Data Analyst Portfolio

Prepared by:

Arunmaran Elango

Professional Background

- A motivated professional seeking a Data Analyst position, aiming to leverage my skills in SQL, Excel, Power BI, Statistics, and Python.
- With a background as an Application Development Analyst in Accenture for 2 years 10 months and a Bachelor's degree in Computer Science and Engineering, I am well-prepared to contribute effectively.
- I have achieved HackerRank certifications in SQL at Basic, Intermediate, and Advanced levels.
- Conducted analyses using Excel, SQL, and Power BI to generate insightful charts addressing business challenges.

Table of Contents:

- Professional Background
- Table of Contents
- Project 1 - Data Analyst Process
- Project 2 - Instagram User Analytics Project
- Project 3 - Operations Analytics and Investigating Metric Spike Project
- Project 4 - Hiring Process Analytics Project
- Project 5 - IMDB Movie Analysis Project
- Project 6 - Bank Loan Case Study Project
- Project 7 - Impact of Car Analysis Project
- Project 8 - ABC Call Volume Company Project
- Learnings From the Projects

Data Analyst Process

- **Description** : Need to buy a Motorcycle / 2-wheeler for my daily use.

- **Plan** : Include the categories like :

1. Mileage Avg. : Greater than 30.
2. Engine Performance : Should be good.
3. Service costs : Should be nominal.
4. Cost of the bike : Approx. 2 to 3 Lakhs(INR).
5. Usage : Multiple(Touring, Street, Comfy).

- **Prepare** :

1. I have my savings from my work which will contribute to 75% of the bike cost.
2. My parents also pitched in some money which will contribute the rest of it.

- **Process :**

  There are many options for buying a bike(For Example : Touring, Sports, Street rider, etc…)

- **Analyse :**

1. Want to buy a bike which gives nice mileage in overall use, i.e. while riding in the cities as well as riding in the highways.
2. Want the bike to be in trend and is the best in its category.
3. Want to buy the bike which falls in my Budget.
4. Make test drives with all the bikes which falls in all the category mentioned above and make the choice for the bike which suits me.

- **Act :**

  Purchased the Honda Highness 350 Pro, aligning with all the categories I listed in the

  prior slide.

Instagram User Analytics Project

- **Description :**
- User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.
- These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.
- You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.
- **Approach :**
- I went through the datasets provided by the Trainity Instagram analysis project and understood that there were 7 tables in the database. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the MySQL Workbench, I did solve the SQL queries and provided the result as expected.

- **Tech-Stack Used :**

  MySQL Workbench – To answer the queries with the help of SQL language in the tool.

- **Insights :**

1. With the help of the SQL queries, I found out many insights which include –
2. Found the oldest 5 users of Instagram (People who knew more about Instagram).
3. Found people who haven’t posted any photo (People who just installed the app just for updating themselves.)
4. Found the user with more likes (People who use Instagram as per the Algorithm).
5. Found commonly used Hashtags (What people like to do in their free time / Hobby).
6. Found the day on which the users registered more (Can use this information to schedule the ad campaign on this day hence it reached more users).
7. Found out how people post their photos on an average basis (Can find out when they are free and how often they use Instagram).
8. Found the total number of users and the total number of photos posted in Instagram.

Result :

- Through this project I was able to understand my SQL skills and how to use SQL to find the answer for the queries being addressed by the business stake holders.
- Have made some analysis for the Instagram on the provided situations.
- Have made me think how the business might think from their point of view and how they can find this Analysis helpful in growing their product.
- I have achieved the end result and I think I have contributed my full support into the Analysis.
- I hope this project helps the Analysis and it achieves what it was tend to achieve

Operations Analytics and Investigating Metric Spike Project

- **Description** :
- Operational Analytics is a crucial process that involves analyzing a company's end-to-end operations. This analysis helps identify areas for improvement within the company. One of the key aspects of Operational Analytics is investigating metric spikes.
- This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales. Your task will be to derive insights from this data to answer questions posed by different departments within the company.
- Your goal is to use your advanced SQL skills to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics
- **Approach** :
- I went through the datasets provided by the Trainity Job Data and Investigation Metric Spike and understood that there were datasets each having tables and tasks for its own analysis.
- Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the MySQL Workbench, I did solve the SQL queries and provided the result as expected.

- **Tech-Stack Used :**
- MySQL Workbench – To answer the queries with the help of SQL language in the tool.
- Microsoft Word – To prepare the reports for presenting to the leadership team.

- **Insights :**
- Case Study 1 (Job Data Analysis):

1. Found the number of jobs reviewed per hour for each day in November 2020.
2. Found the 7-day rolling average of throughput (number of events per second).
3. Calculated the percentage share of each language over the last 30 days.
4. Checked if there are any duplicate rows in the table.

- Case Study 2 (Investigating Metric Spike):

1. Found the weekly user with the event type as engagement.
2. Calculated the user growth of the product by using the sub-queries.
3. Calculated the weekly retention of users based on their sign-up cohort by joining the user and event table.
4. Found the weekly engagement per device from the events table.
5. Calculated the email engagement metrics in the email events table.

Result :

- Through this project I was able to understand my SQL skills and how to use Advanced SQL functions like Sub-Queries, Windows function to find the answer for the queries being addressed by the business stake holders.
- Have made some analysis for the Job Data and the Metric Spike. Have made me think how the business might think from their point of view and how they can find this Analysis helpful in growing insights from the raw dataset.
- I have achieved the end result and I think I have contributed my full support into the Analysis. I hope this project helps the Analysis and it achieves what it was tend to achieve

Hiring Process Analytics Project

- **Description** :
  - The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department. These insights are then used by the hiring team to know how many candidates are hired/rejected from the process and how many are working in various departments across the company.

- **Approach** :
  - The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department. These insights are then used by the hiring team to know how many candidates are hired/rejected from the process and how many are working in various departments across the company.

- **Tech-Stack Used :**
  - Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.
- **Insights :**
  - Task A - Hiring Analysis:

With the help of the Excel formula “countifs”, found the count of hired people for Male as well as Female.

And the total number of people hired was calculated by the “sum” formula.

```
=COUNTIFS('Raw Data'!C2:C7169,"Hired",'Raw Data'!D2:D7169,"Male")
=COUNTIFS('Raw Data'!C4:C7171,"Hired",'Raw Data'!D4:D7171,"Female")
=SUM('Tasks 1,2,3'!C4,'Tasks 1,2,3'!C6)
```

Male Hired - 2562

Female Hired - 1855

Total Hired - 4417

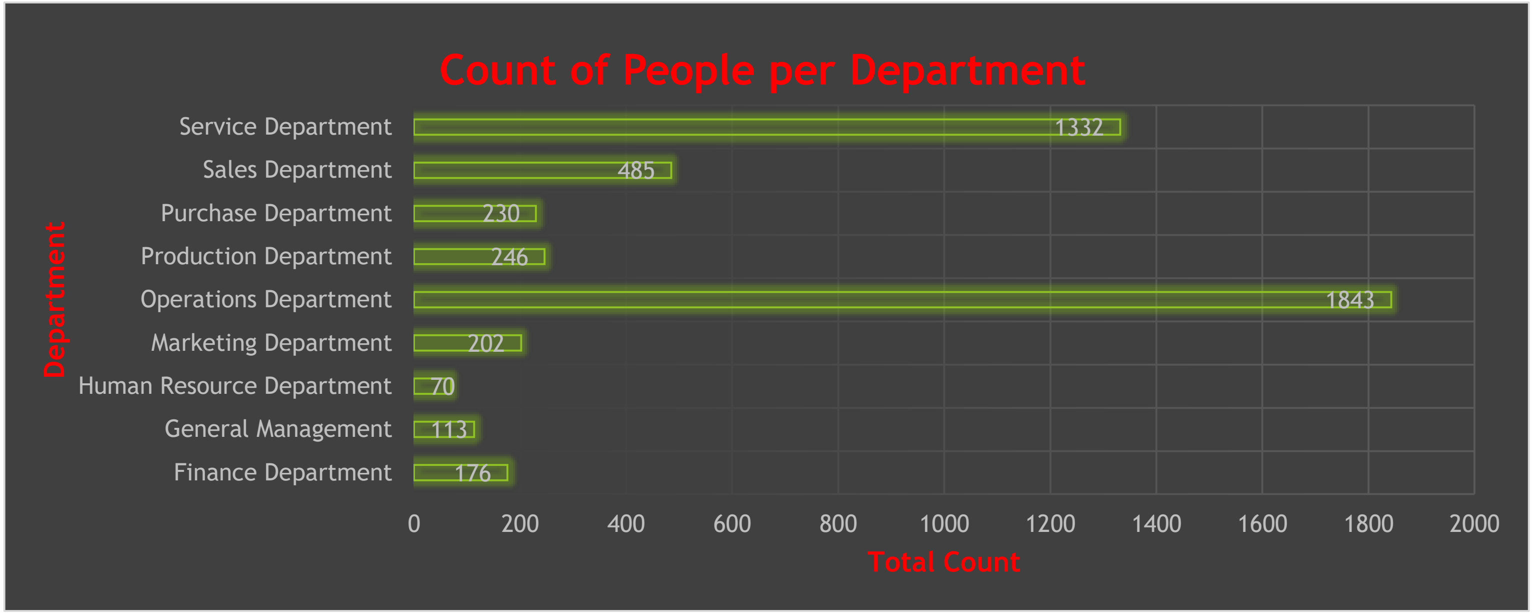
Task B - Salary Analysis:

- What is the average salary offered by this company?
- Use Excel functions to calculate this. The average of the salaries was found by using the Excel formula “average”.
- =AVERAGE('Raw Data'!G2:G7169)
- Average Salary - 49978.1486

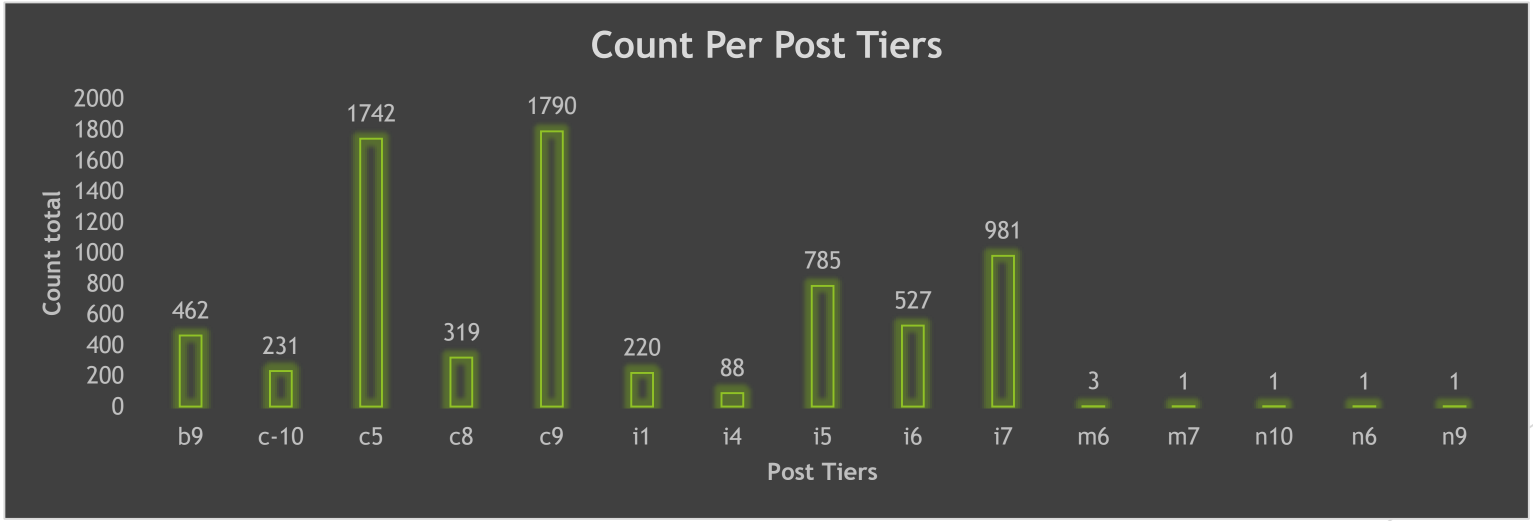
Task C - Salary Distribution:

- Create class intervals for the salaries in the company. This will help you understand the salary distribution. Got the count of the people who falls in the category by “countifs” formula as used below:
- =COUNTIFS('Raw Data'!$G$2:$G$7169,">99", ‘Raw Data'!$G$2:$G$7169, “<10000”)
- It is evident that a greater number of people fall in the Salary category 40,000 to 50,000. And the least number of people fall above 1,00,000.

Task D - Departmental Analysis:



Task E - Position Tier Analysis :



Result :

- Through this project I was able to understand the formulas being used in the Excel which can be used to find the Statistical measures such as Mean, Median, mode and so on. I got used to the Excel formulas and how to convert the Raw Data into meaningful insights.
- And the steps which I used are – cleansing the data and using the formulas to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.
- I have achieved the end result and I think I have contributed my full support into the Analysis. I hope this project helps the Analysis and it achieves what it was tend to achieve.

IMDB Movie Analysis Project

- **Description :**

IMDb (Internet Movie Database) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb began as a fan-operated movie database in 1990, and moved to the Web in 1993. Since 1998, it has been owned and operated by IMDb.com, Inc., a subsidiary of Amazon.

- **Approach :**

I went through the Excel data provided by the Trainity IMDB Movie Analysis project and understood that there were columns related to the movie in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected

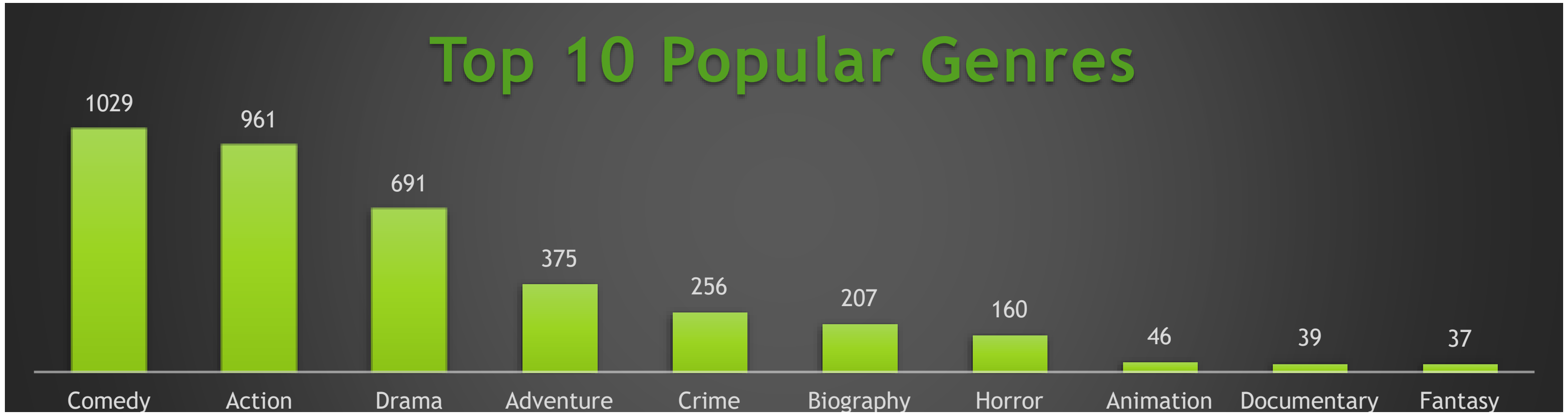
Tech-Stack Used :

- Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

Insights :

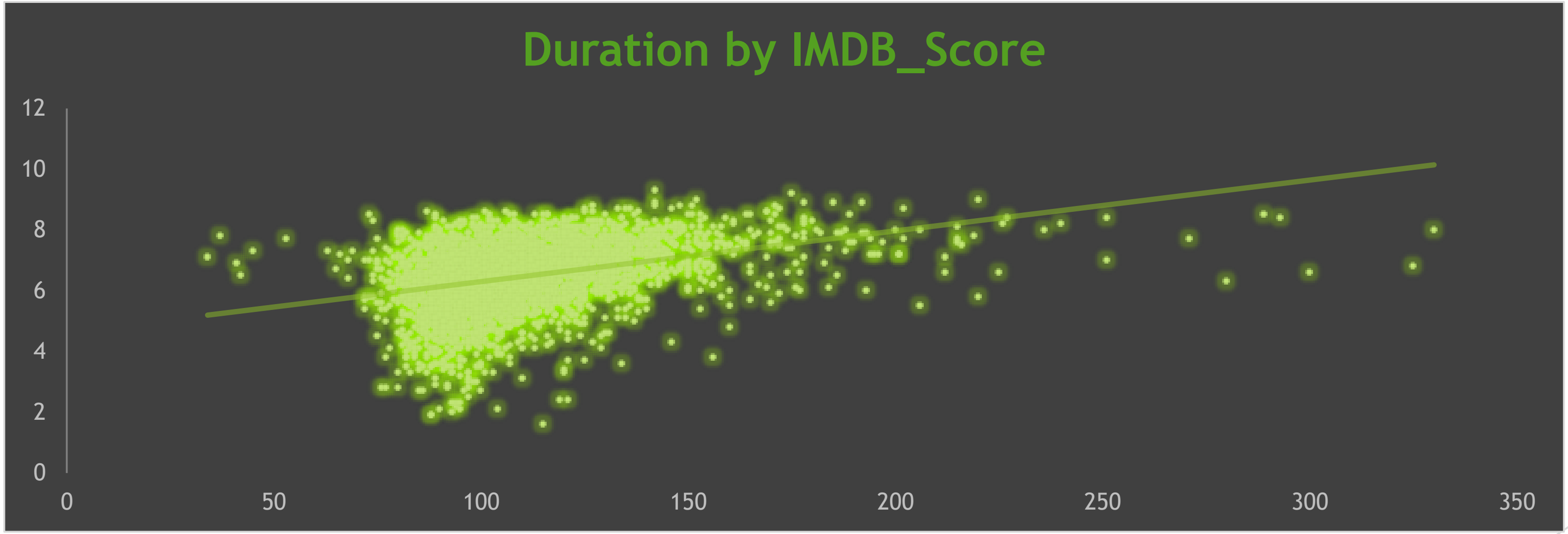
- Removing null values.
- Removed the columns which we don’t use for the analysis.
- Removing the Duplicate rows.
- Have used the in-built formulas in excel for the descriptive analysis such as:
  Mean – average()
- Median – median()
- Mode – mode()
- Max – max()
- Min – min()
- Variance – VAR.P()
- Standard Deviation - STDEV.P()

