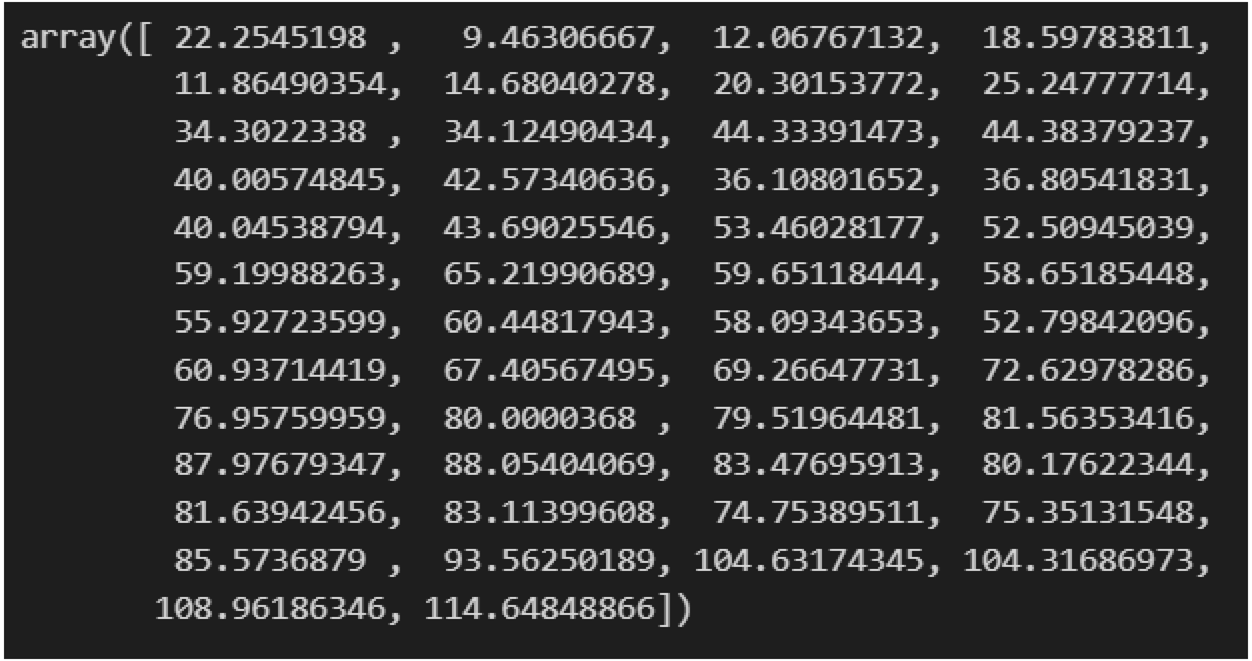
Data Visualizations

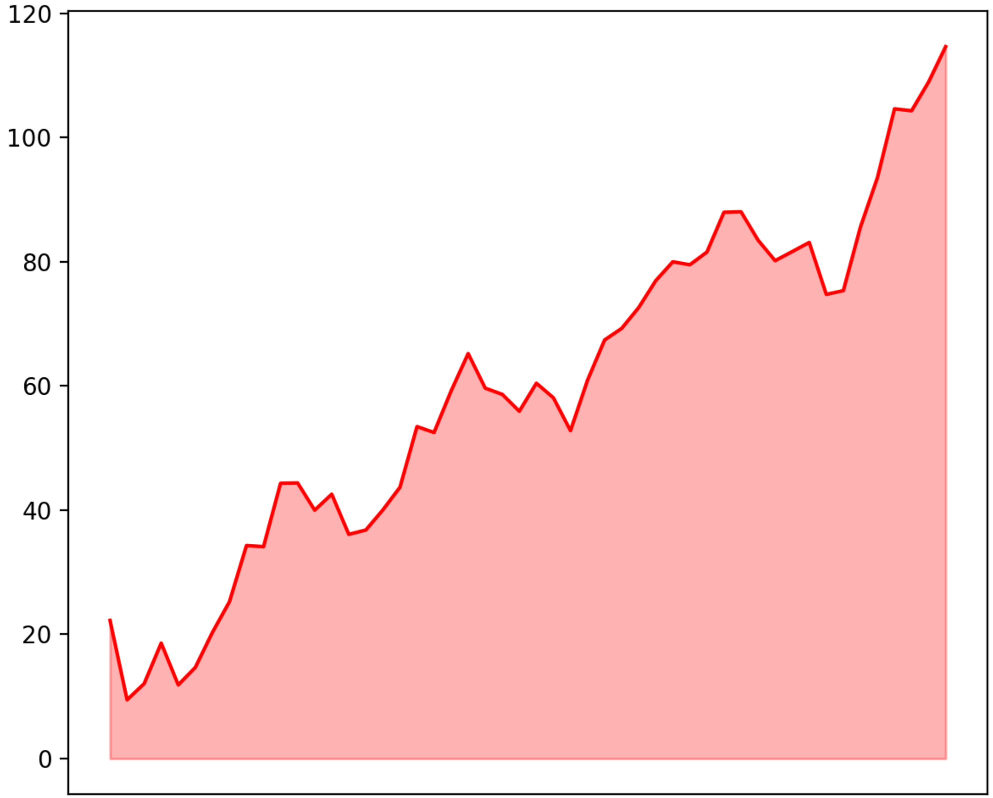
Why do we need visualizations

Before



```
array([ 22.2545198 ,   9.46306667,  12.06767132,  18.59783811,
        11.86490354,  14.68040278,  20.30153772,  25.24777714,
        34.3022338 ,  34.12490434,  44.33391473,  44.38379237,
        40.00574845,  42.57340636,  36.10801652,  36.80541831,
        40.04538794,  43.69025546,  53.46028177,  52.50945039,
        59.19988263,  65.21990689,  59.65118444,  58.65185448,
        55.92723599,  60.44817943,  58.09343653,  52.79842096,
        60.93714419,  67.40567495,  69.26647731,  72.62978286,
        76.95759959,  80.0000368 ,  79.51964481,  81.56353416,
        87.97679347,  88.05404069,  83.47695913,  80.17622344,
        81.63942456,  83.11399608,  74.75389511,  75.35131548,
        85.5736879 ,  93.56250189, 104.63174345, 104.31686973,
       108.96186346, 114.64848866])
```

