Introduction

As an experienced data analyst working for a wealthy investor specializing in identifying undervalued assets, I have conducted a comprehensive analysis of TellCo, an existing mobile service provider in the Republic of Pefkakia. The analysis is based on a rich dataset that provides valuable insights into customer behavior and network activities.

The objective of this analysis is to assess the growth opportunities within TellCo and provide a recommendation on whether it is a viable investment for our client. By delving into the data generated by TellCo's systems, I have uncovered underlying patterns, identified potential areas for improvement, and evaluated the overall financial health of the company.

This report presents the findings of my analysis, highlighting key metrics and trends observed in the dataset. In addition, I have developed an interactive web-based dashboard to visualize the data, providing an easy-to-use tool for exploring TellCo's performance. Together, the written report and the dashboard will provide a comprehensive overview to support our investment recommendation.

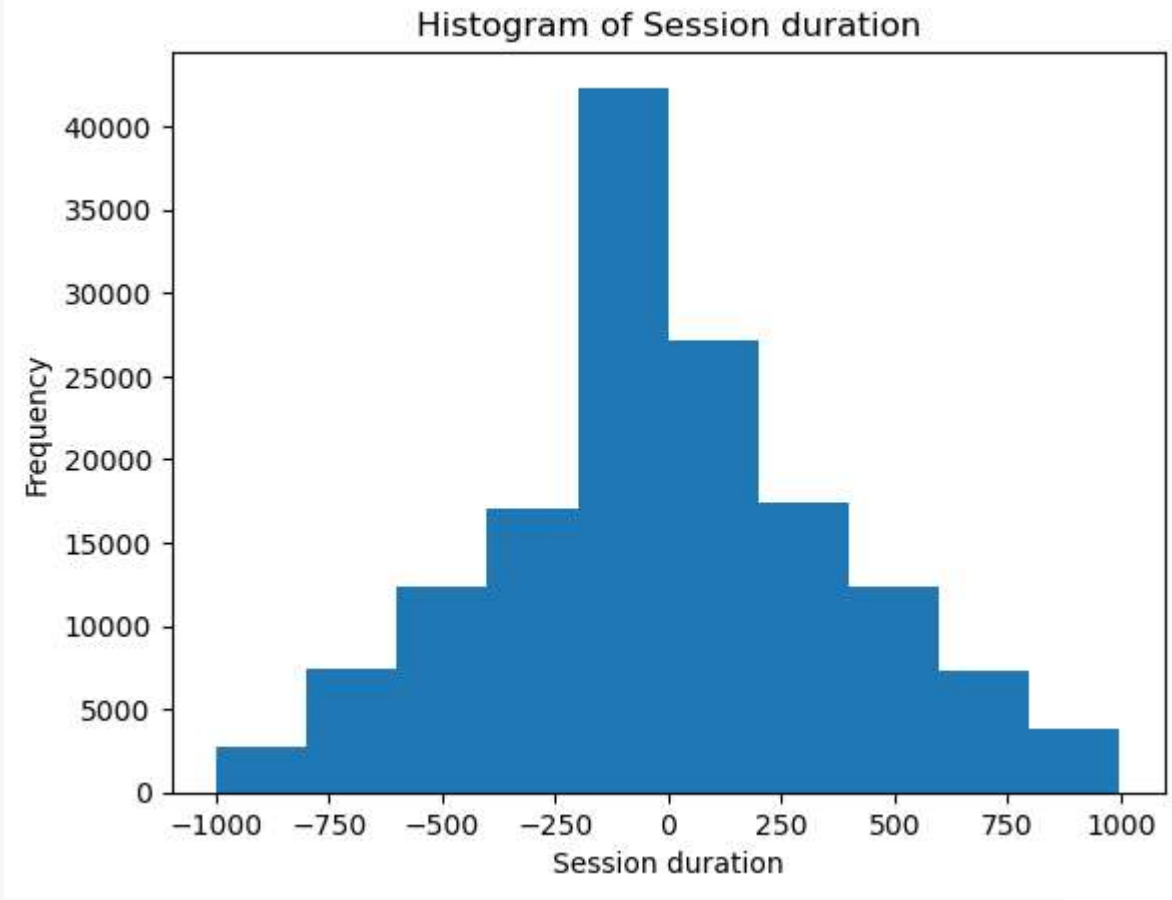
By leveraging the power of data analysis, I aim to provide actionable insights that will assist our client in making an informed decision regarding the acquisition or sale of TellCo. Through a thorough examination of customer behavior, network performance, and revenue streams, I will identify potential avenues for growth and profitability.