- **Task A - Genre Analysis :**



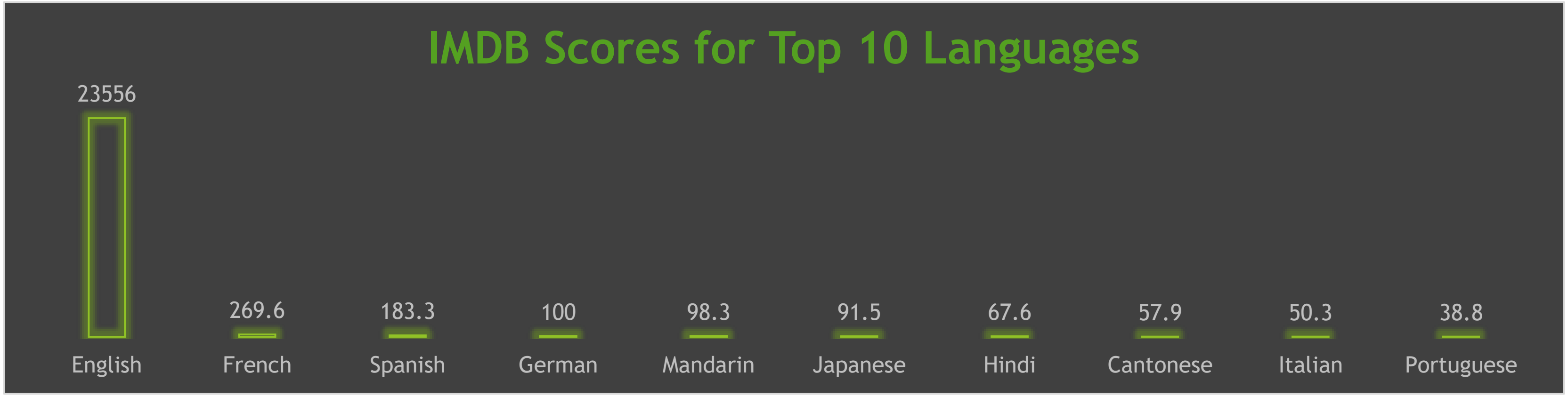
We could see that the most popular genre is **Comedy** and followed by the other genres.

- **Task B – Movie Duration Analysis :**



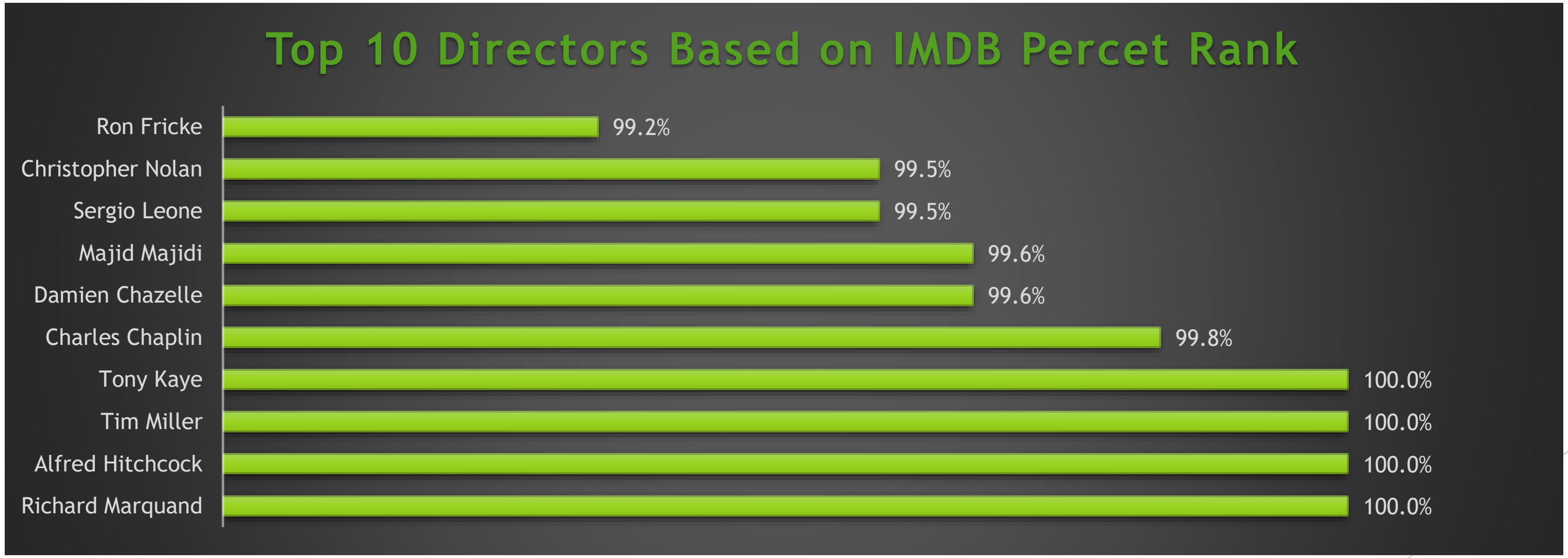
But mostly the IMDB score is more when the duration is between 80 mins to 150mins.

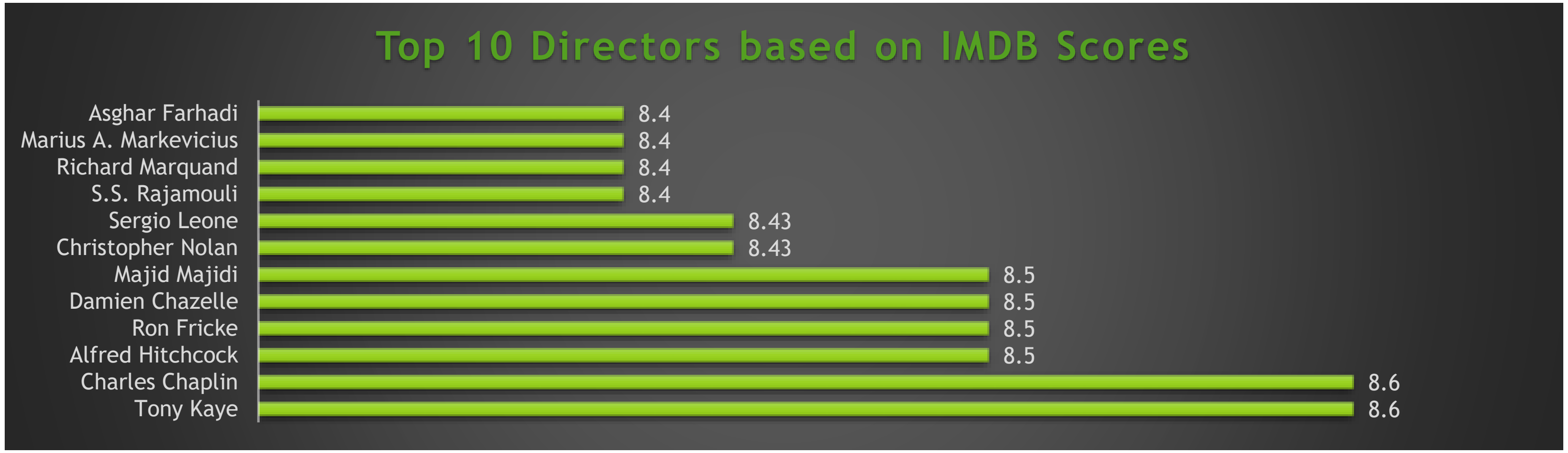
- **Task C – Language Analysis :**



The language “**English**” is the most common language.

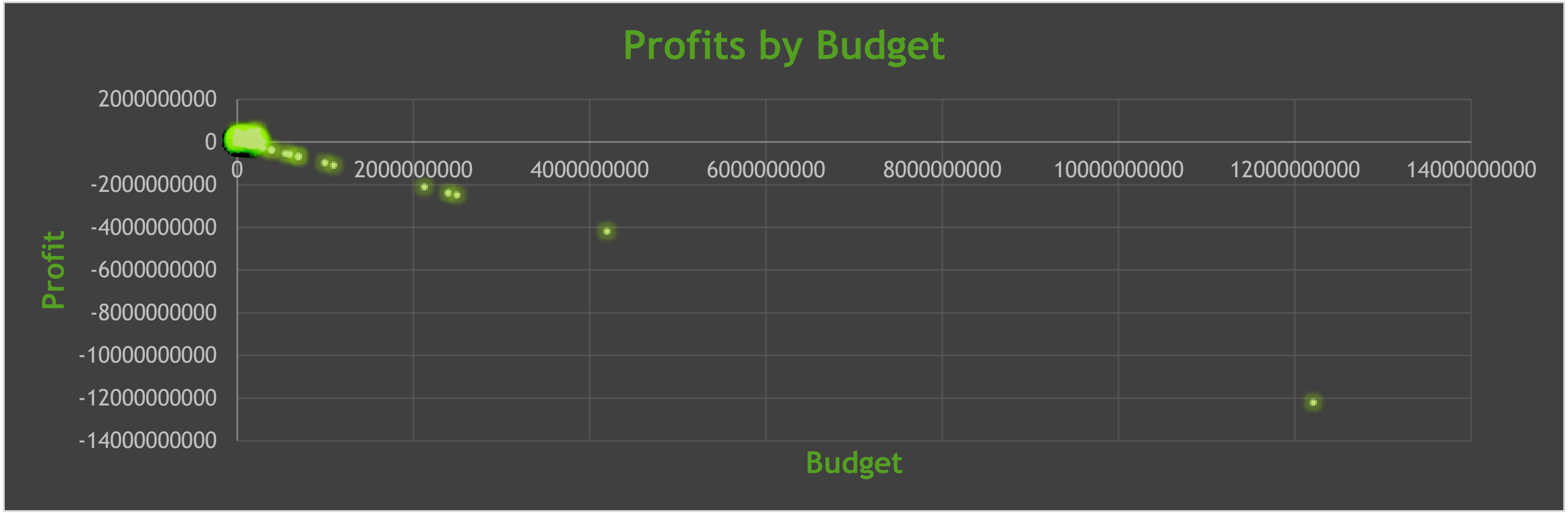
- **Task D – Director Analysis :**





The top directors with more IMDB scores as well as with the percent rank based on the IMDB scores.