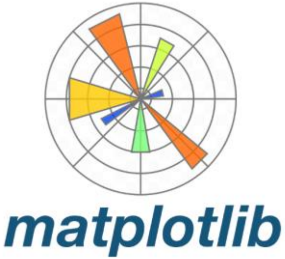
After



Main libraries

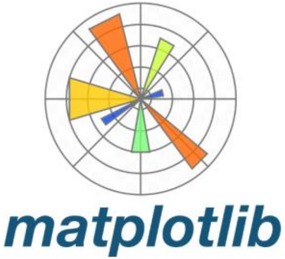
Main libraries

Fundamental



Main libraries

Fundamental



Main libraries

Fundamental

matplotlib

seaborn

Interactive

plotly

Main libraries

Fundamental

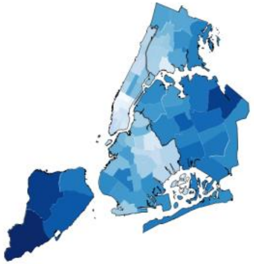
matplotlib

seaborn

Interactive

plotly

Geospatial



Geoplot

Main libraries

Fundamental

matplotlib

seaborn

Interactive

plotly

Geospatial

Geoplot

Folium

Other visualization apps

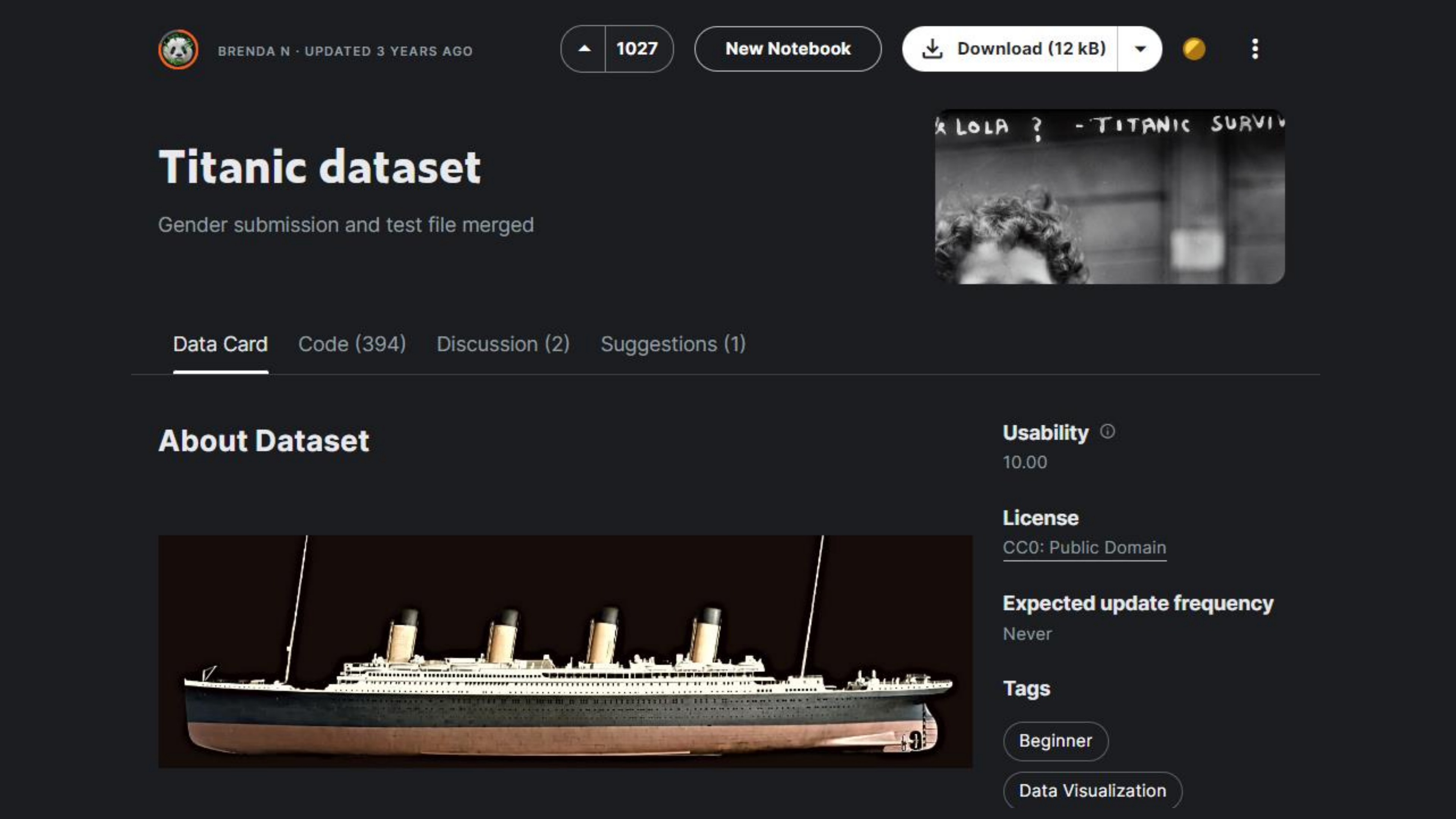
Other visualization apps

Power BI

Excel

tableau

How to visualize data distribution



Titanic dataset

Gender submission and test file merged

About Dataset

Usability

10.00

License

CC0: Public Domain

Expected update frequency

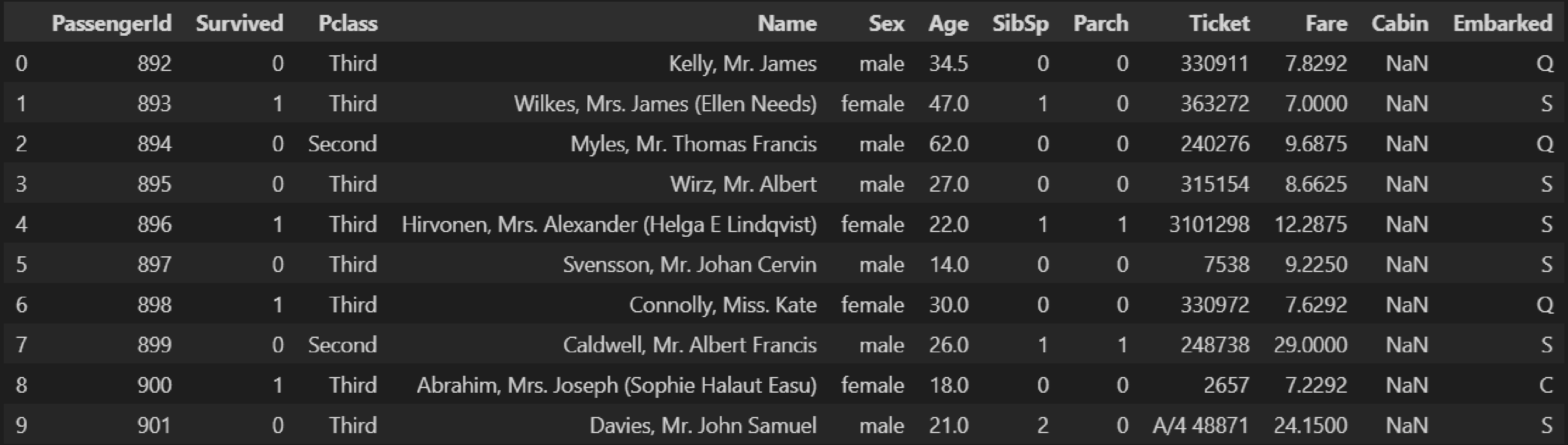
Never

Tags

Beginner

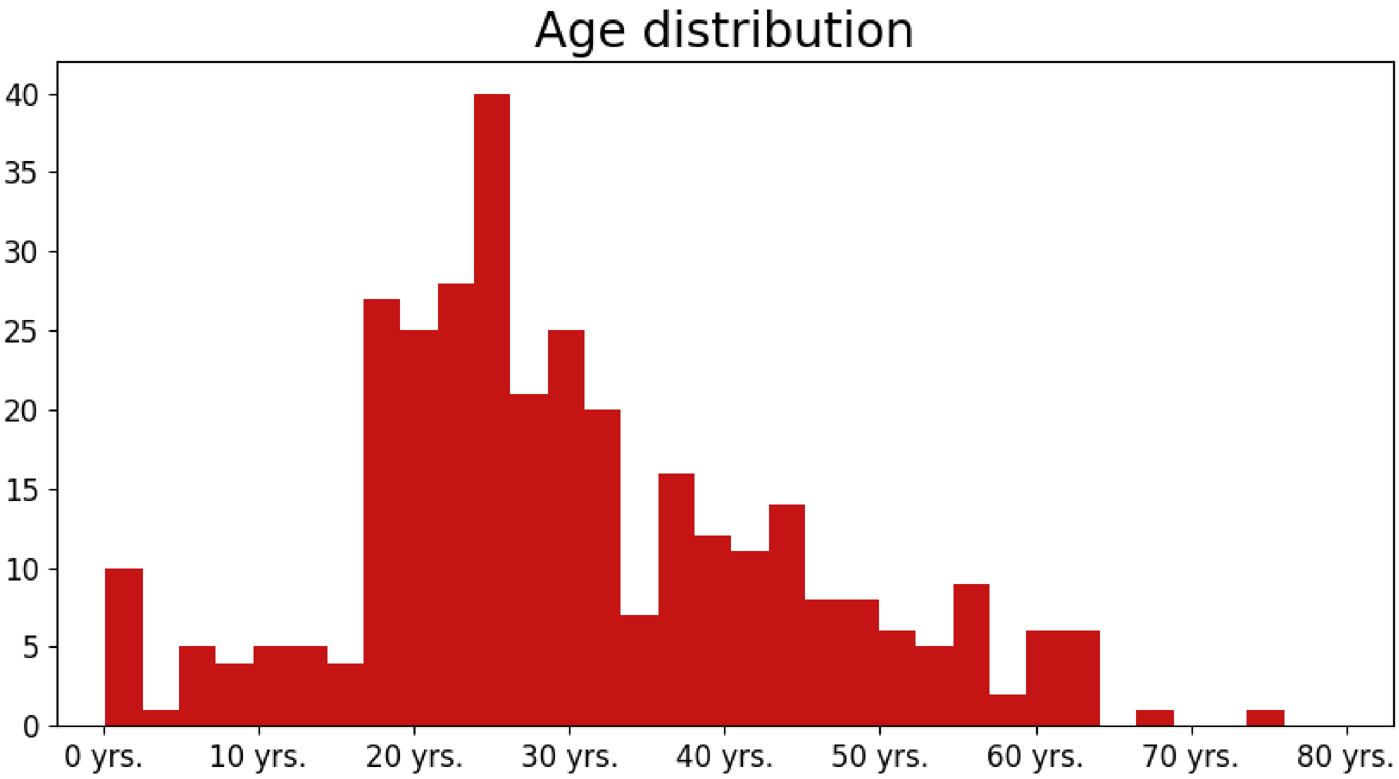
Data Visualization

HISTORICAL ACCURACY
KAGGLE COMPETITORS
imgflip.com

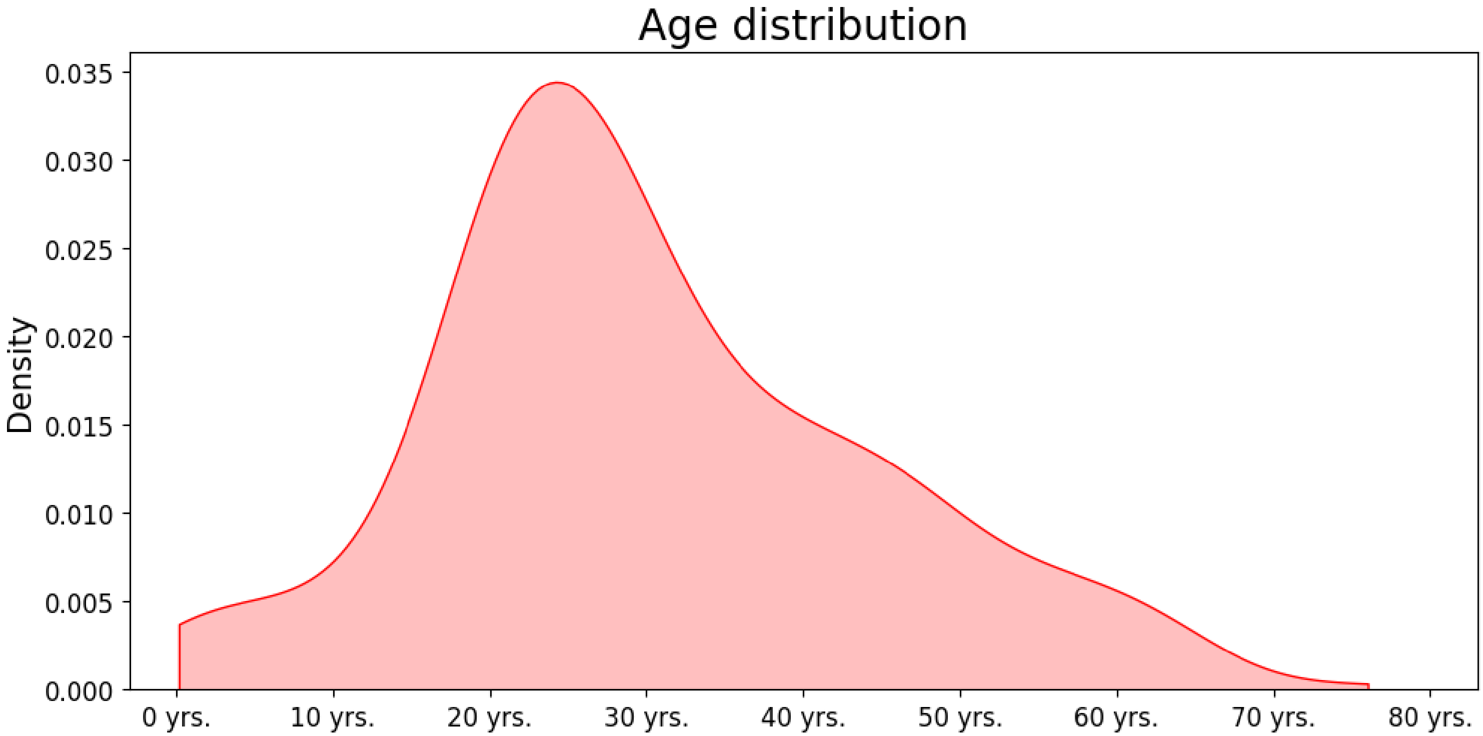


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	Third	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	Third	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	Second	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	Third	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	Third	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
5	897	0	Third	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S
6	898	1	Third	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
7	899	0	Second	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S
8	900	1	Third	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C
9	901	0	Third	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S

Histogram


Age distribution
0
5
10
15
20
25
30
35
40
0 yrs.
10 yrs.
20 yrs.
30 yrs.
40 yrs.
50 yrs.
60 yrs.
70 yrs.
80 yrs.

Density plot

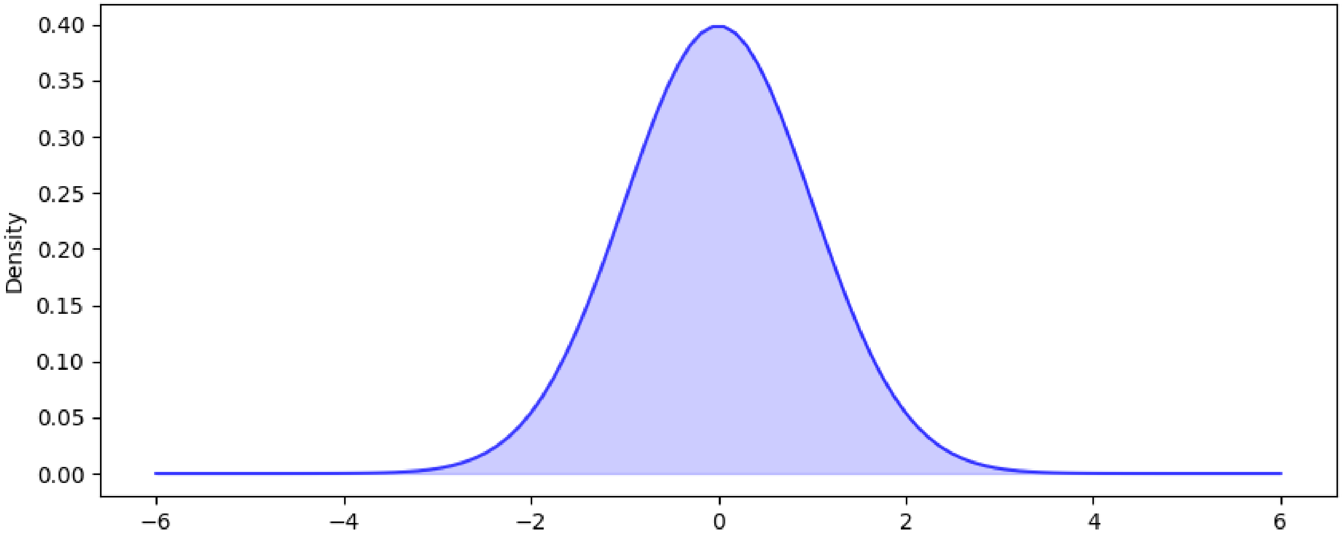


KDE

KDE - Kernel Density Estimator

Density plot

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$



Density plot

$$K(x - x_i)$$

Density plot

$$K(x - x_i)$$

$$K\left(\frac{x - x_i}{h}\right)$$

Density plot

h - kernel bandwidth

$$K(x - x_i)$$

$$K\left(\frac{x - x_i}{h}\right)$$

Density plot

h - kernel bandwidth

$$K(x - x_i)$$

$$K\left(\frac{x - x_i}{h}\right)$$

$$\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

Density plot

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Density plot

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\frac{1}{h} K \left(\frac{x - x_1}{h} \right)$$

Density plot

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\frac{1}{h} K \left(\frac{x - x_1}{h} \right) + \frac{1}{h} K \left(\frac{x - x_2}{h} \right)$$

Density plot

$$f(x) = \frac{1}{2}\left[\frac{1}{h}K\left(\frac{x-x_1}{h}\right) + \ \frac{1}{h}K\left(\frac{x-x_2}{h}\right)\right]$$

Density plot

$$f(x) = \frac{1}{2}\left[\frac{1}{h}K\left(\frac{x-x_1}{h}\right) + \ \frac{1}{h}K\left(\frac{x-x_2}{h}\right)\right] =$$

$$= \frac{1}{2h}\left[K\left(\frac{x-x_1}{h}\right) + \ K\left(\frac{x-x_2}{h}\right)\right]$$

Density plot

$$f(x) = \frac{1}{2}\left[\frac{1}{h}K\left(\frac{x-x_1}{h}\right) + \ \frac{1}{h}K\left(\frac{x-x_2}{h}\right)\right] =$$

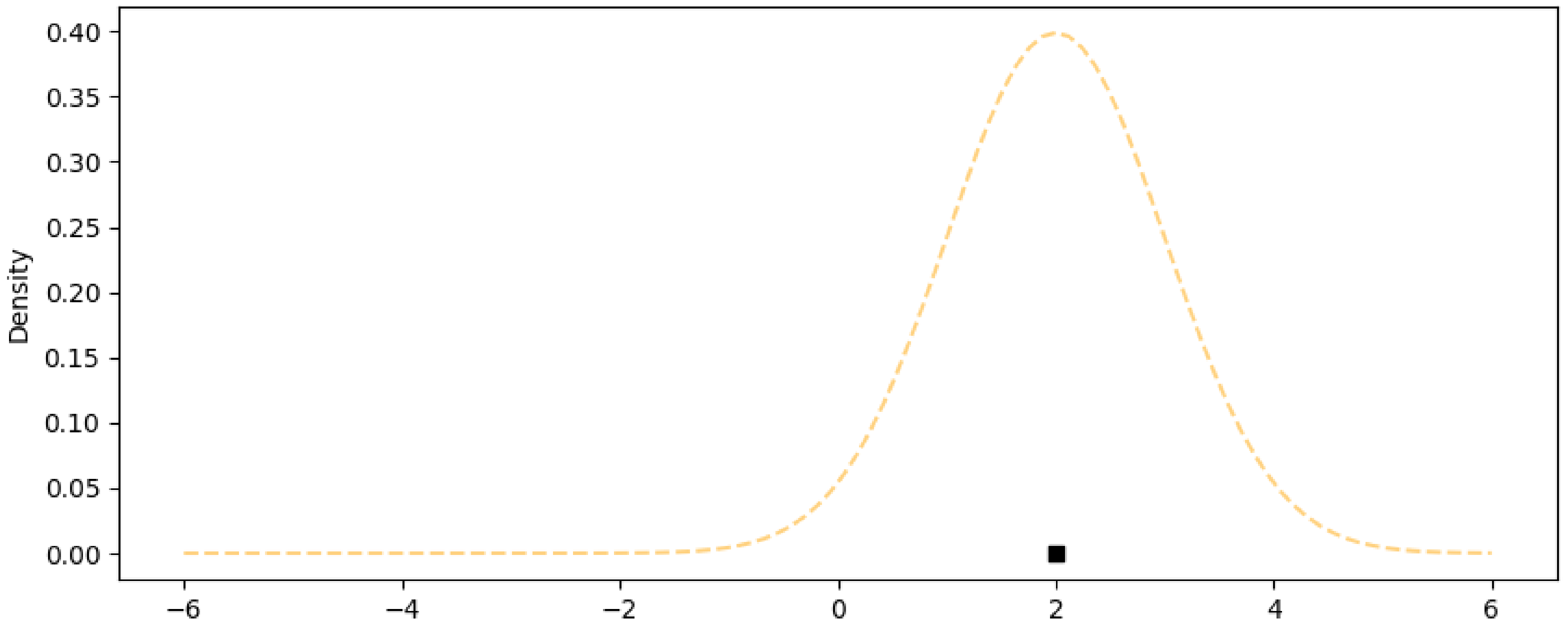
$$= \frac{1}{2h}\left[K\left(\frac{x-x_1}{h}\right) + \ K\left(\frac{x-x_2}{h}\right)\right] =$$