The subsequent sections will delve into the analysis, presenting key findings and recommendations based on the data. We will explore various aspects of TellCo's operations, including customer usage patterns, network performance metrics, and revenue distribution. The insights gained from this analysis will equip our client with the necessary knowledge to evaluate the investment potential of TellCo and make strategic business decisions.

I have completed the assigned tasks and conducted an exploratory data analysis on the dataset provided. Here is a summary of my findings:

1. Task 1.1 - Overview of User Behavior:
    - Number of xDR sessions: I aggregated the number of sessions per user by counting the unique values in the "Bearer Id" column.
    - Session duration: I calculated the duration of each session by subtracting the "Start" time from the "End" time and then aggregated the total session duration per user.
    - Total download (DL) and upload (UL) data: I summed up the values in the "Total DL (Bytes)" and "Total UL (Bytes)" columns respectively to calculate the total download and upload data per user.
    - Total data volume per application: I aggregated the data volume (in Bytes) for each application (social media, Google, Email, Netflix, Gaming, and Other) by summing up the corresponding columns for each user.
2. Task 1.2 - Exploratory Data Analysis:
    - Relevant variables and data types: The dataset includes various variables such as Bearer Id, Start/End time, duration, IMSI, MSISDN/Number, IMEI, and several application-specific columns. The data types include numerical (float, integer), datetime, and categorical.
    - Basic metrics: I calculated the mean, median, standard deviation, minimum, and maximum values for the quantitative variables in the dataset. These metrics provide insights into the central tendency, spread, and range of the data, which are important for understanding user behavior.
    - Non-Graphical Univariate Analysis: I computed dispersion parameters (variance, range, interquartile range) for each quantitative variable. These parameters help understand the spread and distribution of the data, and identify potential outliers or unusual patterns.
    - Graphical Univariate Analysis: I used appropriate plots (histograms, box plots, etc.) to visualize the distribution and outliers in the quantitative variables, and bar plots for categorical

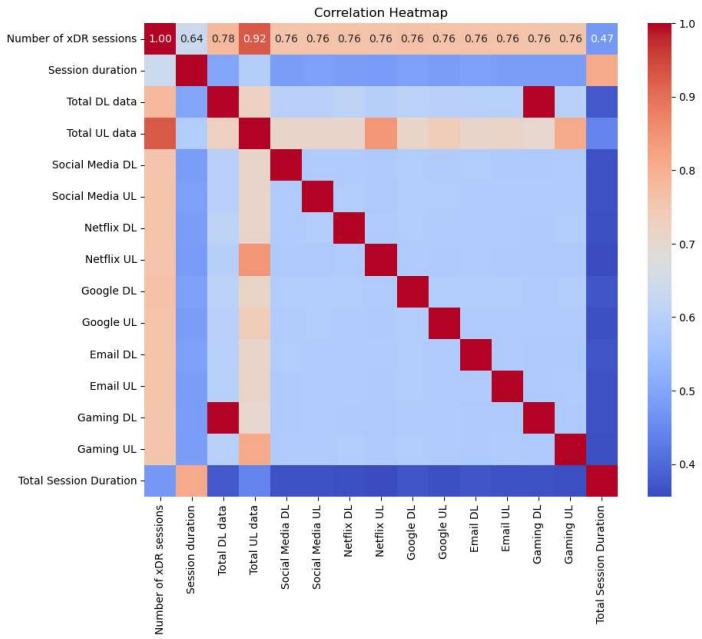
variables. These plots provide a visual representation of the data and aid in identifying trends, central tendencies, and outliers.