- **Task E – Budget Analysis :**



The most profitable movie is **Avatar** with a profit of **523,505,847 Dollars (5.2 billion Approx).**

Result :

- Through this project I was able to understand the formulas being used in the Excel which can be used to find the Statistical measures such as Mean, Median, Mode, Max, Min, Variance and Standard Deviation.

- I got used to the Excel formulas and how to convert the Raw Data into meaningful insights. And the steps which I used are – cleansing the data and using the formulas to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.

- I have achieved the end result and I think I have contributed my full support into the Analysis. I hope this project helps the Analysis and it achieves what it was tend to achieve

Bank Loan Case Study Project

Description :

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

- When a customer applies for a loan, there are four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

Approach :

I went through the Excel data provided by the Trainity Bank Loan Case Study and understood that there were columns related to the Bank loan in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected

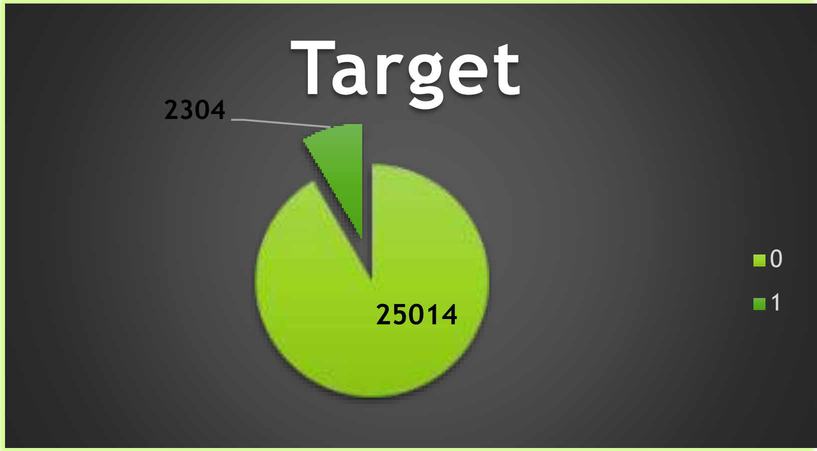
- **Tech-Stack Used :**

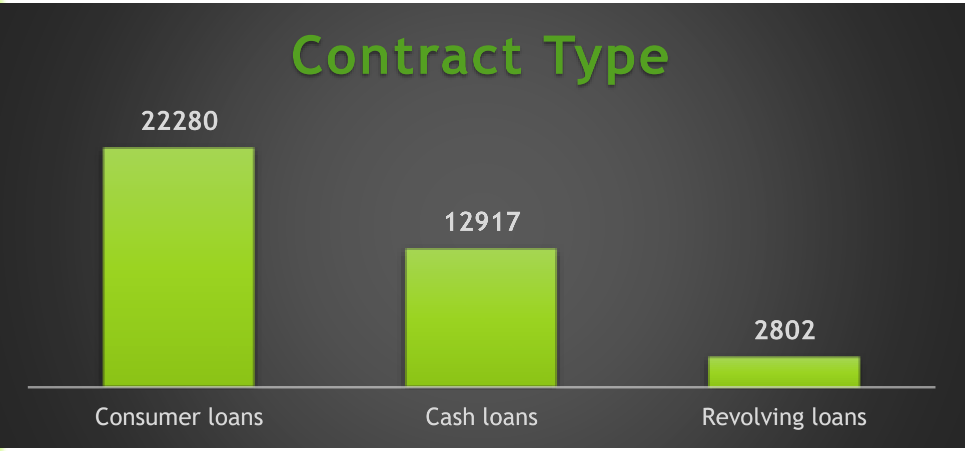
Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

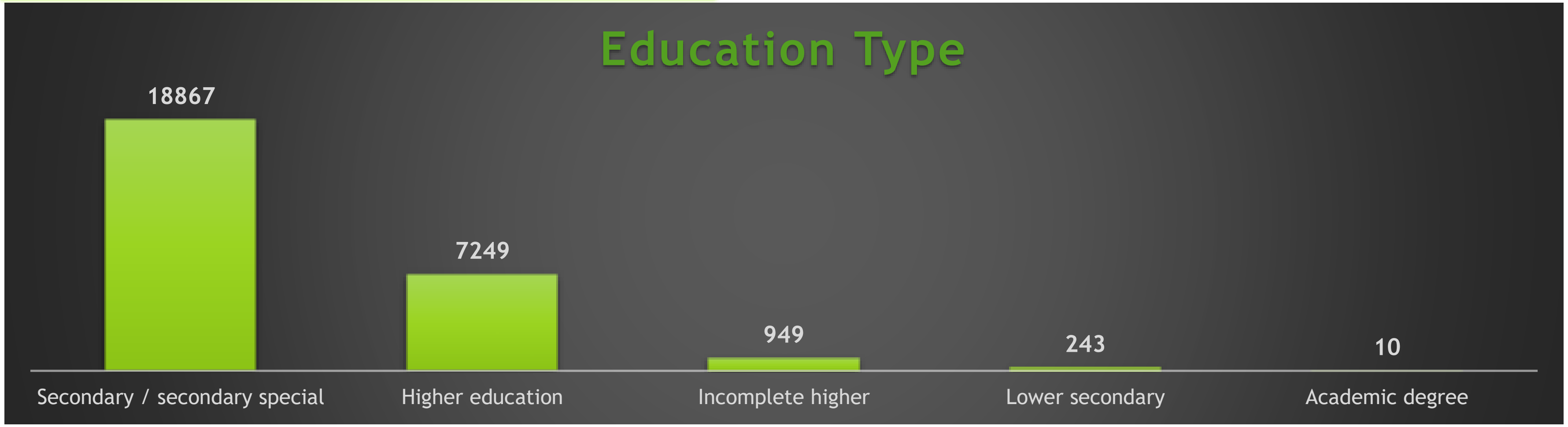
- **Insights :**
- Did the data cleaning like:
  - Removing null values.
  - Removed the columns which we don’t use for the analysis.
  - Removing the Duplicate rows.
- Before the Data Cleaning the column number for the Excel:
- application_data – 126, After cleaning now we have 77 columns.
- previous_application – 37, After cleaning now we have 26 columns.

- **Task A - Identify Missing Data and Deal with it Appropriately :**
- Used the formula counta formula to find out the total number of values in each column:
- =COUNTA(B4:B50002)
- Then to find out the Percentage of Null values used this formula:
- =1-B2/B2
- Also decided to remove the columns which had more than 35% of null values since they don’t contribute much to our analysis.
- **Task B – Identify Outliers in the Dataset**
- For this task, we need to find out the Quartile, Inter Quartile Range (IQR), Upper Limit, Lower Limit. Hence, I used the excel inbuilt formulas such as:
- Quartile 1: =QUARTILE.INC(B2:B27319,1)
- Quartile 3: =QUARTILE.INC(B2:B27319,3)
- Inter Quartile: Q3 – Q1
- Upper Limit: Q3 + (1.5 * IQR)
- Lower Limit: Q3 - (1.5 * IQR)

Task C – Analyse Data Imbalance :

