**Data Mining Exercise 1**

Magdalena König

01455794

Derive source code here: <https://github.com/Magdalena-code/DataMiningSoSe2024/tree/main/UE1>

1. Marketing Data

* Ein Bild, das Text, Diagramm, parallel, Reihe enthält.

  Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Schrift, Zahl enthält.

  Automatisch generierte BeschreibungTask 1: Calculate the age and get the distribution within each marital status
* Task 2 - What is the distribution of the education?

Ein Bild, das Text, Schrift, Screenshot, Design enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Diagramm, Screenshot, Kreis enthält.

Automatisch generierte Beschreibung

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

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

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

Ein Bild, das Text, Screenshot, Schrift, Karte Menü enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

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

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

* Other analysis:
  + Countries x Store/Overall Purchases

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

* + Income x Purchases

Ein Bild, das Reihe, Text, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Reihe, Screenshot, Diagramm enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Reihe, Screenshot, Diagramm enthält.

Automatisch generierte Beschreibung

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

Ein Bild, das Text, Diagramm, parallel, Rechteck enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Diagramm, parallel, Rechteck enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Reihe, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung

* + 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.

Ein Bild, das Text, Screenshot, Diagramm, Rechteck enthält.

Automatisch generierte Beschreibung

* 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.

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

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

Ein Bild, das Text, Diagramm, parallel, Plan enthält.

Automatisch generierte Beschreibung

Ein Bild, das Diagramm, Text, Reihe enthält.

Automatisch generierte Beschreibung

* Task 3 – Boxplot vs. Violine
  + Violine:
  + Ein Bild, das Origami, Diagramm, Reihe enthält.

    Automatisch generierte BeschreibungBoxplot: 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

Ein Bild, das Diagramm, Screenshot, Reihe, Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Reihe, Diagramm, Zahl enthält.

Automatisch generierte Beschreibung

* 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.

Ein Bild, das Text, Schrift, Screenshot, Zahl enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Schrift, Design enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Schrift, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Diagramm, Text, Plan, Reihe enthält.

Automatisch generierte Beschreibung

* + Correlations binned life quality x crime rating

Ein Bild, das Text, Schrift, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Schrift, Rechteck enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Schrift, Screenshot enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Screenshot, Rechteck, Diagramm enthält.

Automatisch generierte Beschreibung

* 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

Ein Bild, das Diagramm, Reihe, Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Reihe, Diagramm, Screenshot enthält.

Automatisch generierte Beschreibung

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!
  + Tesla:

Ein Bild, das Text, Screenshot, Schrift, Karte Menü enthält.

Automatisch generierte Beschreibung

* + Voest:

Ein Bild, das Text, Screenshot, Schrift, Karte Menü enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Diagramm, Screenshot, Reihe enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.

Automatisch generierte Beschreibung

* Z-Normalized for correlations:
  + Pearson Correlation Coefficient for Tesla: 0.4796294 (med. positive Correlation)
  + Pearson Correlation Coefficient for Voest: -0.2532613 (slight negative correlation)
* Index-Normalized for correlations:
  + Pearson Correlation Coefficient for Tesla: 0.47962938631789226
  + Pearson Correlation Coefficient for Voest: -0.25326131091706927
* Correlation naturlly remains the same because the connection to each other isn’t different when rescaled
* Covariance for Tesla: 20550.16523464813
* Covariance for Voest: -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
  + Voest shows a lower volitility and less peaks and therefore is generally "safer" to invest, than Tesla which seems to have had more periods of high price fluctuations, especially recently

Ein Bild, das Text, Schrift, Screenshot, Reihe enthält.

Automatisch generierte Beschreibung

5. Data Sampling

Ein Bild, das Text, Schrift, Screenshot, Reihe enthält.

Automatisch generierte Beschreibung

* 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.