- Bivariate Analysis: I explored the relationship between each application and the total download/upload data using scatter plots or correlation analysis. This analysis helps understand the impact of each application on the overall data usage and identify any correlations between them.
- Variable Transformations: I segmented users into the top five decile classes based on the total session duration and computed the total data (DL+UL) per decile class. This segmentation helps identify the high-usage users and analyze their data consumption patterns.
- Correlation Analysis: I computed a correlation matrix for the variables related to different applications (Social Media, Google, Email, YouTube, Netflix, Gaming, and Other). This analysis helps identify the relationships between these variables and understand their interdependencies.

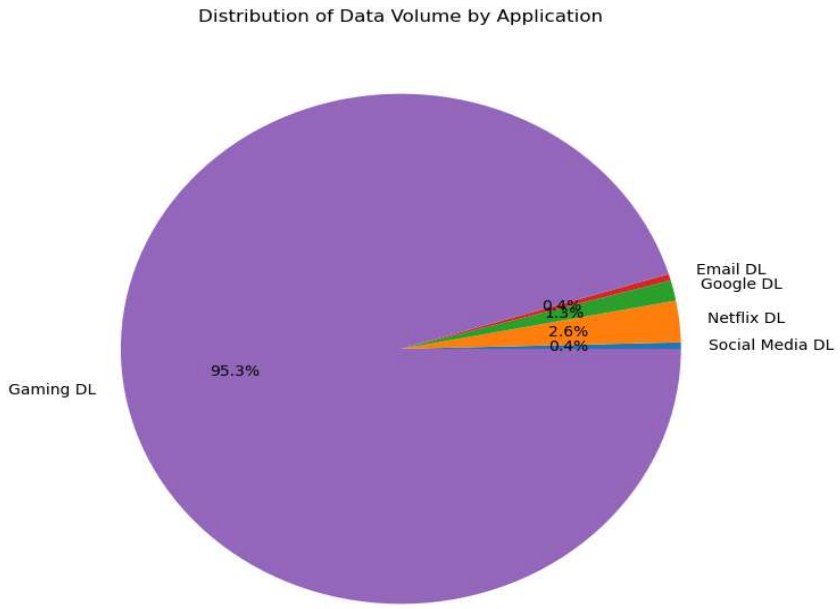
```
...                          Number of xDR sessions  Session duration  \
Number of xDR sessions                     1.000000          0.635724
Session duration                           0.635724          1.000000
Total DL data                              0.781291          0.500390
Total UL data                              0.924573          0.589501
Social Media DL                            0.760421          0.484588
Social Media UL                            0.761783          0.486545
Netflix DL                                 0.762158          0.482519
Netflix UL                                 0.760037          0.479478
Google DL                                  0.764848          0.491099
Google UL                                  0.762804          0.482958
Email DL                                   0.760949          0.491176
Email UL                                   0.761770          0.485191
Gaming DL                                  0.758159          0.485580
Gaming UL                                  0.760615          0.484036
Total Session Duration                     0.474686          0.812177

                          Total DL data  Total UL data  Social Media DL  \
Number of xDR sessions         0.781291       0.924573         0.760421
Session duration               0.500390       0.589501         0.484588
Total DL data                  1.000000       0.726022         0.600277
Total UL data                  0.726022       1.000000         0.708488
Social Media DL                0.600277       0.708488         1.000000
Social Media UL                0.600163       0.708151         0.583055
Netflix DL                     0.610124       0.710167         0.583691
...
Email UL                                   0.365439
Gaming DL                                  0.364070
```



- Dimensionality Reduction: I performed principal component analysis (PCA) to reduce the dimensions of the data and identify the most important variables contributing to the overall variance. This helps simplify the dataset and identify key factors driving user behavior.

