**APPENDIX B**

**(Netflix Movies Recommendations Engine)**

**RESULTS AND GRAPHS**

**Ratings Spread Pie Chart:**

On the complete dataset, we calculated what rating is most common in the total dataset. Here is the Pie-Chart which shows that most common Rating in the dataset is 4, which takes almost 33.6% of the total ratings:

A picture containing umbrella

Description automatically generated

**Ratings Per Weekday and Ratings per Month**

Due to an extremely large dataset, the study and the building of the model required working on different data files separately.

So, when joining the Movie Titles file with the Customer Files, each file was distinctly joined and 4 spark dataframes were produced, containing ratings for:

1. Movie Id from 1 to 4499
2. Movie Id from 4500 to 9210
3. Movie Id from 9211 to 13367
4. Movie Id from 13368 to 17770

Following are the Plots from **the Ratings by Day of Week** and **Ratings by Month of the Year** for the Datasets 2,3 and 4 (Results from Dataset 1 are shown in the report):

Dataset 2 Results:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Dataset 3 Results:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Dataset 4 Results:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

All the above Graphs are in the same direction, giving a clear idea that mostly movies are rated from Mondays to Thursdays during the week. And starting from June till November, the number of ratings is more than all the other moths of the year.

This study can help Netflix plan, what time of the year is best to upload new content to get maximum reviews and viewership. Also, months of February to May, is the time when the viewership is low. So, they should plan their content in a better way to increase their audience at that time.

**RECOMMENDATION MODEL RESULTS for each Dataset:**

Dataset 2: Confusion Matrix and Violin Plot

A screenshot of text

Description automatically generated

A close up of a logo

Description automatically generated

Like Dataset 1 model’s performance, this model also performed well in predicting ratings 3 and 4 and performed bad on the rest of the ratings.

The RMSE for this Dataset is 0.87 and the Accuracy is 32.28%.

Dataset 3: Confusion Matrix and Violin Plot

A screenshot of a cell phone

Description automatically generatedA close up of a logo

Description automatically generated

This model also behaved in the same fashion as the other 2 models.

The RMSE for this dataset is 0.89 and Accuracy is 32.30%.

Dataset 3: Confusion Matrix and Violin Plot

A screenshot of text

Description automatically generatedA close up of a logo

Description automatically generated

The RMSE for the Dataset 4 is 0.88 and the Accuracy is 31.30%.

From all the above plots this can easily be inferred that if we could build a model on the complete dataset, it would perform in this manner only.