$$= \frac{1}{2h}\sum_{i=1}^{2} K\left(\frac{x - x_i}{h}\right)$$

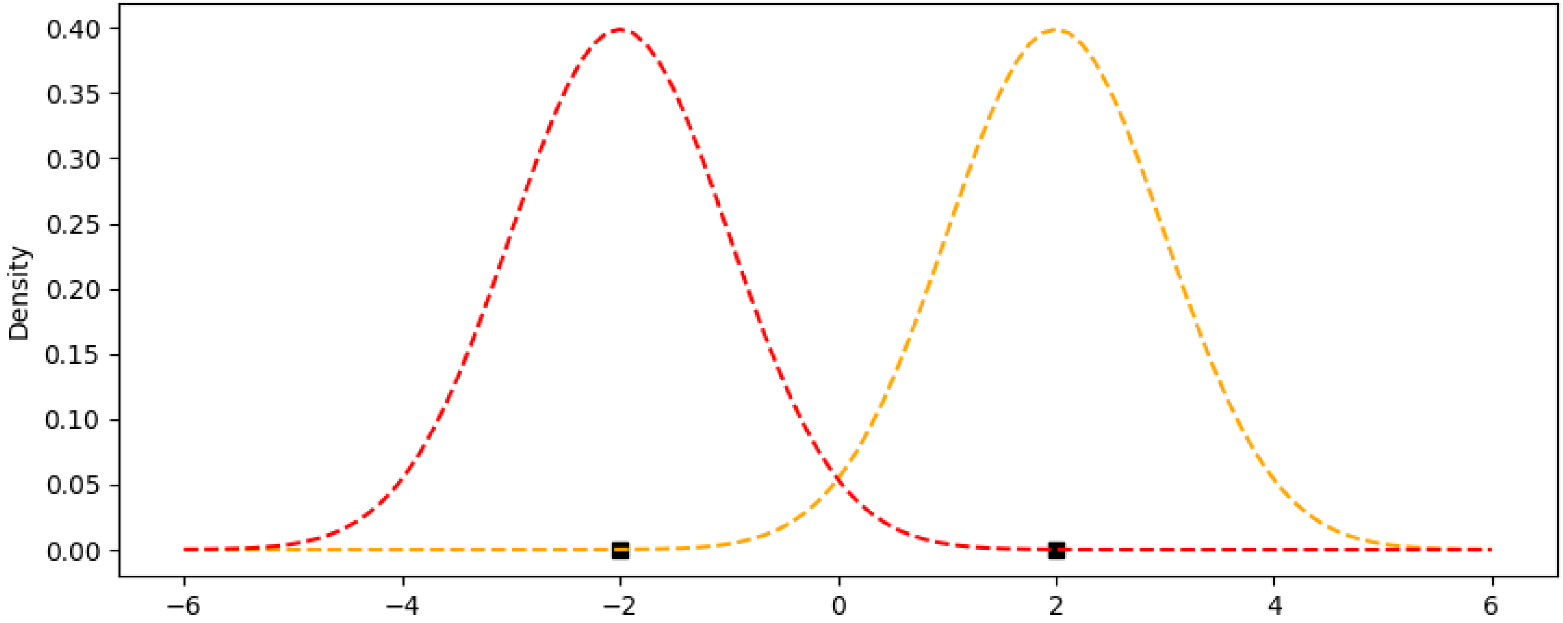
Density plot

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

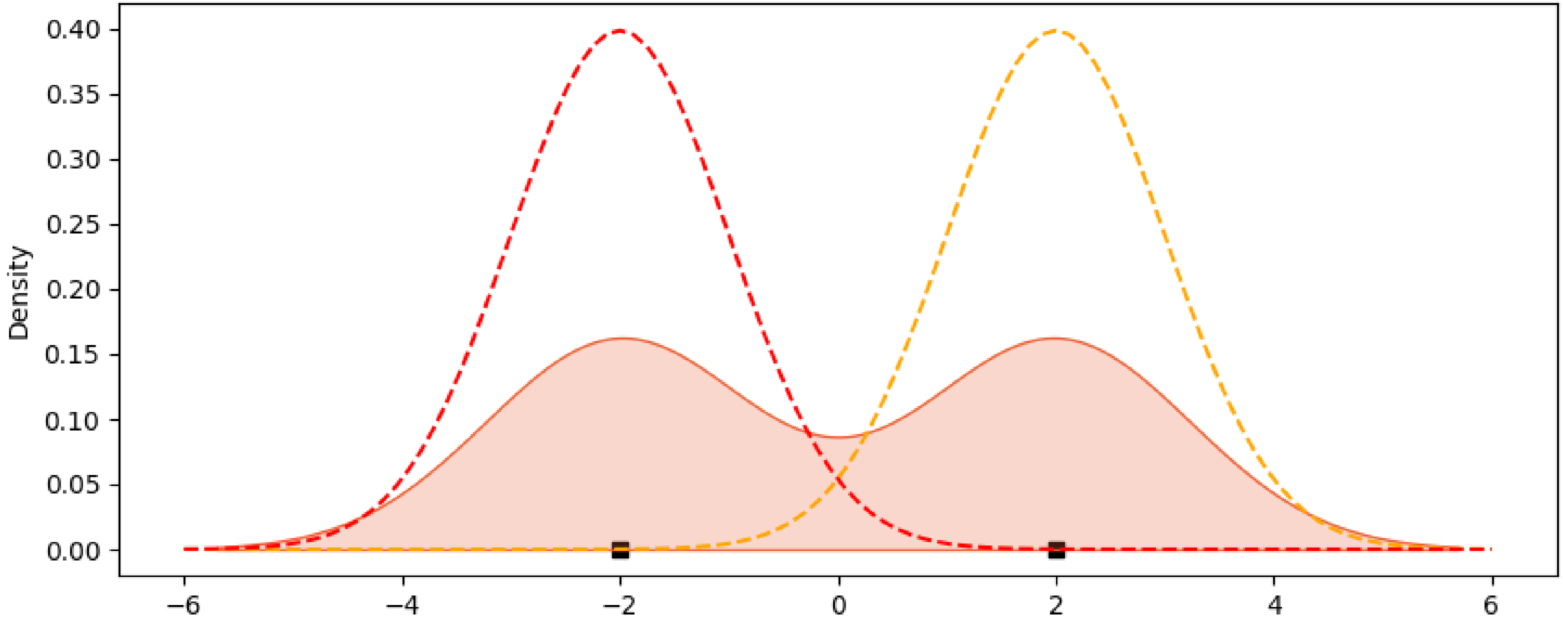
Density plot



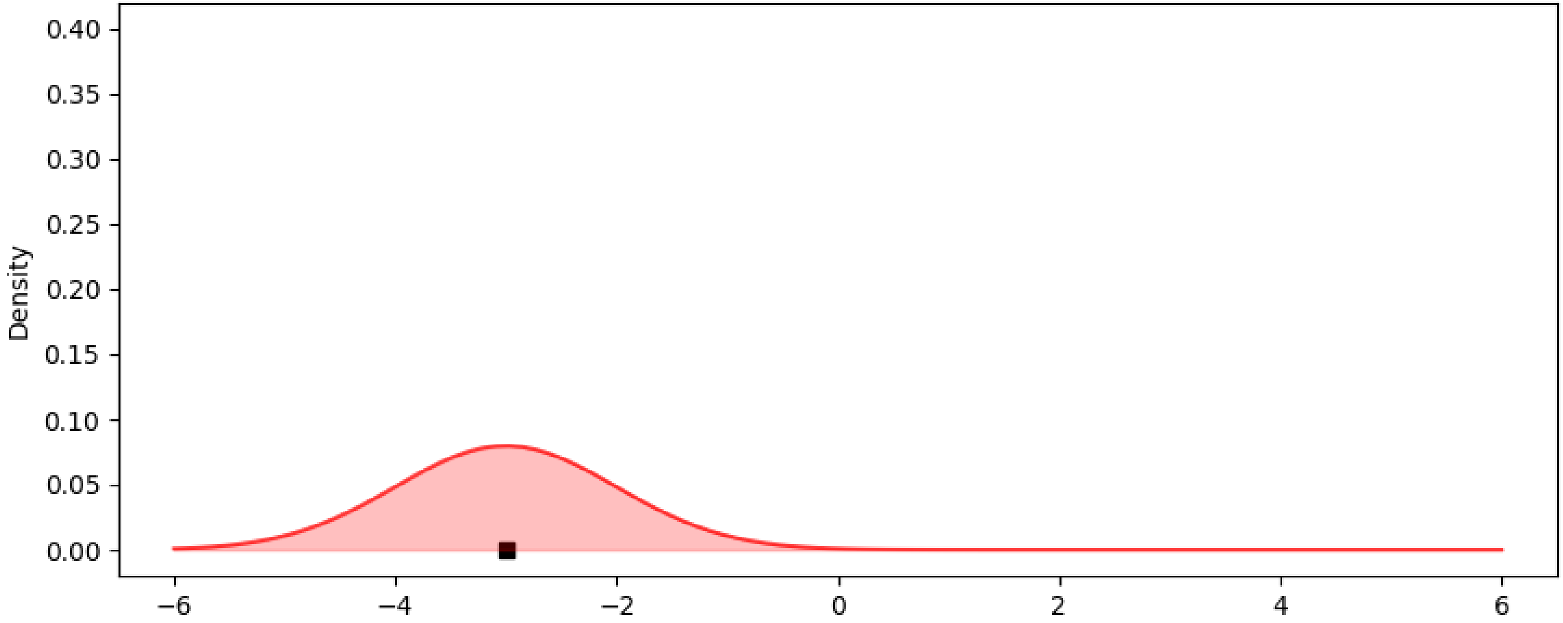
Density plot



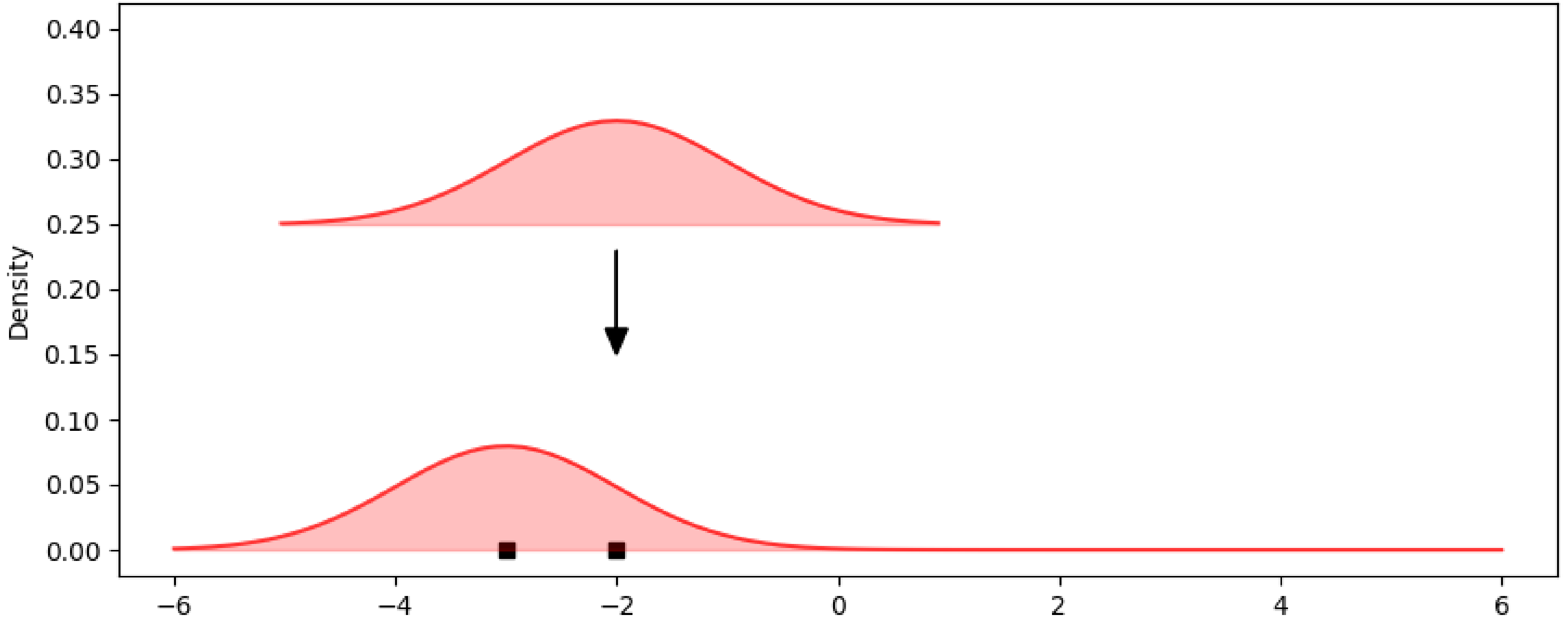
Density plot



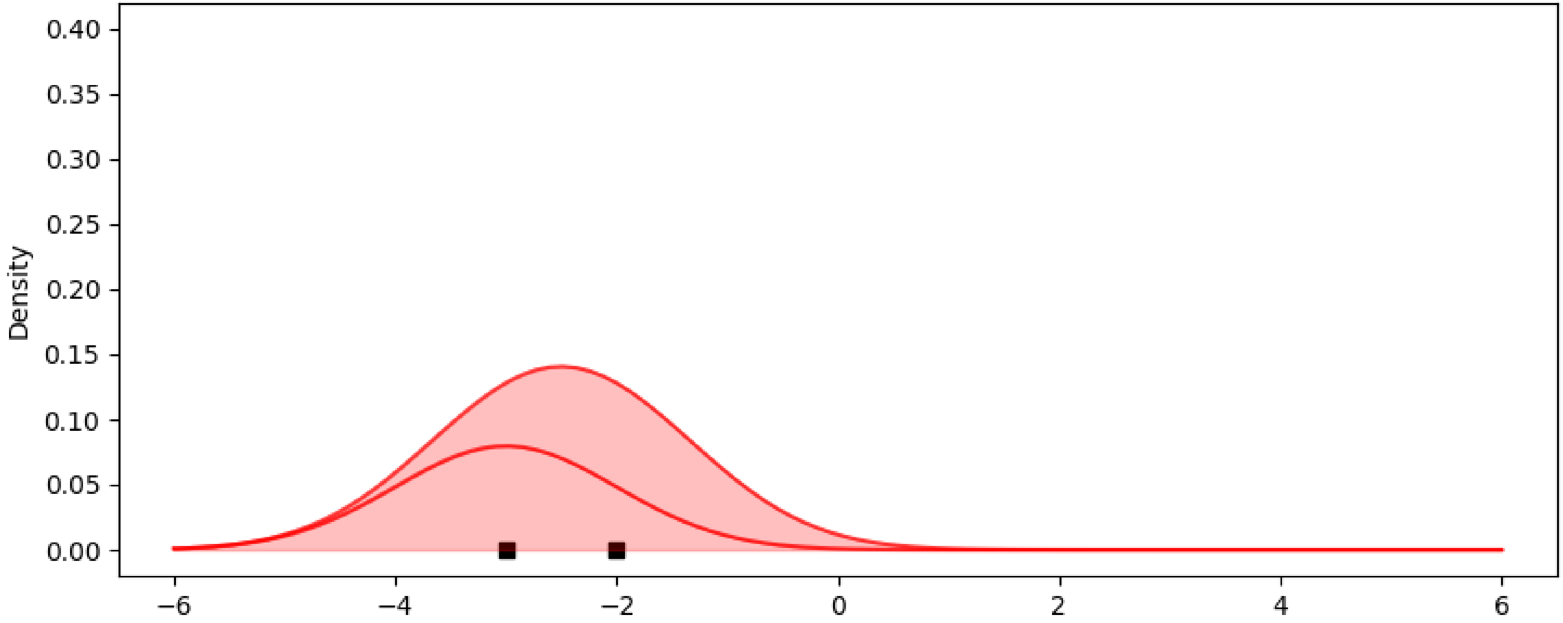
Density plot



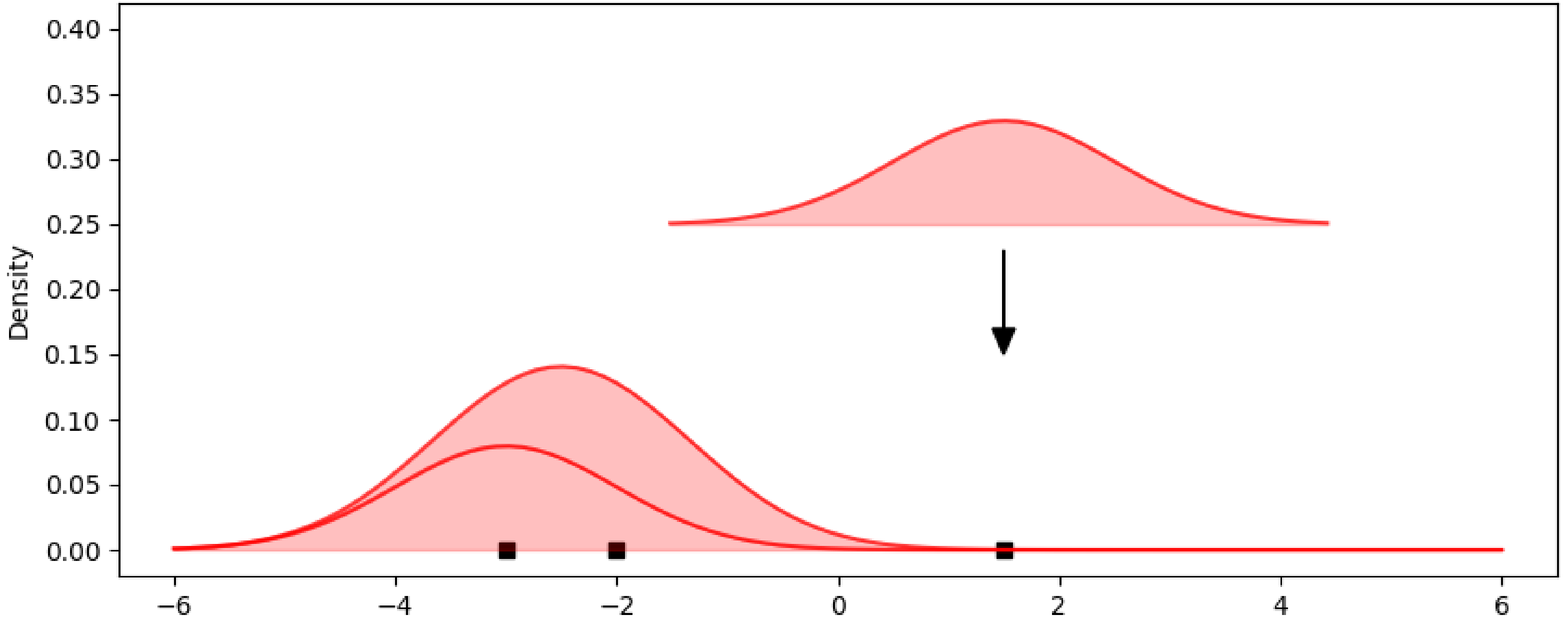