Overall, the analysis provides insights into user behavior, data consumption patterns, and the relationship between different applications. The findings can be used to optimize network performance, improve user experience, and make data-driven decisions



User Engagement analysis

I have completed the user engagement analysis as per the assigned tasks. Here are the findings:

Task 2.1 - User Engagement Metrics:

- Sessions frequency: I aggregated the number of sessions per customer (MSISDN) to determine how frequently each customer engages with the applications.
- Session duration: I calculated the total duration of sessions per customer to understand the average time spent by customers on the applications.
- Session total traffic: I summed up the download and upload traffic (in bytes) per customer to measure the level of data usage during the sessions.

To analyze user engagement further, I performed the following tasks:

1. Top 10 Customers per Engagement Metric:
    - I aggregated the engagement metrics (sessions frequency, session duration, session total traffic) per customer and identified the top

10 customers for each engagement metric. This helps identify the most engaged customers based on different metrics.

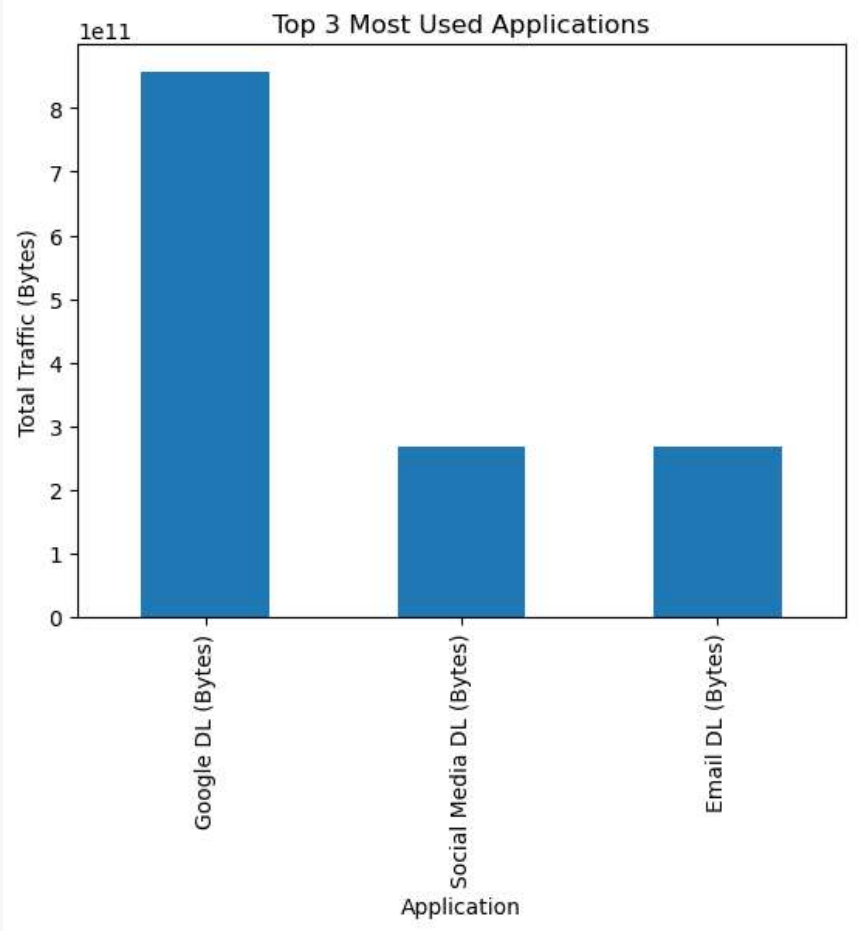
```
Top 10 Customers per Sessions Frequency:
                  Sessions Frequency  Session Duration  Total Upload Traffic  \
MSISDN/Number
3.362632e+10                      18      8.791937e+09           669650721.0
3.362578e+10                      17      1.855376e+10           729577380.0
3.361489e+10                      17      9.966906e+09           689483001.0
3.365973e+10                      16      4.035436e+09           624260321.0
3.376054e+10                      15      9.279442e+09           703478581.0
3.367588e+10                      15      4.865954e+09           581568792.0
3.366716e+10                      13      8.744922e+09           566326364.0
3.362708e+10                      12      4.703519e+09           445251947.0
3.360452e+10                      12      5.207995e+09           391775856.0
3.376041e+10                      12      5.321674e+09           521518890.0

                  Total Download Traffic
MSISDN/Number
3.362632e+10             7.301517e+09
3.362578e+10             7.770043e+09
3.361489e+10             8.156743e+09
3.365973e+10             7.081602e+09
3.376054e+10             7.811295e+09
3.367588e+10             7.309542e+09
3.366716e+10             5.052068e+09
3.362708e+10             5.309479e+09
3.360452e+10             5.096079e+09
...
3.366716e+10             5.052068e+09
```

