Data Mining Exercise 1

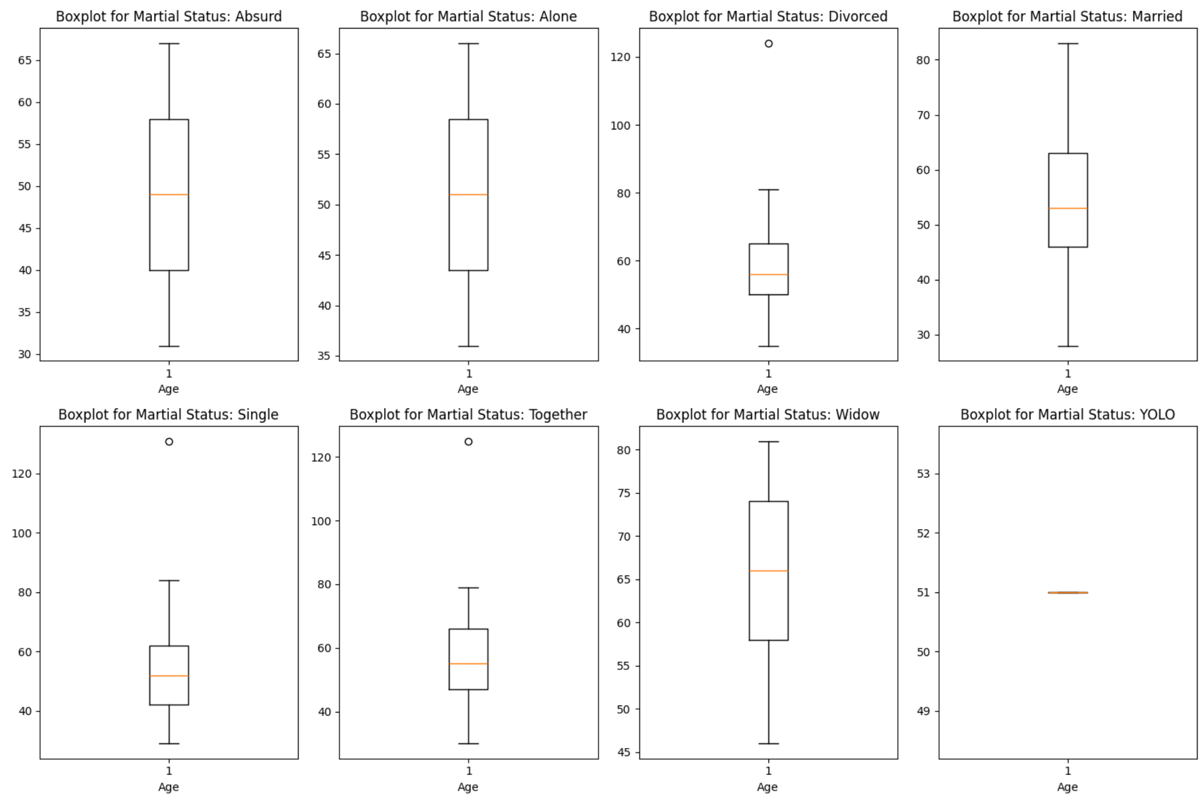
Magdalena König

01455794

1. Marketing Data

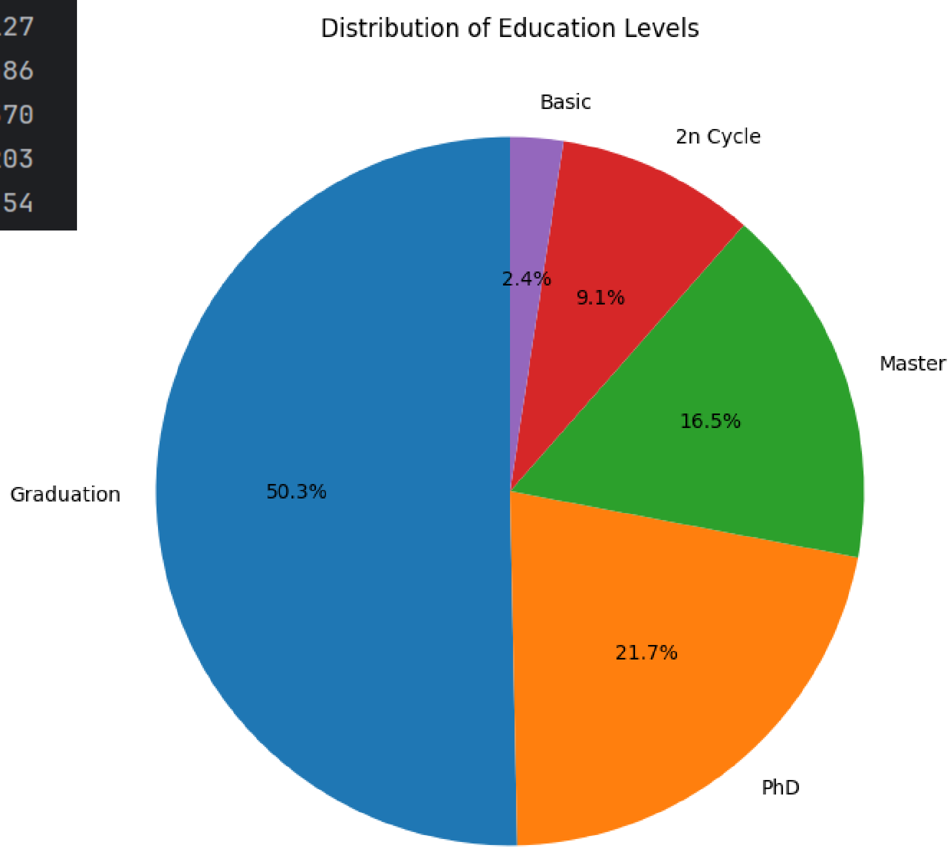
- Task 1: Calculate the age and get the distribution within each marital status

	count	mean	std	min	25%	50%	75%	max
marital_status								
Absurd	2.0	49.000000	25.455844	31.0	40.0	49.0	58.0	67.0
Alone	3.0	51.000000	15.000000	36.0	43.5	51.0	58.5	66.0
Divorced	232.0	57.724138	10.686874	35.0	50.0	56.0	65.0	124.0
Married	864.0	54.420139	11.404421	28.0	46.0	53.0	63.0	83.0
Single	480.0	52.510417	12.872098	29.0	42.0	52.0	62.0	131.0
Together	580.0	56.253448	11.863337	30.0	47.0	55.0	66.0	125.0
Widow	77.0	65.441558	9.335125	46.0	58.0	66.0	74.0	81.0
YOLO	2.0	51.000000	0.000000	51.0	51.0	51.0	51.0	51.0

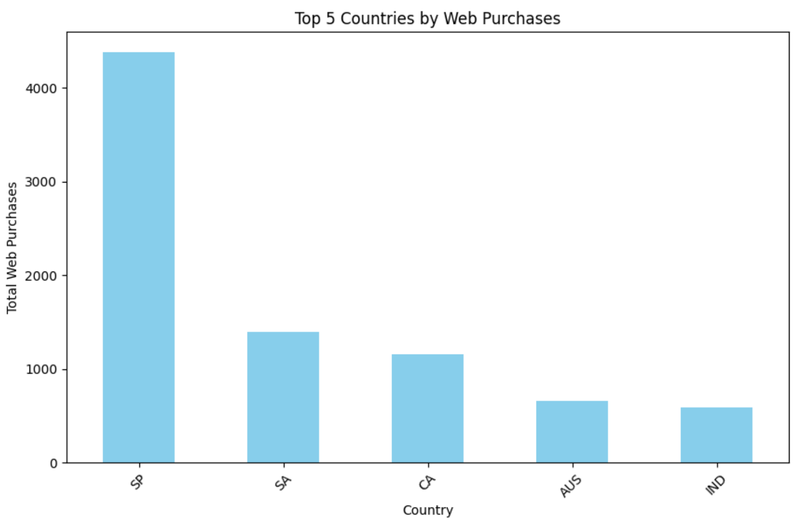


- Task 2 - What is the distribution of the education?

```
education
Graduation    1127
PhD            486
Master         370
2n Cycle       203
Basic           54
```

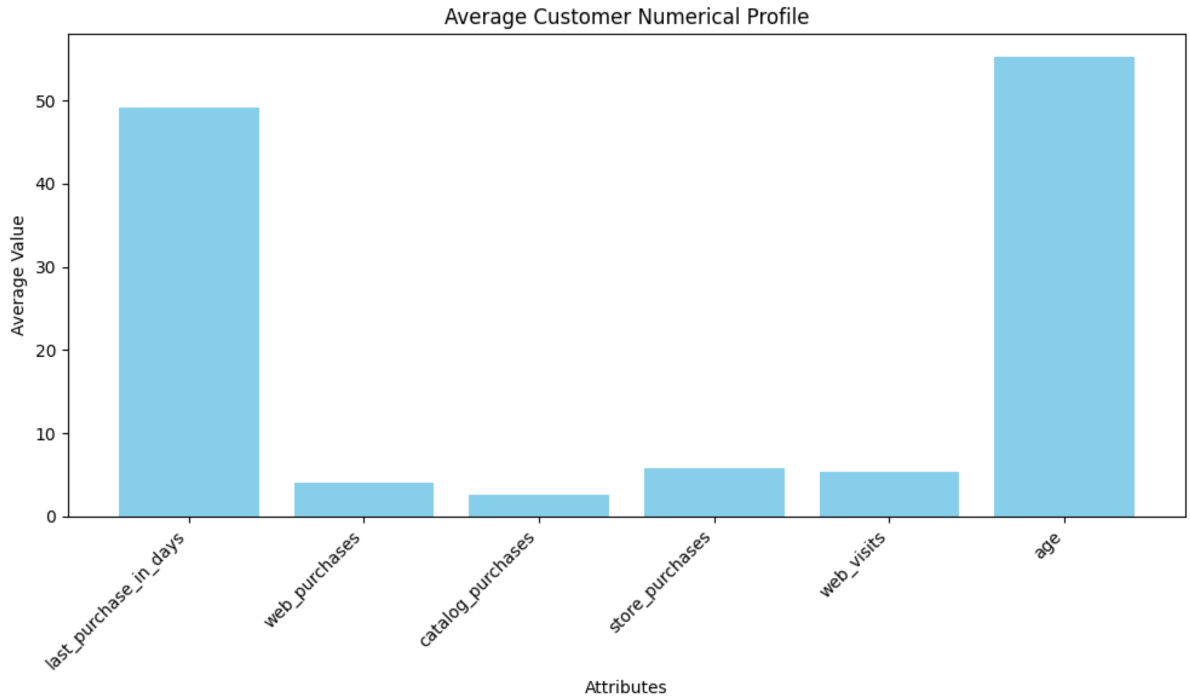


- Task 3 - Which country has the most web purchases?
    - Spain with 4382 purchases



