Sameer Kumar Mohapatra
IS6030 – Data Management

EDA of Movies on OTT Media Services (Netflix, Prime Video, Hulu, and Disney+)

Data Background and Overview

The dataset contains movie titles present in over the top media platforms such as Netflix, Prime Video, Hulu and Disney+ till early 2020. It has movie titles and indicators whether it is present on any of the four above mentioned streaming services. The dataset further contains details of Genre, Director, Age group, run time, country, and language.

The data frame has 16744 rows and 16 columns

The Data Columns are:

```
"ID"    "Title"          "Year"           "Age"            "IMDb"
"Rotten Tomatoes"       "Netflix"        "Hulu"           "Prime Video"
"Disney"                "Type"           "Directors"      "Genres"
"Country"               "Language"       "Runtime"
```

Column Description:

Column	Data Type (R)	Description
ID	Int	Primary Key ID
Title	Factor	Name of movie
Year	Int	Year produced
Age	Factor	Target age group
IMDb	Num	IMDb rating
Rotten Tomatoes	Num	Rotten Tomatoes rating
Netflix	Int	Netflix platform binary indicator
Hulu	Int	Hulu platform binary indicator
Prime Video	Int	Prime Video platform binary indicator
Disney	Int	Disney+ platform binary indicator
Type	Int	Additional column only
Directors	Factor	Name of director(s) of the movie
Genres	Factor	Genre type of movie
Country	Factor	Country of the movie
Language	Factor	Language of the movie
Runtime	Num	Total movie time in minutes

Data Quality

The first column from the data with no header name is dropped before loading the data for analysis because that column was a duplicate. This is done using Excel

In the dataset, the column Country has country of origin as well as other countries where the movie was produced, similarly the column language has primary language of the movie as well as dubbed language details. To get the primary country and language two new columns: “PrimaryCountry” and “PrimaryLanguage” is created where the only the first written Country name and language is populated respectively. This is done using SQL.

The raw data also has many Null values. The count is mentioned below:

ID	Title	Year	Age	IMDb	RottenTomatoes
0	0	0	9390	571	11586
Netflix	Hulu	PrimeVideo	Disney	Type	Directors
0	0	0	0	0	726
Genres	Country	Language	Runtime	PrimaryCountry	PrimaryLanguage
275	435	599	592	385	599

The analysis will be done based on non-null values. For movie ratings, RottenTomatoes column has 11586 Null values so, it will not be preferred for analysis. Hence, only IMDb ratings will be used.

Summary Statistics:

Total Movie Count by Platform:

NetflixCount	HuluCount	PrimeCount	DisneyCount
3560	903	12354	564

Top 10 Primary Countries by movie count:

	PrimaryCountry	MovieCount
1	United States	9507
2	United Kingdom	1328
3	India	1093
4	Canada	803
5	France	358
6	Italy	311
7	Australia	254
8	Germany	215
9	Hong Kong	202
10	Japan	195

Top 10 Primary Language by movie count:

	PrimaryLanguage	MovieCount
1	English	12528
2	Hindi	661
3	Spanish	380
4	French	291
5	Italian	227
6	Mandarin	224
7	Japanese	188
8	German	144
9	Korean	128
10	Tamil	122

Movie Count by Age (Null values are omitted):