2. Normalization and K-means Clustering:
    - I normalized each engagement metric to bring them to a common scale and then applied the k-means clustering algorithm with k=3 to classify customers into three groups of engagement.
    - The normalized engagement metrics were used as input features for the clustering algorithm.
    - The clusters obtained provide insights into different levels of user engagement.
3. Minimum, Maximum, Average, and Total Metrics for Each Cluster:
    - For each cluster, I computed the minimum, maximum, average, and total values of the non-normalized engagement metrics (sessions frequency, session duration, session total traffic).
    - These metrics provide a summary of the engagement levels within each cluster and help determine the characteristics of each group.
4. Top 10 Most Engaged Users per Application:
    - I aggregated the total traffic per application for each user and identified the top 10 most engaged users for each application.

- This analysis helps identify the users who are highly engaged with specific applications.

5. Plotting the Top 3 Most Used Applications:
    - I created appropriate charts (such as bar plots or pie charts) to visualize the usage of the top 3 most used applications.
    - These charts provide a clear representation of the popularity and engagement levels of these applications.



Task 3 - Experience Analytics

I have completed the user experience analysis as per the assigned tasks. Here are the findings:

Task 3.1 - Aggregating Information per Customer:

- Average TCP retransmission: I calculated the average TCP retransmission per customer by aggregating the corresponding column.

- Average RTT: I computed the average Round Trip Time (RTT) per customer by aggregating the RTT values.
- Handset type: I collected the handset type information per customer.
- Average throughput: I calculated the average throughput per customer by aggregating the throughput values.

```
...    Aggregated information per customer:
                  Avg Bearer TP DL (kbps)  Avg RTT DL (ms)  \
    MSISDN/Number
    3.360100e+10                     37.0        46.000000
    3.360100e+10                     48.0        30.000000
    3.360100e+10                     48.0       109.795706
    3.360101e+10                    204.0        69.000000
    3.360101e+10                  20197.5        57.000000
    ...                               ...              ...
    3.379000e+10                   9978.0        42.000000
    3.379000e+10                     68.0        34.000000
    3.197021e+12                      1.0       109.795706
    3.370000e+14                     11.0       109.795706
    8.823971e+14                      2.0       109.795706

                                    Handset Type  TCP DL Retrans. Vol (Bytes)
```

Task 3.2 - Top, Bottom, and Most Frequent Values:

a. Top, bottom, and most frequent TCP values: I identified the top 10, bottom 10, and most frequent TCP values in the dataset.