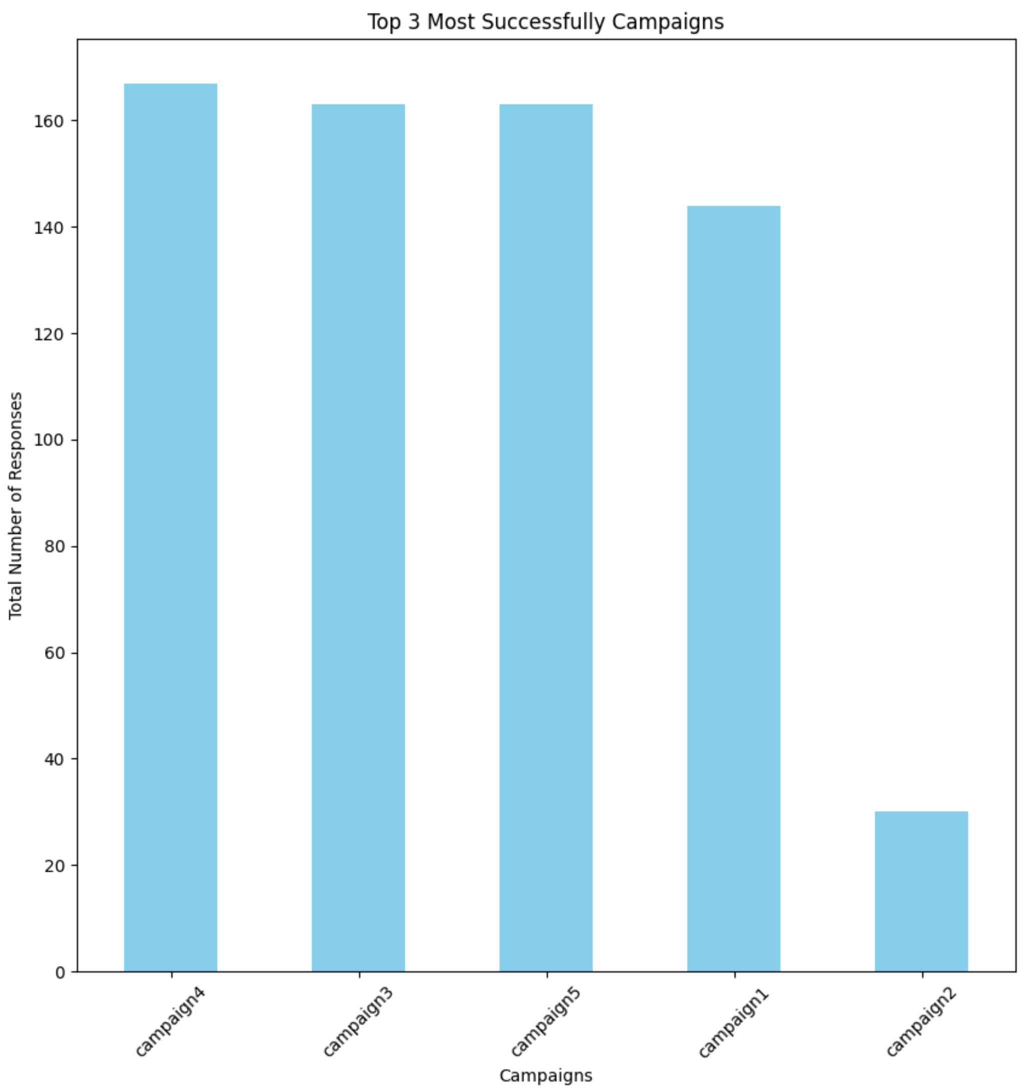
- Task 4 - How does the average customer look like?

kids	0.444196
teens	0.50625
last_purchase_in_days	49.109375
web_purchases	4.084821
catalog_purchases	2.662054
store_purchases	5.790179
web_visits	5.316518
campaign3	0.072768
campaign4	0.074554
campaign5	0.072768
campaign1	0.064286
campaign2	0.013393
age	55.194196
education	Graduation
marital_status	Married
country	SP

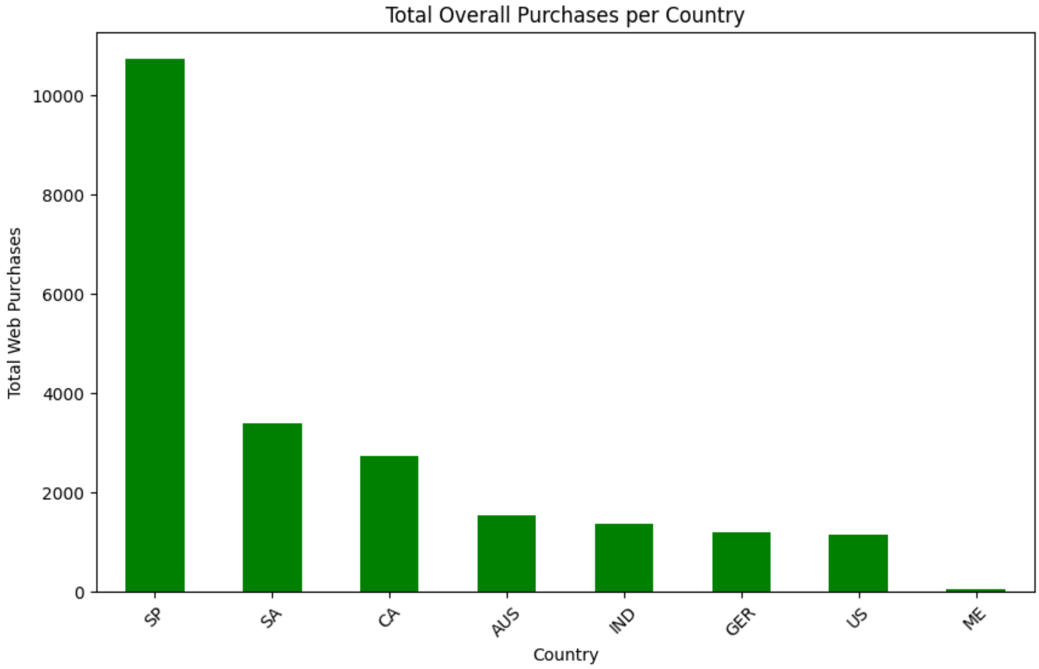


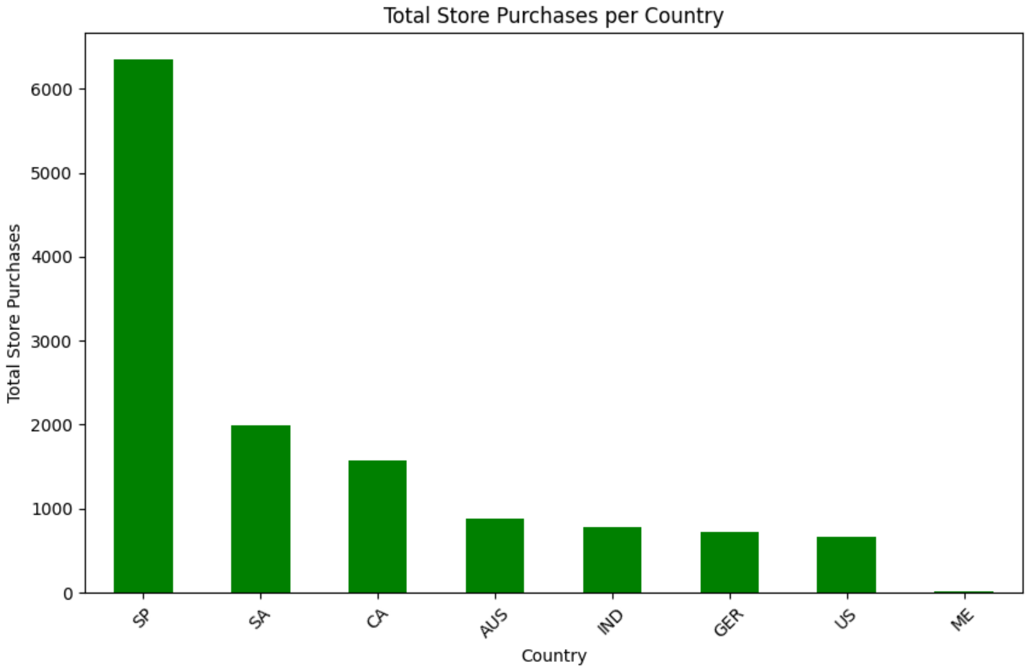
- Task 5 - Which previous marketing campaign was most successful?
    - Campain 4 is the most sucessfull

```
Most Successful Campaign: campaign4
Number of Responses: 167
campaign4    167
campaign3    163
campaign5    163
campaign1    144
campaign2     30
```