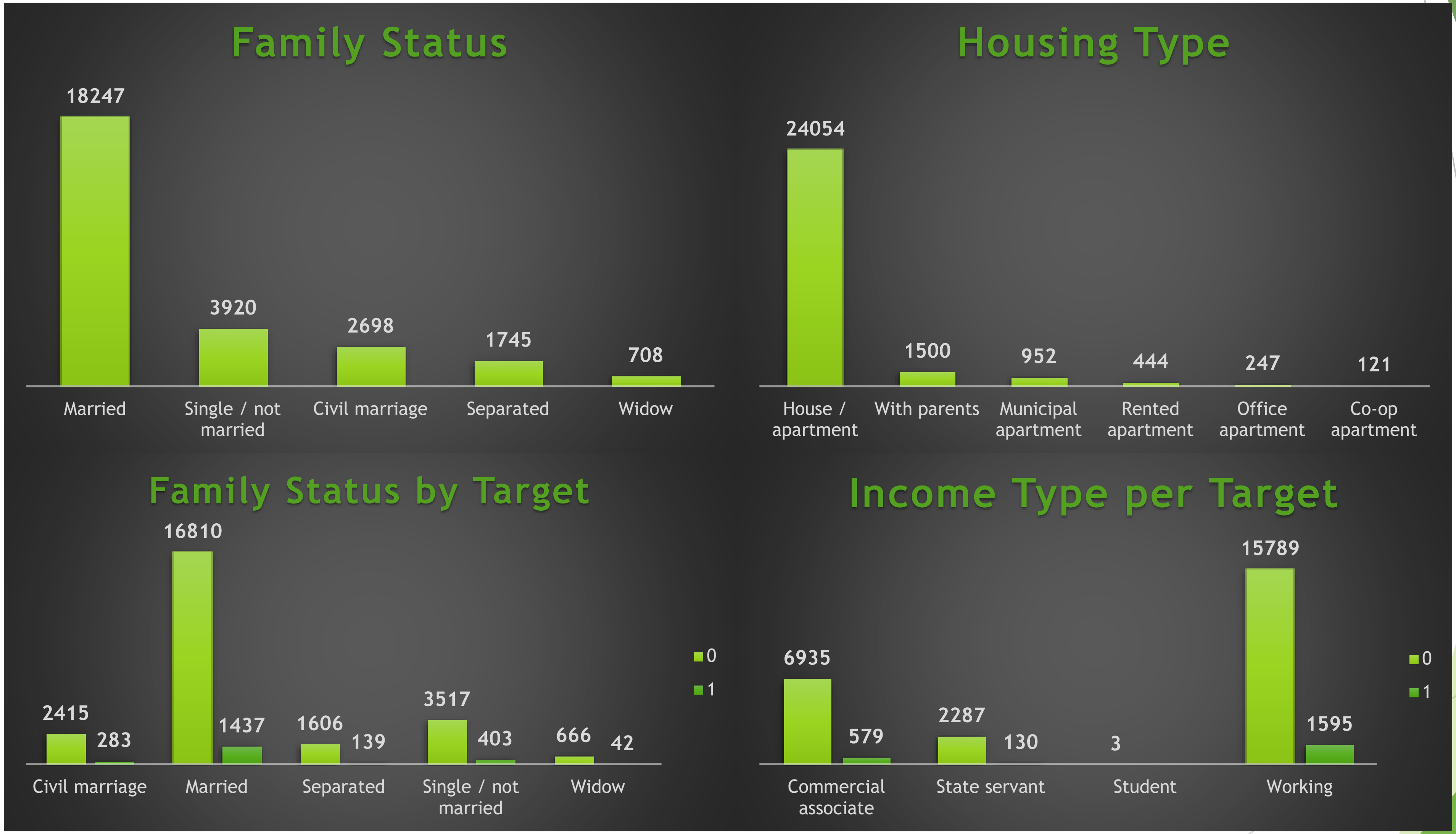




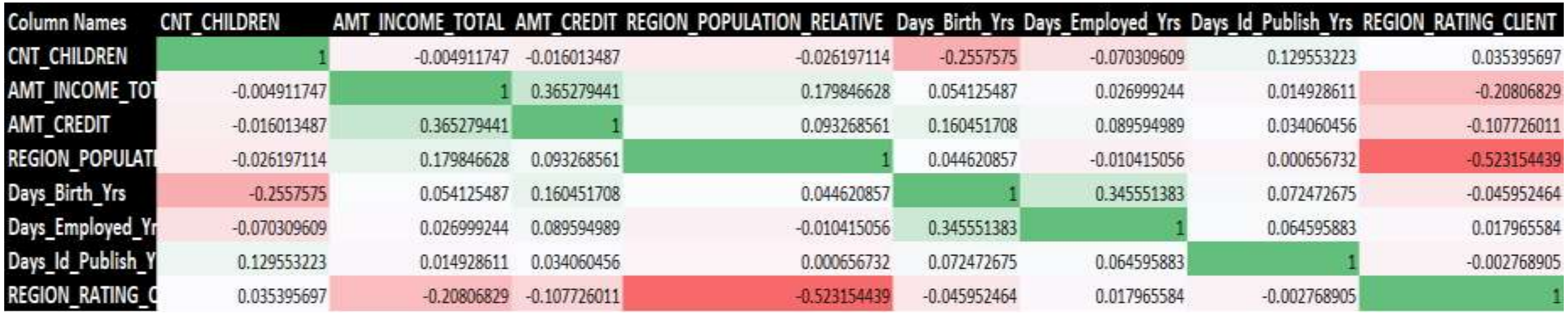


- The target 1 is so less when compared to the target value 0.
- Consumer Loans are more in number than the Cash loans and revolving loans.
- Secondary / Secondary Special use more loans than other categories in education type.

- **Task** D – Perform Univariate, Segmented Univariate, and Bivariate Analysis :



- **Task E – Identify Top Correlations for Different Scenarios :**



Column Names	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	Days_Birth_Yrs	Days_Employed_Yrs	Days_Id_Publish_Yrs	REGION_RATING_CLIENT
CNT_CHILDREN	1	-0.004911747	-0.016013487	-0.026197114	-0.2557575	-0.070309609	0.129553223	0.035395697
AMT_INCOME_TOT	-0.004911747	1	0.365279441	0.179846628	0.054125487	0.026999244	0.014928611	-0.20806829
AMT_CREDIT	-0.016013487	0.365279441	1	0.093268561	0.160451708	0.089594989	0.034060456	-0.107726011
REGION_POPULATI	-0.026197114	0.179846628	0.093268561	1	0.044620857	-0.010415056	0.000656732	-0.523154439
Days_Birth_Yrs	-0.2557575	0.054125487	0.160451708	0.044620857	1	0.345551383	0.072472675	-0.045952464
Days_Employed_Yr	-0.070309609	0.026999244	0.089594989	-0.010415056	0.345551383	1	0.064595883	0.017965584
Days_Id_Publish_Y	0.129553223	0.014928611	0.034060456	0.000656732	0.072472675	0.064595883	1	-0.002768905
REGION_RATING_C	0.035395697	-0.20806829	-0.107726011	-0.523154439	-0.045952464	0.017965584	-0.002768905	1

Columns	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	CNT_PAYMENT
AMT_ANNUITY	1	0.825528512	0.818147682	0.825564271	0.394100654
AMT_APPLICATION	0.825528512	1	0.993466353	0.999901663	0.663825921
AMT_CREDIT	0.818147682	0.993466353	1	0.993444101	0.69334727
AMT_GOODS_PRICE	0.825564271	0.999901663	0.993444101	1	0.663684765
CNT_PAYMENT	0.394100654	0.663825921	0.69334727	0.663684765	1

Result :

- Through this project I was able to understand the formulas being used in the Excel which can be used to find the Correlation and various charts on how to use them. I got used to the Excel formulas and how to convert the Raw Data into meaningful insights. And the steps which I used are – cleansing the data and using the formulas to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.

- As a result, we could summarize as there is higher possibility for the adults who fall in the category:

1. Married
2. Educated
3. Strong Work Experience
4. Previously Approved Clients

- The people who don't tend to take loan falls in the category:

1. Unemployed
2. Youth
3. Less Work Experience
4. Previously Unapproved Clients

Impact of Car Analysis Project

- **Description :**

- The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

- This problem could be approached by analysing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

Approach:

I went through the Excel data provided by the Trainity Impact of Car Features project and understood that there were columns related to the Car Features in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected.

Tech-Stack Used:

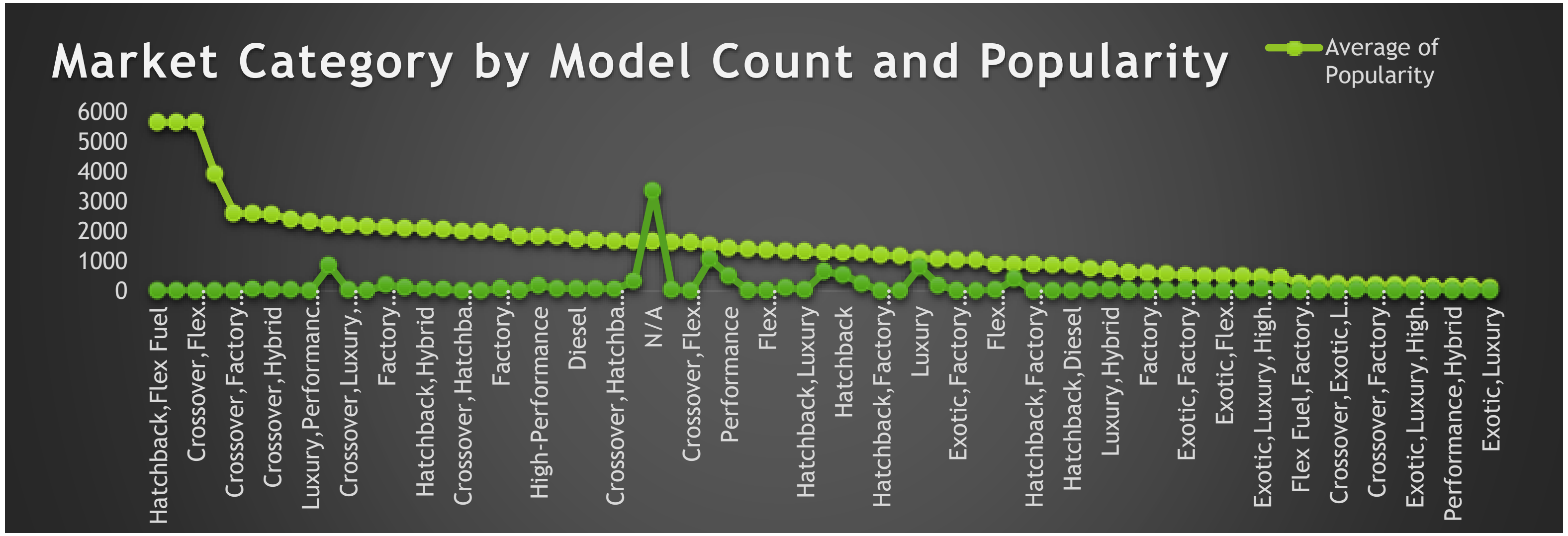
Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

Insights:

Did the data cleaning like:

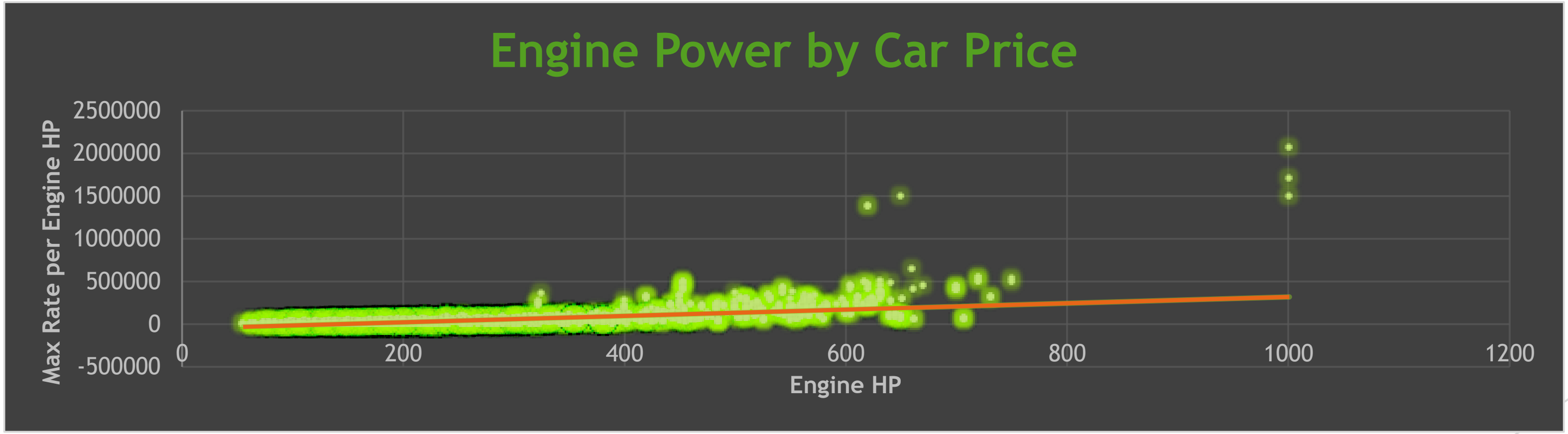
- Removing null values.
- Removed the columns which we don’t use for the analysis.
- Removing the Duplicate rows.
- Before the Data Cleaning the number of columns in the Excel were:
- Car data – 11915, After cleaning now we have 11098 columns.

- **Task A – How does the popularity of a car model vary across different market categories?**



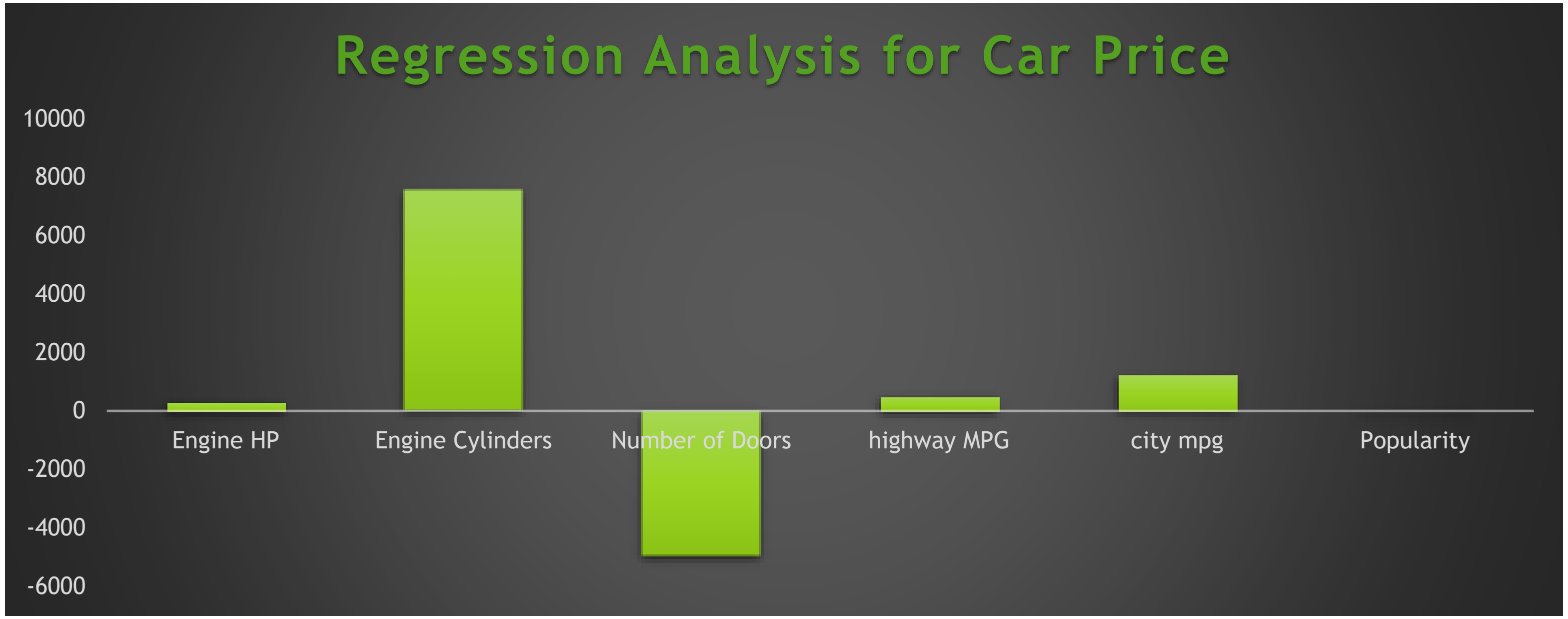
The “Crossover” has a greater number of counts and with the popularity it is: “Hatchback, Flex Fuel, Crossover, Performance” is more.

- **Task B – What is the relationship between a car's engine power and its price?**



With the increase in the Engine HP, the Price of the car also increases.

- **Task C – Which car features are most important in determining a car's price?**

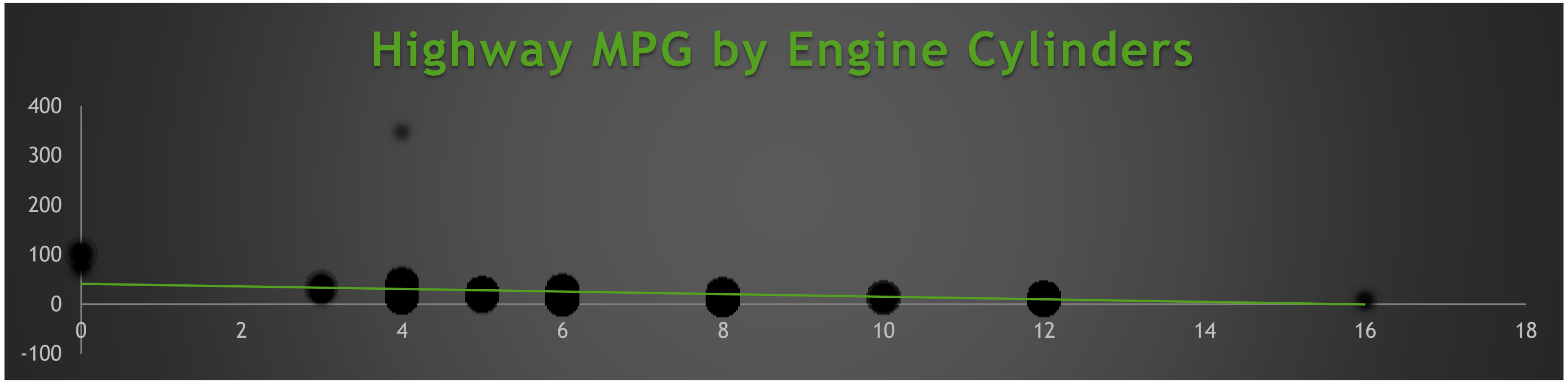


The “Engine Cylinders” are more related to the Car Price and the least related column is “Number of doors”.

- **Task D – How does the average price of a car vary across different manufacturers?**

- Bugatti is having a greater average of MSRP among other brands.
- And Plymouth is having the least average of MSRP.

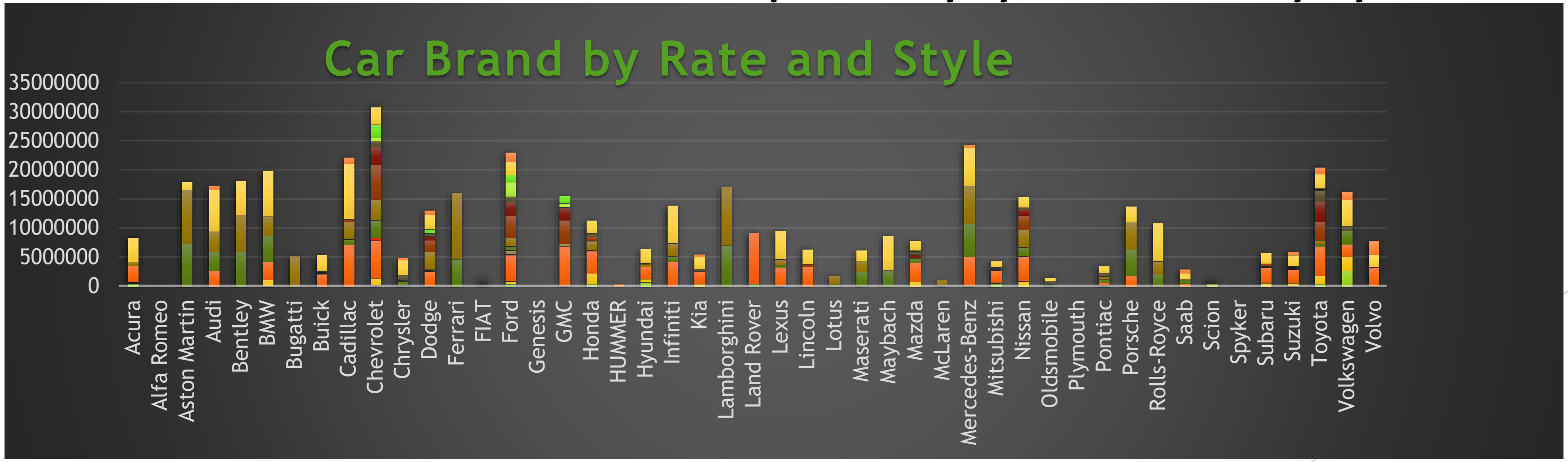
- **Task E – What is the relationship between fuel efficiency and the number of cylinders in a car's engine?**



