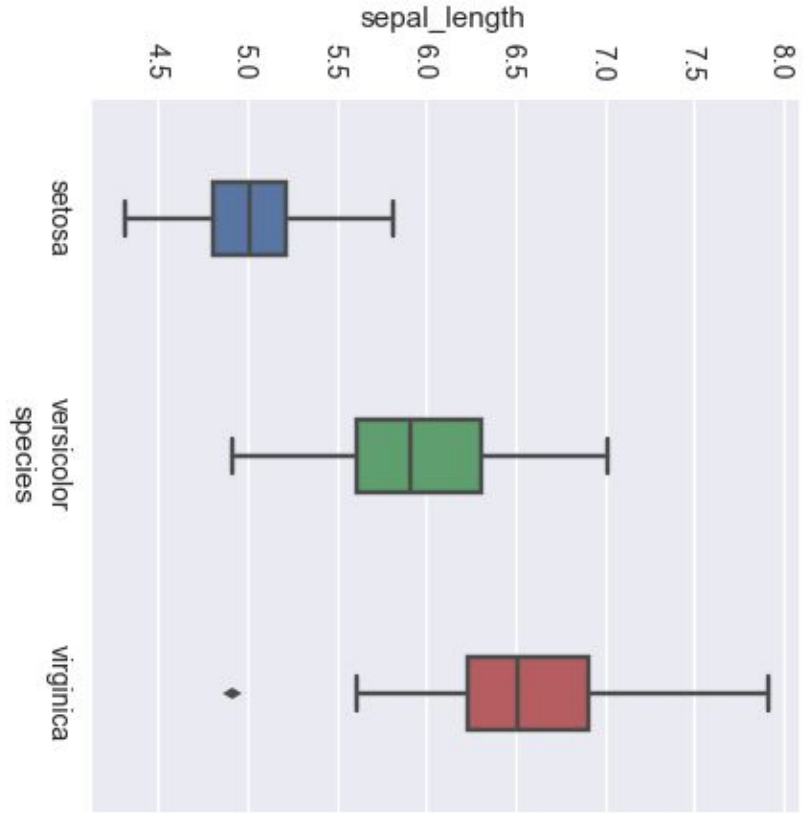

Points Scored Per NBA Game
Points
50
40
30
20
10
Carmelo Anthony
Dwyane Wade
Deron Williams
Brook Lopez
Damian Lillard
Carmelo Anthony
Dwyane Wade
Deron Williams
Brook Lopez
Damian Lillard

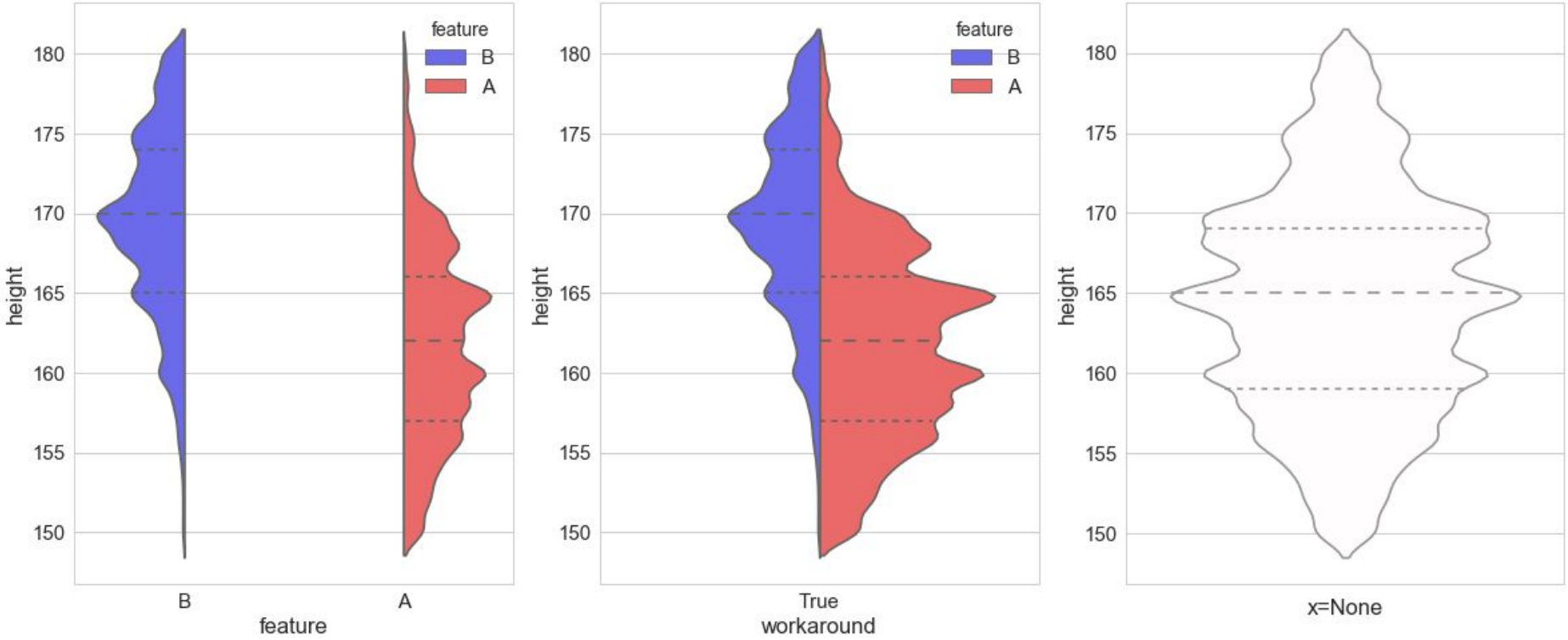

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

Violin plot


setosa
ersicolor
virginica
4
5
6
7
8
sepal_length


sepal_length
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
setosa
versicolor
virginica
species


feature
B
A
height
180
175
170
165
160
155
150
B
A
feature
feature
B
A
height
180
175
170
165
160
155
150
True
workaround
height
180
175
170
165
160
155
150
x=None

