Amirreza Akbari | 810899045
Reza Baghestani | 810899046
Hananeh Jamali | 810899053
Mohammad Yaser Azad | 810800003

Task 1

Question 1)

These methods are useful for quickly getting an overview of the data in a DataFrame and understanding its structure and contents. For example through info() we can identify the missing values to understand which columns are more reliable for our analysis, or describe() generates descriptive statistics for numerical columns in a DataFrame, such as count, mean, standard deviation, minimum, maximum, and various percentiles. It provides a quick overview of the distribution of values in each numerical column.

Question 2)

This cell identifies categorical columns (columns with data type 'object') and numerical columns (columns with data type other than 'object'), converts non-numerical columns to categorical data type, creates a mapping of each unique category in categorical columns to a numerical code and stores it in the label_mapping dictionary. Then it encodes the categorical columns into numerical codes using .cat.codes method and replaces -1 values with None, displays the mapping of category labels to numerical codes for each categorical column, displays the data types of each column after label encoding and finally shows the first few rows of the modified DataFrame.

Question 3)

This cell generates a heatmap visualization of the correlation matrix using the seaborn library. The annot=True attribute adds numerical values to each cell of the heatmap. The cmap='coolwarm' attribute sets the color map for the heatmap. The fmt=".2f" attribute formats the numerical values in the heatmap to two decimal places. Warmer colors indicating a higher positive correlation, cooler colors indicating a higher negative correlation, and values closer to 0 indicating weaker correlation.

Question 4)

This cell retrieves the indexes of the upper triangle of a correlation matrix, identifies columns with a correlation of 1 that are not on the main diagonal and prints the extra columns that are to be deleted based on the required condition. Then it removes the extra columns from the DataFrame and repeats the operations in question 3 for the modified DataFrame and saves it to a new CSV file named 'task1_modified.csv'.

Question 5)

We used .shape[0] method to count number of rows.

Question 6)

Same as question 5.

Question 7)

Printing list and number of survived passengers.

Amirreza Akbari | 810899045
Reza Baghestani | 810899046
Hananeh Jamali | 810899053
Mohammad Yaser Azad | 810800003

Question 8)

Strategy for handling missing values is imputation. We fill missing numerical values with the mean and missing categorical values with the mode. For age it worked well, but regarding the deck, since the number of missing values is three times the known values, by replacing the missing values with the mode, we are biasing the data with respect to this variable. In this regard, we can perform analyses only on passengers whose deck we know and disregard the rest.

Question 9)

Average age of all passengers: 29.70 years

Average age of male passengers: 30.51 years

Average age of female passengers: 28.22 years.

No significant points.

Question 10)

The correlation between ticket price and survival is 0.26, which can be considered not very significant. By examining and comparing the statistics and the first, second, and third quartiles between the two datasets (all passengers and survivors), we find that although they do not have a similar distribution, the median of the survivors is still higher than the third quartile of the overall dataset.

Question 11)

0.24 of third class, 0.47 of second class and 0.63 of first class have been survived, comparing to overall surviving rate which is 0.38, second and first class actually had a greater chance to survive.

Question 12)

Except for children under ten years old, in general, we expect that the survival rate would be lower than the non-survival across all age groups (as indicated by the average of 0.38). However, near the average age of the passengers on the ship (30 years), a very noticeable difference between the survivors and the non-survivors is observed.

Question 13)

As we know from previous analyses, as we move towards the upper side of the graph (more expensive tickets), it becomes noticeably more blue (higher ratio of survivors to non-survivors)

Question 14)

The table shows us that although the percentage of survivors is higher in the higher classes, there is a stark difference between men and women. 50% of women in the third class have survived, whereas this number is only 36% for men in the first class.

Amirreza Akbari | 810899045
Reza Baghestani | 810899046
Hananeh Jamali | 810899053
Mohammad Yaser Azad | 810800003

Question 15)

As we move to lower classes, the difference between the average ticket price of survivors and non-survivors decreases, although the average ticket price of survivors of $512 from the two passengers from first class has had an impact on average fare of first-class survivors.

Task 2

This report presents an analysis of data scientist salaries across different regions from 2020 to 2024. The analysis encompasses data pre-processing, including handling duplicates, missing values, standardizing currencies, and converting salaries to a single currency. Furthermore, it explores the distribution of job titles and identifies the top 10 most popular job titles and highest salaries. Visualization techniques are utilized to provide insights into salary distribution, job title frequency, and remote work ratios.

1. Data Pre-processing:

Identifying and Removing Duplicates:

Duplicate rows in the dataset were identified and removed to ensure data integrity and avoid skewing analysis results.

Identifying and Handling Missing Data:

Missing values were identified and handled by dropping rows containing any missing values, ensuring the dataset's completeness and reliability.

Standardizing Salaries to a Single Currency:

Salaries were standardized to a single currency to facilitate meaningful analysis by filtering out data associated with currencies represented fewer than ten times.

Converting Currencies to USD:

Exchange rates for each currency to USD were fetched, and salaries were converted to USD based on these rates. The 'Salary Currency' column was updated to reflect the conversion to USD.

2. Analysis:

Identifying the Top 10 Most Popular Job Titles:

The analysis identified the top 10 most popular job titles based on their frequency in the dataset.

Identifying the Top 10 Highest Salaries:

The top 10 highest salaries, along with their corresponding job titles, were identified by sorting the dataset by salary in descending order.

Amirreza Akbari | 810899045
Reza Baghestani | 810899046
Hananeh Jamali | 810899053
Mohammad Yaser Azad | 810800003

3. Visualization:

Visualizations were generated to provide insights into various aspects of the dataset:

Salary Distribution (Histogram): Shows the distribution of salaries across different ranges.

Job Title Distribution (Bar Plot): Illustrates the frequency of job titles, highlighting the top 10 most popular titles.

Remote Work Ratio Distribution (Violin Plot): Provides insights into the distribution of remote work ratios among data scientists.