Density plot



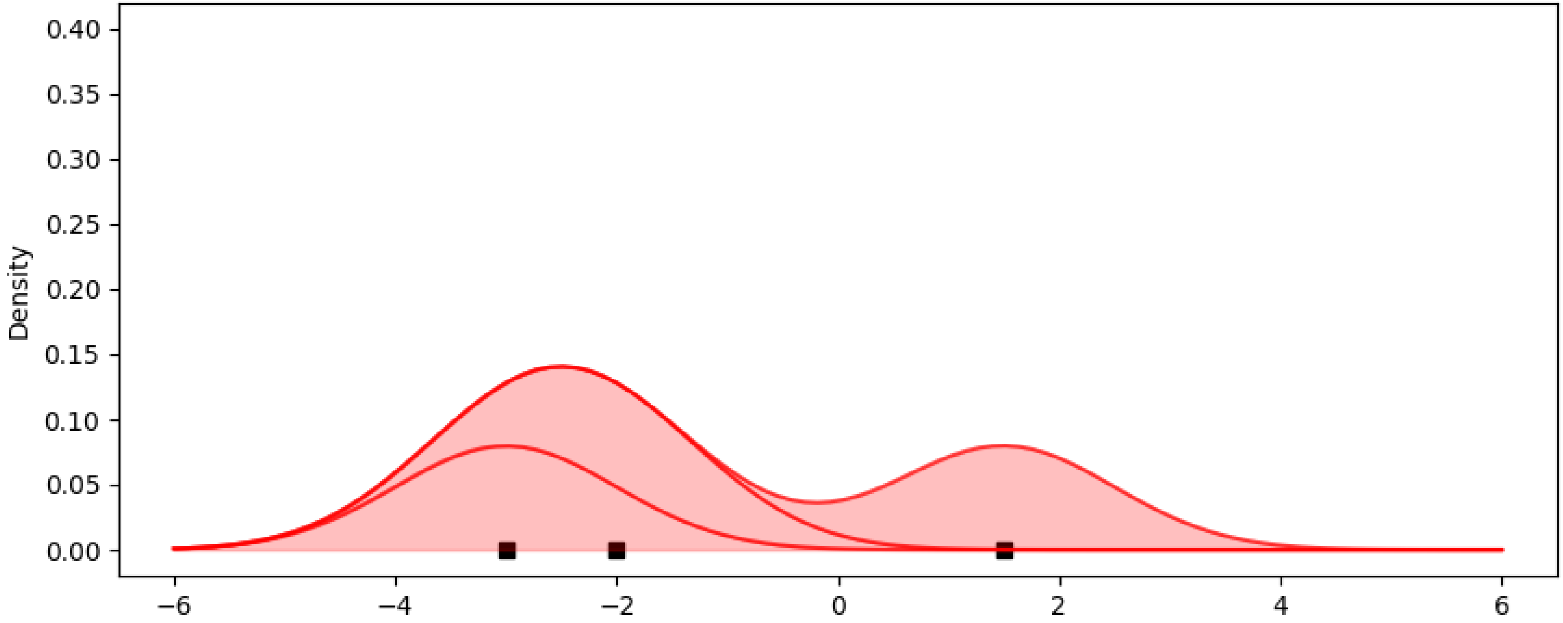
Density plot



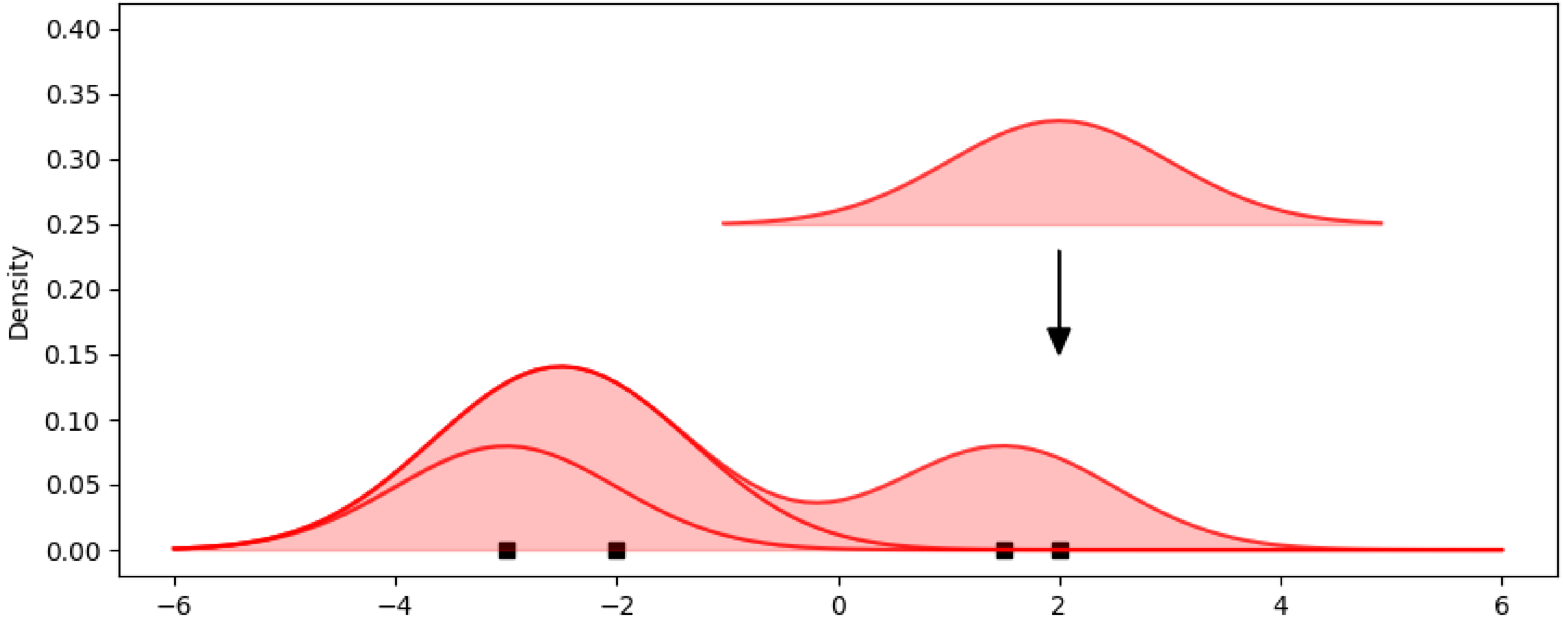
Density plot



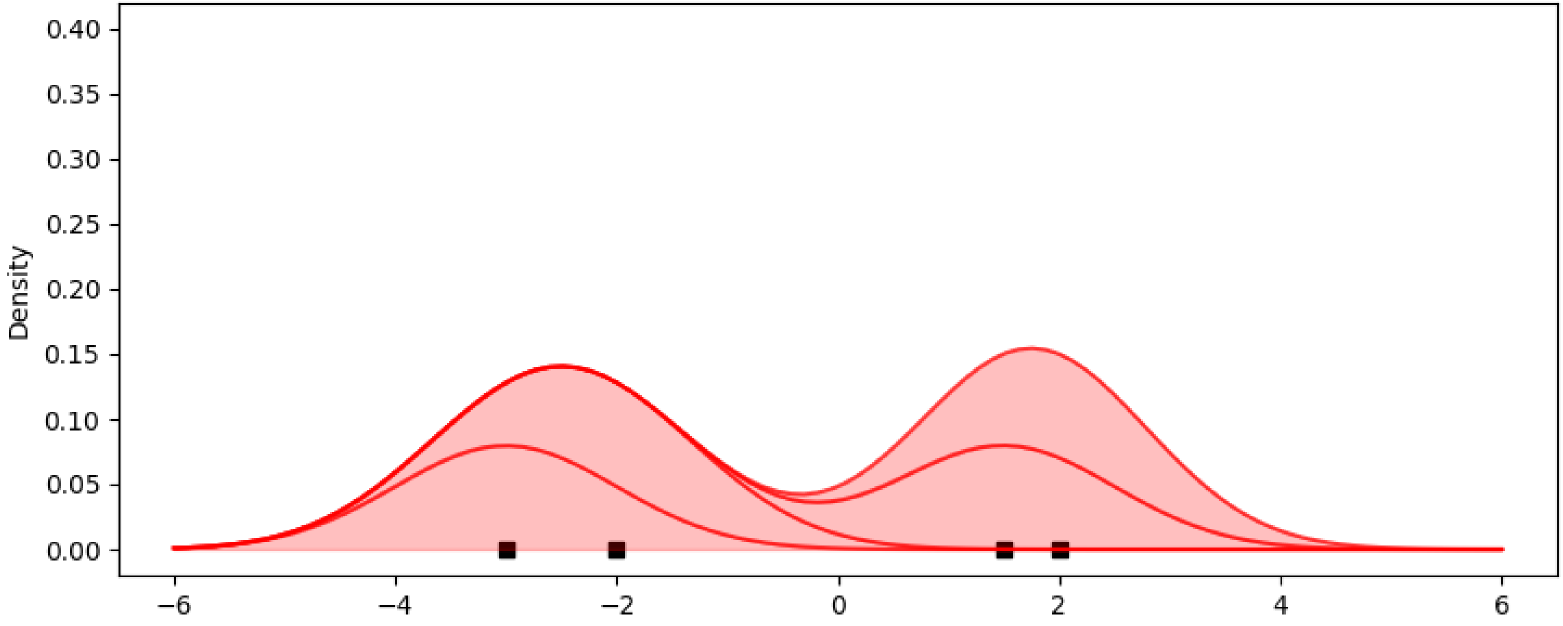
Density plot



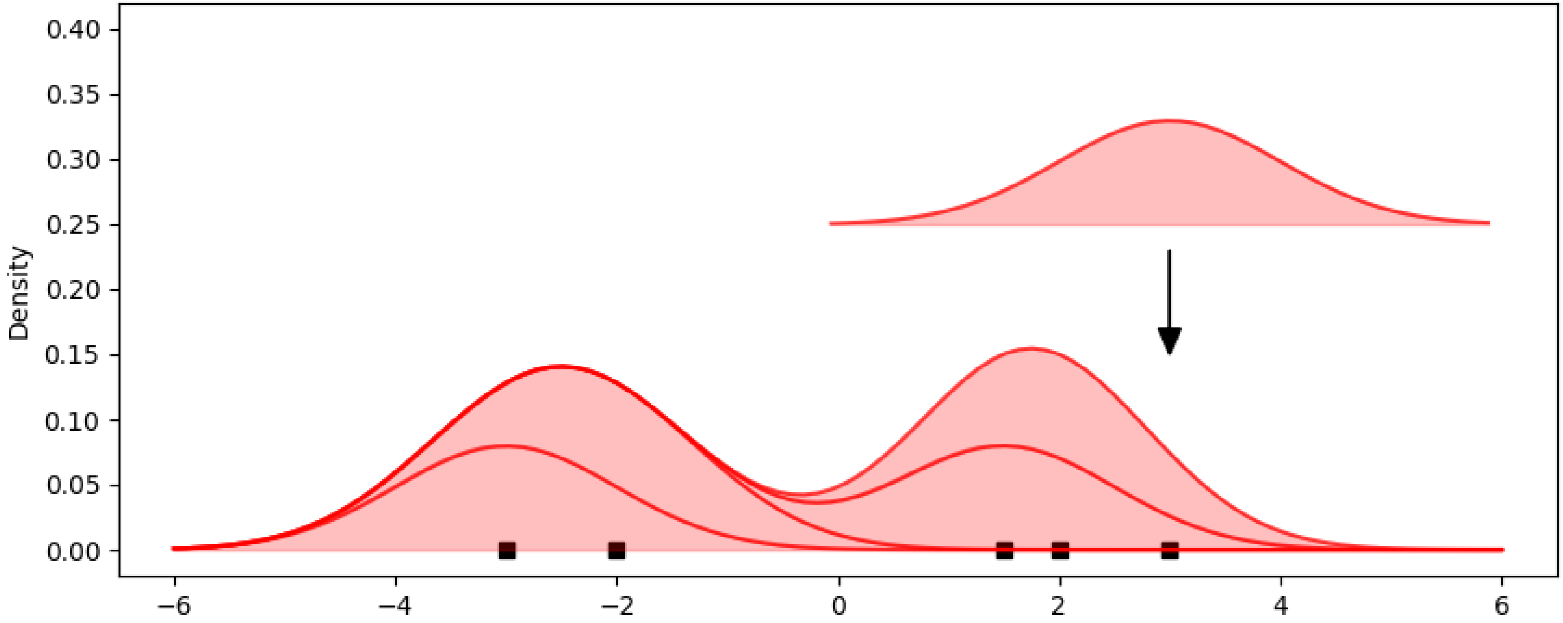
Density plot



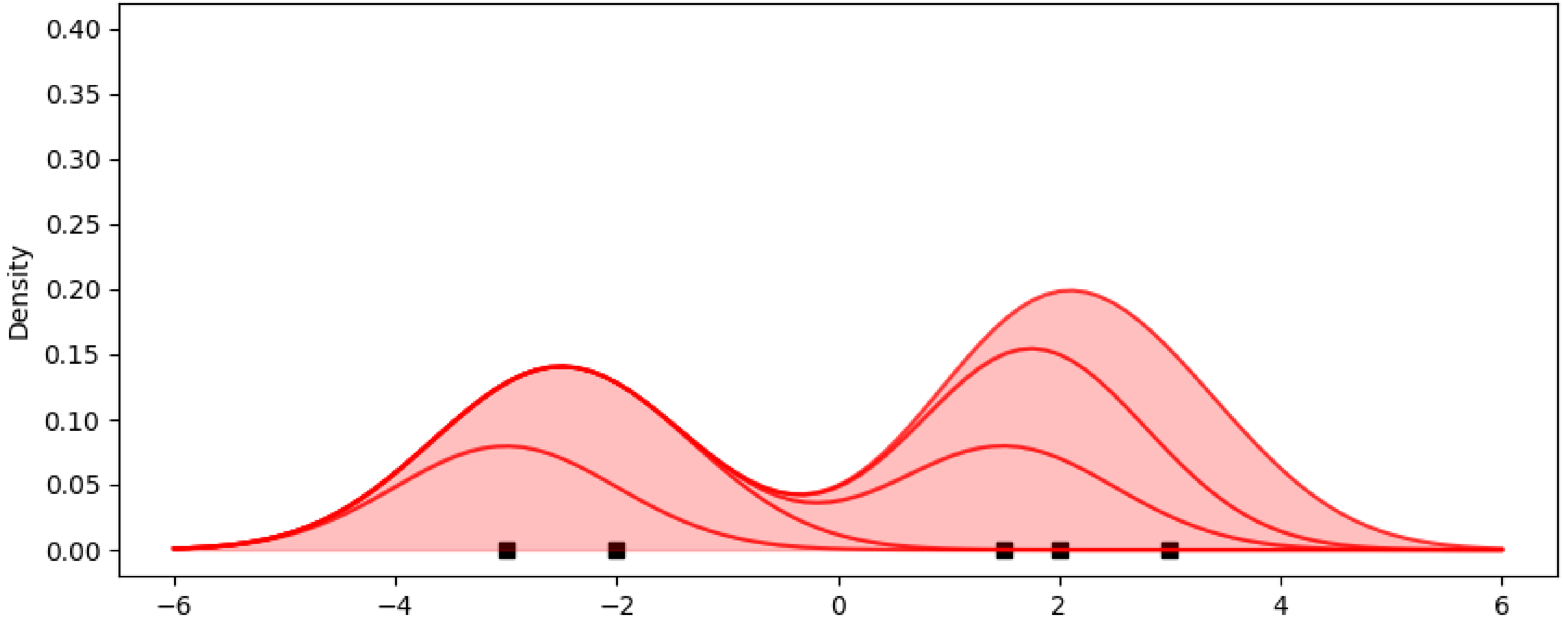
Density plot



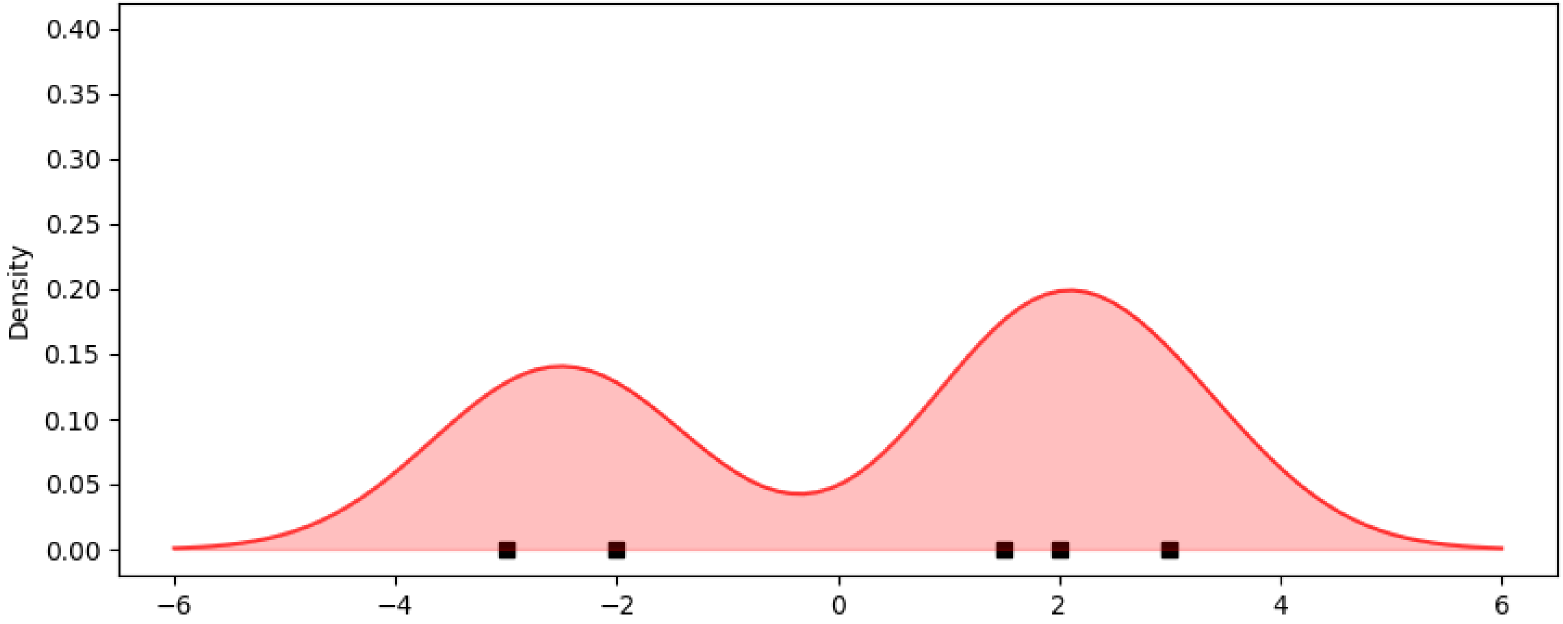
Density plot



Density plot

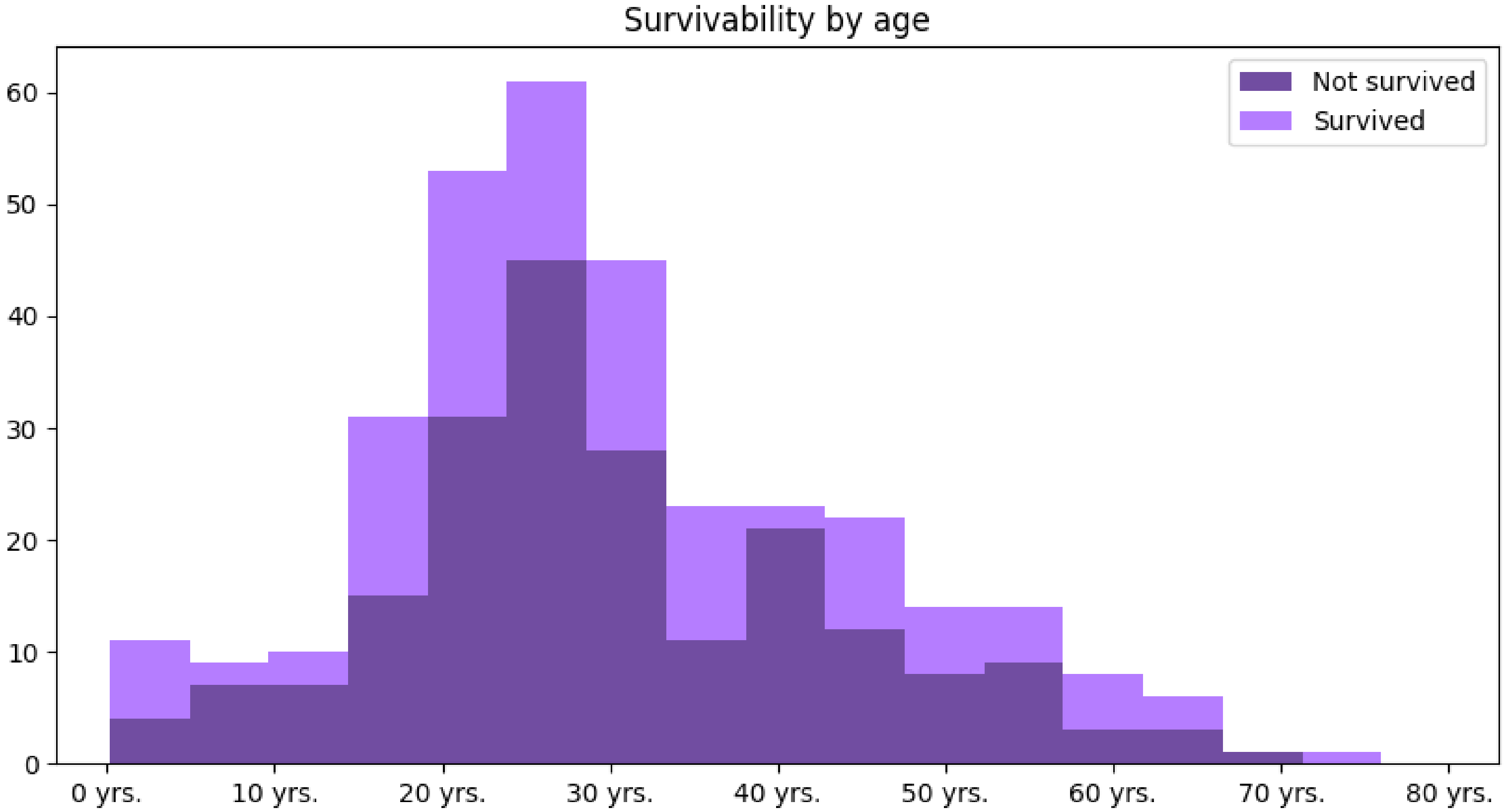