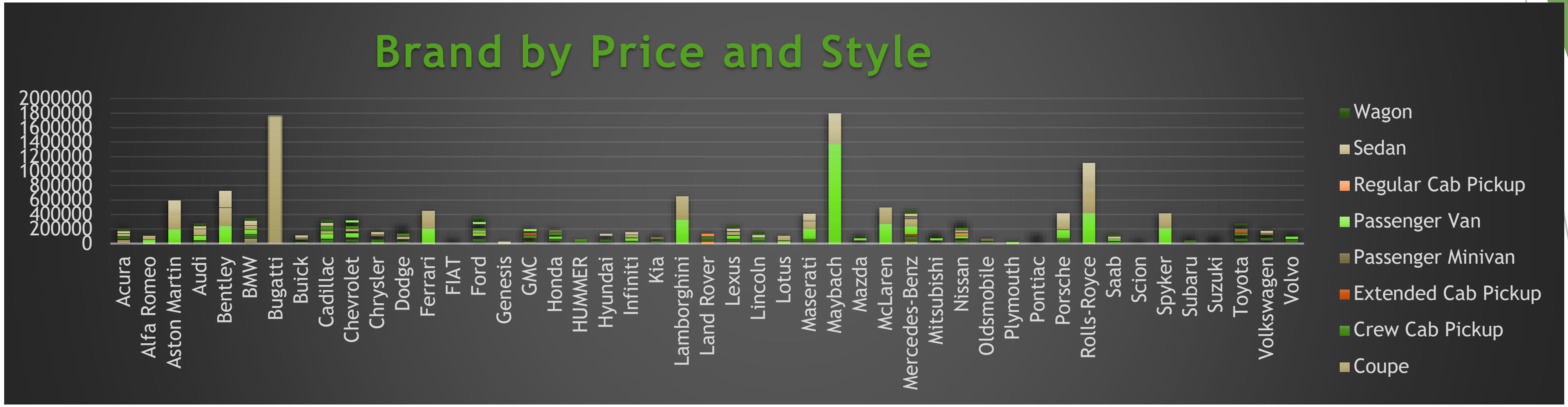
The Trendline is coming down when there is an increase in the Engine Cylinders.

Building a Dashboard:

- **Task 1: How does the distribution of car prices vary by brand and body style?**

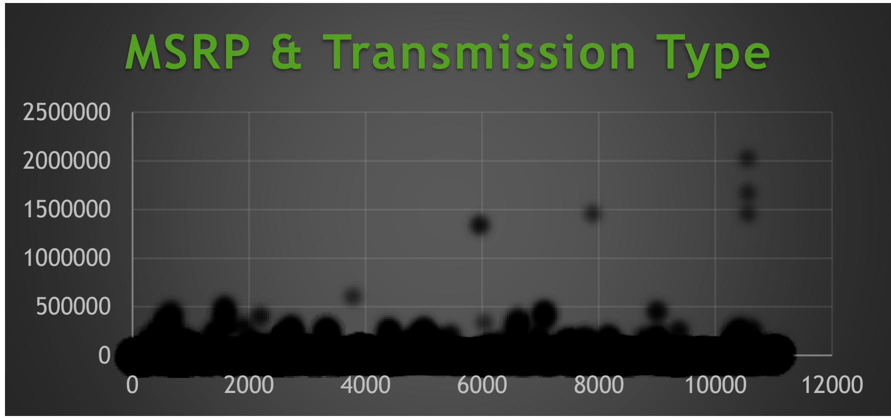


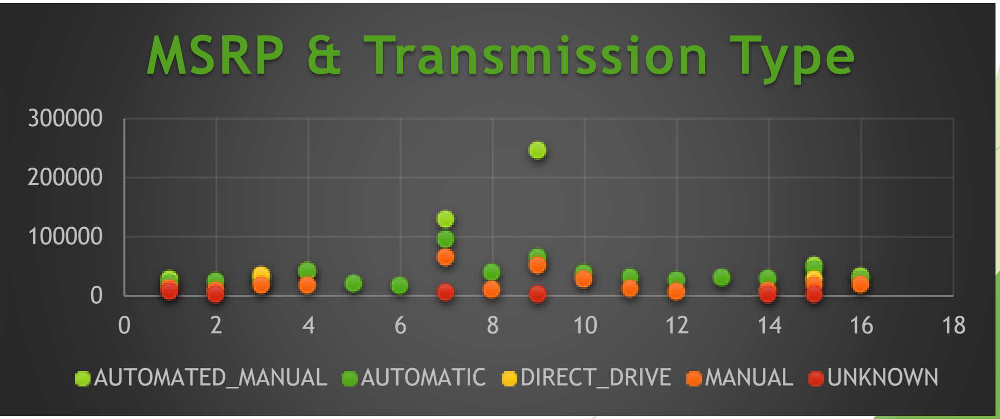
- **Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?**



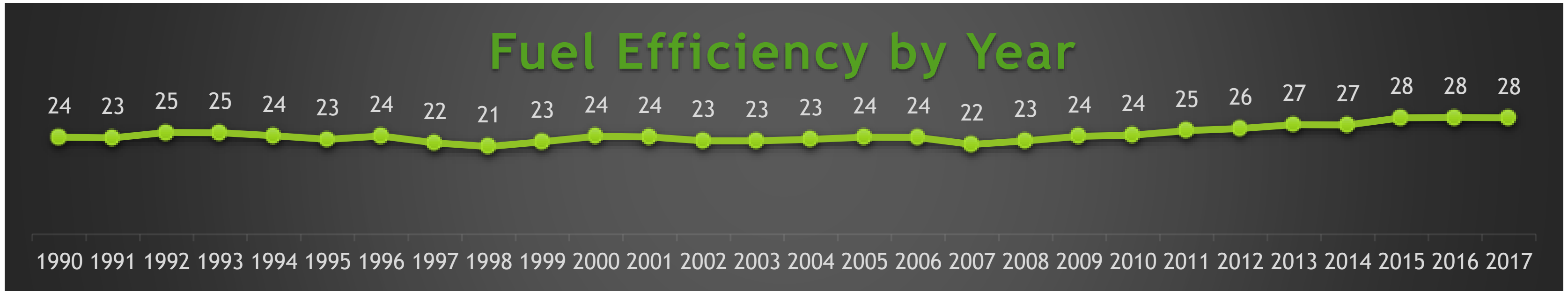
Bugatti is the car brand with Highest Average and Plymouth is the car brand with Lowest Average.

- **Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?**



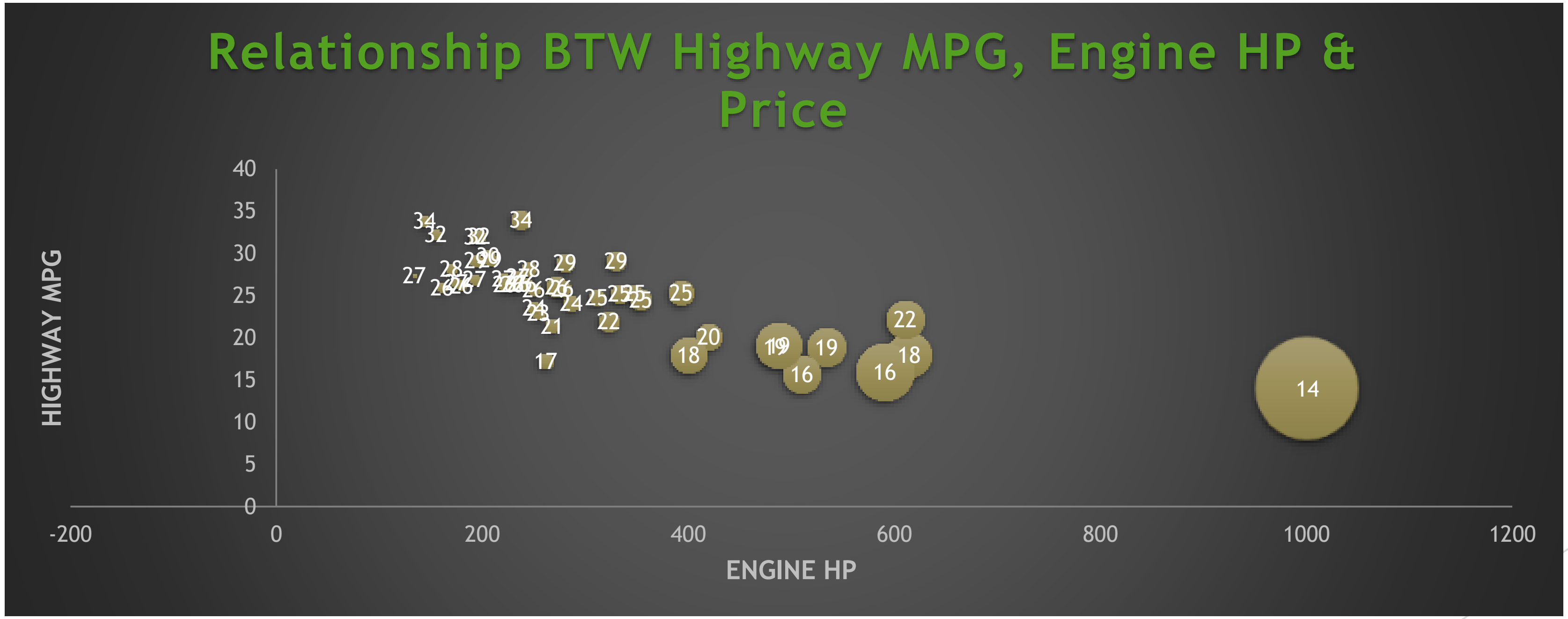


- **Task 4: How does the fuel efficiency of cars vary across different body styles and model years?**



There was a dip in the year 2007, the Fuel efficiency is slowly increasing year by year.