Pie chart

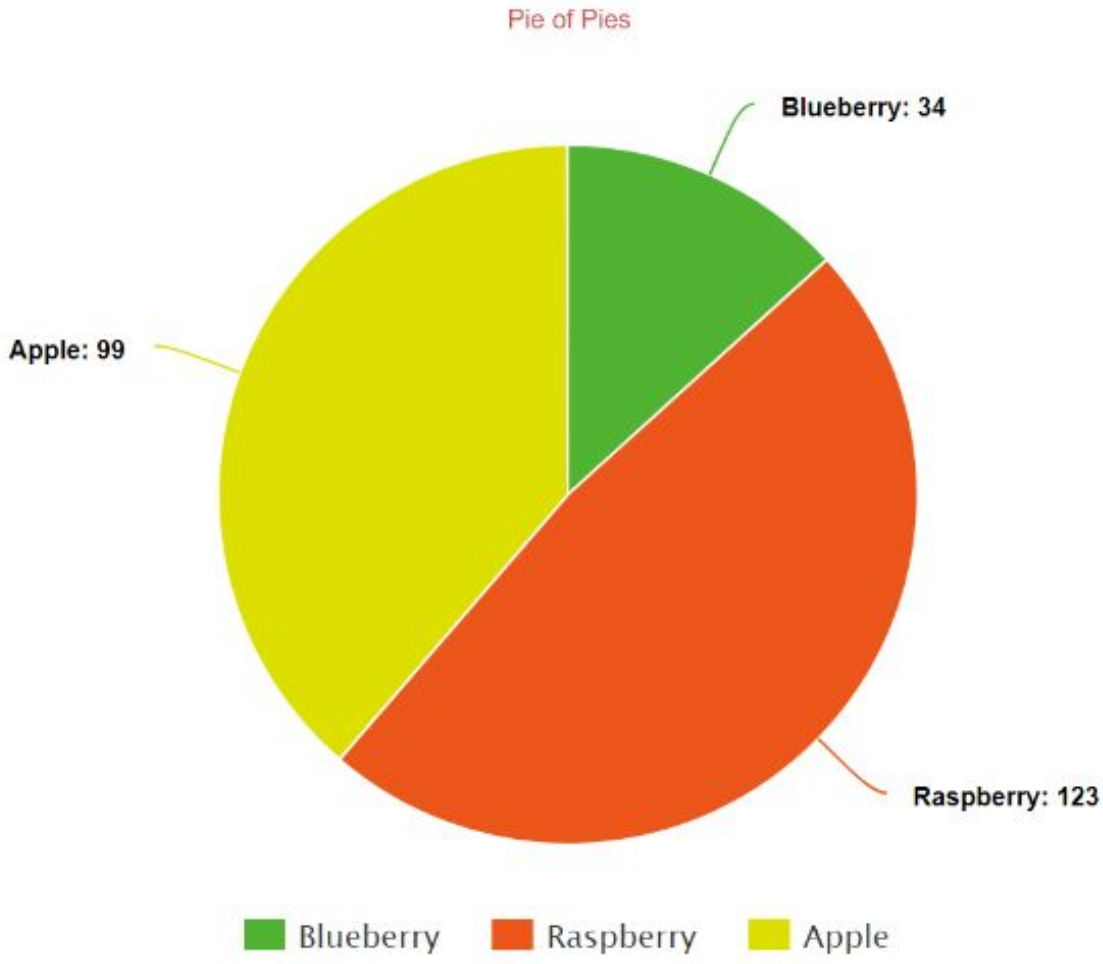
Пицца – это круговая диаграмма, показывающая, сколько у тебя осталось пиццы.

101 6

Пирог тоже
today at 8:42 2

Умно
today at 8:46

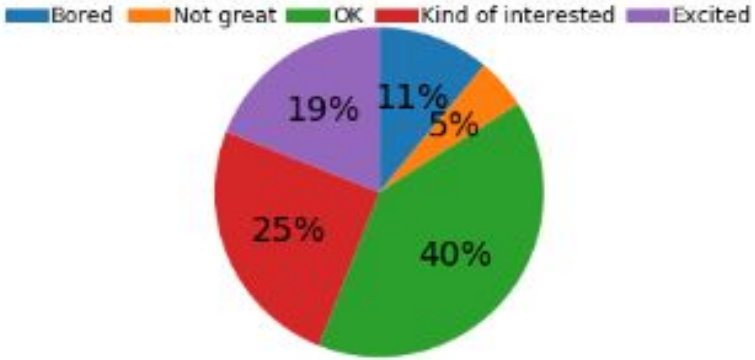
Вадим, пирог не показывает, сколько у тебя осталось пиццы. Так что - нет, не то-же.
today at 8:46 to Vadim 24


Pie of Pies
Blueberry: 34
Apple: 99
Raspberry: 123
Blueberry
Raspberry
Apple

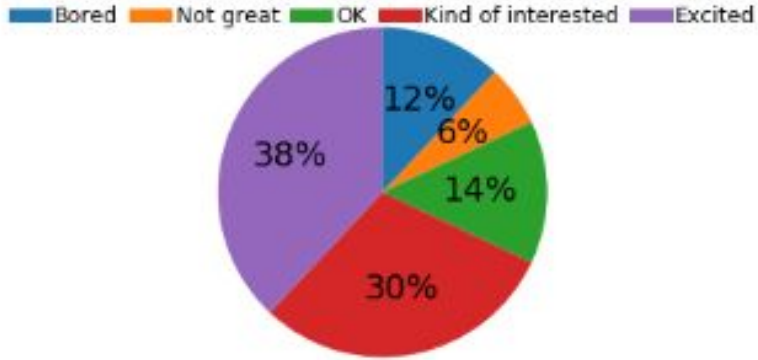
Pie charts are the worst

Survey Results

PRE: How do you feel about doing science?