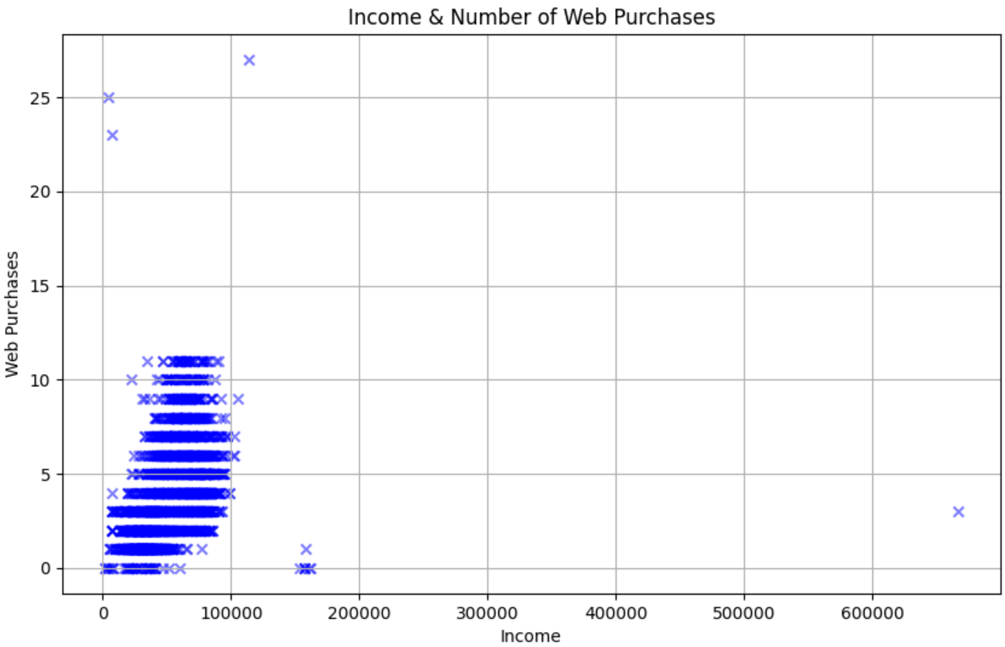


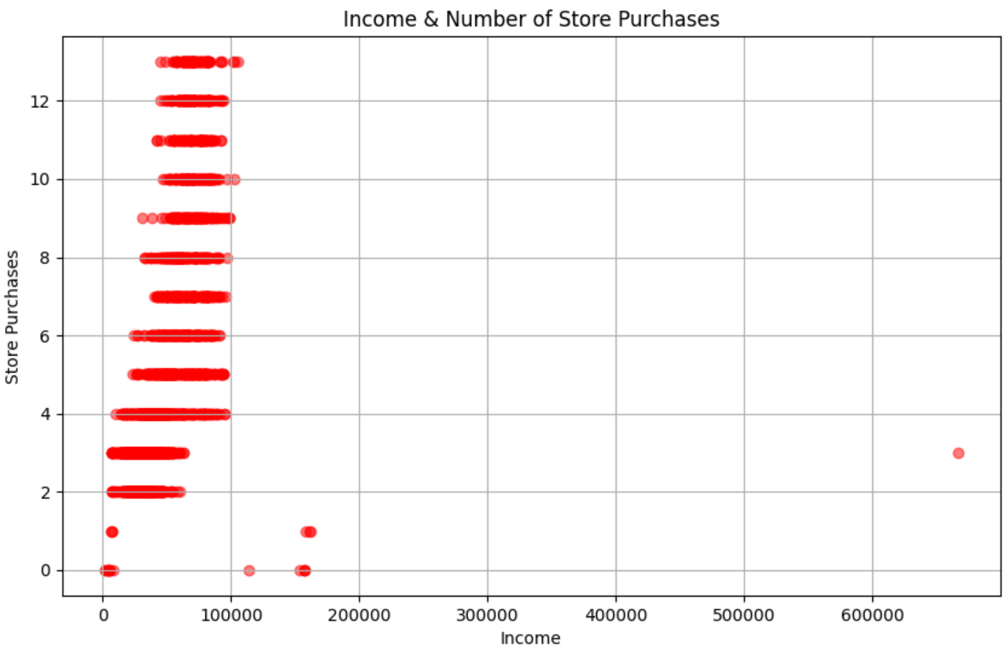
- Other analysis:
  - Countries x Store/Overall Purchases