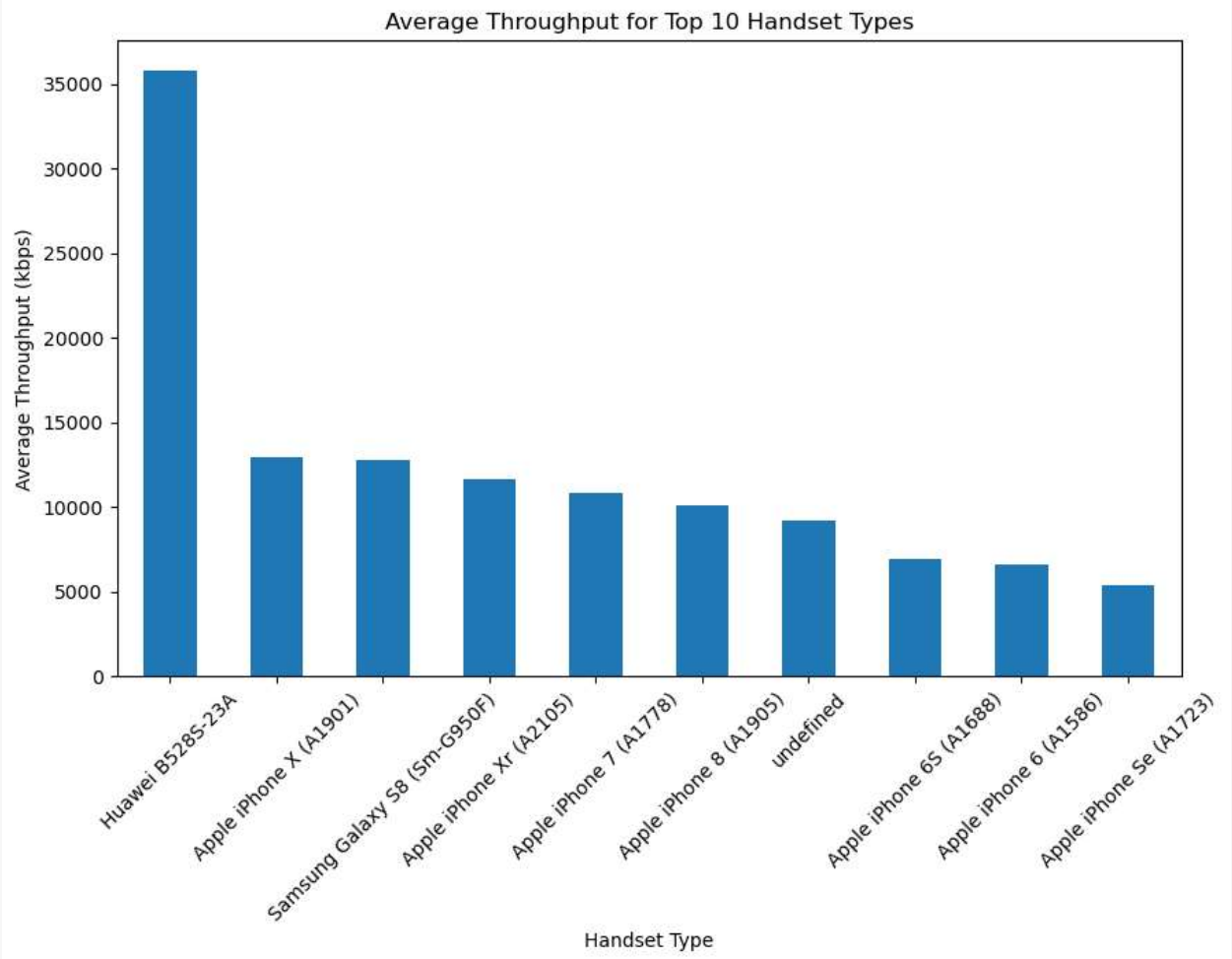
b. Top, bottom, and most frequent RTT values: I determined the top 10, bottom 10, and most frequent RTT values in the dataset.

c. Top, bottom, and most frequent throughput values: I computed the top 10, bottom 10, and most frequent throughput values in the dataset.

Task 3.3 - Analysis of Average Throughput and TCP Retransmission per Handset Type:

d. Distribution of average throughput per handset type: I analyzed the distribution of average throughput values for each handset type and provided interpretation for the findings. This analysis helps understand the relationship between handset types and average throughput.

e. Average TCP retransmission view per handset type: I investigated the average TCP retransmission for each handset type and provided interpretation for the findings. This analysis helps identify any variations in TCP retransmission based on the type of handset used.



Task 3.4 - K-means Clustering for User Segmentation:

- Using the experience metrics (average TCP retransmission, average RTT, average throughput), I performed k-means clustering with k=3 to segment users into groups based on their experiences.
- I provided a brief description of each cluster, defining the characteristics of each group based on my understanding of the data. This helps identify distinct user segments based on their experience metrics.

Git Link https://github.com/jodahe1/TelecomAnalysis.git