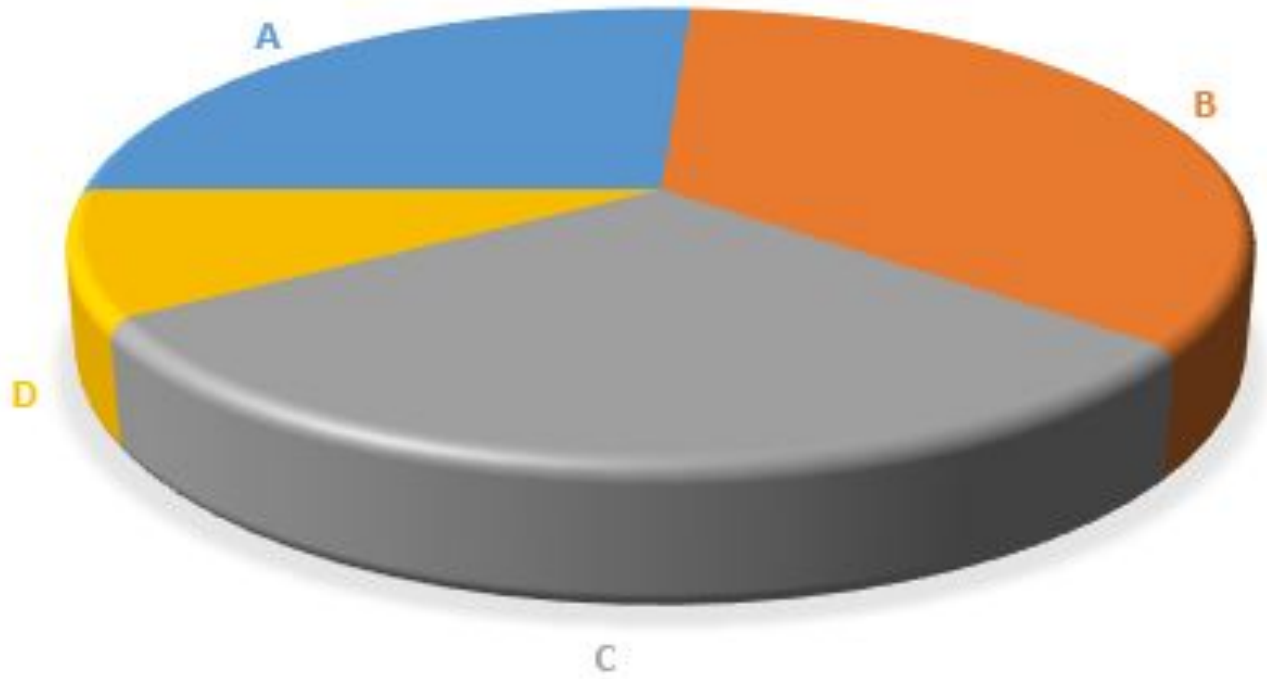


POST: How do you feel about doing science?

