|  |  |
| --- | --- |
| **Name** | **Mamoon Haseeb**  **M.Mursaleen** |
| **Reg.No** | **NUM-BSCS-2022-28**  **NUM-BSCS-2022-32** |

**Phase 2:**

**Data Import and Initial Exploration**

In Phase 2, we started with importing essential Python libraries such as pandas, numpy, matplotlib.pyplot and seaborn to facilitate data handling and visualization. We then loaded the dataset into a pandas DataFrame for analysis. To understand the structure and quality of the data, we examined the first few rows, checked data types and generated summary statistics. Additionally, we assessed the dataset for missing values to determine any need for preprocessing in later stages.

**Missing Values Visualization**

As part of our exploratory data analysis in Phase 2, we visualized missing data using a heatmap created with Seaborn. This technique helps us quickly identify patterns and locations of missing values across the dataset. In the plot, missing values are highlighted using a color gradient (viridis), while non-missing values are left blank. This visual overview aids in deciding whether imputation or removal of missing data is required in the preprocessing phase.

**Histograms of Hourly Rates and Budget**

In this section, we create histograms to visualize the distribution of numeric columns: hourly rates (both low and high) and budget. Each histogram shows how frequently different values occur within these columns. The x-axis represents the value ranges, while the y-axis indicates the number of occurrences. By using 30 bins, we can see the distribution more clearly. This visualization helps us understand the overall trends in hourly rates and budgets for jobs, giving insights into common pricing structures in the dataset.

**Boxplots of Hourly Rates and Budget**

In this section, we create boxplots for the numeric columns: hourly rates (low and high) and budget. Each boxplot displays the distribution of values, highlighting the median, quartiles, and potential outliers. The box represents the interquartile range (IQR), while the lines (whiskers) extend to show the rest of the data, except for outliers, which are plotted as individual points. This visualization helps us identify the central tendency and variability of the data, as well as any outliers that may need further investigation.

**Initial Observations**

1. **Missing Values Visualization**:
   * The heatmap reveals distinct patterns of missing data across various columns.
   * Columns such as "Is\_hourly," "budget," and "cetegory" exhibit significant missing values, which may require imputation or removal in the preprocessing phase.
   * Conversely, columns like "title" "country," and "link" appear to contain fewer or none missing values, indicating a more complete dataset for these features.
2. **Histograms of Hourly Rates and Budget**:
   * The histograms for "hourly\_low" and "hourly\_high" show a right-skewed distribution, suggesting that most jobs have lower hourly rates, with fewer jobs commanding higher rates.
   * The histogram for "budget" also exhibits a right-skewed distribution, indicating that most budget values are concentrated in the lower range, with few projects having significantly higher budgets.
   * The use of 30 bins allows for a clearer understanding of the distribution and frequency of values in these columns, revealing trends in pricing structures.
3. **Boxplots of Hourly Rates and Budget**:
   * The boxplots for "hourly\_low," "hourly\_high," and "budget" highlight the presence of outliers, particularly in the higher ranges of hourly rates and budgets.
   * The median values are marked within each box, showing the central tendency, while the interquartile range (IQR) provides insights into the variability of these features.
   * The presence of outliers suggests that while most jobs fall within a certain price range, there are exceptional cases that may need further investigation.

**Phase3:**

**Handling Missing Values**

In this section, we address the missing values in the dataset. For each column, we check if there are any missing values. If a column is categorical (with text data), we fill the missing entries with the most common value (mode). If a column is numeric, we fill the missing values with the median. This approach ensures that we retain as much information as possible while cleaning the data. After completing this process, we confirm that there are no remaining missing values in the dataset, enhancing its quality for analysis.

**Checking and Removing Duplicates**

In this section, we check for duplicate rows in the dataset. We count how many duplicates are present and display that number. If duplicates are found, we remove them to ensure each entry in the dataset is unique. After removing the duplicates, we confirm the new shape of the DataFrame, which indicates the number of rows and columns remaining. This step is crucial for maintaining the integrity of the data and ensuring accurate analysis.

**Identifying and Removing Outliers**

In this section, we first identify the numeric columns in the dataset. For each numeric column, we create a boxplot to visualize the distribution and identify potential outliers. Next, we use the Interquartile Range (IQR) method to remove these outliers. We calculate the first quartile (Q1) and third quartile (Q3) to determine the IQR, then establish lower and upper bounds. Any rows with values outside these bounds are filtered out from the dataset. Finally, we display the shape of the DataFrame before and after removing the outliers, ensuring our dataset is clean for analysis.

**Encoding Categorical Variables and Scaling Numeric Features**

In this section, we first identify the categorical and numeric columns in the dataset. Categorical columns are converted into numerical format using Label Encoding, which assigns a unique number to each category. This transformation allows the model to understand the categorical data better. Next, we scale the numeric columns using StandardScaler, which standardizes the values to have a mean of 0 and a standard deviation of 1. This step is important for ensuring that all features contribute equally to the model. Finally, we display a preview of the transformed dataset to confirm the changes.