# Data Analysis Portfolio

Prepared By:-
Akshay Gosavi

# Professional Background

I am delighted to introduce myself as Akshay Gosavi, a individual, with a strong academic background and a genuine interest in data analysis. I hold a Bachelor of Science degree in Information Technology from K.M. Agrawal College in Kalyan, where I achieved a CGPI of 9.73.

During my time at K.M. Agrawal College I not gained a theoretical foundation but also honed my analytical and problem solving skills. Through coursework I mastered subjects such as Python, SQL and Advanced Excel equipping me with the tools to effectively handle complex data analysis tasks.

My ambition is to apply my accomplishments to contribute practically in the field of data analysis. With my belief in the insights derived from data driven approaches I am thrilled to utilize my skill set to extract meaningful patterns detect trends and provide actionable recommendations based on raw data. Ultimately my goal is to assist organizations in making decisions streamlining processes and achieving their objectives through data strategies.

# Table of contents


## 1. ABC Call Volume Trend

- Call Volume Analysis by Time Buckets
- Manpower Planning for Abandoned Calls
- Night Shift Manpower Planning

## 2. Impact of Car Features on Price and Profitability

- Exploring Car Features and Pricing Trends Over Time
- Predicting Car Prices Based on Features and Categories

## 3. Bank Loan Case Study

- Identifying Missing Data with Excel Functions
- Utilizing Statistical Measures and Conditional Formatting

## 4. IMDB Movie Analysis

- Identifying Top Directors and Their Impact on Ratings
- Revealing Movie Success Patterns by Choosing Popular Genre

## 5. Hiring Process Analytics

- Analyzing Company's Hiring Process
- Providing Insights for Improved Hiring Processes

## 6. Operation and Metric Analytics

- Deriving Insights from Operational Data for Decision-Making
- Leveraging Advanced SQL Skills for Data Analysis and Insights

## 7. Instagram User Analytics

- Provide insights for business growth
- Influencing development of a major social media platform

## 8. Data Analytics In Everyday Life

- Using real life situation to link it with the data analytics process

## 9. Appendix: Links To All Projects

# 1. ABC Call Volume Trend

## Project Description

- We have a call centre data with attributes like agent name, agent id, customer phone number, call duration  and so on.
- We need to find the average call duration, how to use all the agents to their full potential so that 90% of the calls are picked among all calls
- First I looked into the data to see if it needs to be cleaned or not but the given data was proper and doesn't need any cleaning or filtering
- I have also created charts wherever necessary for better visuals and understanding
- All the tasks are divided as per the sheets and also labelled properly

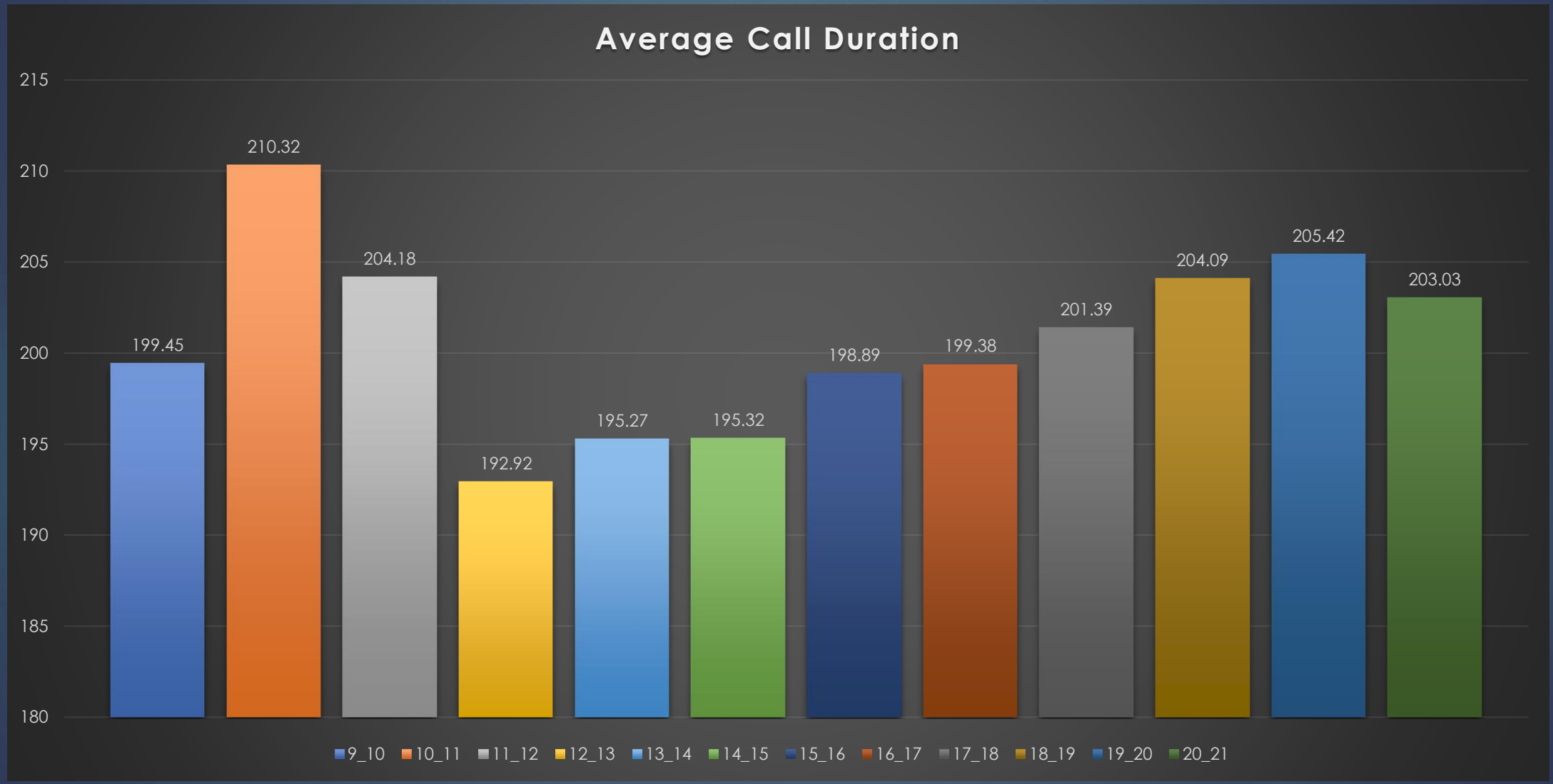
# Approach

- In this project I have used in-built formulas as well as pivot tables equally
- I have used averageifs, unique, sumproduct, countifs and done statistics to get the manpower required during night shift and pick up the 90% of the total calls
- Firstly I have put the data into table for better visuals
- One agent takes approximately 136 calls per day to get this I have divided Working hours in seconds by average answer time and rounded it to zero