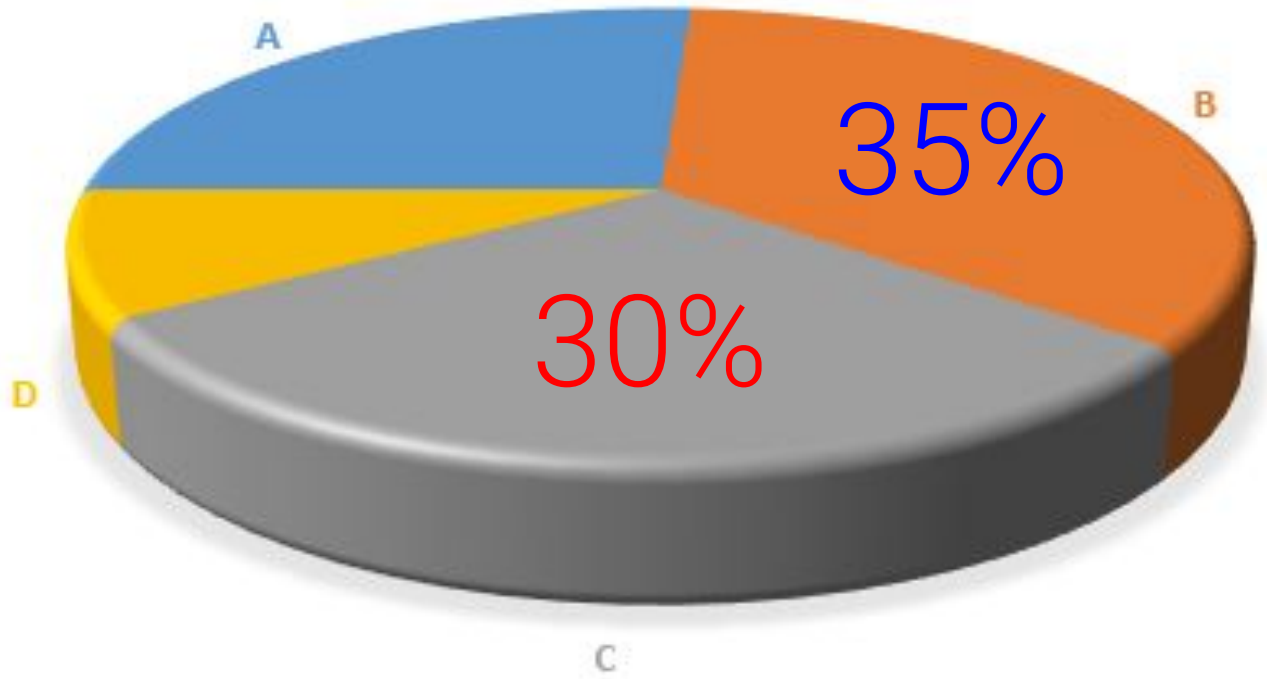


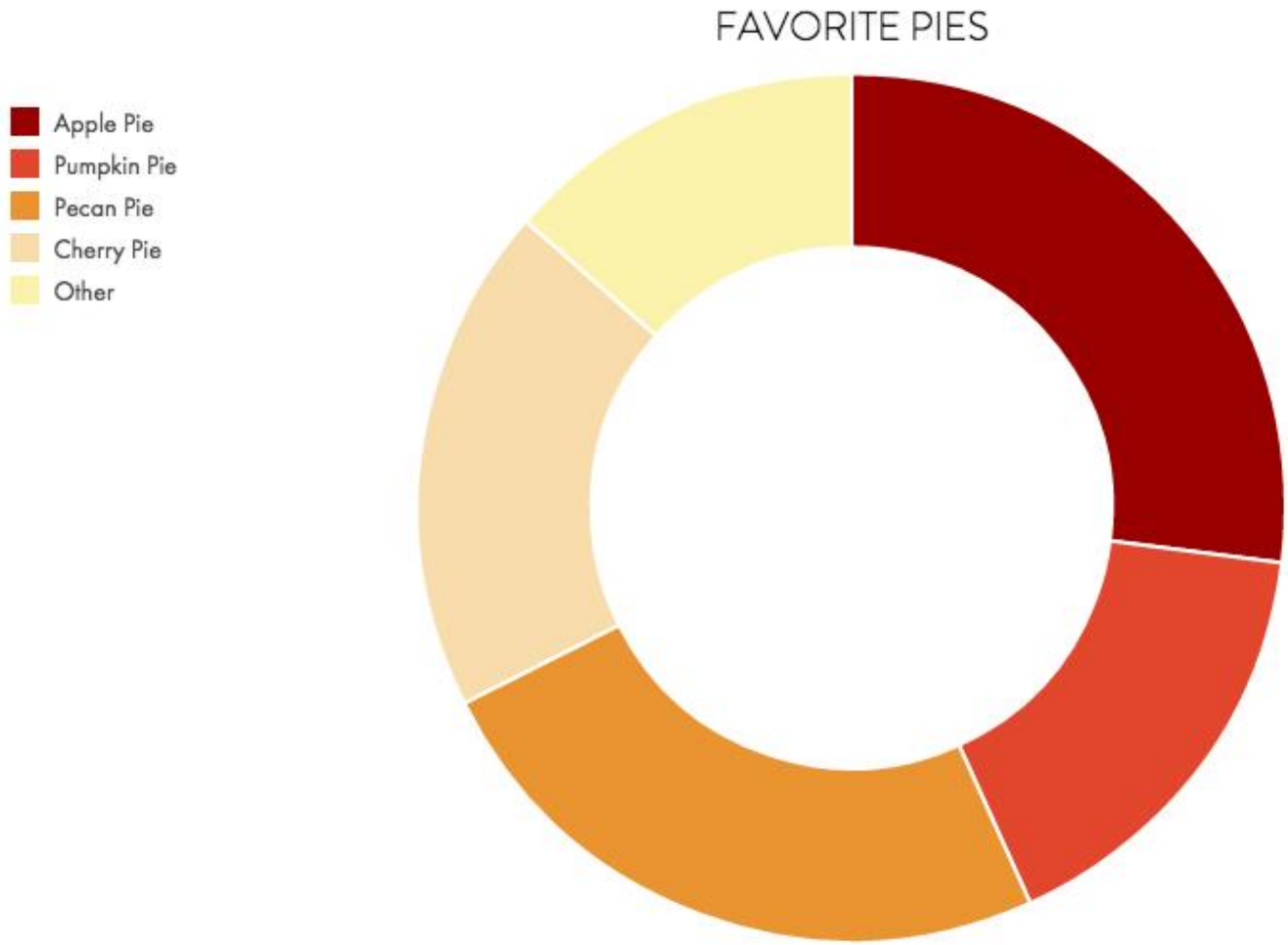
3D Pie charts are evil

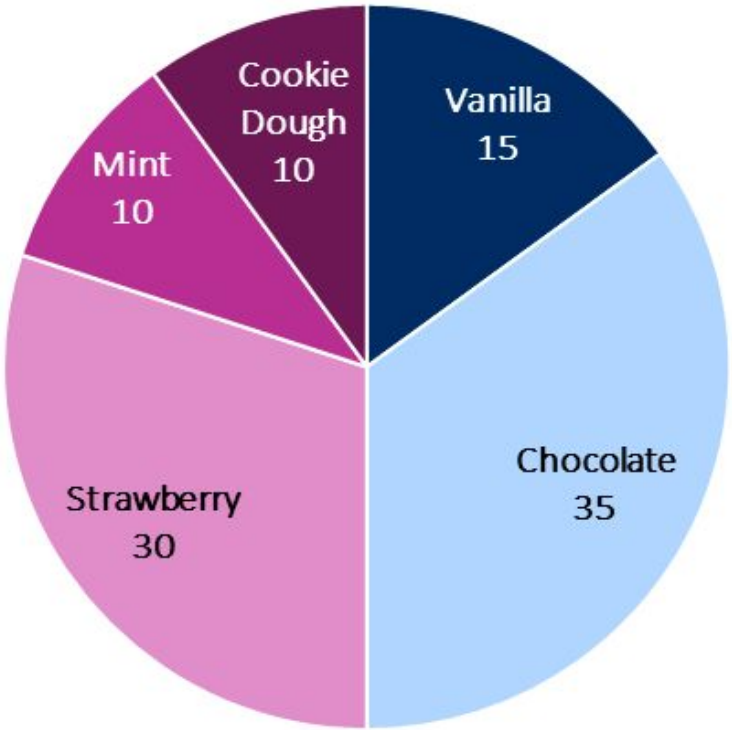
Which is bigger C or B?

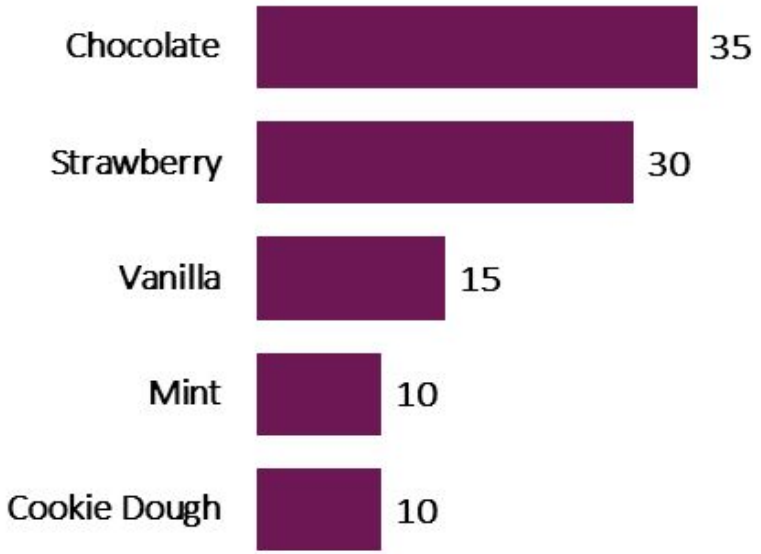


Which is bigger C or B?