Density plot

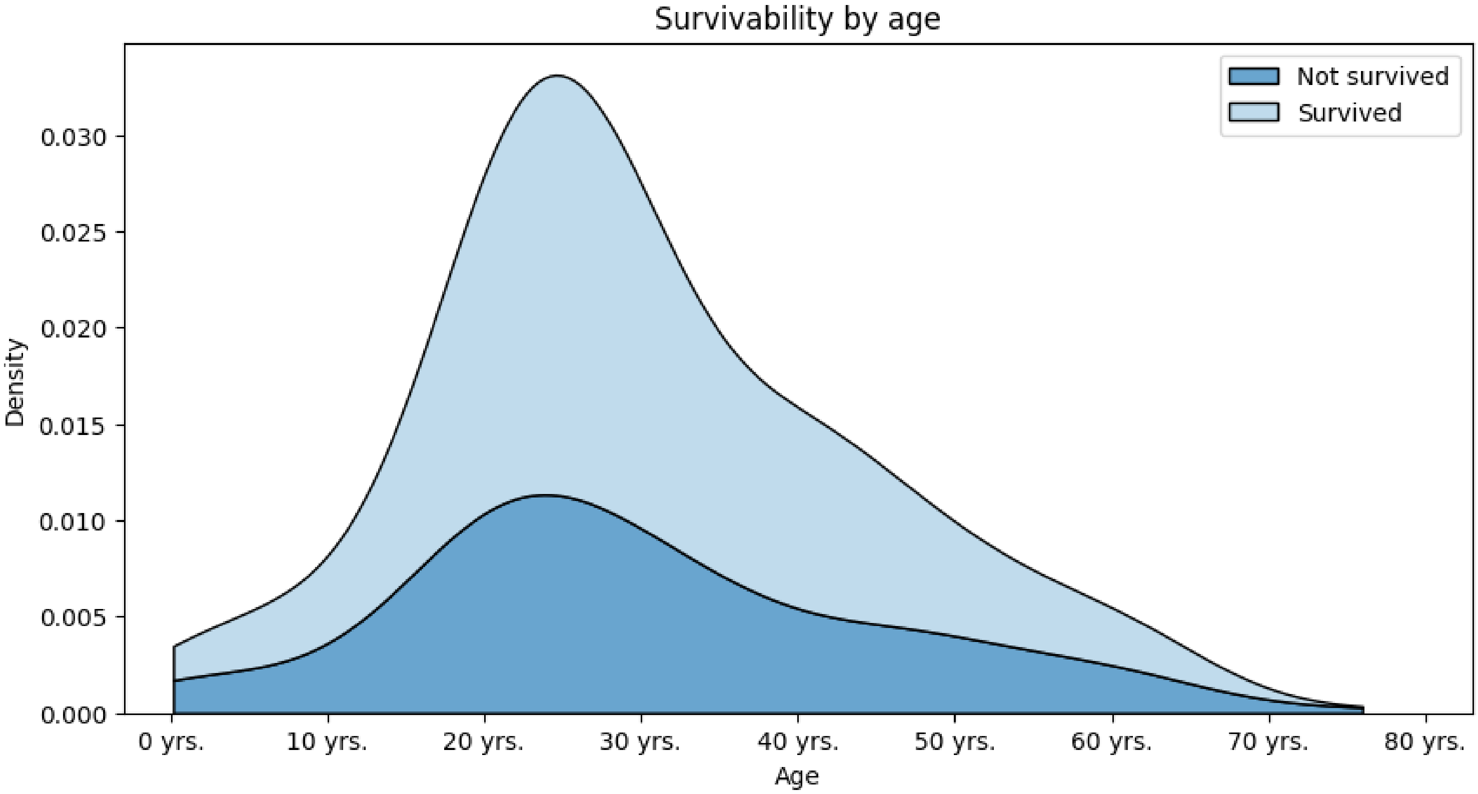


Comparing multiple distributions

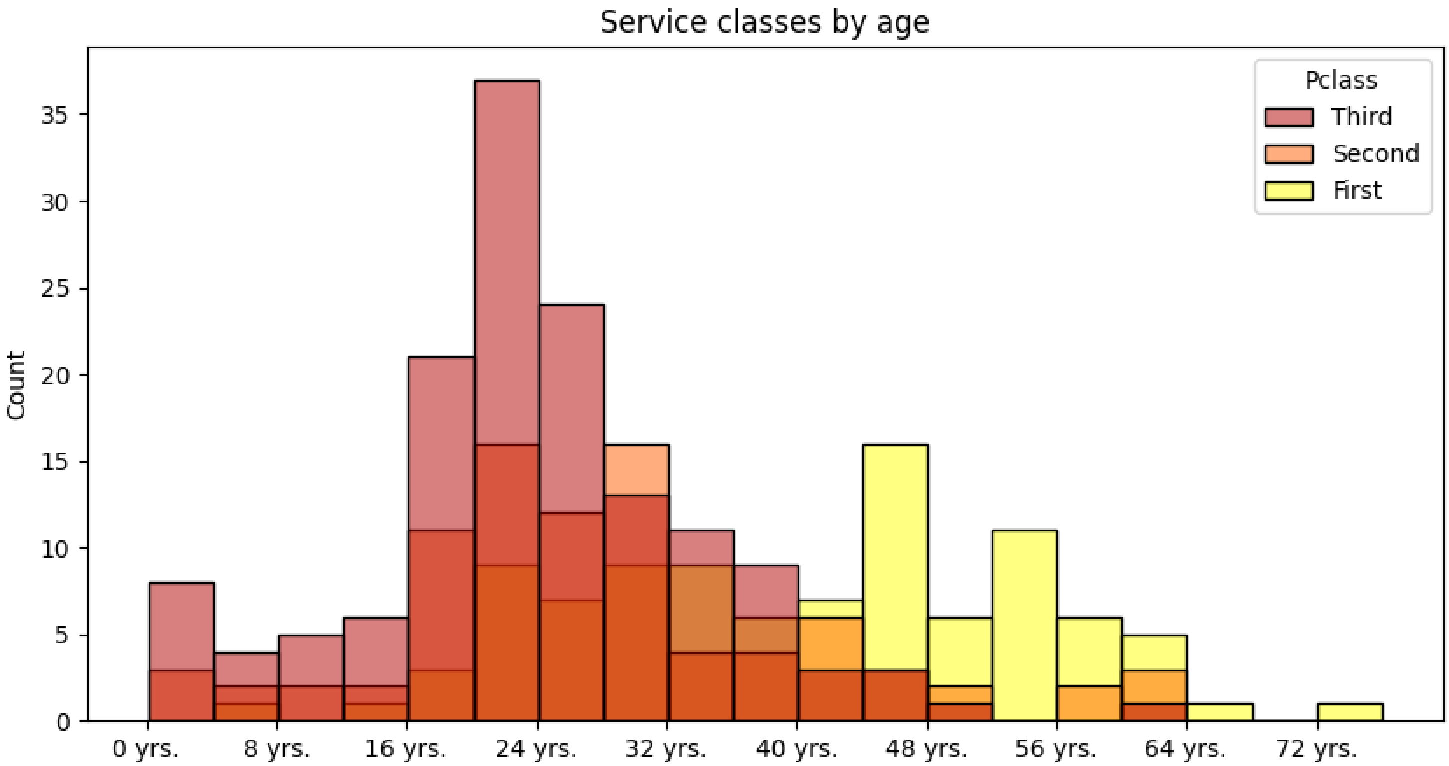
Stacked histogram



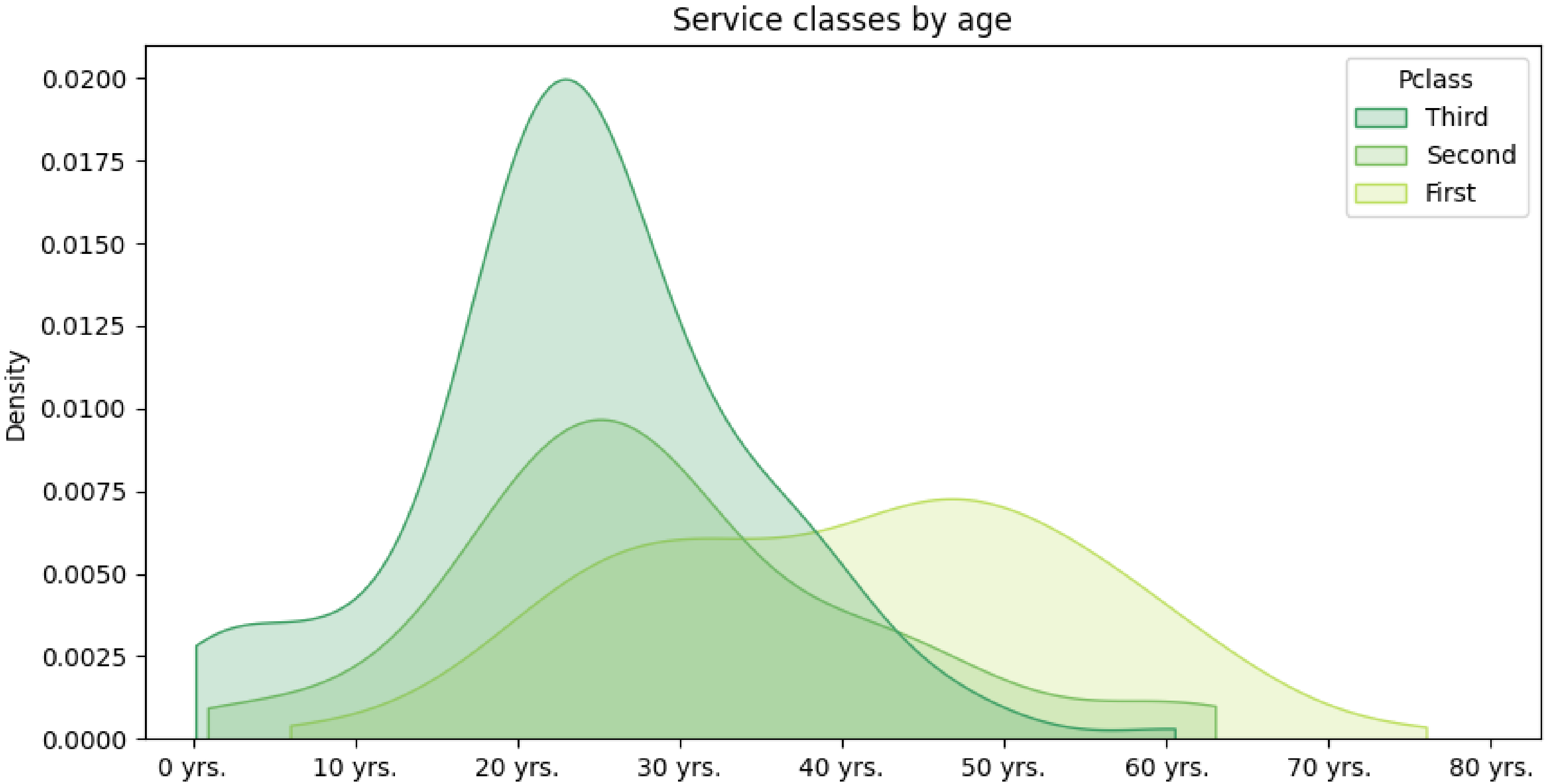
Stacked density plot



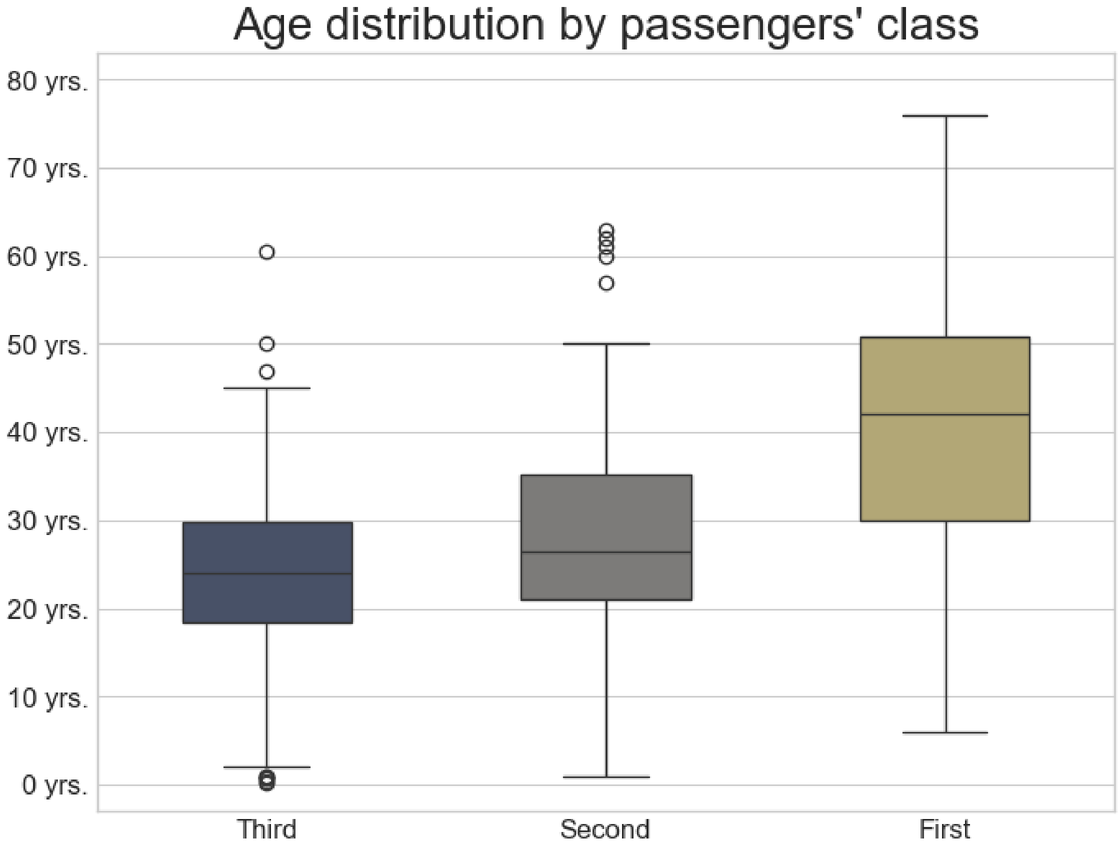
Overlapping density plot



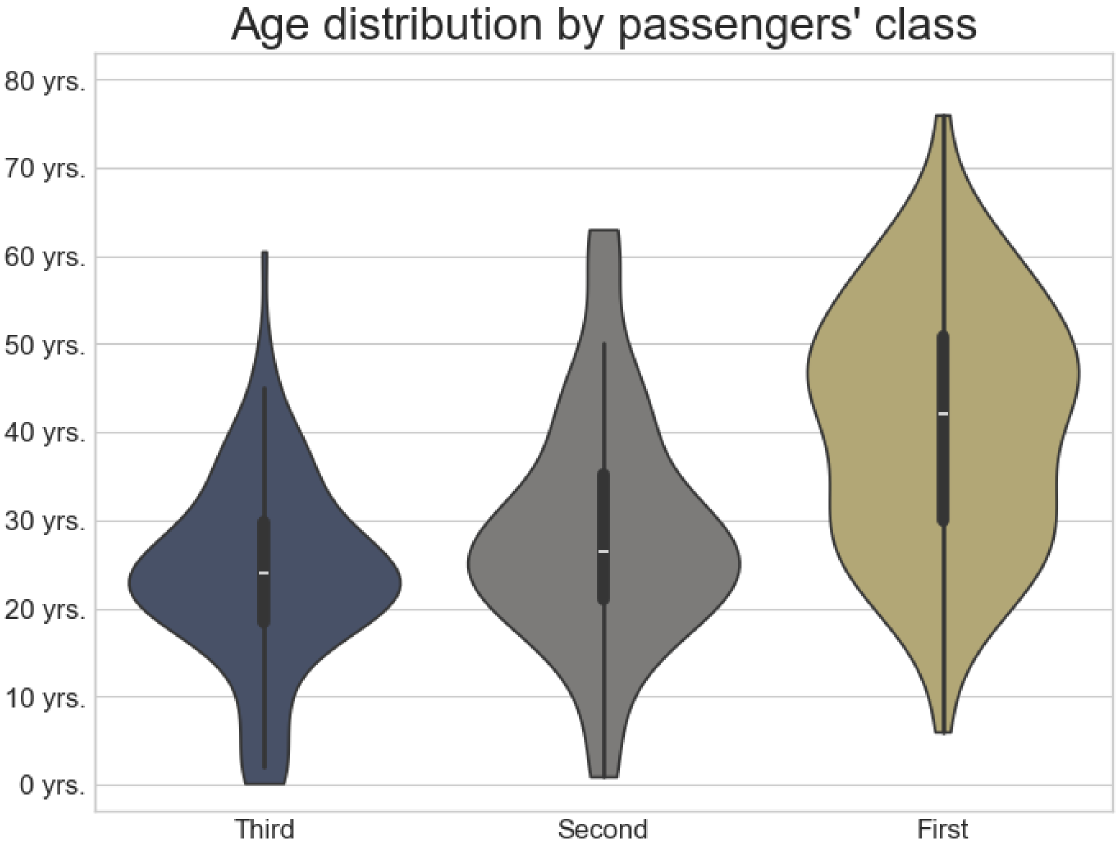
Overlapping density plot



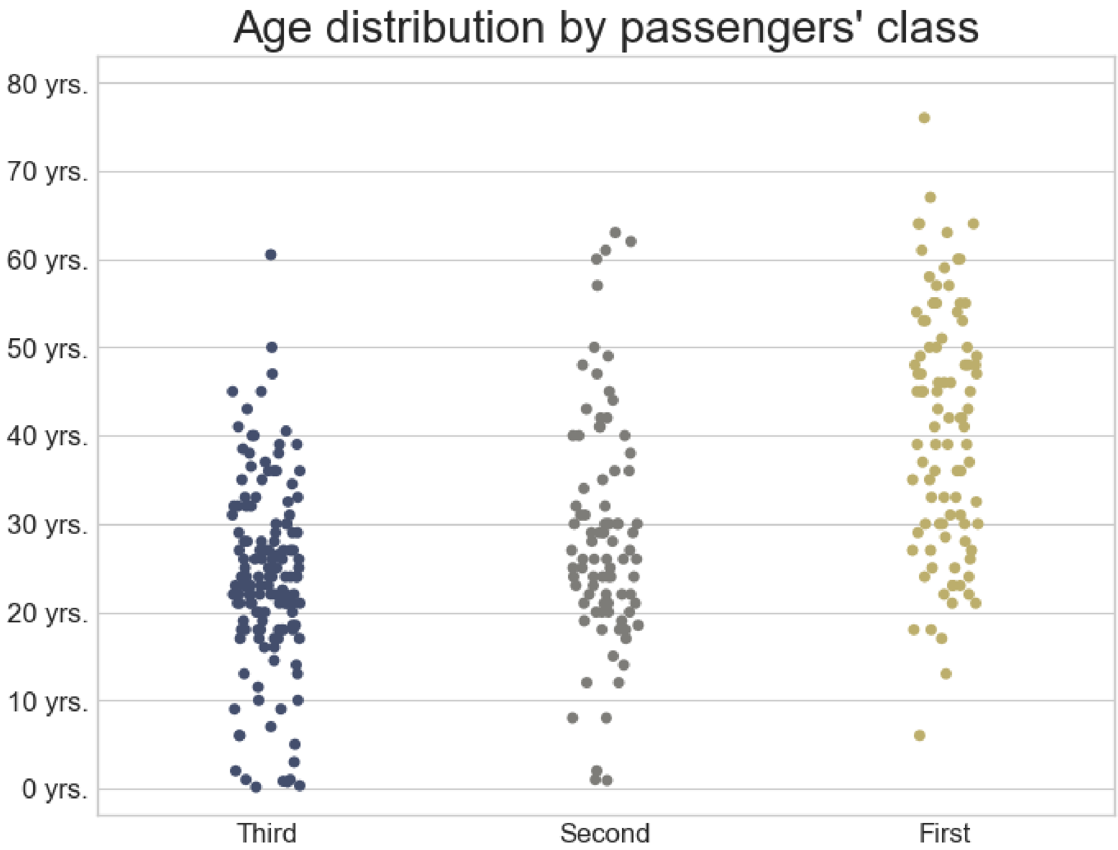
Box plots


Age distribution by passengers' class