	Age	MovieCount
1	18+	3474
2	7+	1462
3	13+	1255
4	all	843
5	16+	320

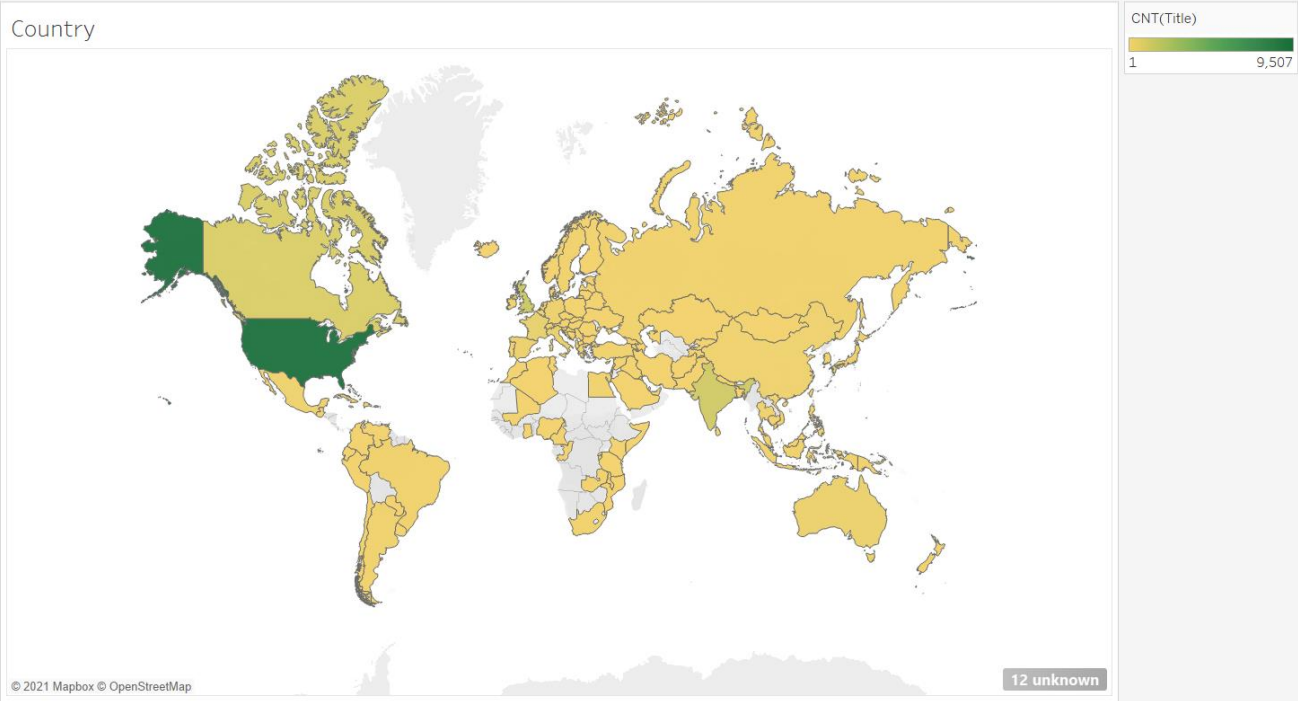
Movie generation (20th Century vs 21st Century) based on platform:

century	platform	quantity
20th century	Disney	205
21st century	Disney	327
20th century	Hulu	90
21st century	Hulu	788
20th century	Netflix	229
21st century	Netflix	3331
20th century	PrimeVideo	3516
21st century	PrimeVideo	8258

From the above, summary statistics we can clearly see that Prime Video has highest number of movies followed by Netflix. The count of movies based on different categorical variables are shown above. Further visualizations can be done on these to understand the variations in movies and platforms.

Visualizations:

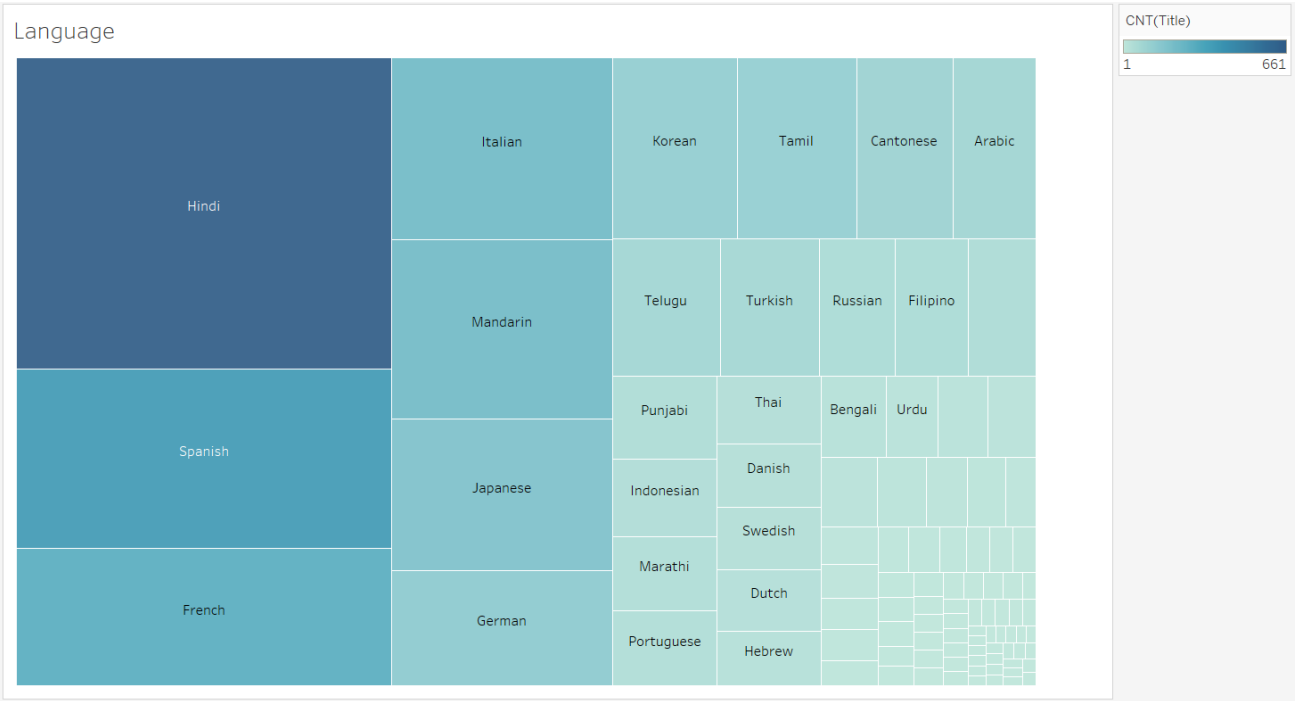
- Movies by primary country of production:



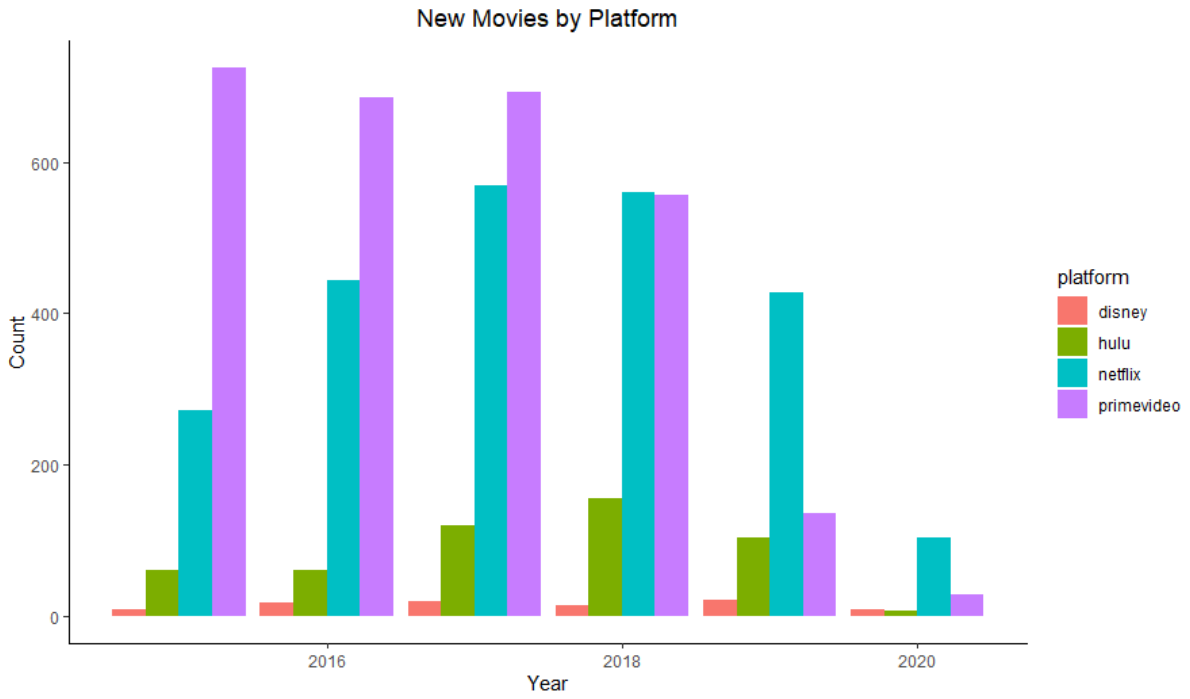
From the above map it can be clearly seen that the United States is having very high number of movie production

From summary statistics it was clear that English is the most common primary language of movies. We can further visualize the next commonly occurring languages after English.