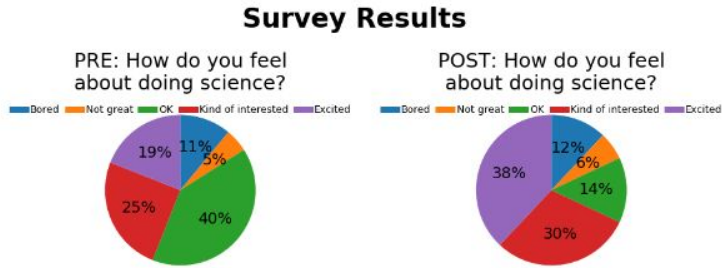



FAVORITE PIES
Apple Pie
Pumpkin Pie
Pecan Pie
Cherry Pie
Other

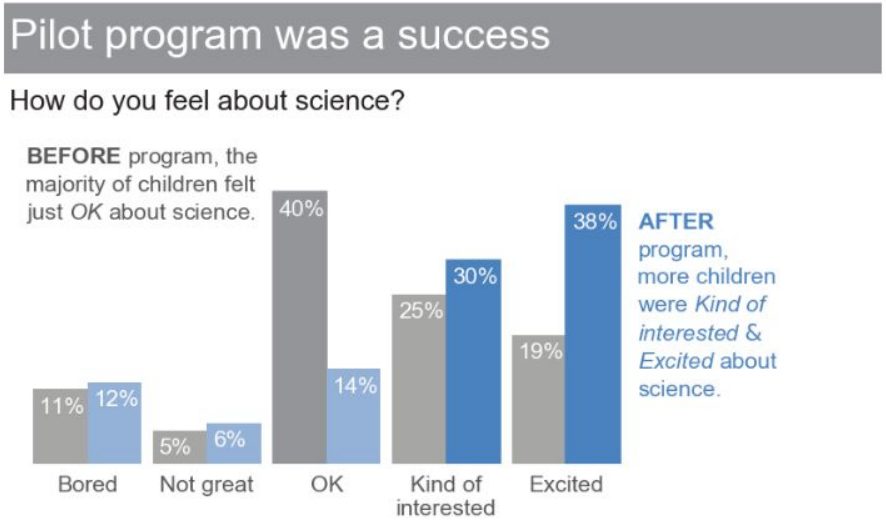

Cookie Dough 10
Vanilla 15
Mint 10
Chocolate 35
Strawberry 30


Chocolate 35
Strawberry 30
Vanilla 15
Mint 10
Cookie Dough 10

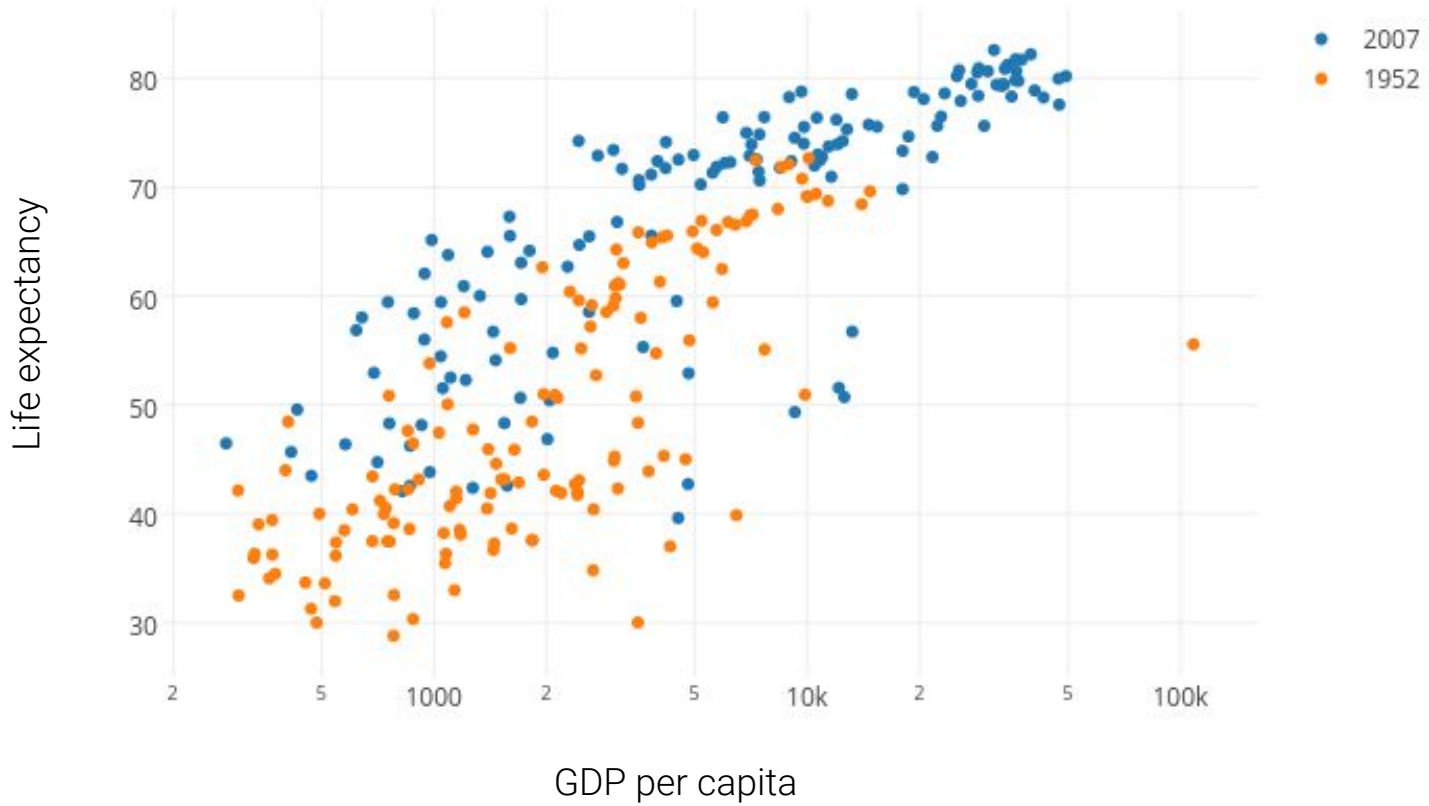
Было:

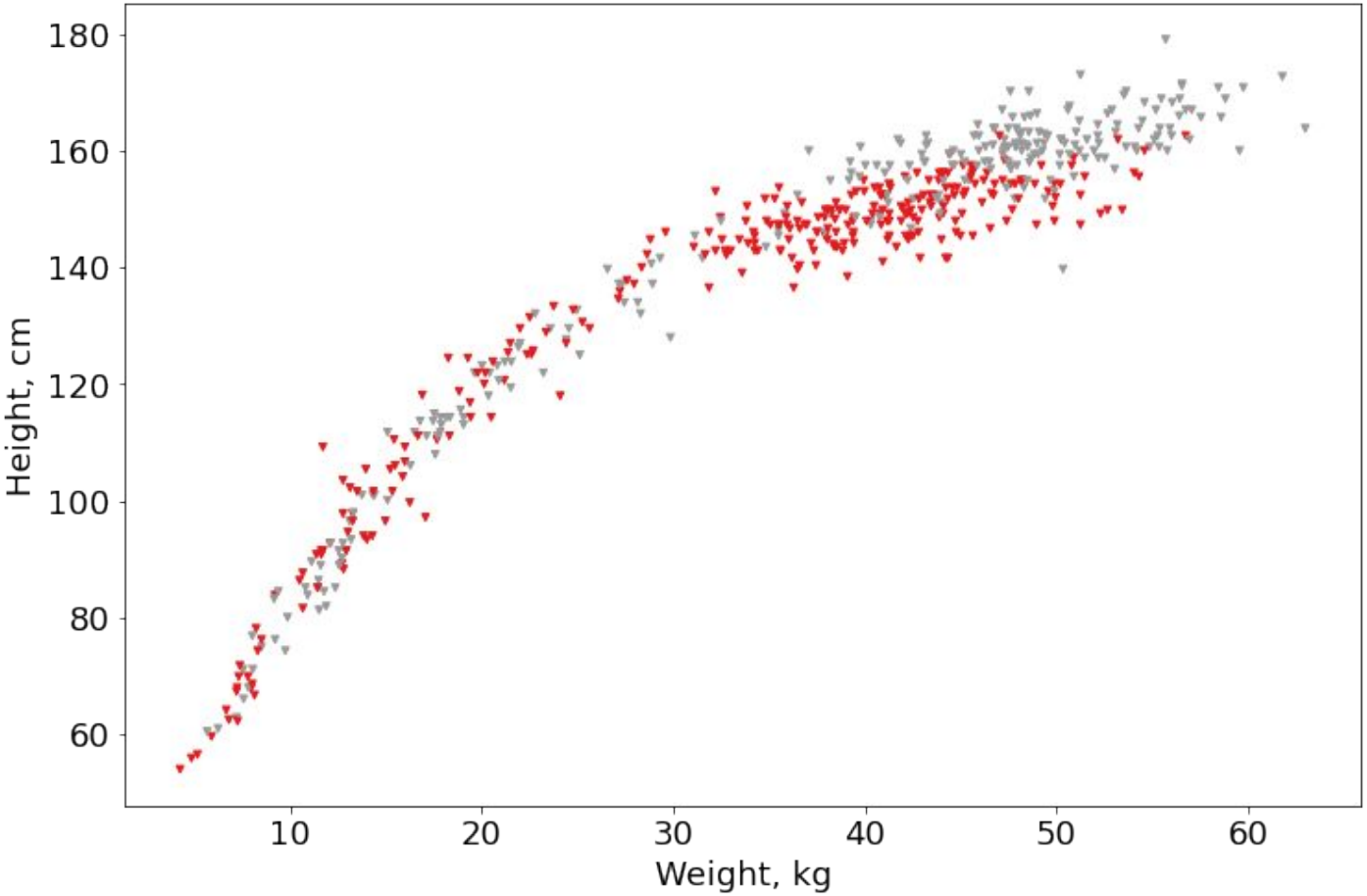

Survey Results
PRE: How do you feel about doing science?
Bored
Not great
OK
Kind of interested
Excited
11%
5%
40%
25%
19%
POST: How do you feel about doing science?
Bored
Not great
OK
Kind of interested
Excited
12%
6%
14%
30%
38%

Стало:


Pilot program was a success
How do you feel about science?
BEFORE program, the majority of children felt just OK about science.
AFTER program, more children were Kind of interested & Excited about science.
11%
12%
5%
6%
40%
14%
25%
30%
19%
38%
Bored
Not great
OK
Kind of interested
Excited

Scatter plot


Life expectancy
80
70
60
50
40
30
2
5
1000
2
5
10k
2
5
100k
GDP per capita
2007
1952