80 yrs.
70 yrs.
60 yrs.
50 yrs.
40 yrs.
30 yrs.
20 yrs.
10 yrs.
0 yrs.
Third
Second
First

Violin plot

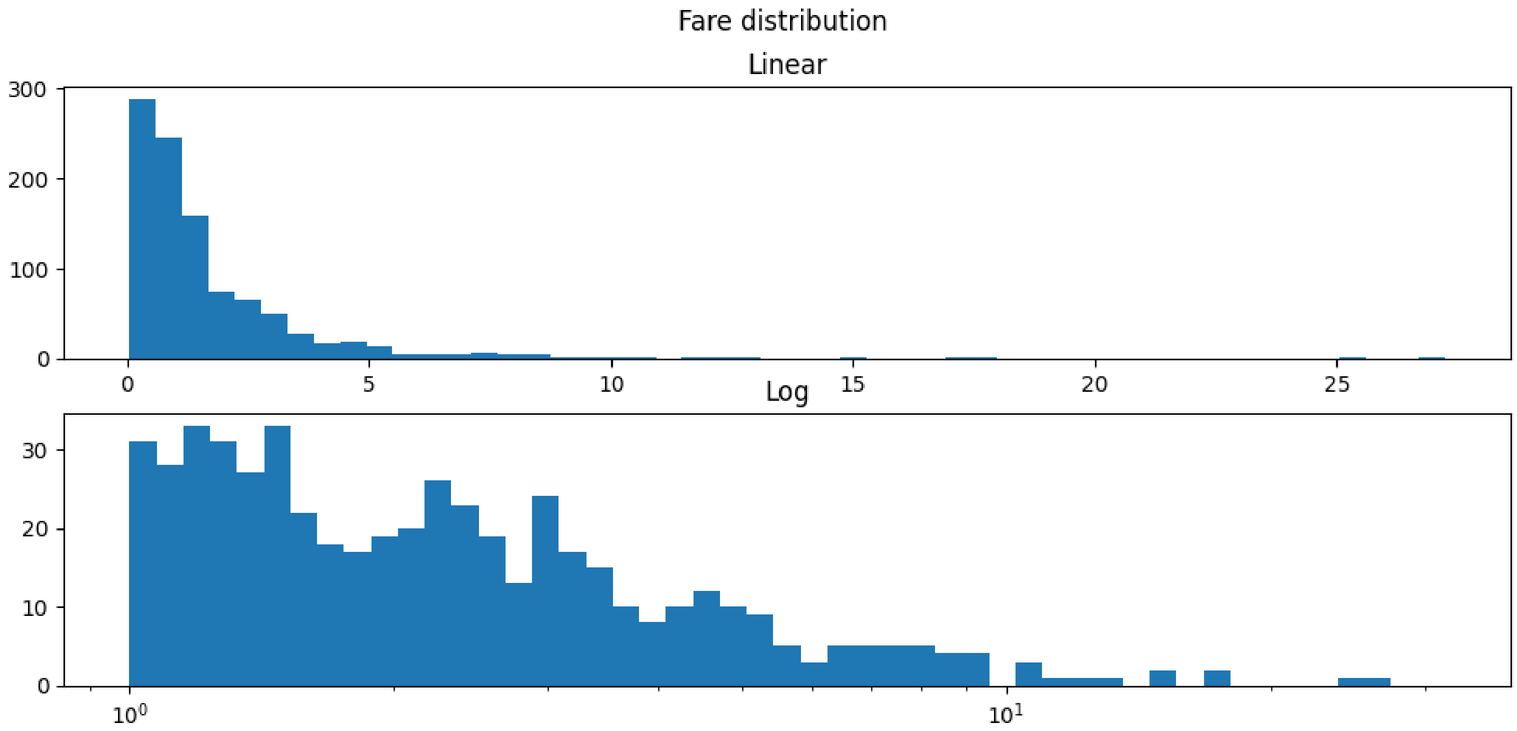

Age distribution by passengers' class
80 yrs.
70 yrs.
60 yrs.
50 yrs.
40 yrs.
30 yrs.
20 yrs.
10 yrs.
0 yrs.
Third
Second
First

Strip plot


Age distribution by passengers' class
80 yrs.
70 yrs.
60 yrs.
50 yrs.
40 yrs.
30 yrs.
20 yrs.
10 yrs.
0 yrs.
Third
Second
First

