**DSA210 Project Report**

**Exploratory Data Analysis Part**

**Cell 1:**

Data from YouTube, Netflix and Spotify came from user data in these platforms. For YouTube, it was an Excel file, for Netflix, it was a CSV, and for Spotify, it was Excel file. First 5 row for each dataset illustrated after those files were turned in to pandas DataFrames. Columns in YouTube dataset contain Video Title, Creator and Timestamp. Columns in Netflix dataset that includes features that are Profile Name, Start Time, Duration and Device Type. There are some fields in the Spotify which are endTime, artistName, trackName, and msPlayed.

A screenshot of a computer

Description automatically generatedA computer screen shot of a list

Description automatically generated

**Cell 2:**

Text data from Spotify, YouTube, and Netflix was standardized in that part to make it more consistent and easy to use. As part of the standardization process, all capital letters were replaced to the lowerCell, non alpha characters removed, and extra spaces removed.

A white screen with text

Description automatically generated

**Cell 3:**

That part fulfilled empty numbers which exist in the YouTube, Netflix, and Spotify platforms’ datasets with placeholders that had already been set. Each dataset was given function MyFiller, that filled in any empty spaces with "unknown" string value for Netflix, "unknown track/artist" string value for Spotify, and "unknown" string value YouTube. Function was then used and the first 5 rows of each dataset were printed to ensure missing numbers were handled without error.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated

**Cell 4:**

That part sorts YouTube data into groups based on some certain keywords found in video title names. There is list that contains prepared wortds like Education, Gaming, Entertainment, Technology, News, Music, Sports, Kids, Cooking, Travel, and Lifestyle which helped me to categorize with title column. Number of movies in each area calculated and printed. Result illustrates how the videos are spread out among the different categories. The most common category is Entertainment, that followed Education and Gaming. That sorting helps us find what kinds of things are in collection list and can be used for more research and train model.

A screenshot of a computer

Description automatically generated

**Cell 5:**

In cell 5 with .info function, I got informations about each data like number of entites, data types etc.

A screenshot of a computer program

Description automatically generated

**Cell 6:**

In this cell I got values like count, mean, std etc. with using decribe method for each data.

A screenshot of a computer

Description automatically generated

**Visualization Part**

**Cell 7:**

In order to find Spotify's monthly total length, endTime value was changed to datetime, and msPlayed values changed from milli-seconds to seconds mode. Info was make together by month and total amount of time spent playing in each month was succefully founded. Answer was then changed to minutes so that it could be read easily by user. As lastly, plot was made to show how much time was spent on Spotify platform over course of year. That showed clear trends of usage over time. Line graph illustrates how much time people spent on Spotify platform each month in 2024. y-axis shows time in minutes. There are clear changes in usage shown by that plot in the below, with big peak in November and big drop around January. With considering data, it looks like more I used Spotify around university times.

A graph with blue lines and dots

Description automatically generated  
  
  
  
**Cell 8:**  
From July to November 2024, below graph illustrates how much time was spent on Netflix per each month. Time spent in seconds is shown on y value. In August, there is a clear high point and then there is significant drop in the next months. That change could mean that I couldn’t find good movies or series in these times.

A graph with a line going up

Description automatically generated

**Cell 9:**Pie chart, I got it by via calculating morning, afternoon, evenning and night from my youtube timestamp. After that I print pie chart with using matplotlib. Below pie chart showed that %82.9 I watch youtube at the night.

A pie chart with different colored circles

Description automatically generated

**Cell 10:**

Heatmap demostrates how I watch YouTube videos at different times of weekend’s day and on different days of the week. Each day's "Afternoon," "Evening," "Morning," and "Night" categories illustrated, along with number of videos played during each time. For example , Wednesdays have the most videos watched in afternoon, but Fridays and Saturdays have a lot of watchers in evening time.

A screenshot of a video game

Description automatically generated

**Cell 11:**

In here, For Netflix data, I create pie chart for device type that I open netflix each time.

Mostly I opened Netflix by PS5 to watch.

A red circle with numbers and a blue line

Description automatically generated

**Cell 12:**

Also, with using value\_counts I counted top 10 artists I listen in Spotify. I plot them using histogram via holding playing amount for each artist. Below results showed that maximumly I listen The Blaze Artist.

A graph with yellow squares

Description automatically generated

**Cell 13:**

With using wordcloud library I form a word cloud to show frequently listened artists.

A close up of words

Description automatically generated

**Cell 14:**

I plot histogram for categories I generated from youtube data, that plot illustrates how YouTube video categories are spread out based on amount of videos in each category.

A graph of blue bars with black text

Description automatically generated

**Machine Learning Part**

**Cell 15:**

RandomForestRegressor model that was used on the Spotify data has Mean Absolute Error (MAE) of 80.38 that means model's estimates are off by about 80.38 units from real values. That result illustrates that model worked well.

A screenshot of a computer program

Description automatically generated

**Cell 16:**

Time of day (hour), day of the week and month are used to find how long a Spotify song will play in milliseconds. Following are expected numbers for future playtime:

A screenshot of a computer

Description automatically generated

**Cell 17:**

In this part I use a HistGradientBoostingRegressor model to guess how long I will watch Netflix based on things like hour, day of the week, month, type of device and length of session. Model was trained with the default hyperparameters at first, that cause to RMSE of 166.21 and an MAE of 44.11.  
  
Later, I used GridSearchCV to find best values for hyperparameters like learning rate, maximum depth and number of repetitions. Once everything was tuned just right, learning rate was set to 0.05, maximum depth to 10, number of cycles to 100 and minimum sample size per leaf to 20. On the other hand, adjusted model's RMSE went up to 166.90 and its MAE went up to 44.77 illustrating that while tuning did improve some things, changes were not very much noticeable.  
  
That model shows how changing hyperparameters can help make predictions making much more accurate, but for more complicated datasets.

A screenshot of a computer

Description automatically generated

**Cell 18:**

I guess how I will use Netflix based on data from future. Future data in that case is list of inputs that contains hour, day of the week, month, type of device and amount of session for six different times. GridSearchCV-tuned HistGradientBoostingRegressor model is then used to guess how long those future data points' sessions will last.

Result is:

A screenshot of a computer

Description automatically generated