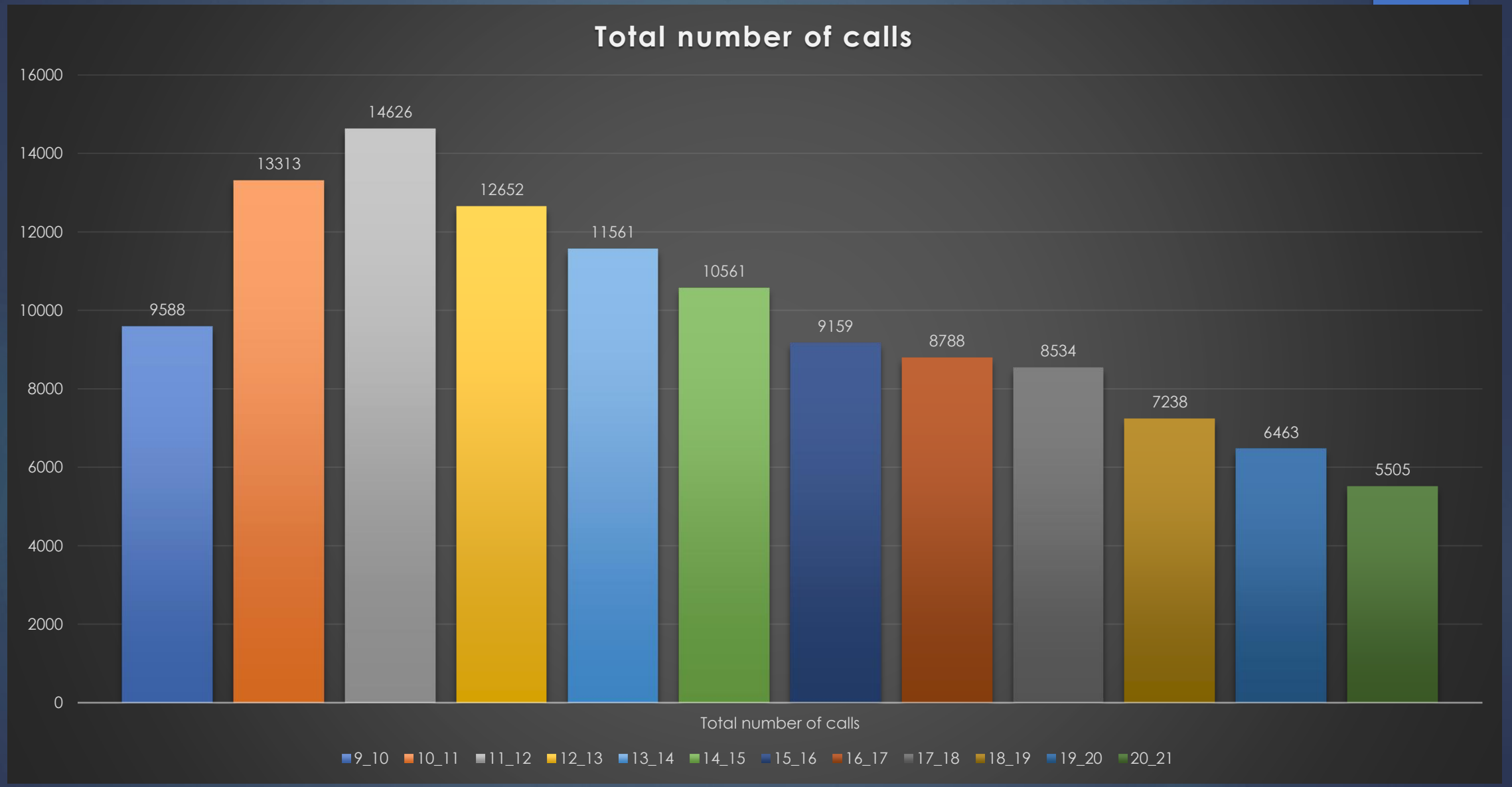
# Tech Stack Used

- Excel 2021
- Powerpoint 2021

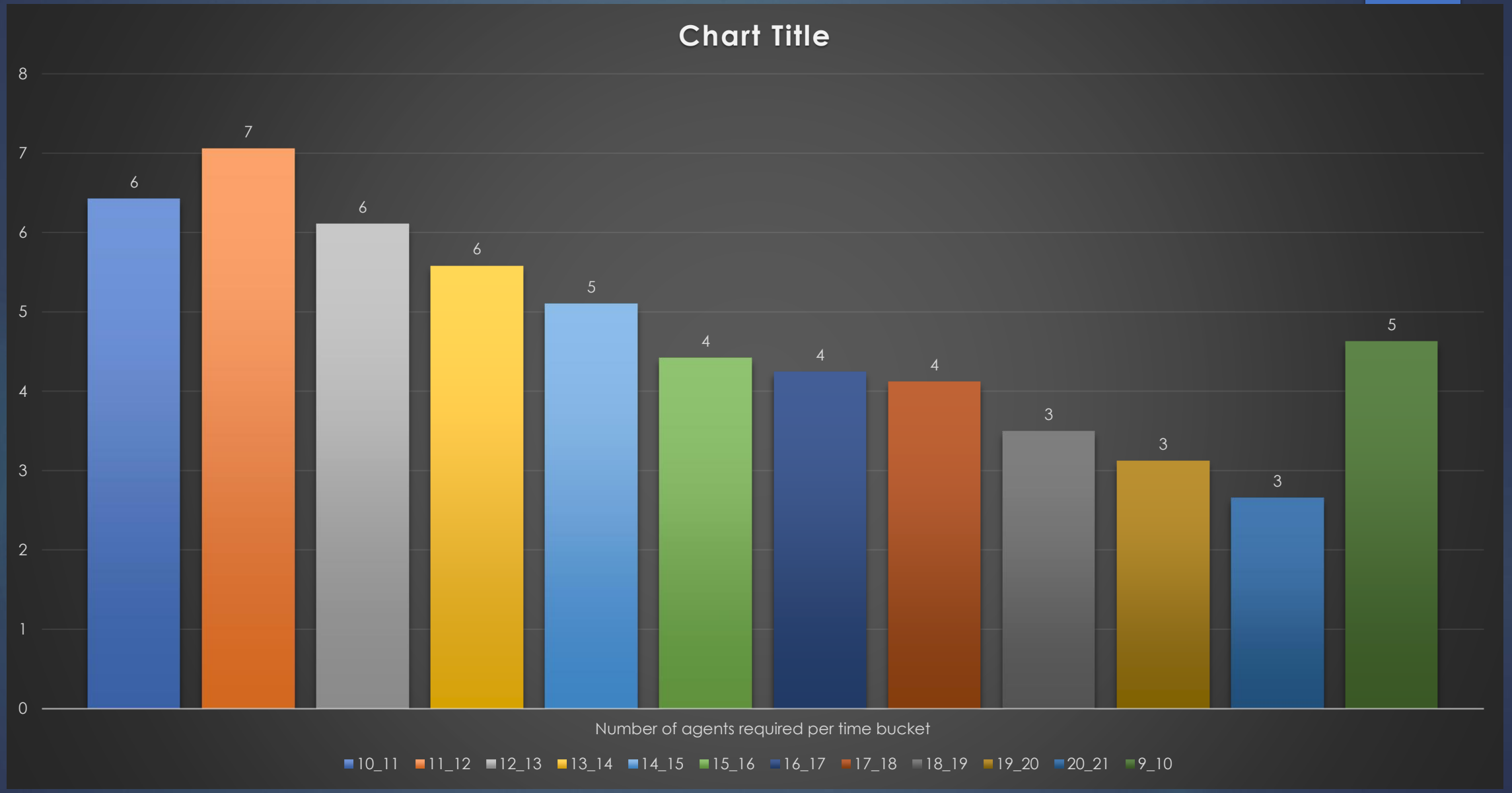
# Task 1: What is the average duration of