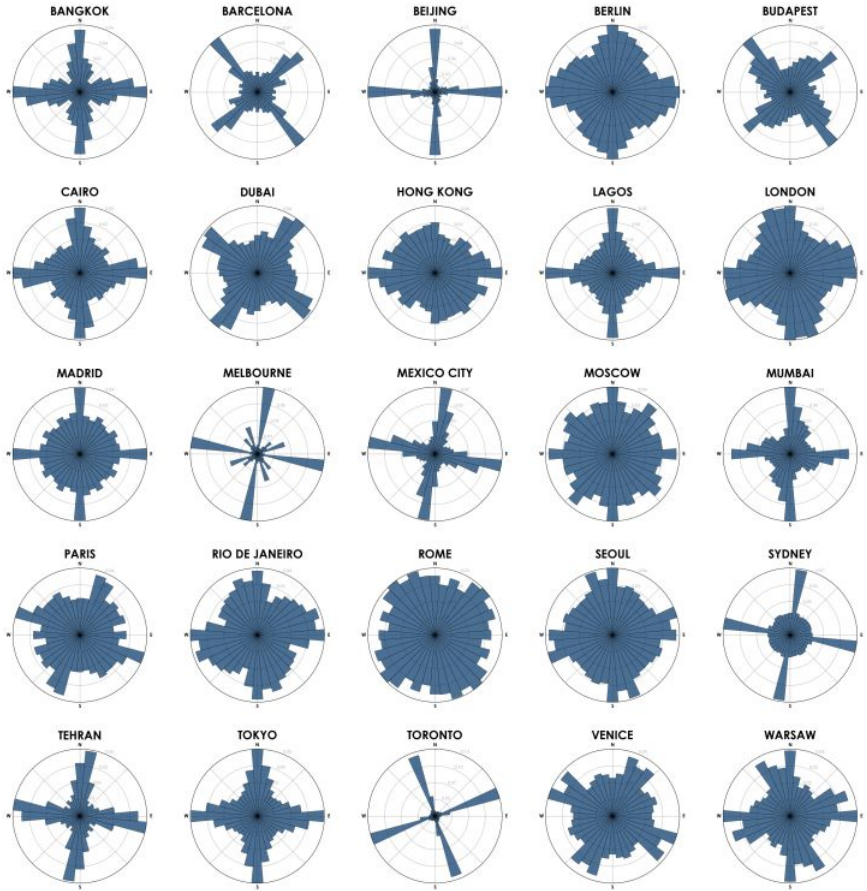

180
160
140
120
100
80
60
Height, cm
10
20
30
40
50
60
Weight, kg

What else?

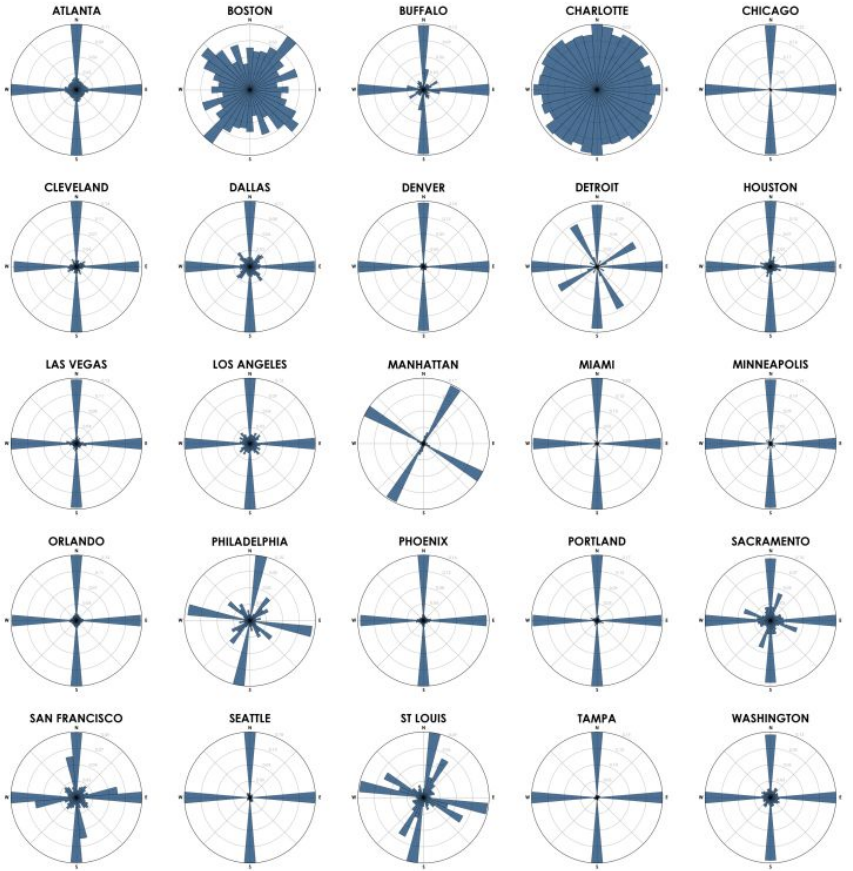
Other types of visualizations

1. Slope chart
2. Venn diagrams
3. Geo maps
4. Bubble chart
5. Tag cloud
6. Stacked graphs
7. Stem and Leaf plot
8. Networks (social, information, etc)
9. ...