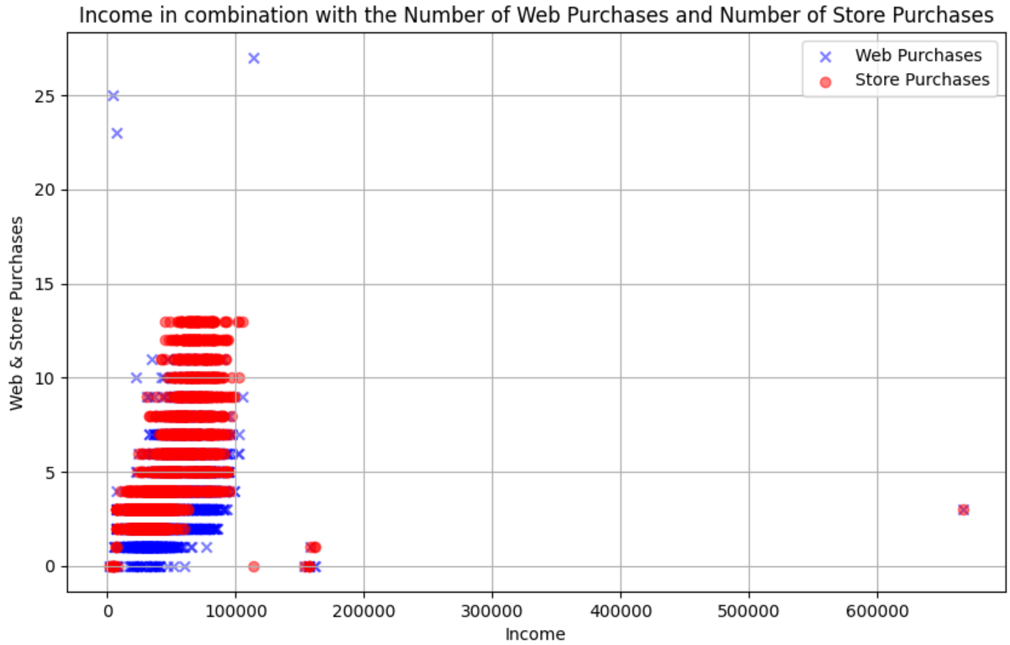




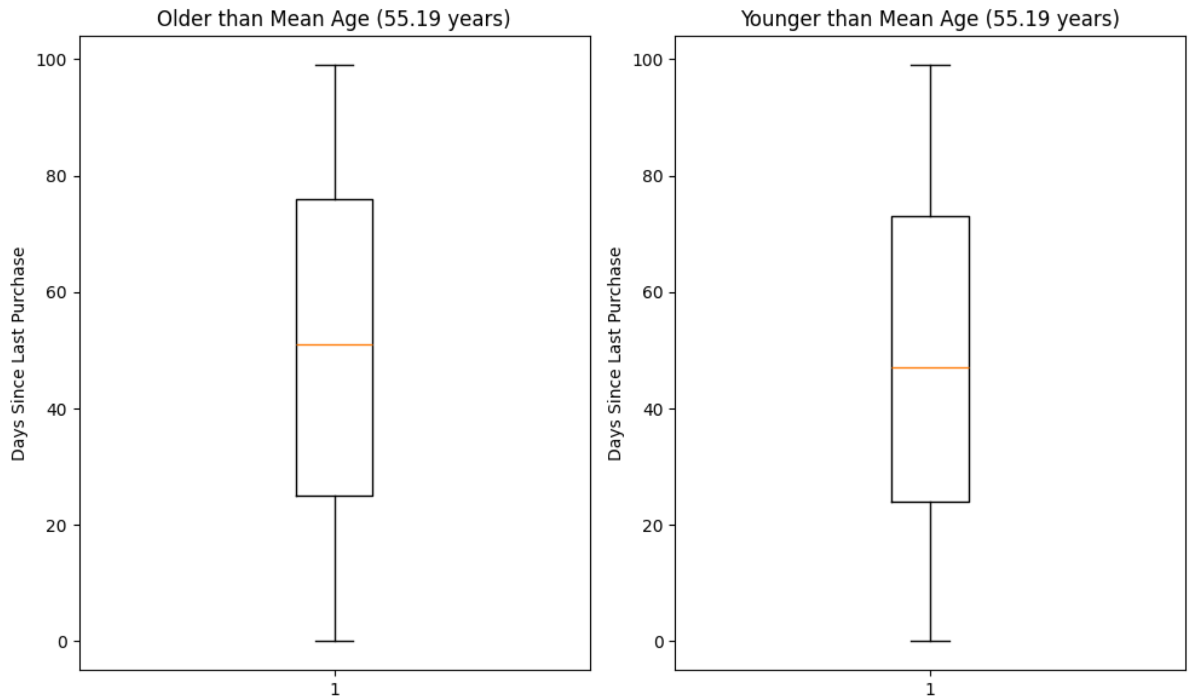
- Income x Purchases

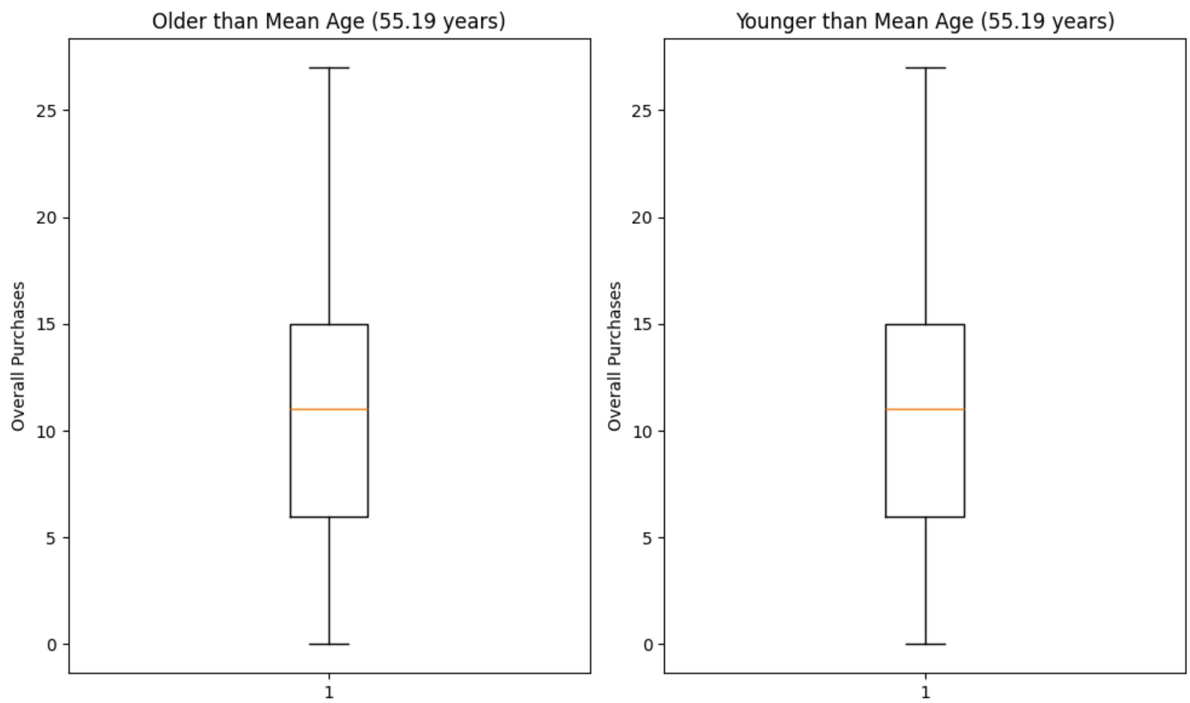

Income & Number of Web Purchases
Web Purchases
Income
0
5
10
15
20
25
100000
200000
300000
400000
500000
600000

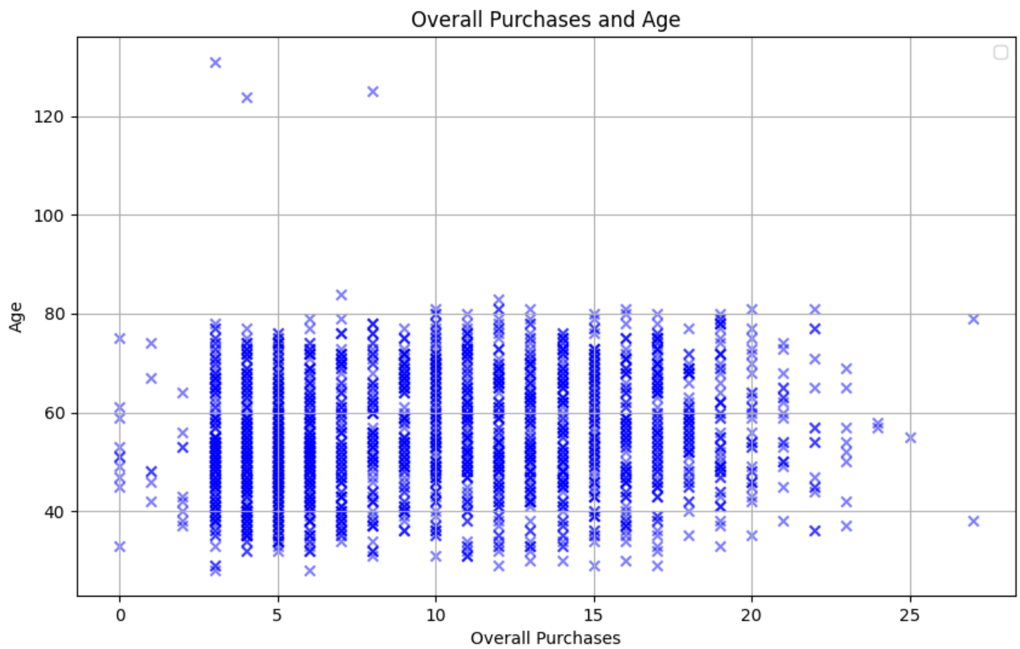

Income & Number of Store Purchases
Store Purchases
Income
0
2
4
6
8
10
12
100000
200000
300000
400000
500000
600000



- Age Analysis – for Marketing reasons to establish a well fit marketing strategy accoring to the age of the customers





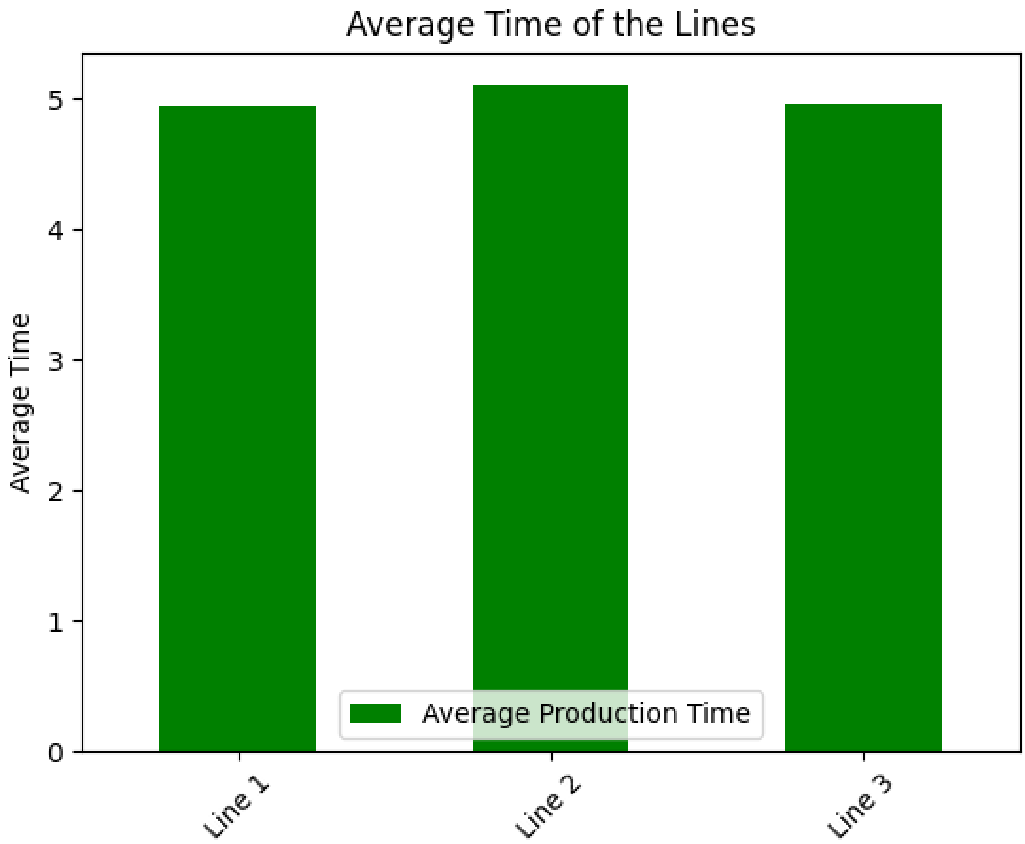


- When it comes to age the quality of data is questionable – there are people older than 120 years which is highly doubtable.

2. Production Lines

- Task 1 - If the order has to be produced as fast as possible, which of the production lines do you choose? Why?
  - The average speed of production will be used here:
    - Line 1: 4.9374328310594775
    - Line 2: 5.100732470496723
    - Line 3: 4.960518336826572

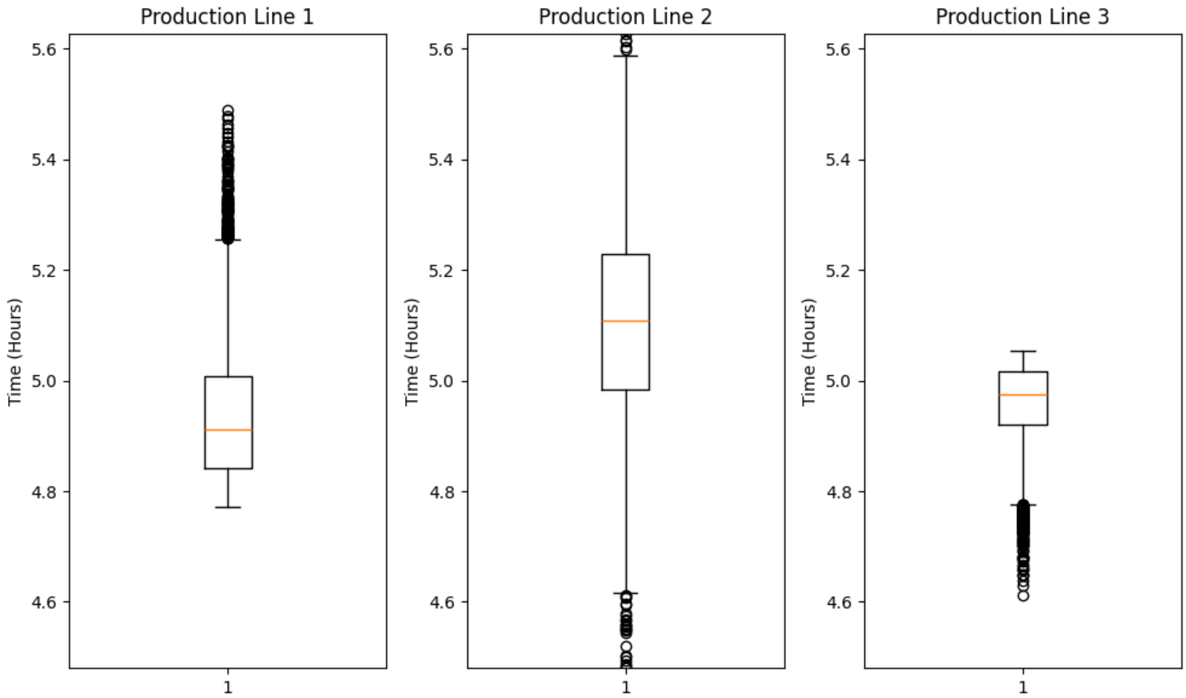
- Production Line 1 will be chosen.