- **Task 5: How does the car's horsepower, MPG, and price vary across different Brands?**



Increase in the Engine HP, the MRSP also increases and there is a decrease in the Highway MPG.

Result :

- As a result, we could summarize as:

1. The “Crossover” has a greater number of counts and with the popularity it is: “Hatchback, Flex Fuel, Crossover, Performance”.
2. With the increase in the Engine HP, the Price of the car also increases.
3. The “Engine Cylinders” are more related to the Car Price.
4. Bugatti is having a greater average of MSRP and Plymouth is having the least average.
5. The transmission type Automated Manual is expensive.
6. The Fuel efficiency is slowly increasing year by year.
7. When there is an increase in the Engine HP, the MRSP also increases and there is a decrease in the Highway MPG.

ABC Call Volume Company Project

- ## Description :

We will be provided with a dataset that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

In this project, I'll be using my analytical skills to understand the trends in the call volume of the CX team and derive valuable insights from it.

- ## **Approach** :

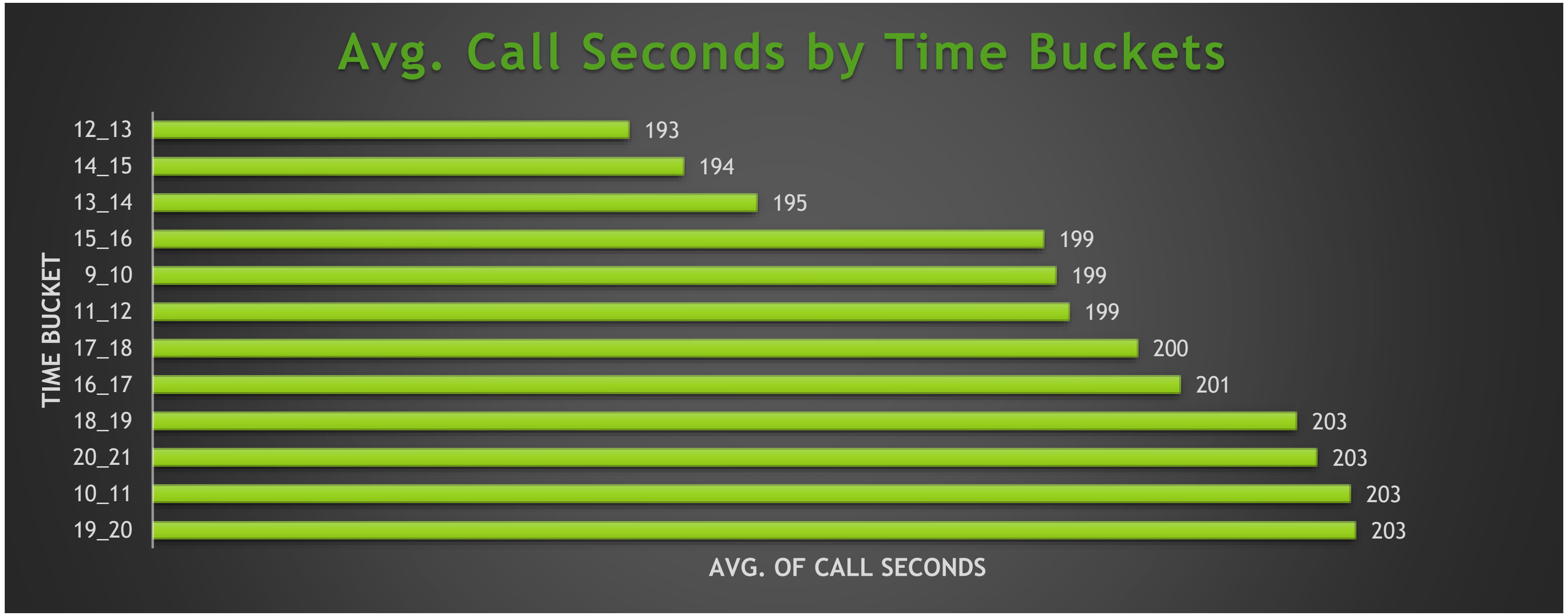
I went through the Excel data provided by the Trainity ABC Call Volume Trend Analysis project and understood that there were columns related to the Call Features in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected

- **Tech-Stack Used:**

Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

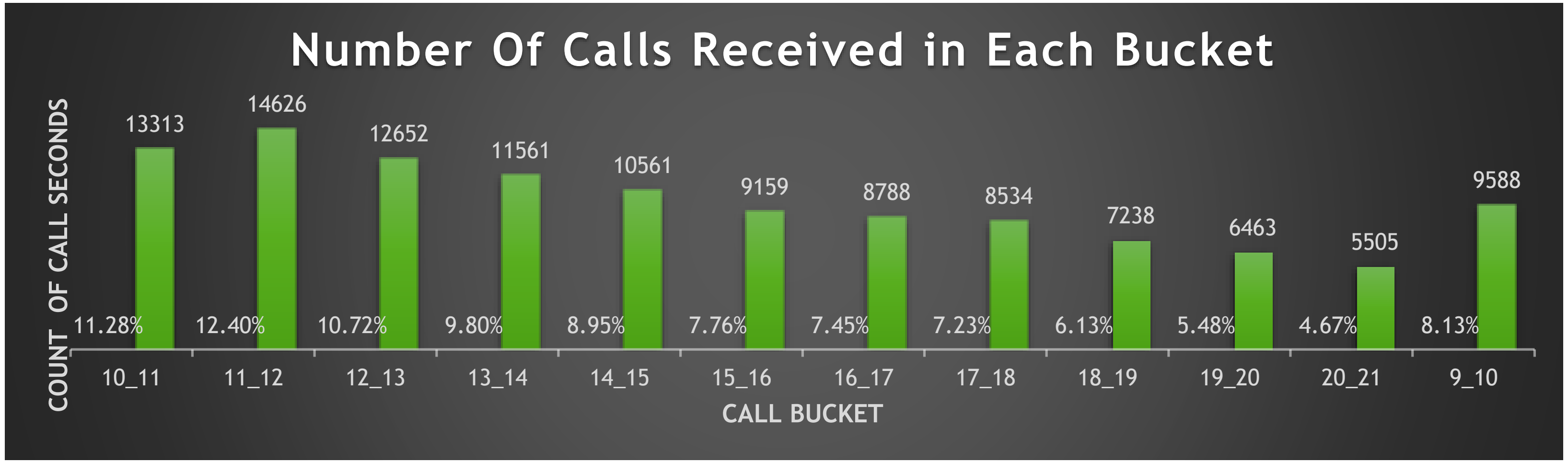
- **Insights:**

- **Task A – Average Call Duration: What is the average duration of calls for each time bucket?**



The time bucket from 18pm – 21pm have a greater number of customers who answer the call.

- **Task B – Call Volume Analysis: Can you create a chart or graph that shows the number of calls received in each time bucket?**



The time bucket 11am -12pm is having a greater number of calls with count 14626.

- **Task C – Manpower Planning: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?**
- Answered: **70%**
- Abandoned: **29%**
- Transferred: **1%**
- We can add a total of 57 agents to reduce the abandon rate from 30% to 10%.

- **Task D – Night Shift Manpower Planning: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.**

- Answered: **82452**
- Abandoned: **34403**
- Transferred: **1133**

Avg. Daily Call	5130
Night Calls 30%	1539
Additional Hours Required	77
Additional Agents Required	15

Night Time Bucket	Call Distribution	Time Distribution	Agent Required
9pm - 10pm	3	10	1.5
10pm - 11pm	3	10	1.5
11pm - 12am	2	15	1
12am - 1am	2	15	1
1am - 2am	1	30	0.5
2am - 3am	1	30	0.5
3am - 4am	1	30	0.5
4am - 5am	1	30	0.5
5am - 6am	3	10	1.5
6am - 7am	4	7.5	2
7am - 8am	4	7.5	2
8am - 9am	5	6	2.5
Total	30		15

- 30 Agents who can be added in the time bucket so that there can be agents who can answer to the query in the night time as well.

Result :

- As a result, we could summarize as:

1. The time bucket from 18pm – 21pm have a greater number of customers who answer the call.
2. The time bucket 11am -12pm is having a greater number of calls with count 14626 when compared with other time buckets.
3. A total of 57 agents to reduce the abandon rate from 30% to 10%.
4. 30 Agents who can be added in the time bucket so that there can be agents who can answer to the query in the night time as well.

Learnings From the Projects

- Learned how to analyse the raw data and how to do the data cleaning in Excel.
- To understand the problem description of the user.
- And to get the questions from the users to answer their needs/requirements.
- Use only the data which is needed for the analysis.
- Learnt various charts and its use cases.
- Learnt how to query in SQL for the analysis.
- Learnt how statistics help in the analysis.
- And learnt how the overall work for a Data Analyst to do in a company and how we can achieve the results with such raw data.