- Movies by primary language after English:

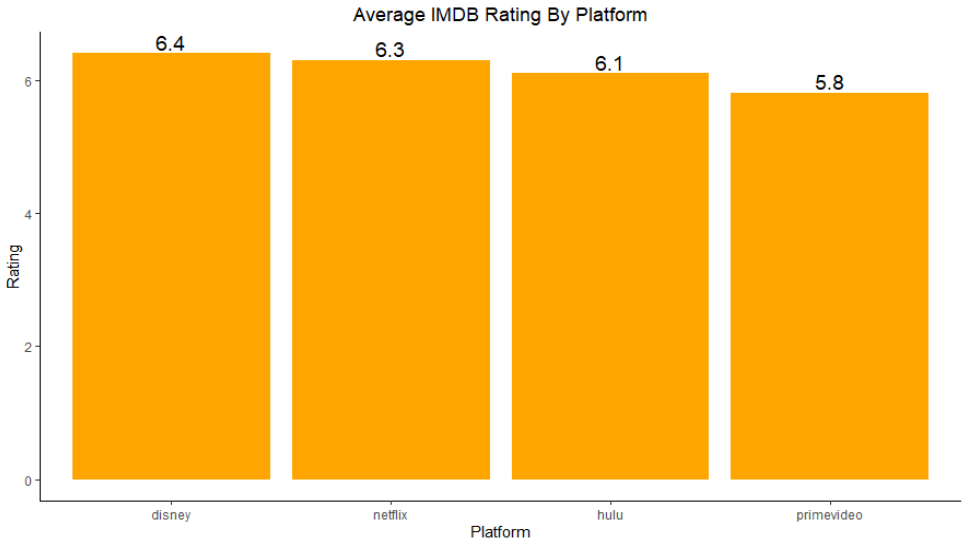


- New movies by platform (2015+):



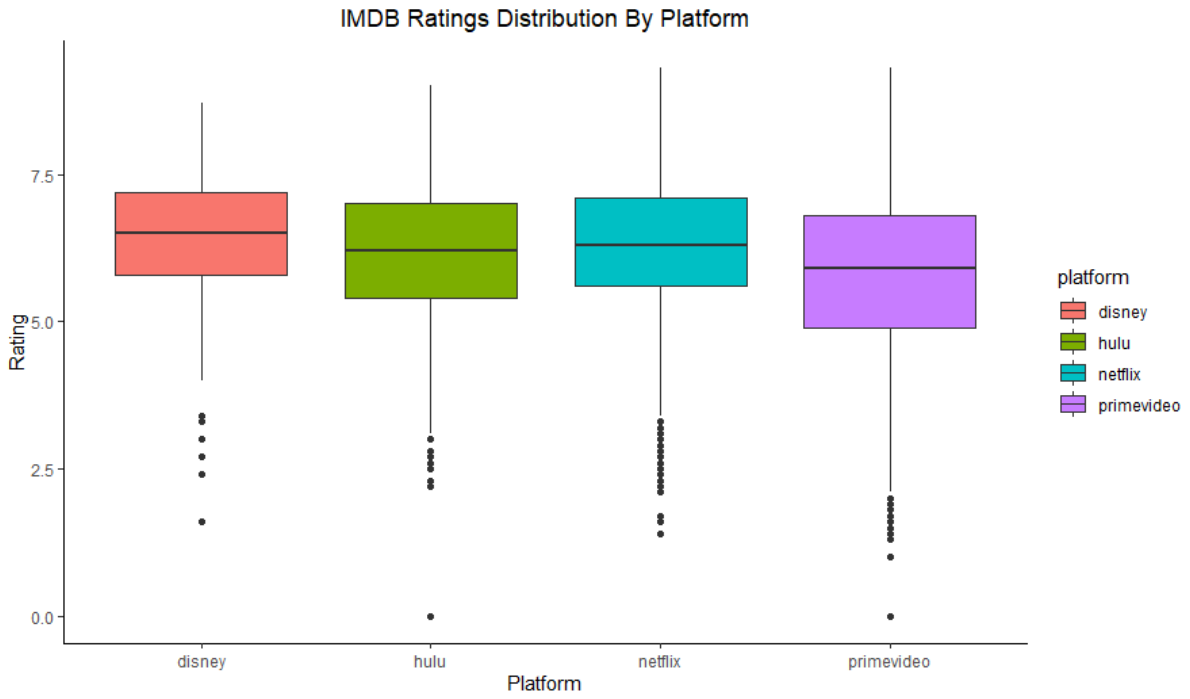
From the chart above, Netflix is having more content than any other platform since 2018. As per overall level Prime Video is having more content but if we need more latest content then Netflix is ahead. Also, the data has only has movies of early 2020 hence, actual 2020 movie count is not shown above.