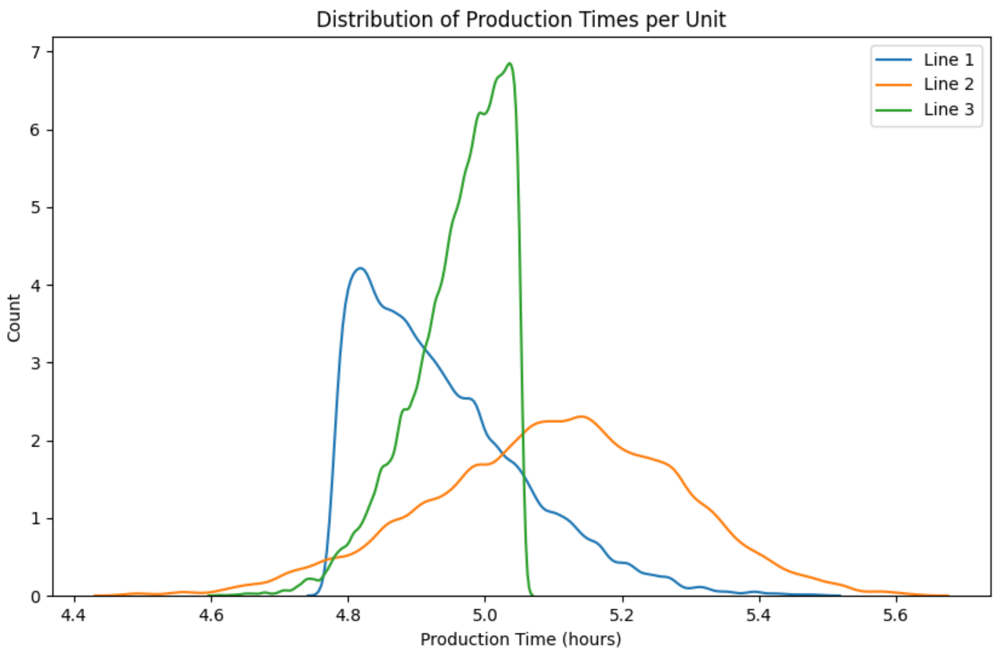


- Task 2 - If the order needs to be produced for just-in-time production, i.e. a reliable estimation of production time is necessary, which of the production lines do you choose? Why?
  - For just in time production not only the average production speed but also the viariety oft he data. Does the production lines has many outliners? How is the distribution in time and varience?
  - Reliable estimation of production time is crucial for JIT production, choosing the right production line involves considering factors beyond just the average production times. It's essential to assess the consistency and predictability of each production line's performance.

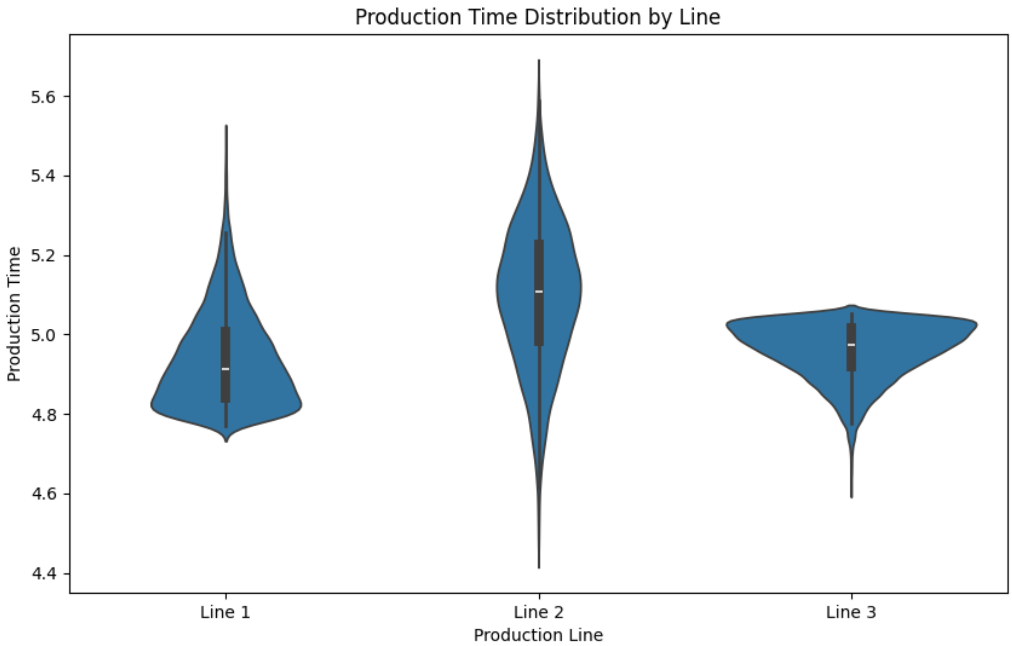
	Production Line	Average Production Time	Standard Deviation
0	Line 1	4.937433	0.119258
1	Line 2	5.100732	0.180232
2	Line 3	4.960518	0.069642

  - For this reason Production Line 3 shall be used when predictable times are the goal





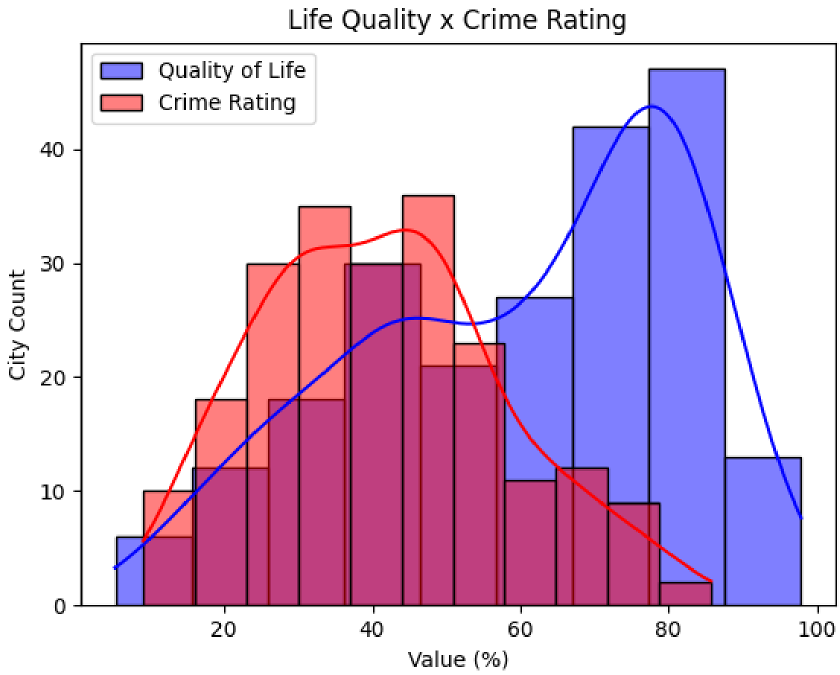
- Task 3 – Boxplot vs. Violine
    - Violine:

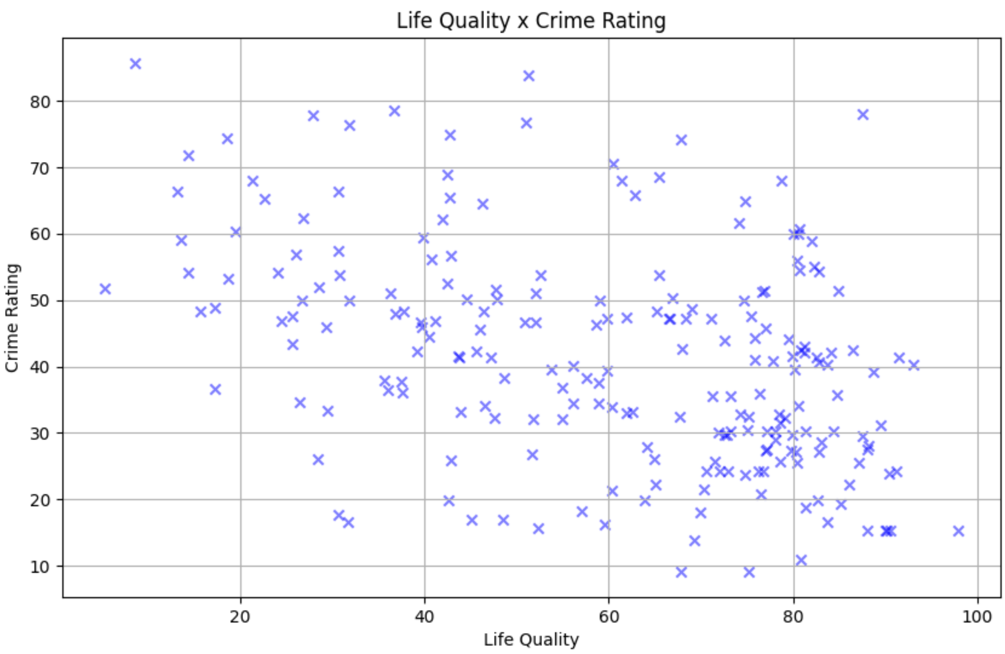


  - Boxplot: see aboth
  - Comparison:
    - Boxplot:
      + see data like distribution, including median, quartiles, and outliers
      - does not show the distribution shape and density
      Usage: straightforward comparison of distributions
    - Violine:
      + can also show the density of the distribution, as well as median and quartiles
      - can be more complex to be interpreted
      Usage: when understanding the distribution's shape and density is important