calls for each time bucket?



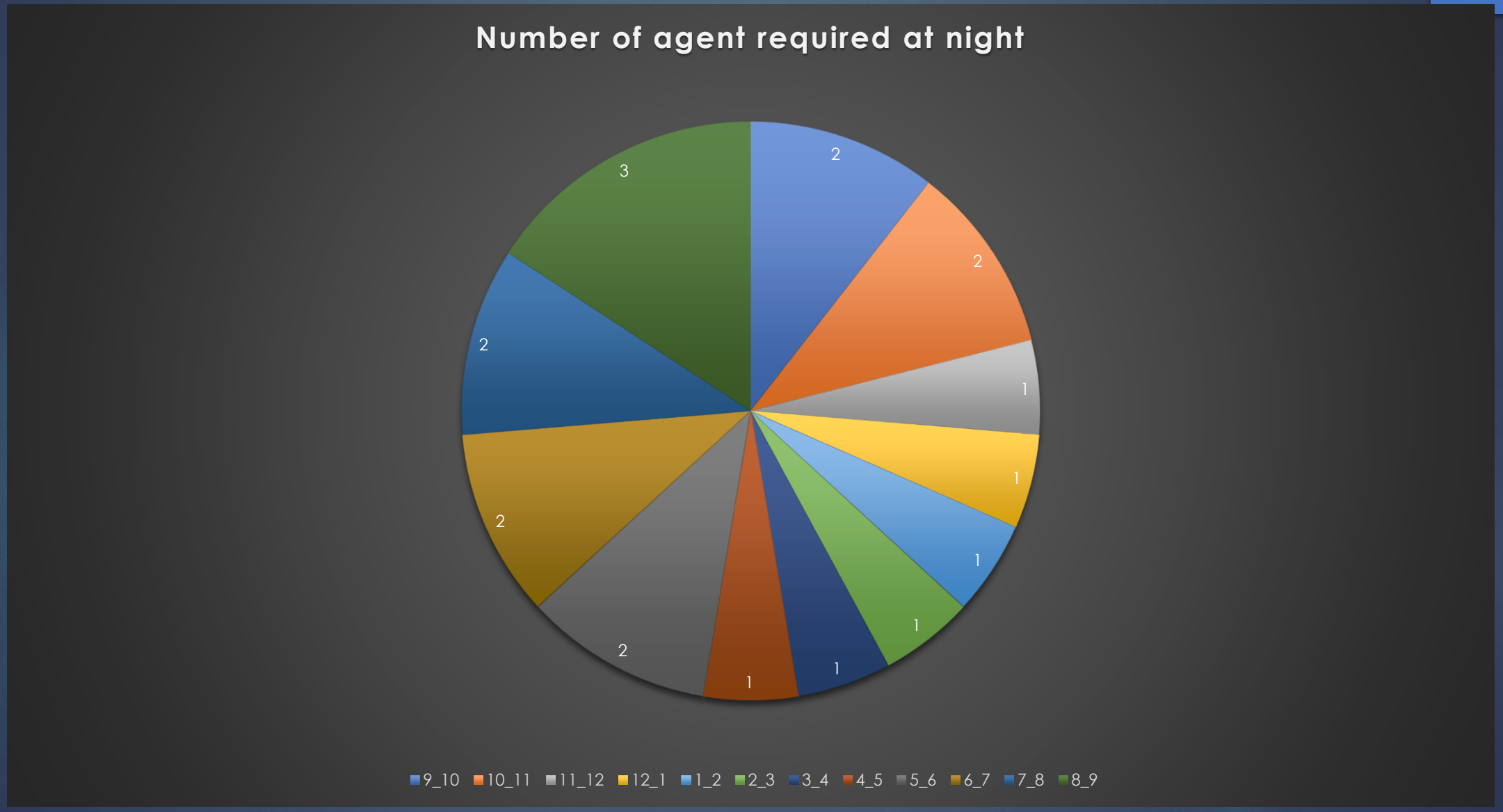
# Task 2: Can you create a chart or graph that shows the number of calls received in each time bucket?