- Average IMDb ratings across the platforms:



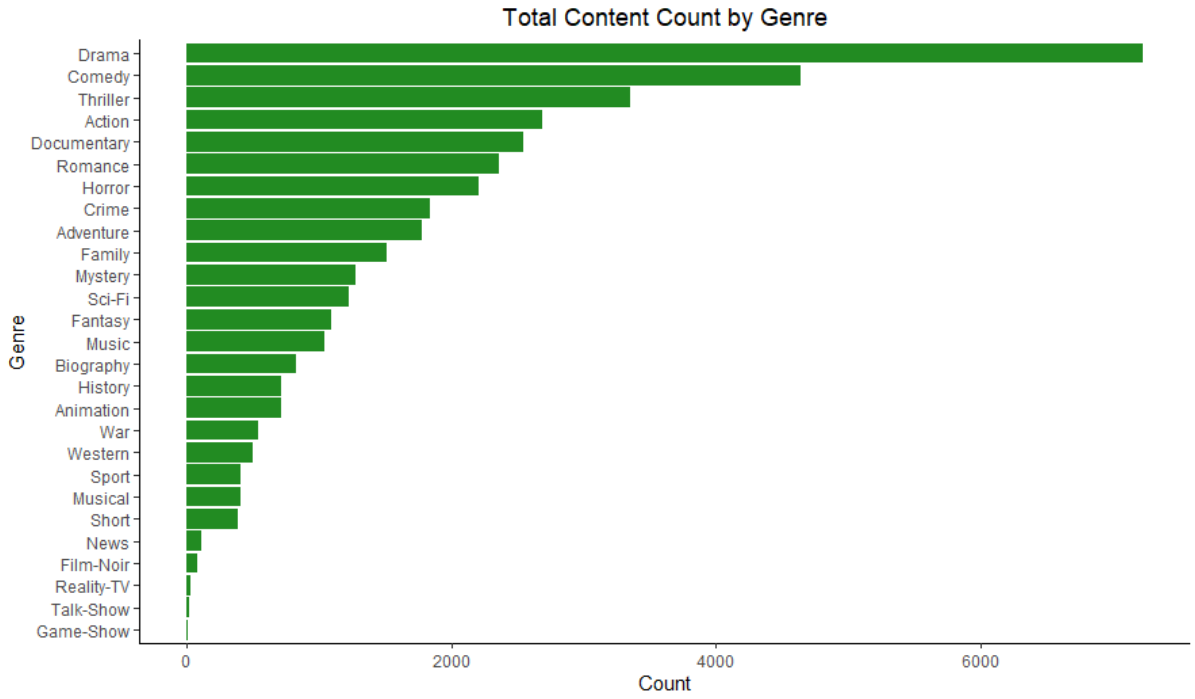
Except Prime Video average IMDB rating is very similar for other platforms. Disney+ has highest average rating

- IMDb ratings box plot for each platform:



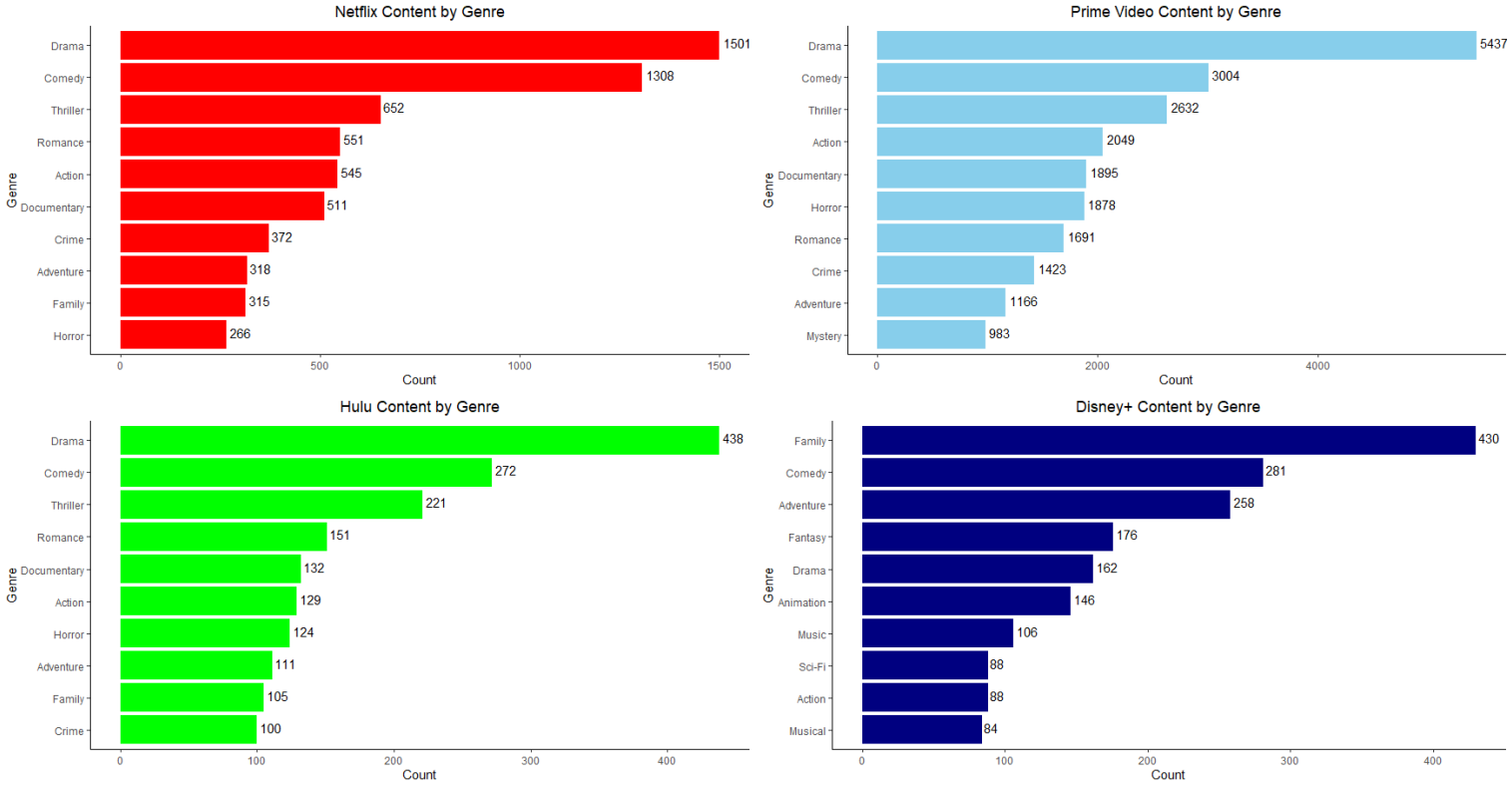
The IMDb ratings distribution is very similar for Disney+, Hulu and Netflix however Prime Video has high Inter Quartile Range and median is lower than other platforms. Disney+ has highest median.

- Content count by Genre:



Drama is most common genre type followed by Comedy and Thriller. One thing to note is that the movies are characterized by more than one genre and Drama is the genre most commonly present in the movie genre description.

- Movie Genre based on platforms:



At platform level Drama, Comedy and Thriller is top three genre for Netflix, Prime Video and Hulu but for Disney+ top three genres are Family, Comedy and Adventure.

Conclusion:

The Exploratory Data Analysis of the movies in OTT platform data provided us with the insights on Movie types and its distribution in different platforms based on various parameters. We could conclude that if we want to have access to large number of movies content then Prime Video is best. If we want more content which is latest, then Netflix is the best option. In terms of genre variety Netflix, Prime Video and Hulu are similar but Disney+ is more oriented towards family movies and is considered good for kids. In general, Movies are majorly produced in United States and are of English language.

Challenges Faced:

Faced two challenges, first was while loading the data set. The SQL server was not taking all data from .xls version rather was truncating the data and was showing some error and for .xlsx version I was getting error while loading due to 64-bit file issue. To resolve this, I tried using “SQL Server 2019 Import and Export data (64-bit)” from START menu and proceeded to load.

Second challenge was while preparing plots in R, as it needed a lot of data manipulation. At the end with some practice and trials I was able to understand.