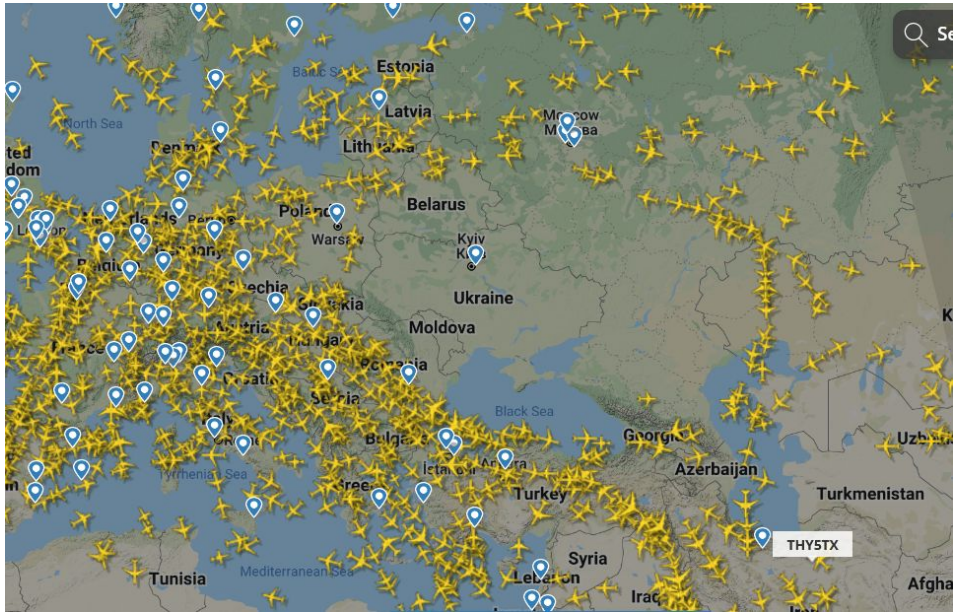
City Street Network Orientation



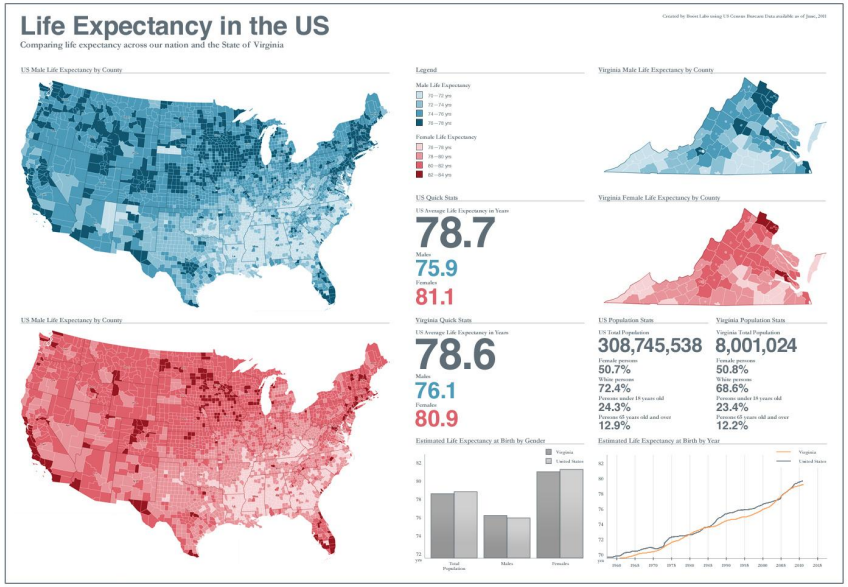
City Street Network Orientation



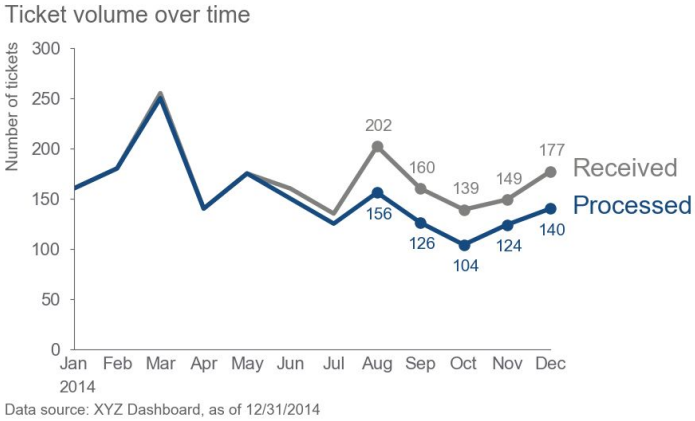
link



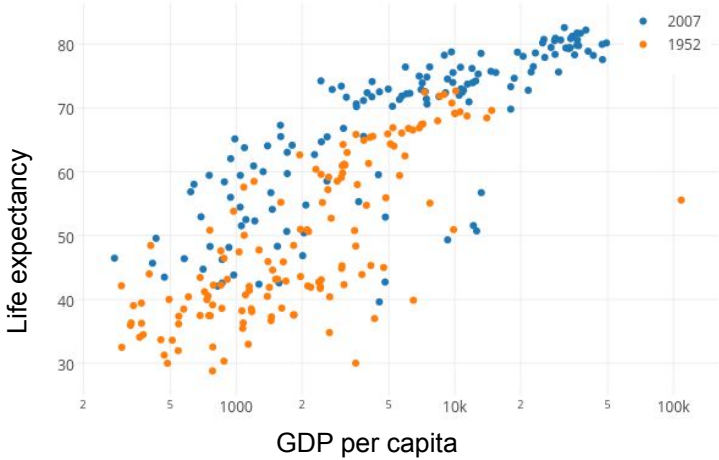
https://www.flightradar24.com

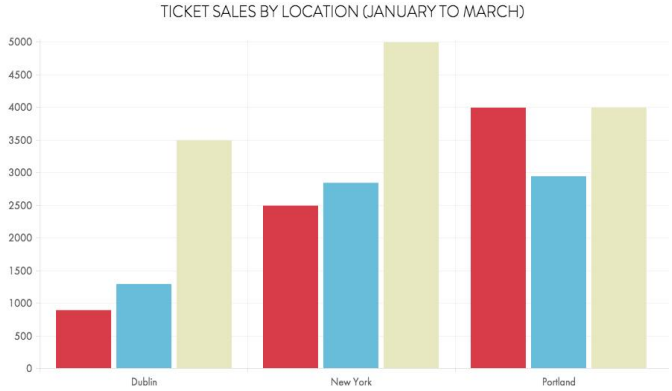


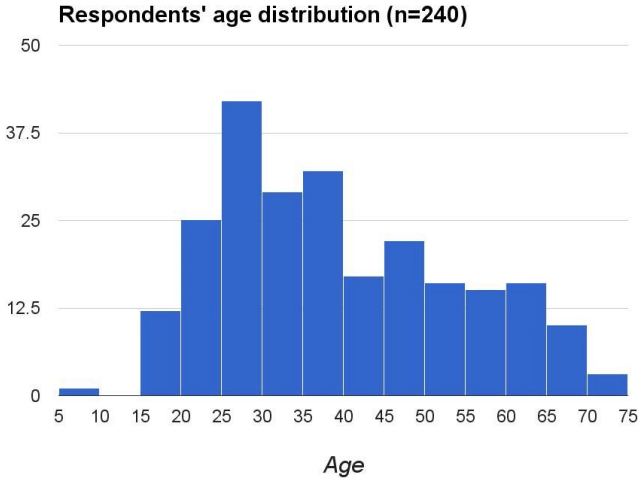
https://blogs.stockton.edu/datavis/2017/11/27/geographical-visualizations/

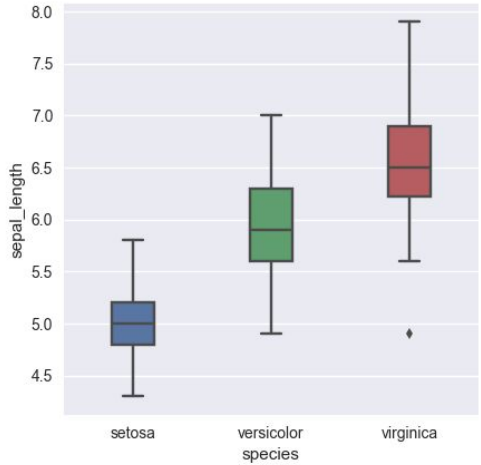


- Index chart
- Scatter plot
- Bar chart
- Histogram
- Box plot









Orange https://orangedatamining.com/

KNIME https://www.knime.com/

Rapid Miner https://rapidminer.com/

Microsoft Azure https://azure.microsoft.com/en-us/