# Task 3: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?



# Task 4: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.



# Results

- I have used formulas as well as pivot table, I could have fully used pivot table but it just makes tasks too easy
- I was mostly confused in the statistics part but I figured it out in the end
- Most task were easy, I just had to read and understand them carefully
- Using all the knowledge gain from statistics videos till now was quite a experience for me
- I will be explaining most of the task in loom videos since it will be quite convenient for me to explain and easy for you to understand

# 2. Impact of Car features

## Project Description

- In this project I have car data along with its various features, its brand, model, mileage, popularity, etc.
- I had to analyse that how the car's horsepower would affect the price of the car also keeping in mind about the car type such as sedan, hatchback, luxury, performance, etc.
- We also had certain columns which were blank so I had to first count them and the removed the whole row to get accurate results
- I have mainly used pivot tables to solve and create visualizations for the given queries

# Approach

- First of all I read the description of data column then I use countblank function and then removed the sheet rows which had blank cell
- Then for most of the queries I have used the pivot table to solve and create data visualisation charts
- I have divided all queries in different sheets for better view and understanding
- For query number 3 I have used excel's built in data analysis tool for regression analysis

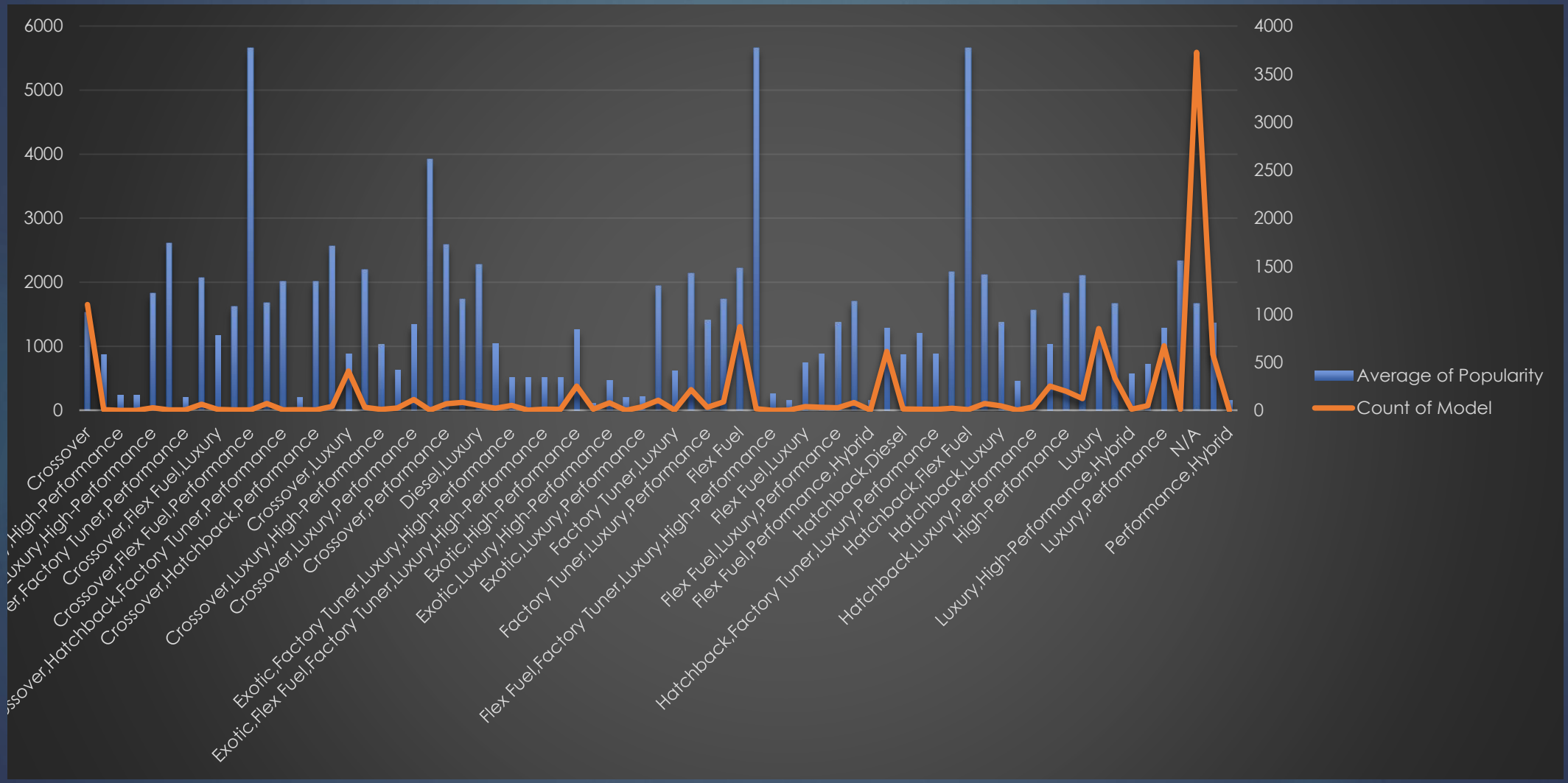
# Tech Stack Used

- Excel 2021
- PowerPoint 2021

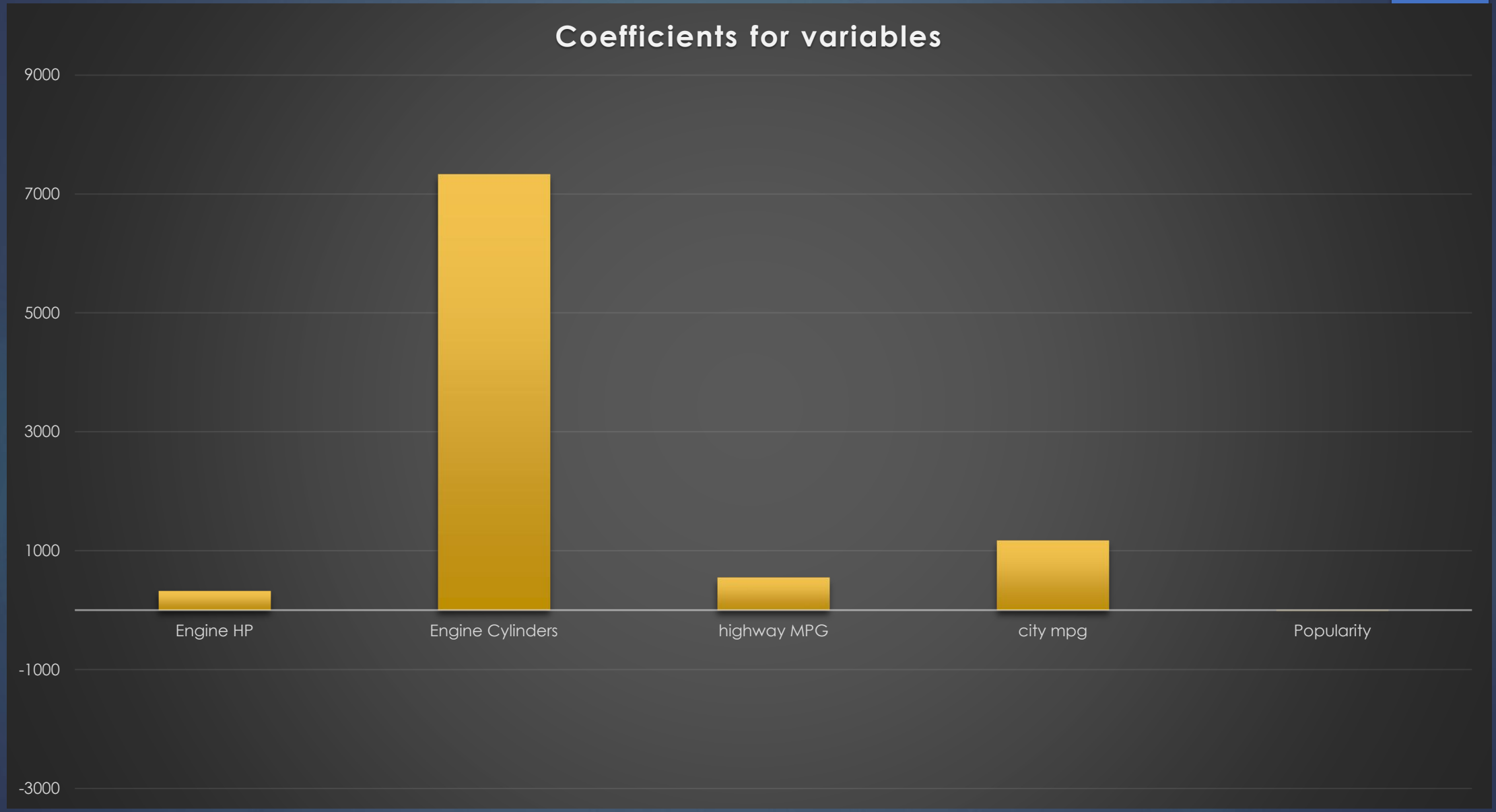
By Akshay Gosavi

Task 1.A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

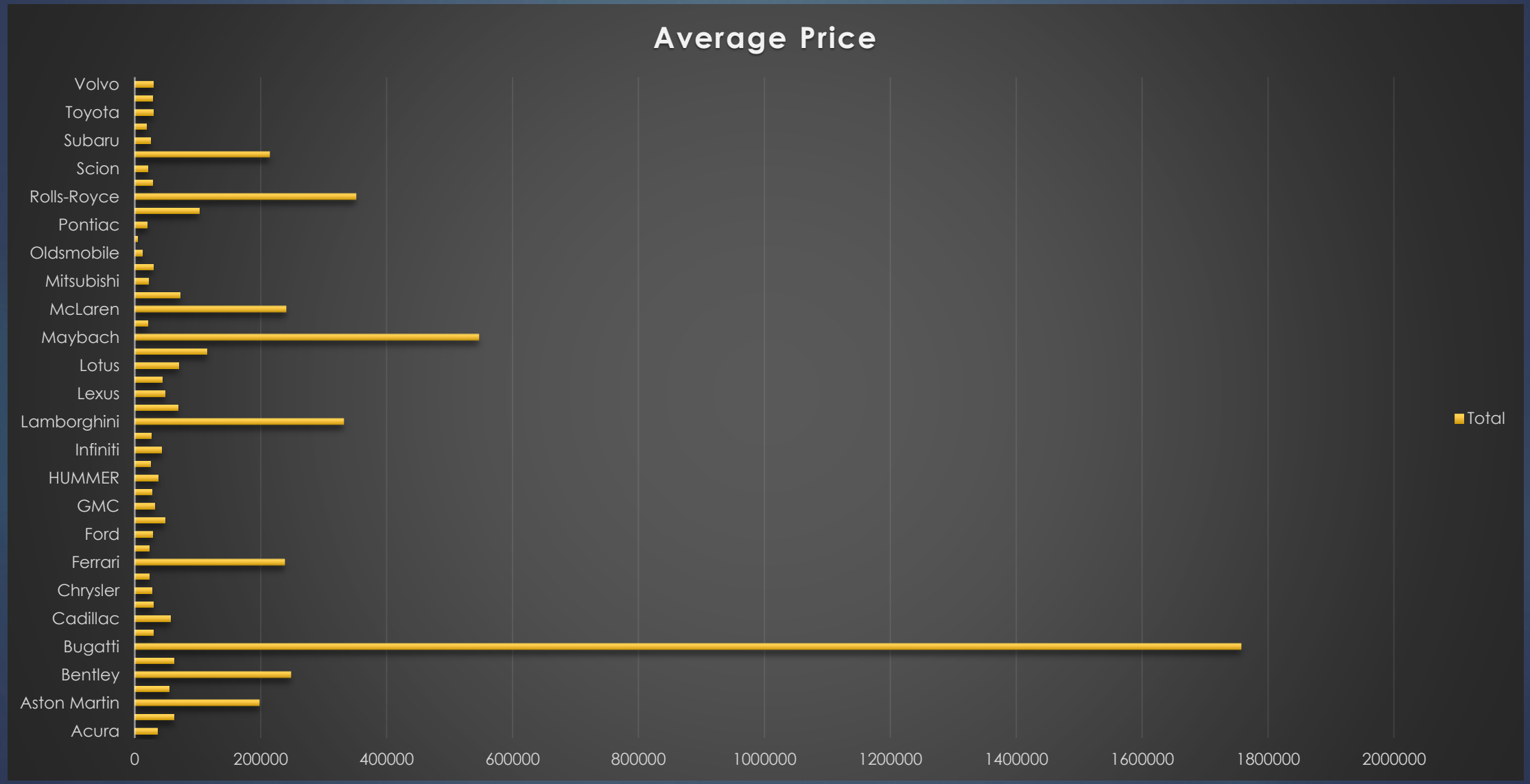
Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.



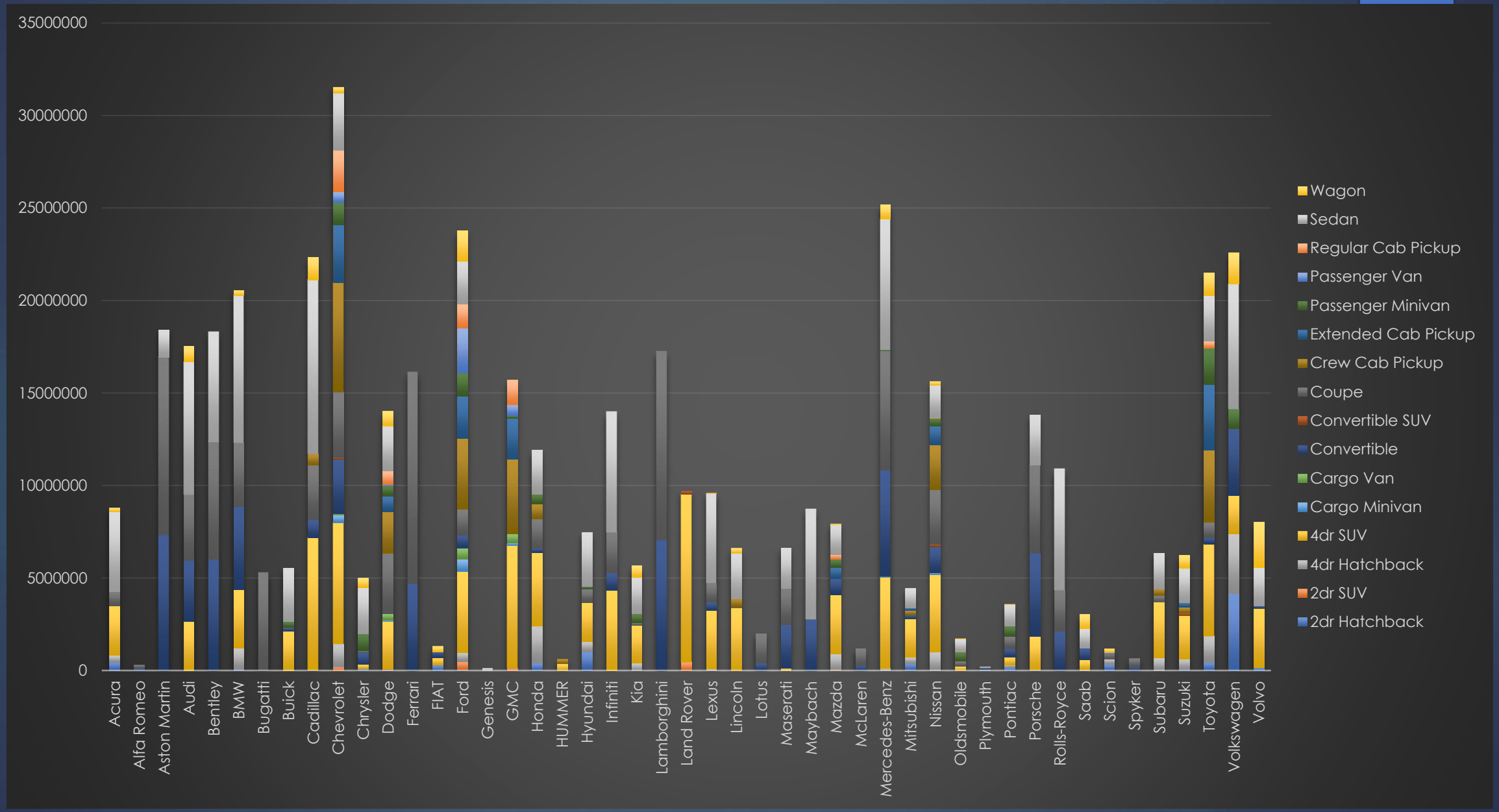
Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price.
Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.



