|  |  |  |  |
| --- | --- | --- | --- |
|  | 2023-04-02 |  | |
|  | |  | |
| Exercise 1)  *Intelligent dataanalys DV1597* | | | |
|  |  | |  |
|  |  | | Gubran Alshekh Ali |

Exercise 1)

Intelligent dataanalys DV1597

1. The WHO Mortality Database is the leading data source for comparative epidemiological studies of mortality by cause. The visualization portal gives the WHO Mortality database unprecedented impact, accessibility and relevance and provides export facilities for cause-of-death data from 1950 to date.



Question about the first dataset:

1. In which region did the largest number of people die?
2. Find out what was the reason for why the largest number of people die in this area(the answer from question a) and if there an influence of an external factor for that?
3. Dataset of songs of various artist in the world and for each song is present:

* Several statistics of the music version on spotify, including the number of streams. Number of views of the official music video of the song on youtube.



Question about the second dataset:

1. Who are the three singers whose songs are listened to the most on Spotify, and what are the numbers of their views on YouTube for the same song?
2. The most listened songs on YouTube, the most comments and likes on the songs?
3. The dataset consist of Top 100 most expensive football transfers of all time represent by [goal.com](https://www.goal.com/en/news/100-most-expensive-football-transfers-all-time/ikr3oojohla51fh9adq3qkwpu)



Question about the third dataset:

1. Who are the ten most expensive deals in the football market?
2. From the answer in question a can we foundout why there is one team how every time buy with allot of mony
3. Why does data cleaning play a vital role in data analysis? Motivate your answer by providing an example
4. Data cleaning is an essential step in data analysis because it ensures that the data used for analysis is accurate, complete, and consistent. Raw data can be noisy, incomplete, or contain errors that can lead to incorrect or biased analysis results if not properly cleaned.

For example, suppose a company wants to analyze customer purchase behavior to improve their marketing strategy. The dataset contains customer purchase records from different sources and formats, including transactional data, web logs, and social media data. However, the dataset also contains missing values, duplicates, and inconsistent data formats that could impact the accuracy of the analysis results.

By performing data cleaning, the company can remove the irrelevant or incomplete data, handle missing values, and standardize the data format, ensuring the data is consistent and accurate. This enables the company to obtain reliable insights on customer behavior, such as the most popular products, customer preferences, and the best channels to reach them.

1. During data analysis, how do you treat missing values in data? Motivate your answer by providing an example.
2. Missing values in a dataset can occur due to a variety of reasons, such as human error, data corruption, or technical issues during data collection. It's essential to handle missing values correctly during data analysis as they can skew analysis results and lead to inaccurate conclusions.
3. The exemple:
   1. Here we have an example list with a missing value 'date': ['2022-01-01', '2022-02-01', None, '2022-04-01', None]
4. Implement a function that extracts the set of hashtags from a data frame of tweets. Hashtags begin with the "#" character and contain any combination of upper and lowercase characters and digits. Assume the hashtag ends where there is a space or a punctuation mark, like a comma, semicolon, or period. For instance, consider the following example data:

This is an #example Tweet for the interesting #DataAnalysis course at #BTH in #2023. The #AIStudents are taking the course for the second year at the campus in #Karlskrona.

Reference:

First Dataset: https://www.kaggle.com/datasets/faroukkadri/world-health-organization-mortality-database

Seconde Dataset: https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Third Dataset: https://www.kaggle.com/datasets/thitiwat/top-100-most-expensive-football-transfers