3. Cities

- Is there a between crime rating and quality of life?
  - Visually this connection isn't really applicable
  - However when calculation the Pearson Correlation Coefficient and Covariance a slight-medium negative correlation can be seen.
    - Pearson Correlation Coefficient Crime Rating x Quality of Life: -0.4271
    - Covariance Crime Rating x Quality of Life: -154.3747



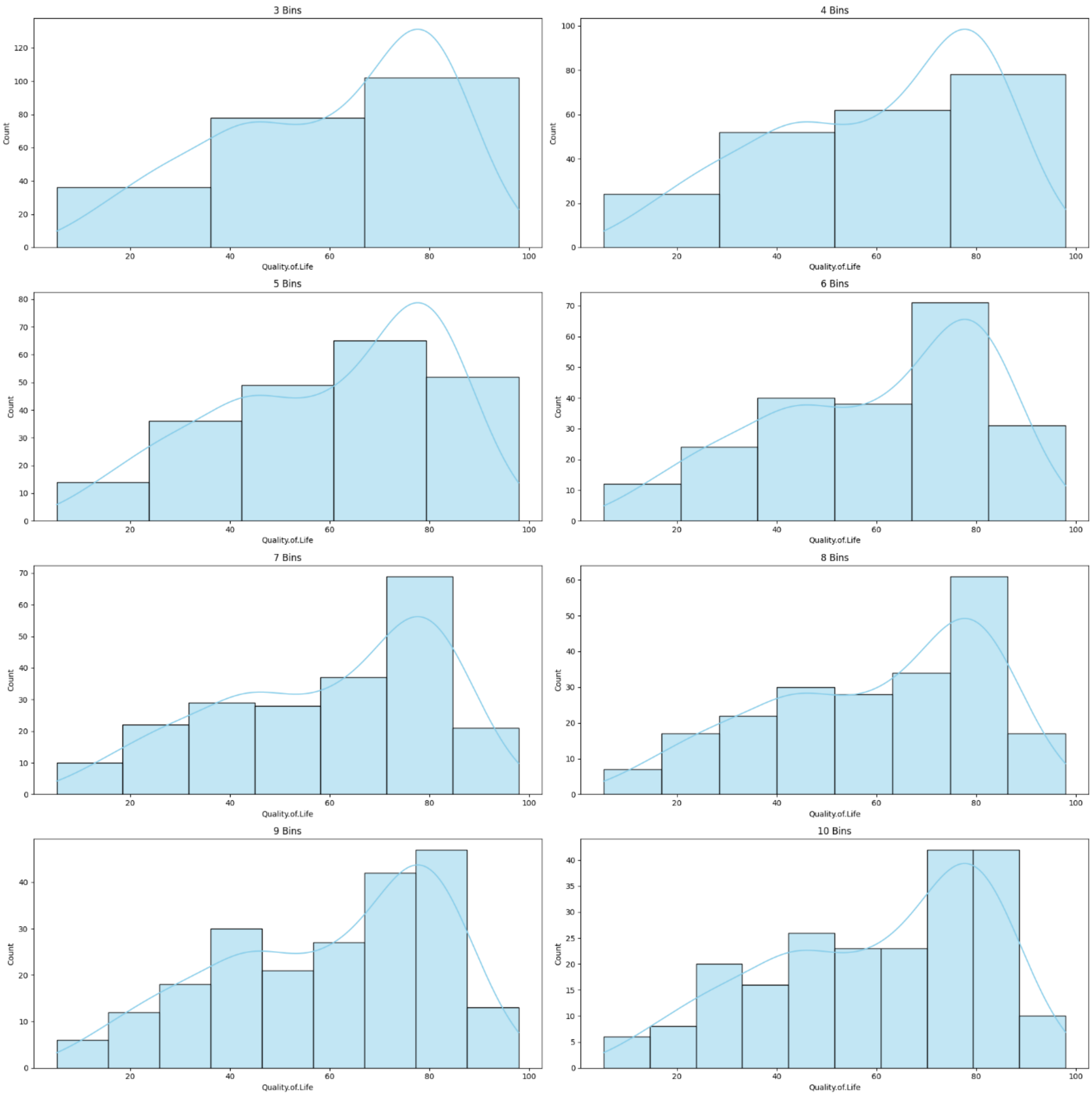


- Binning was performed next in order to combine life quality in quality groups:
  - 3-10 bins were formed (example for the ranges of groups with 3, 5 and bins below). 3, 5 and 10 Bins were also correlated with crime rate. There was no real trend appliceable, only in bin 2 (10 Bins) there was a medium positive correlation with crime rate. Other than that when sorted in 5 bins a connection between the crime rate and life quality is visiable.

```
Range of Life Quality in 5 Bins:

Bin 1: 5.29 to 22.67, Mean: 15.71
Bin 2: 24.05 to 41.88, Mean: 32.89
Bin 3: 42.40 to 60.50, Mean: 50.83
Bin 4: 61.44 to 79.08, Mean: 71.84
Bin 5: 79.58 to 97.91, Mean: 84.51
```

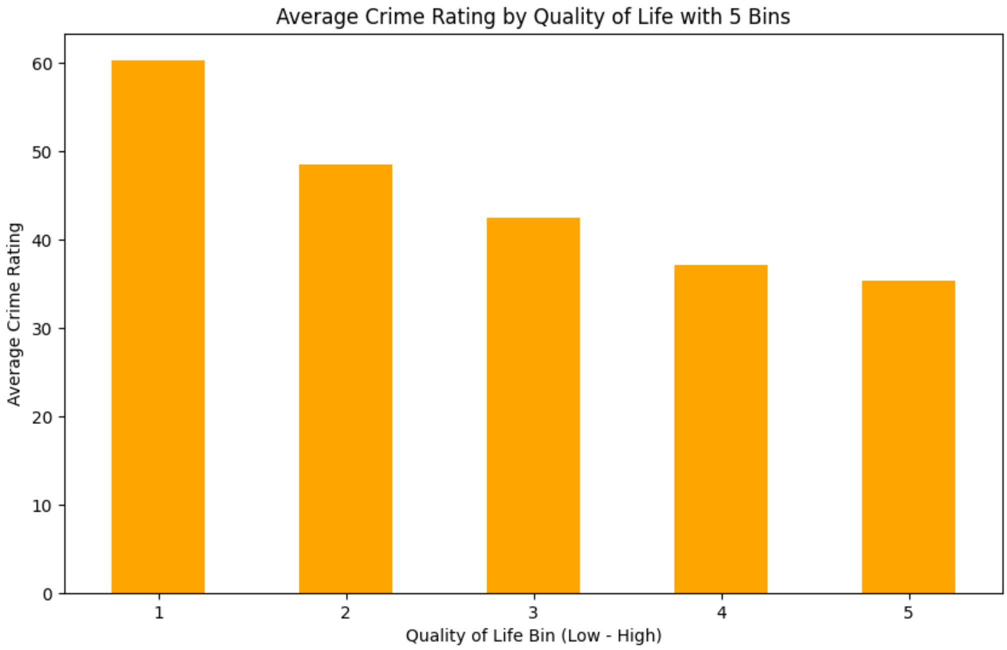
```
Range of Life Quality in 3 Bins:

Bin 1: 5.29 to 36.03, Mean: 23.86
Bin 2: 36.26 to 66.98, Mean: 51.29
Bin 3: 67.72 to 97.91, Mean: 79.40
```

```
Range of Life Quality in 10 Bins:

Bin 1: 5.29 to 14.36, Mean: 11.53
Bin 2: 15.66 to 22.67, Mean: 18.84
Bin 3: 24.05 to 31.87, Mean: 28.36
Bin 4: 35.69 to 41.88, Mean: 38.55
Bin 5: 42.40 to 51.26, Mean: 45.83
Bin 6: 51.66 to 60.50, Mean: 56.48
Bin 7: 61.44 to 69.91, Mean: 65.78
Bin 8: 70.34 to 79.08, Mean: 75.16
Bin 9: 79.58 to 88.27, Mean: 82.89
Bin 10: 88.76 to 97.91, Mean: 91.32
```