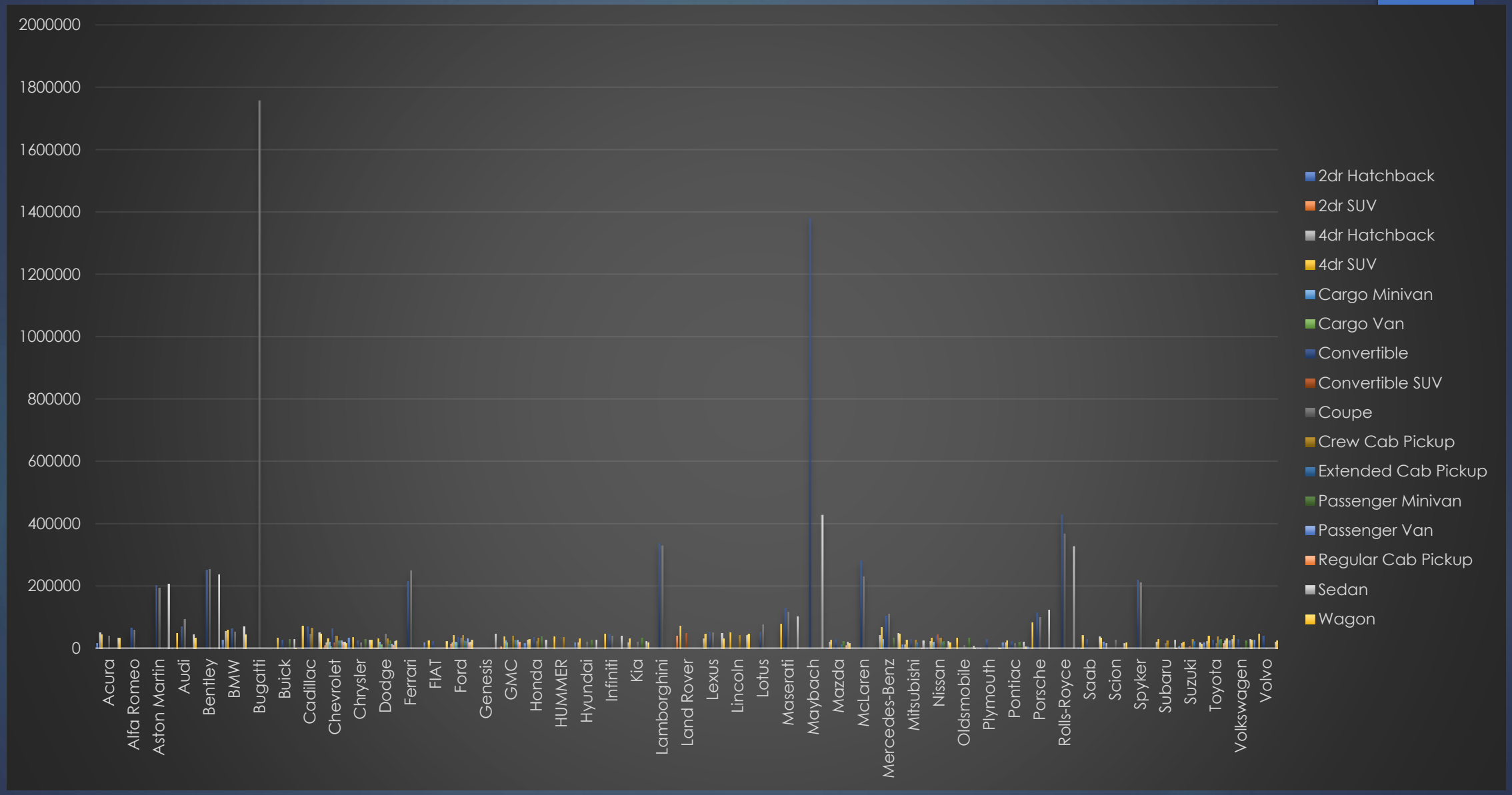
Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.
Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