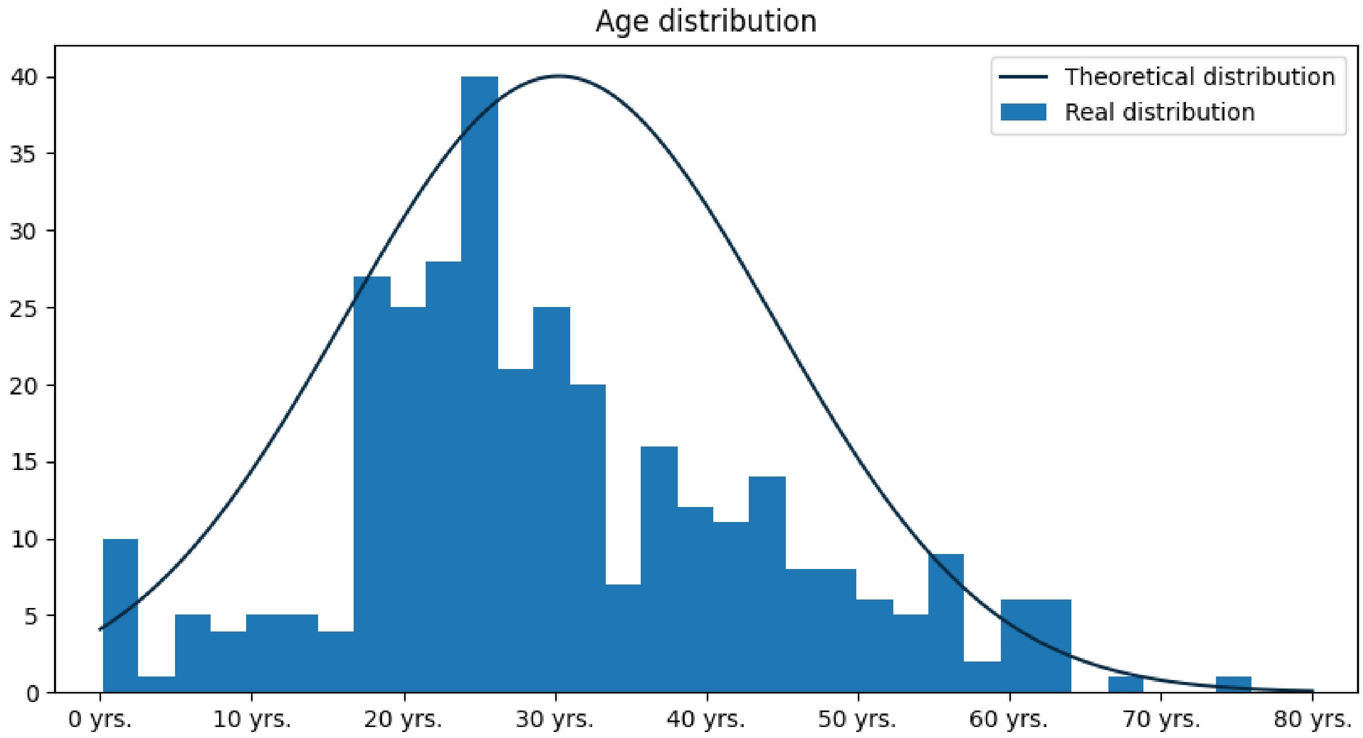
Logarithmic scale

Logarithmic scale

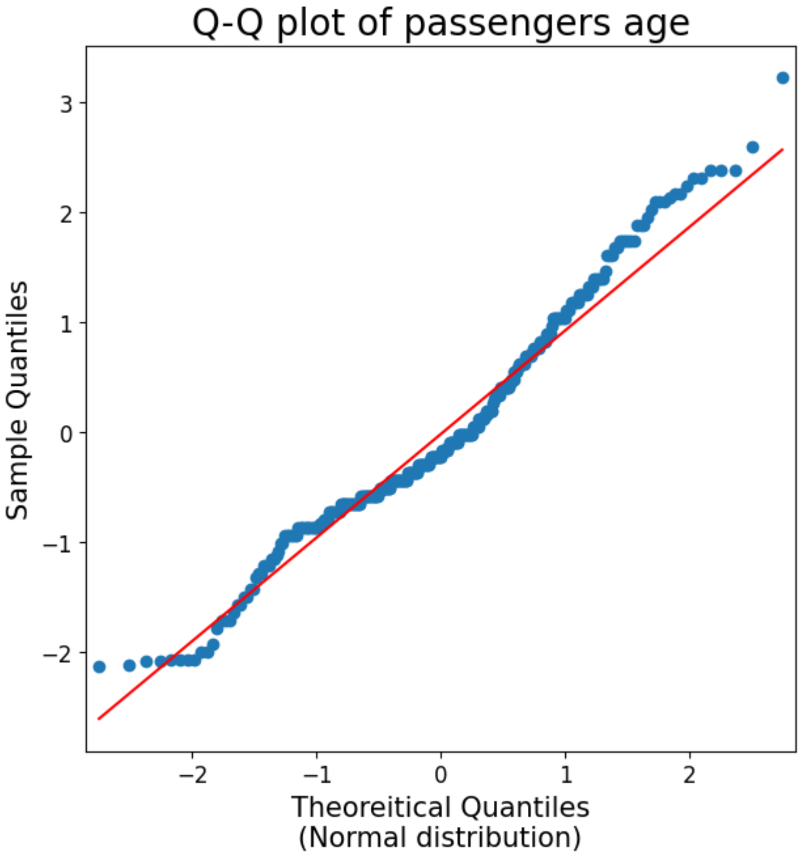


Real vs theoretical distribution

Real vs theoretical distribution

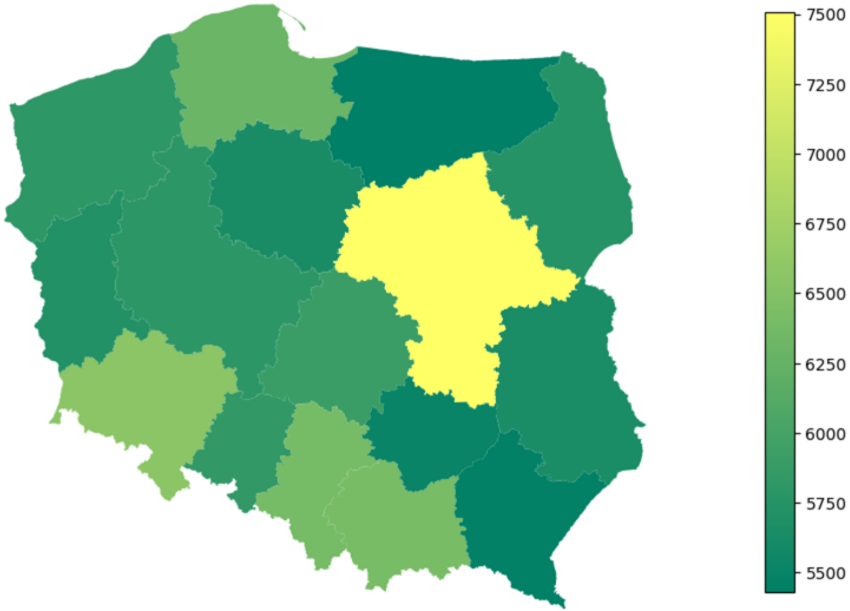


Real vs theoretical distribution

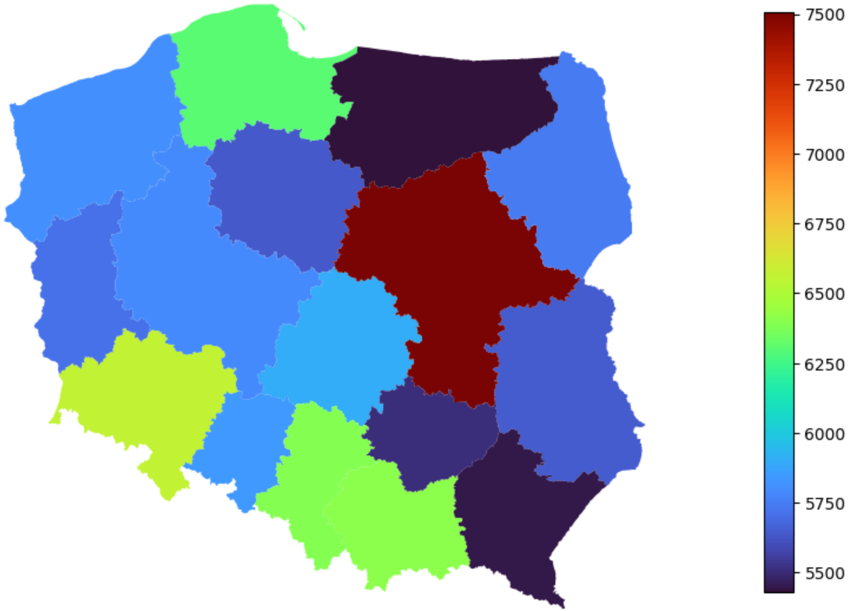


Frequently made mistakes

Incorrect color scale

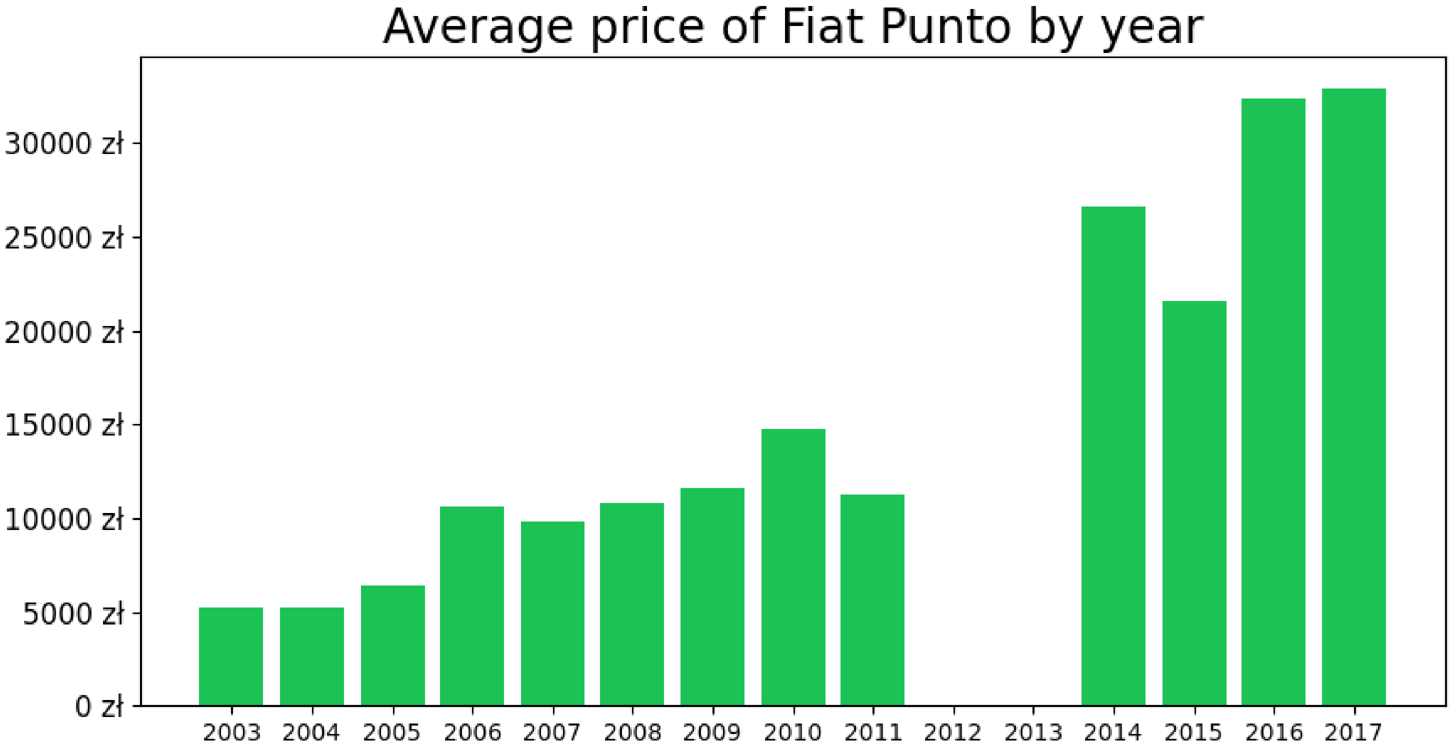


Salary by viovodership



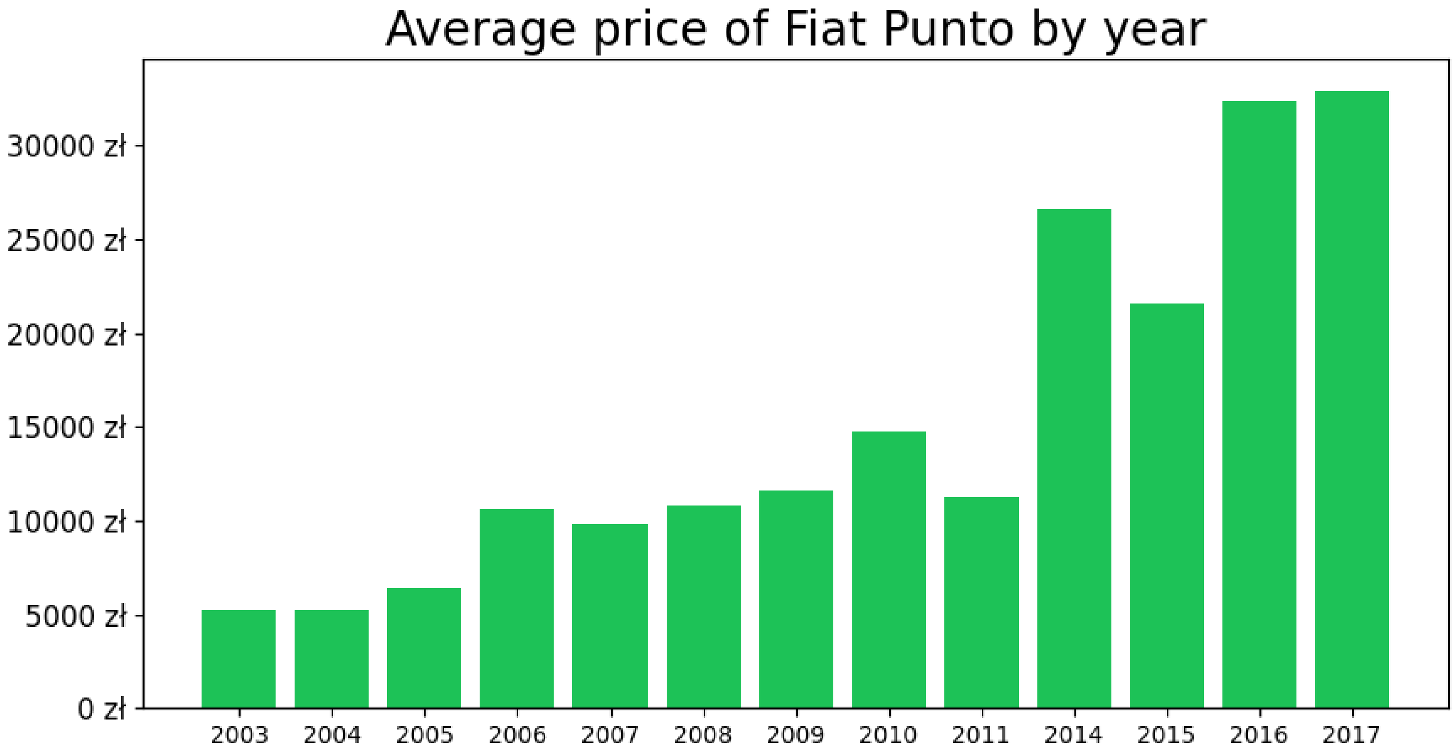
Salary by viovodership

Informing about missing data

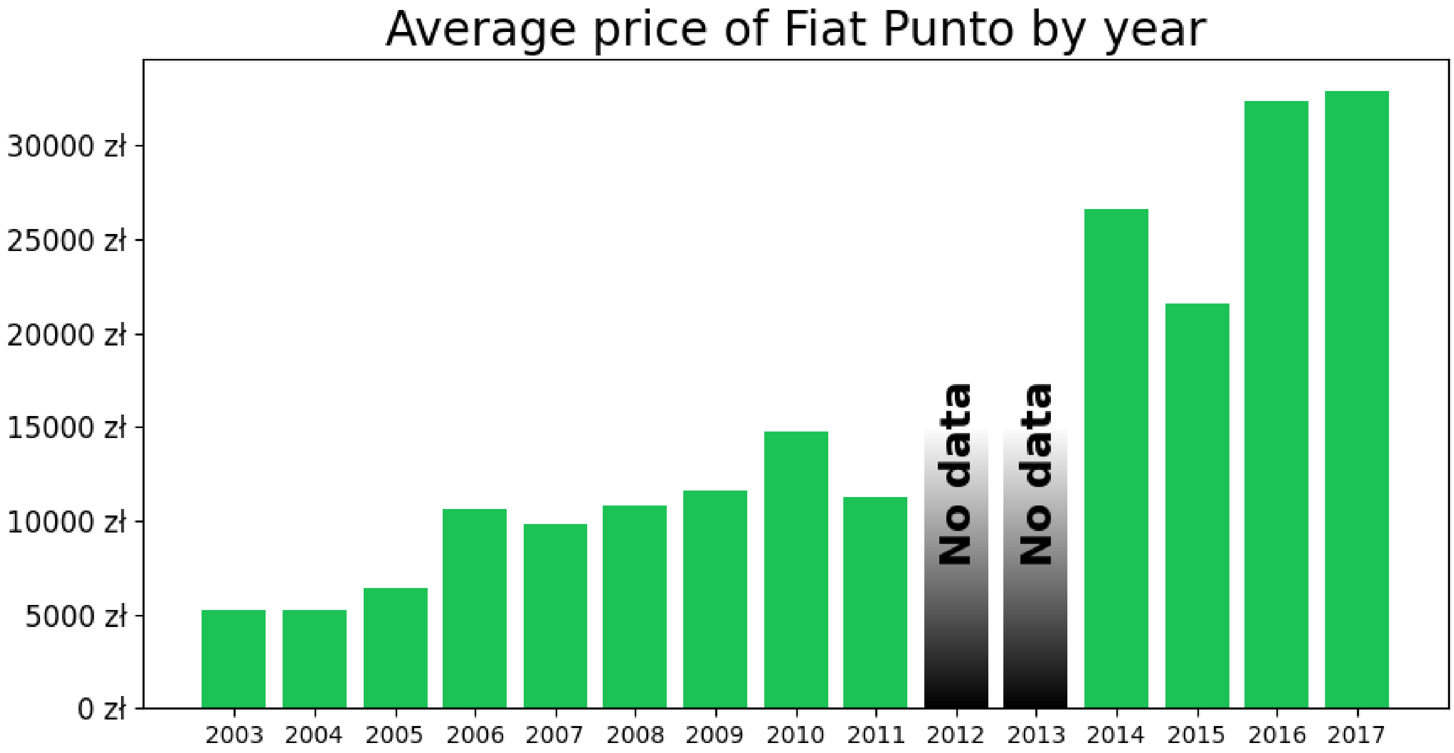


Average price of Fiat Punto by year

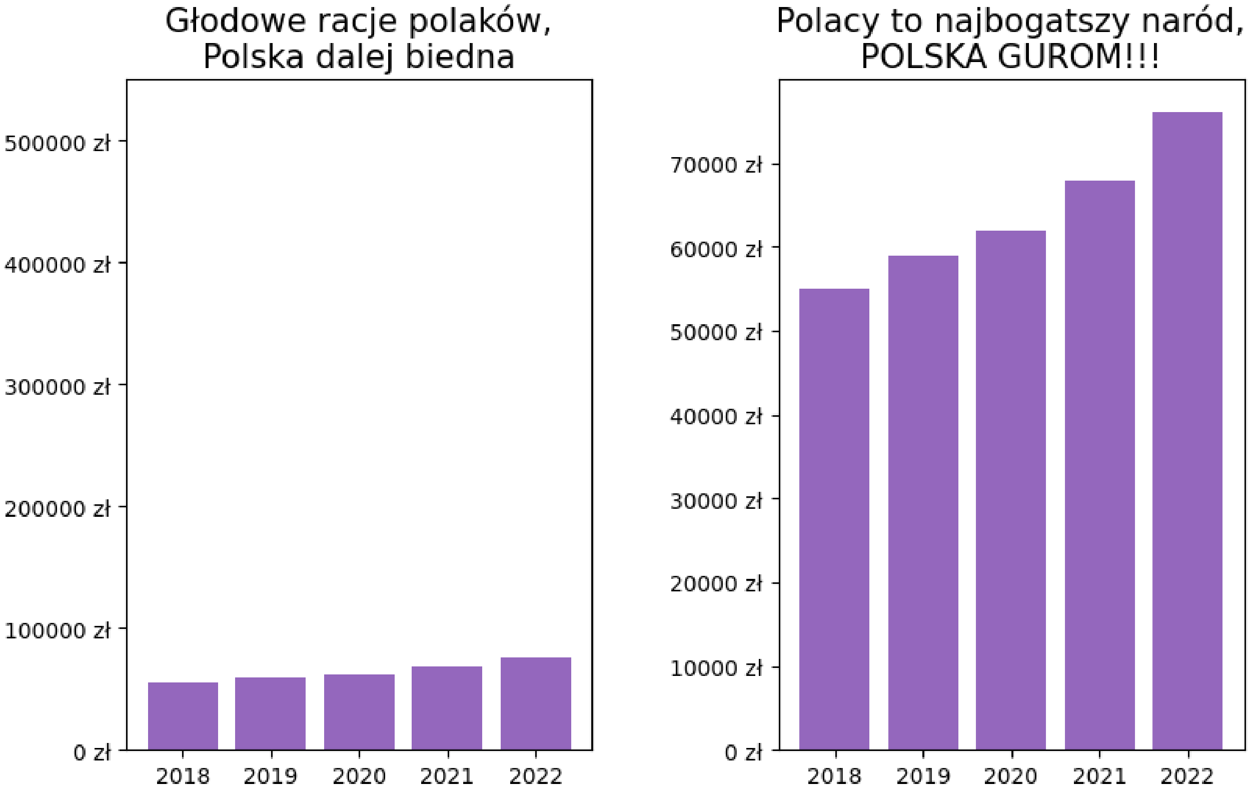
Informing about missing data



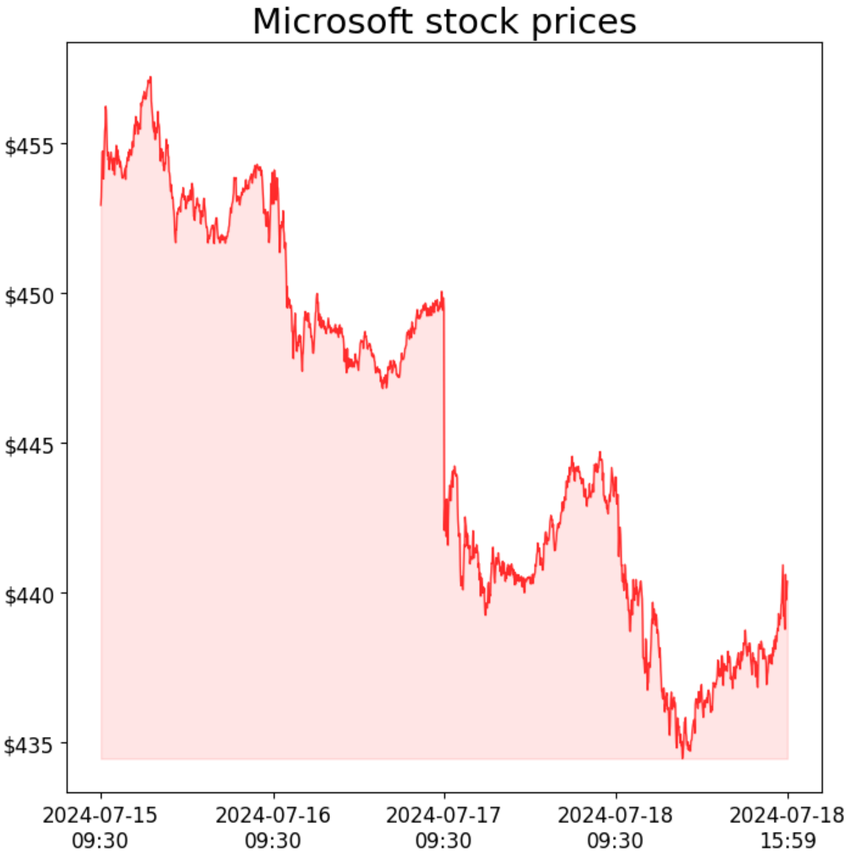
Informing about missing data



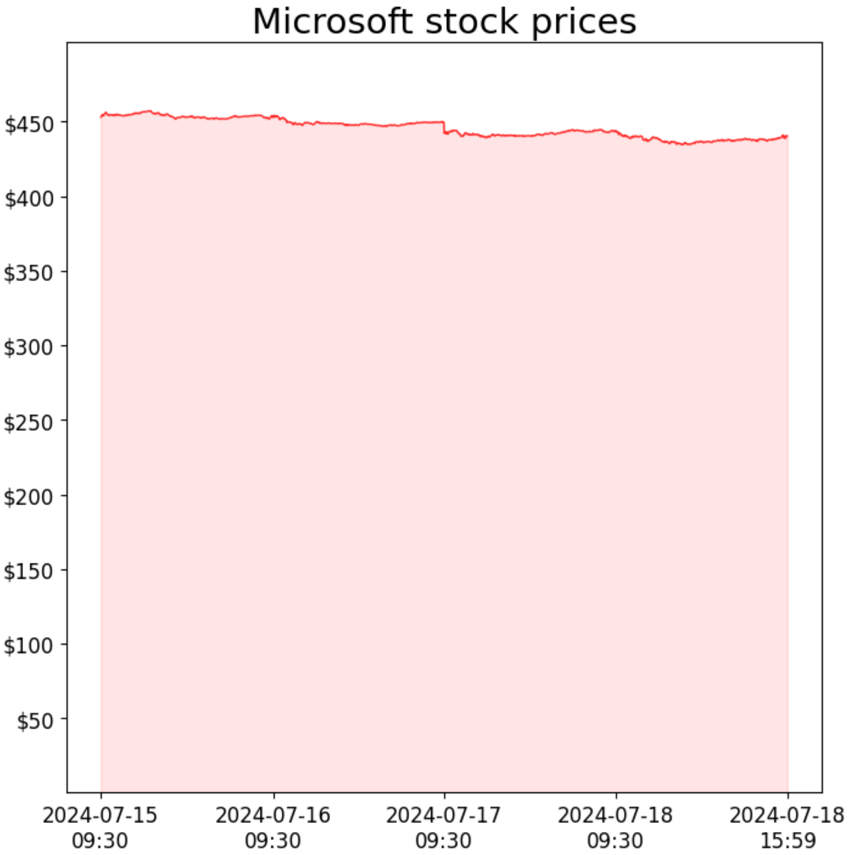
Manipulating scale



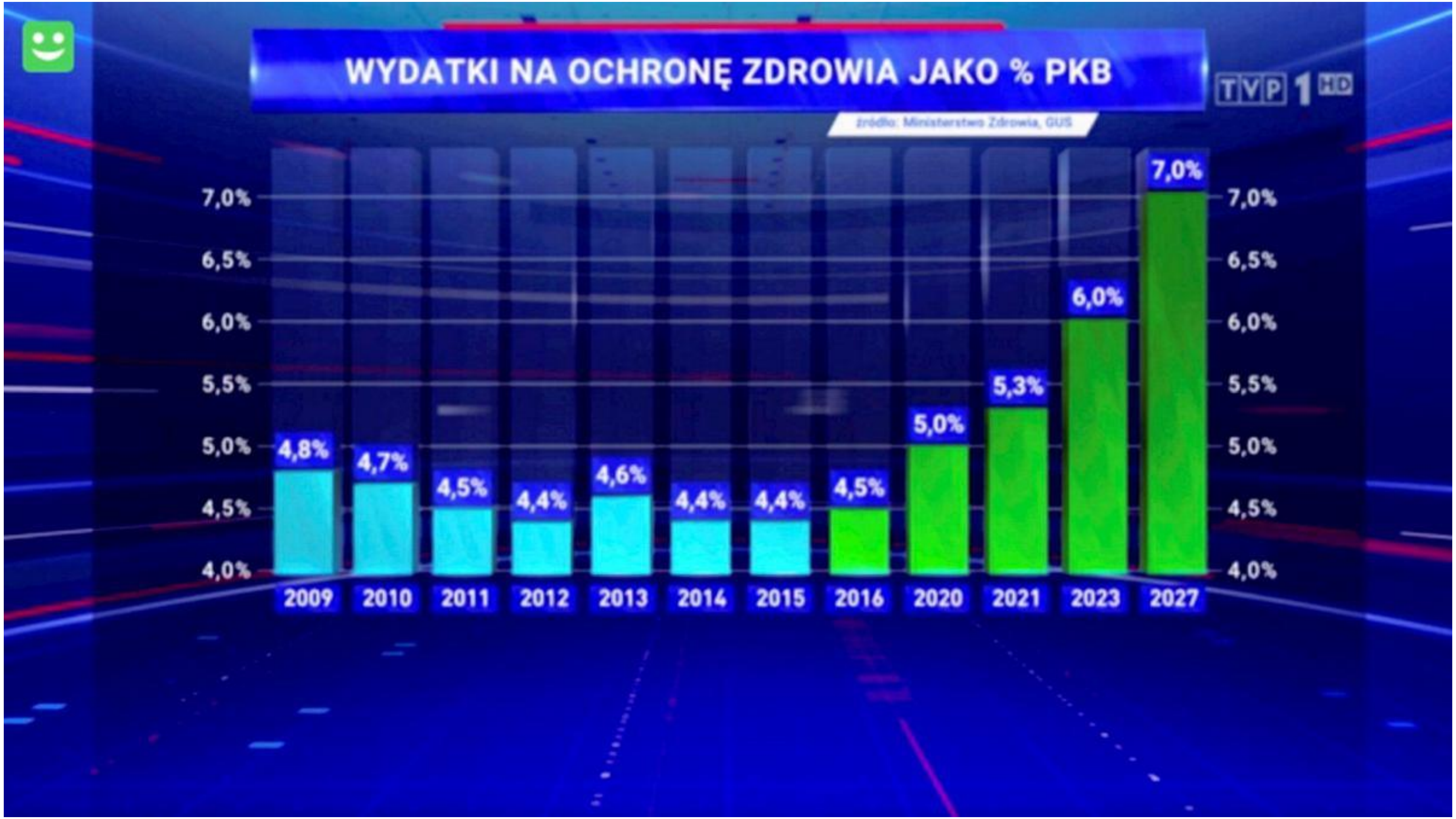
Manipulating scale



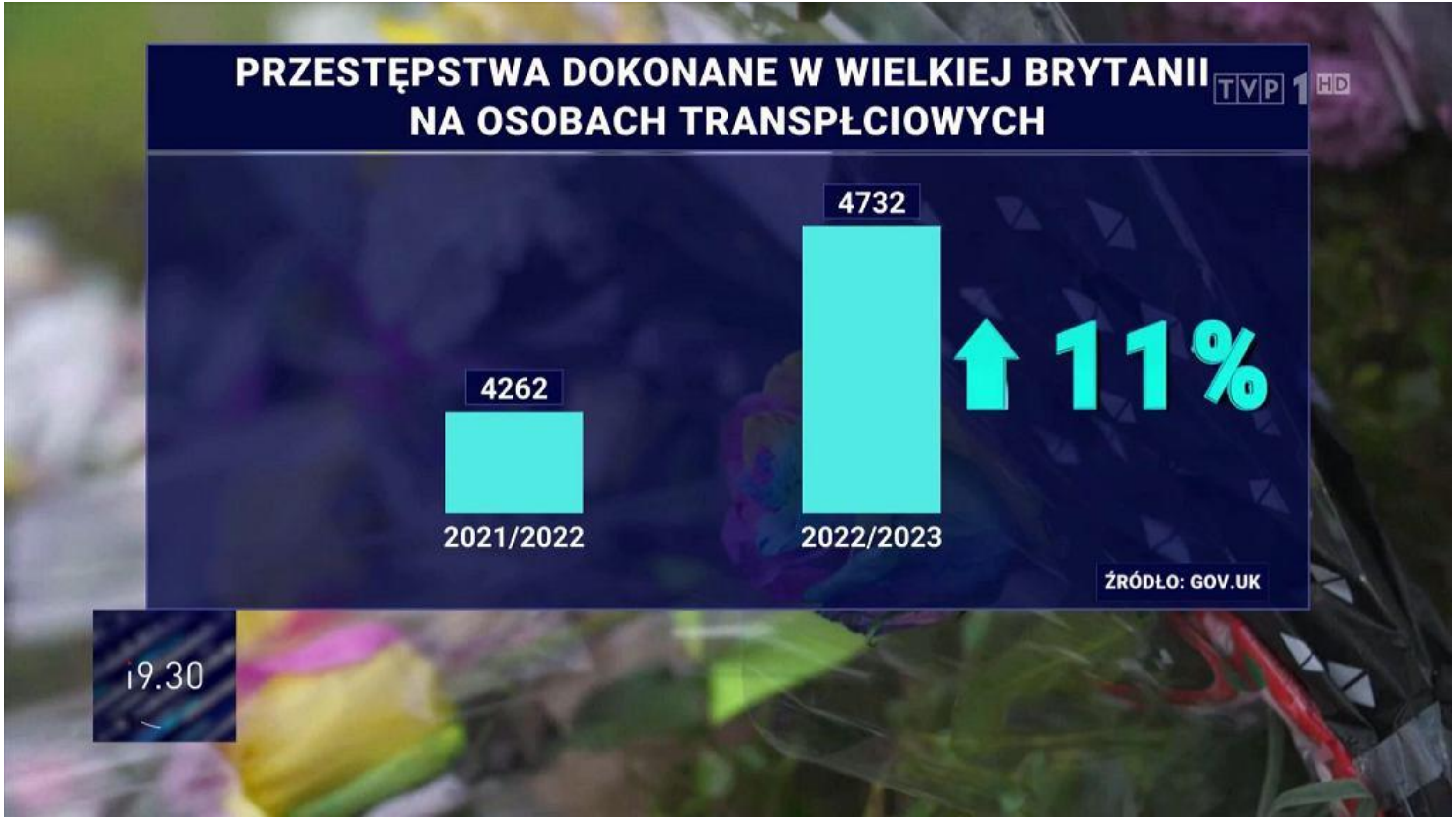
Manipulating scale



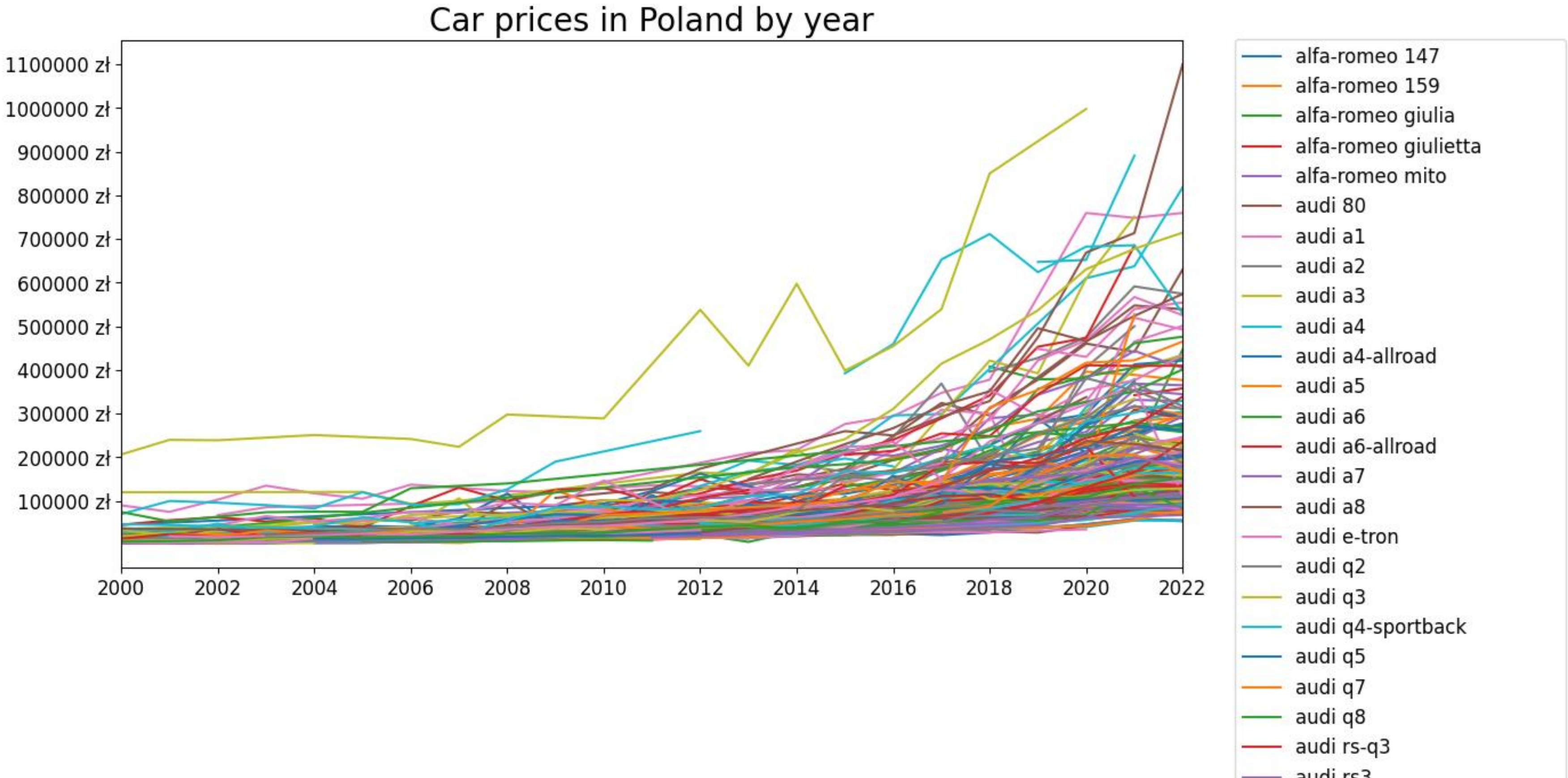
Manipulating scale



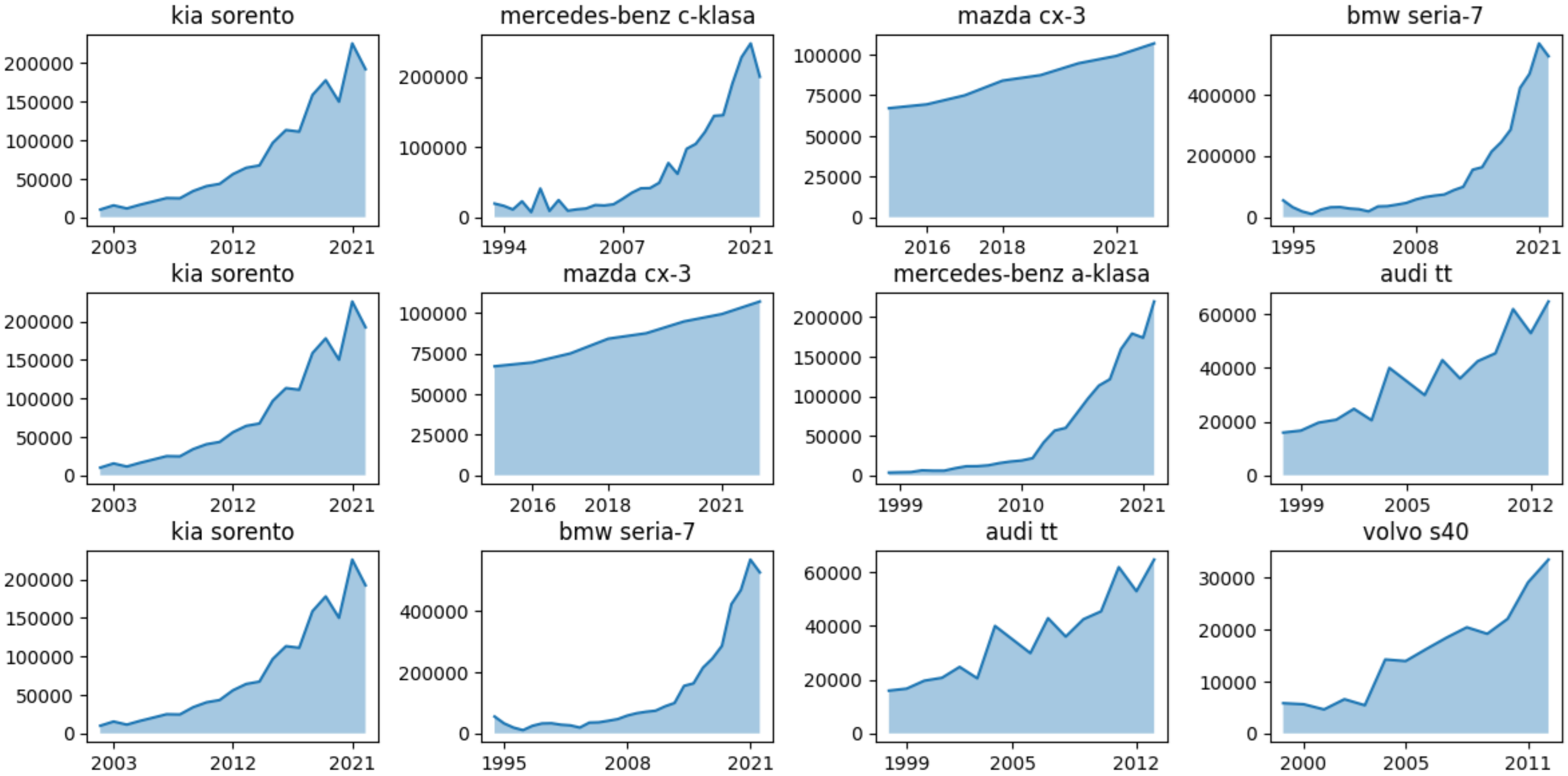
Manipulating scale


PRZESTĘPSTWA DOKONANE W WIELKIEJ BRYTANII
NA OSOBACH TRANSPŁCIOWYCH
TVP 1 HD
4262
2021/2022
4732
2022/2023
11%
ŹRÓDŁO: GOV.UK
19.30

Too many informations



Too many informations



References

Information sources:

[1] Wes McKinney, Python for Data Analysis, 3E (2022), Wes’s Blog

[2] Claus O. Wilke, Fundamentals of Data Visualization (2019), Claus Website

[3] Jarosław Drapala, Kernel Density Estimator explained step by step (2023), Medium - Towards Data Science

[4] 3Blue1Brown (Grant Sanderson), Why π is in the normal distribution (beyond integral tricks) (2023), Youtube

Data sources:

[5] Brenda N, Titanic dataset (2021), Kaggle

[6] Główny Urząd Statystyczny, Obwieszczenie w sprawie wysokości przeciętnego miesięcznego wynagrodzenia brutto w gospodarce narodowej w województwach w 2022 roku (2023), GUS

[7] Aleksandr Glotov, Car Prices Poland (2021), Kaggle

Other:

[8] My private notes about data visualization an examples

Presentation author: Maksymilian Norkiewicz