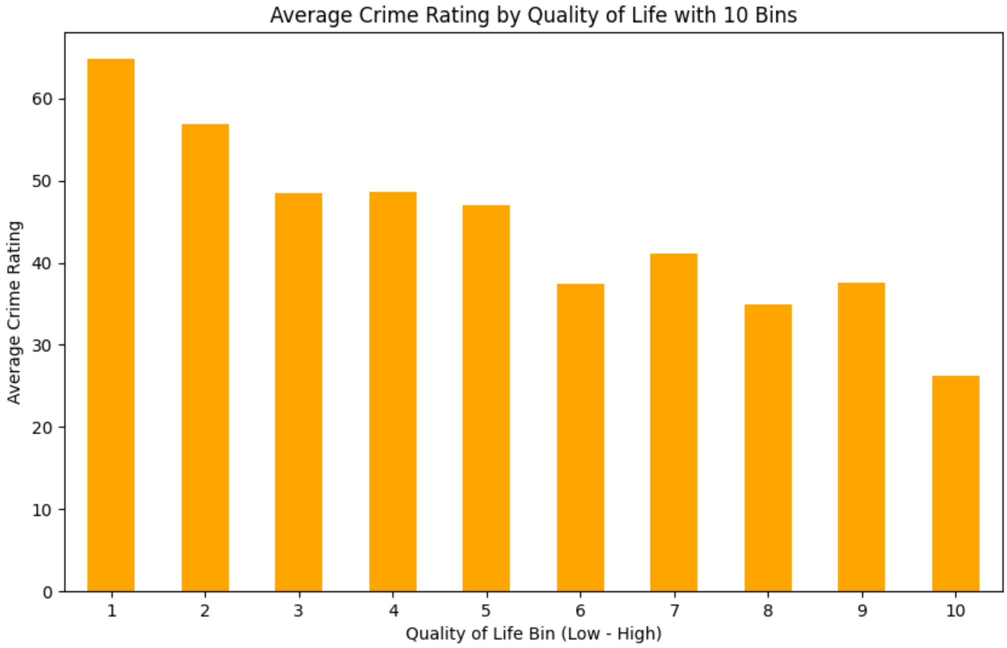


- Correlations binned life quality x crime rating

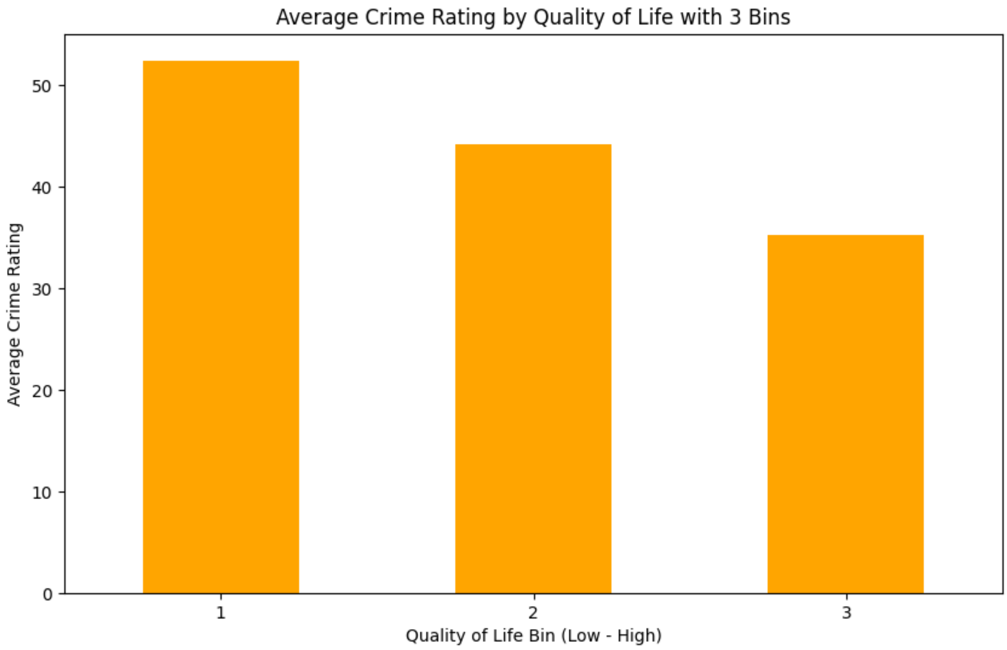
```
5 Bins:
Correlation in Bin 1: -0.08683810182850626
Correlation in Bin 3: -0.2023502524515913б
Correlation in Bin 2: 0.015219090007617501
Correlation in Bin 5: -0.35110751165570575
Correlation in Bin 4: -0.18242099715920645
```




```
10 Bins:
Correlation in Bin 1: 0.006722590645532163
Correlation in Bin 5: 0.1398014296б884357
Correlation in Bin 4: 0.23588613109920598
Correlation in Bin 9: -0.19188557291886862
Correlation in Bin 3: -0.06653386621205397
Correlation in Bin 2: 0.6555966325232102
Correlation in Bin 7: -0.22122214657267475
Correlation in Bin 6: 0.08207587411782162
Correlation in Bin 8: 0.1376419906378364
Correlation in Bin 10: -0.19001697642186663
```

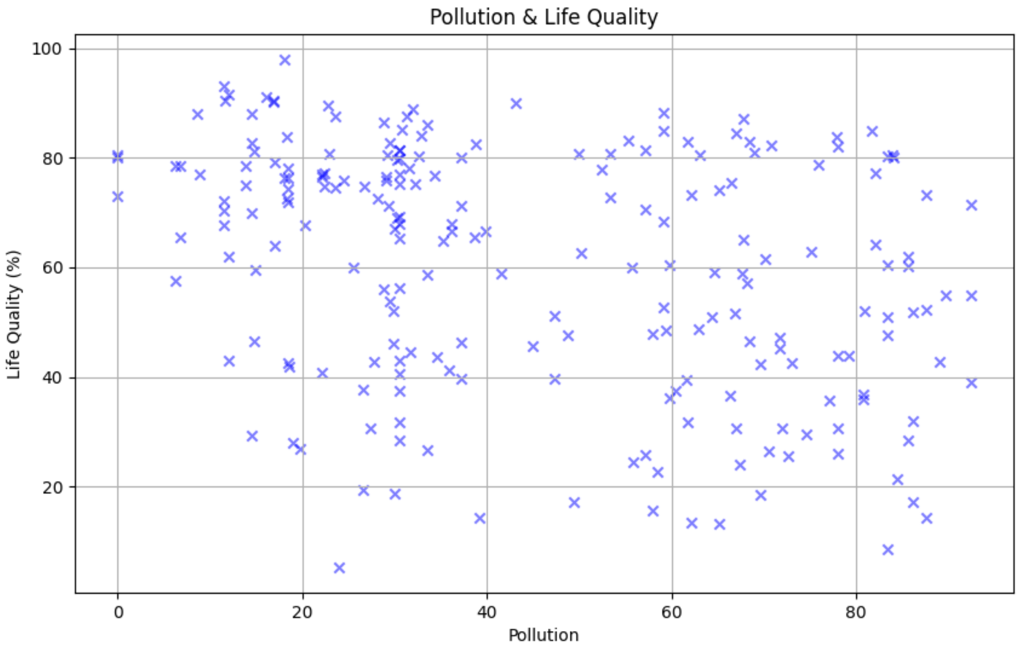


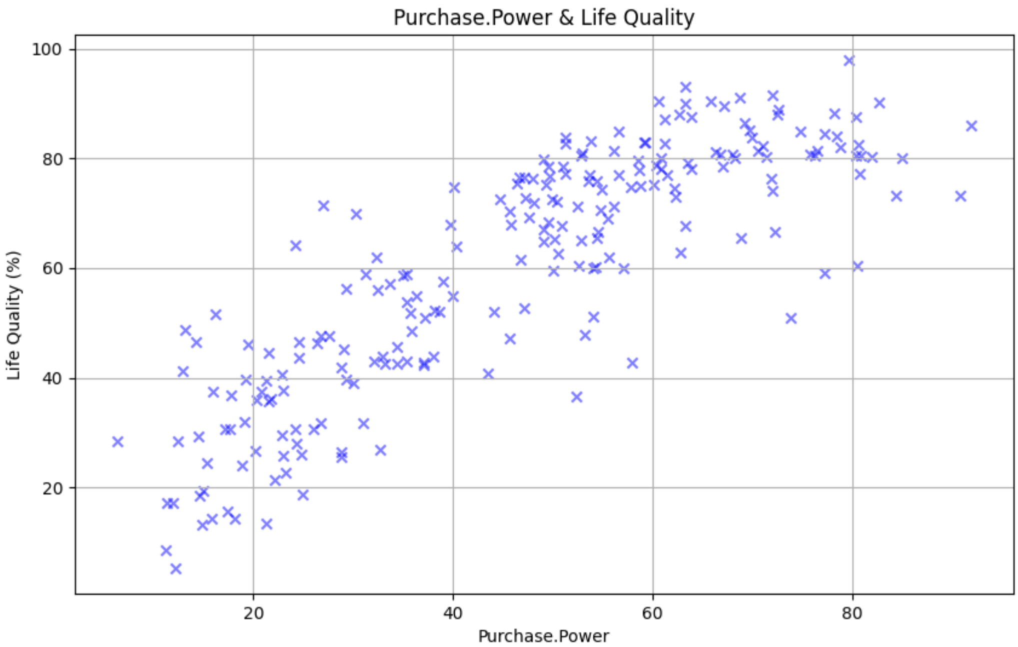

```
3 Bins:
Correlation in Bin 1: -0.4053436850746426
Correlation in Bin 2: -0.18284414235965382
Correlation in Bin 3: -0.10349148389726545
```



- Other analysis
    - Pollution – Life Quality
        - Pearson Correlation Coefficient Pollution x Quality of Life: -0.33496
        - Covariance Pollution x Quality of Life: -187.119051

- Slight negative correlation
- Purchase Power – Life Quality
  - Pearson Correlation Coefficient Purchase Power x Quality of Life: 0.84496
  - Covariance Purchase Power x Quality of Life: 383.541713
  - Strong positive correlation





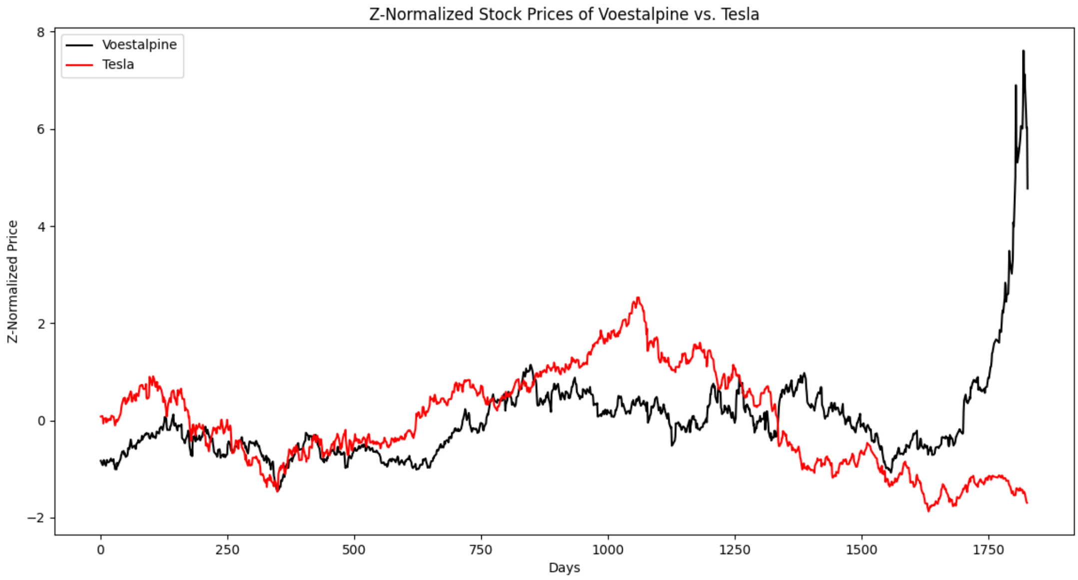
4. Stock Prices

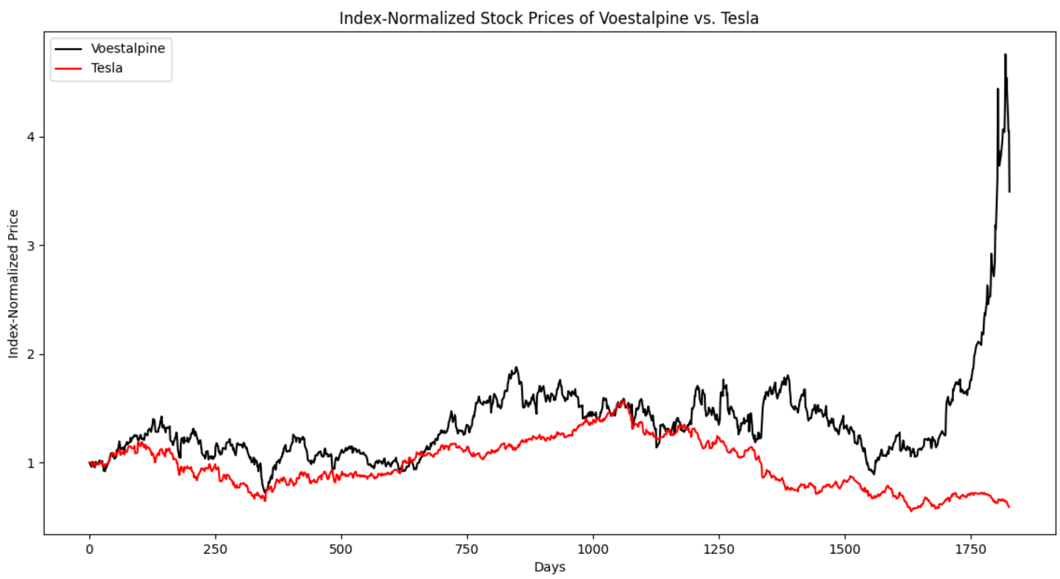
- Normalization:
    - Z- Normalization:
        - Z-score normalization is a technique where you scale the data based on the mean and standard deviation of the dataset.
        - The result is a standard score that indicates how many standard deviations an element is from the mean.
        - Z-score normalization is particularly useful when you want to compare scores from different samples or when the data needs to be normally distributed.
    - **Index- Normalization:**
        - Index normalization is a method where you rescale a dataset relative to a base value or a reference point within the dataset. This base value is typically the value at a particular time point or the initial value.
        - It's often used in time-series data to observe the relative change over time.
        - The result is a percentage that shows how much the value has increased or decreased relative to the base value.
        - So in this case Index Normalization is favored since the development over time is crucial!
    - Voest:

	Days	Close	Z-Normalized	Index-Normalized
0	1	182.850006	-0.830141	1.000000
1	4	176.449997	-0.908727	0.964999
2	5	175.500000	-0.920392	0.959803
3	6	179.199997	-0.874959	0.980038
4	7	184.100006	-0.814792	1.006836
...	...	...	...	...
1263	1821	830.799988	7.126070	4.543615
1264	1824	767.799988	6.352490	4.199070
1265	1825	740.000000	6.011132	4.047033
1266	1826	741.099976	6.024639	4.053049
1267	1827	638.900024	4.769721	3.494121

    - Tesla:

	Days	Close	Z-Normalized	Index-Normalized
0	1	34.930000	0.079680	1.000000
1	4	35.014999	0.090266	1.002433
2	5	34.759998	0.058507	0.995133
3	6	33.779999	-0.063547	0.967077
4	7	34.330002	0.004953	0.982823
...	...	...	...	...
1238	1820	22.469999	-1.472149	0.643287
1239	1821	22.420000	-1.478376	0.641855
1240	1824	21.200001	-1.630321	0.606928
1241	1825	20.809999	-1.678894	0.595763
1242	1826	20.680000	-1.695084	0.592041





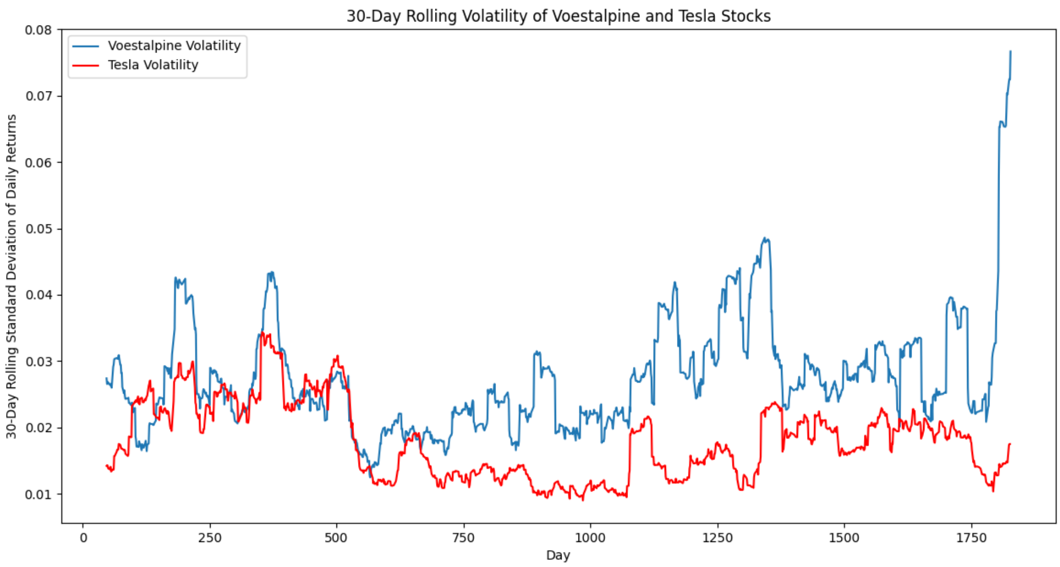
- Z-Normalized for correlations:
    - Pearson Correlation Coefficient for Voest: 0.4796294 (med. positive Correlation)
    - Pearson Correlation Coefficient for Tesla: -0.2532613 (slight negative correlation)
- Index-Normalized for correlations:
    - Pearson Correlation Coefficient for Voest: 0.47962938631789226
    - Pearson Correlation Coefficient for Tesla: -0.25326131091706927

➔ Correlation naturly remains the same because the connection to each other isn't different when rescaled

- Covariance for Voest: 20550.16523464813
- Covariance for Tesla: -1070.9793283084077

➔ Covariance doesn't really reveal much in this case other than both days and the price went into the same direction or not.

- Volatility is the degree of variation of a trading price series over time many changes and high peaks --> lots of changes in the standard veviation and therefore the stock is less "stable" as well as the fluctation higher monthly
    - Tesla shows a lower volitility and less peaks and therefore is generally "safer" to invest, than Voest which seems to have had more periods of high price fluctuations, especially recently



5. Data Sampling

Table 1: Datasets and their distribution of classes

Dataset	Class1	Class2	Class3	Class4
A	30%	34%	1%	35%
B	26%	23%	23%	28%
C	37%	19%	37%	7%

- Dataset A: Classes 1, 2 and 4 are fairly balanced, however Class 3 is not with only 1%. Here a **stratified sampling** technique would be suitable to ensure that Class 3 is represented appropriaetly too. The very low percentage of Class 3 can be problematic so it might not be represented adequatly in a random sample.
- Dataset B: All Classes show a more even distribution across the classes, although Class 3 and Class 4 are a bit less represented. A **simple random sampling or stratified sampling** could work here to ensure that each class is proportionally represented.
- Dataset C: Class 1 and Class 3 have a higher representation compared to Class 2 and Class 4. For this dataset, **stratified sampling** is recommended to ensure that the underrepresented classes (Class 2 and particularly Class 4) are not overlooked in the samples. A problem in this dataset ist the really high percentage in Class 3 and very low in Class 4 so a bias towards Class 3 could be a result.

The primary issue across these datasets when sampling is the significant imbalance in the distribution of classes, especially the extremely low percentage of some classes, which could lead to their underrepresentation in the sample and could hinder a proper analysis result.

Disclaimer: ChatGPT 4 was used to get an idea how to solve ceretain tasks.