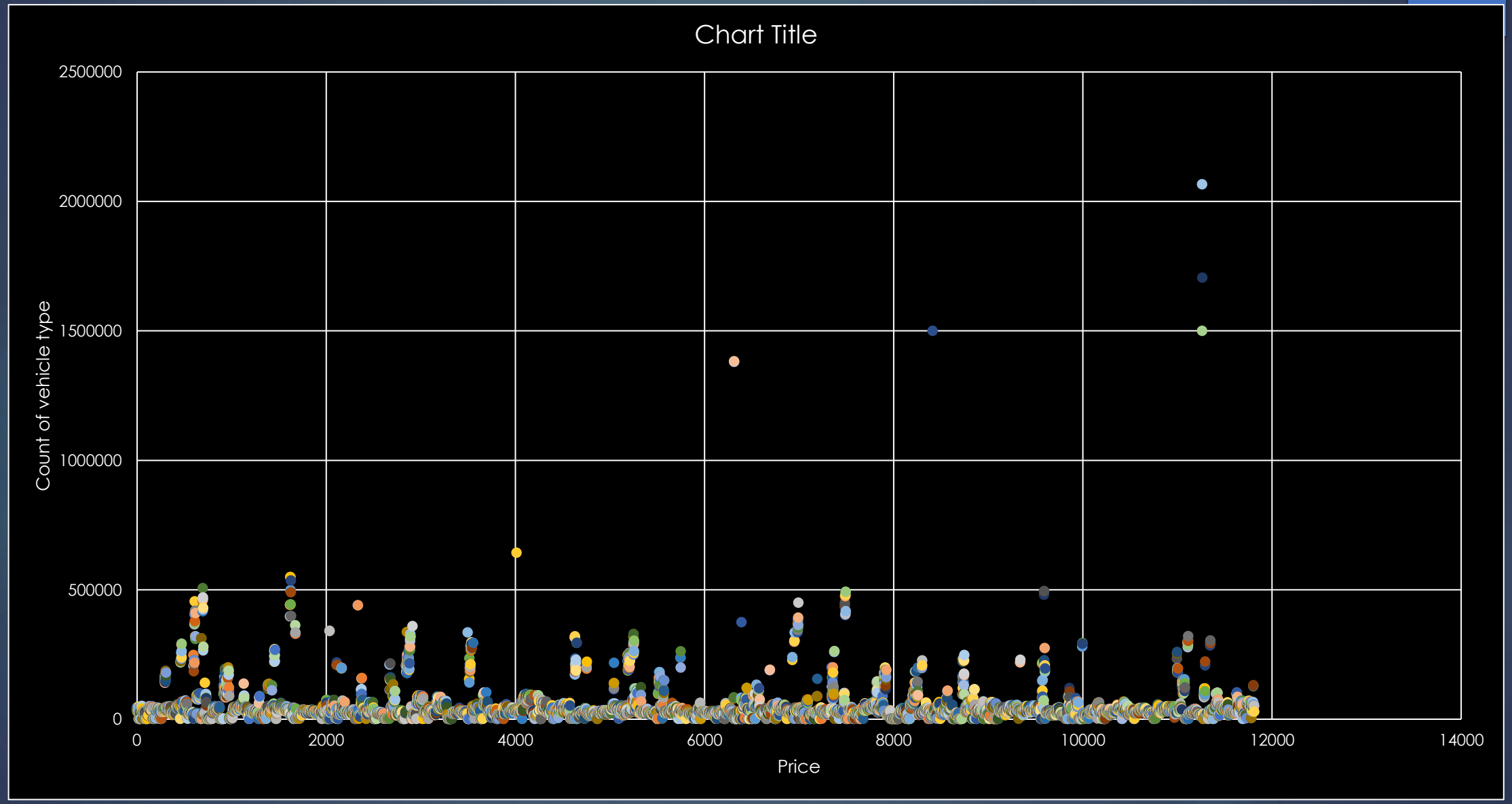
# Task 1: How does the distribution of car prices vary by brand and body style?



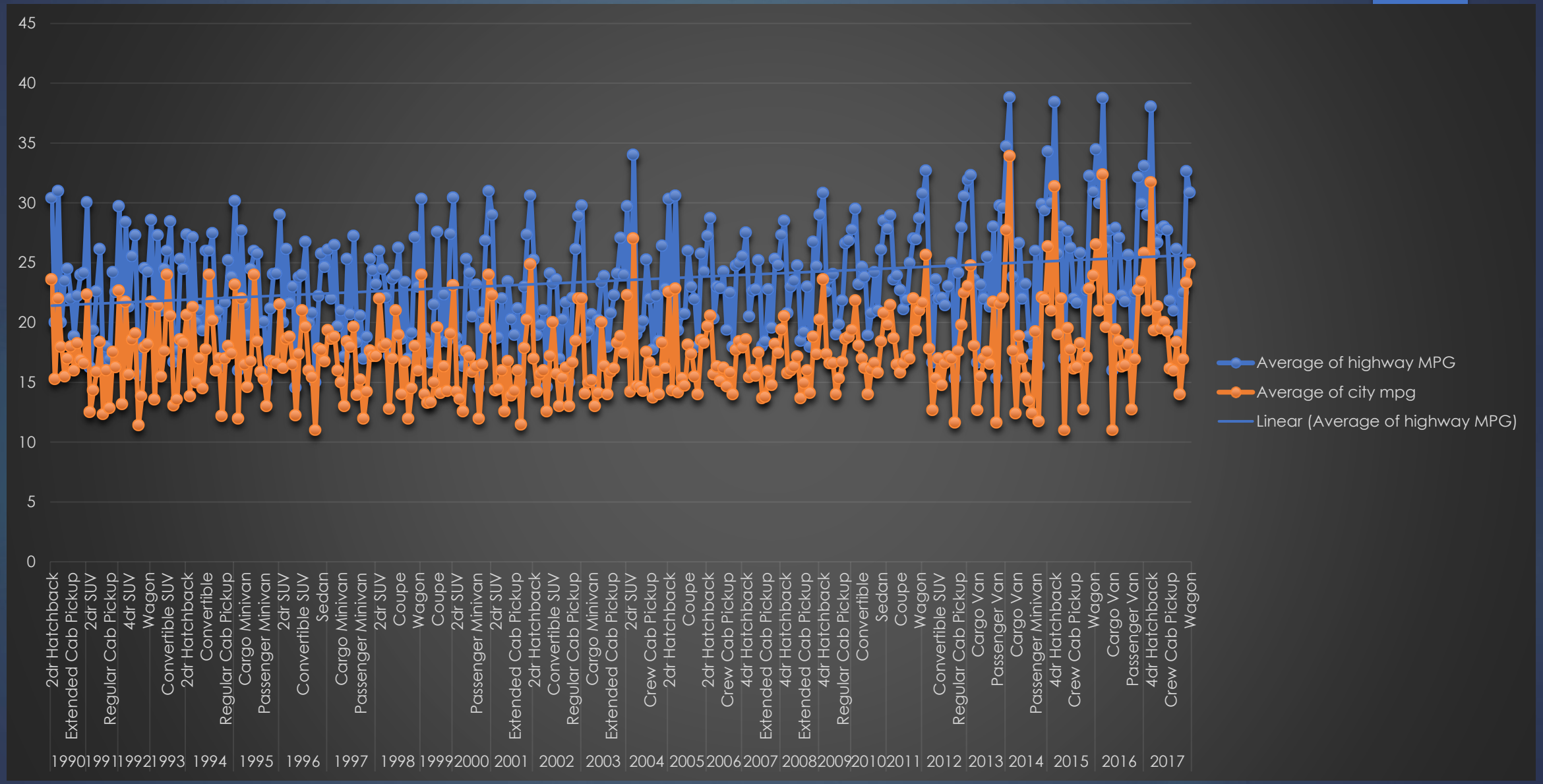
# Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?



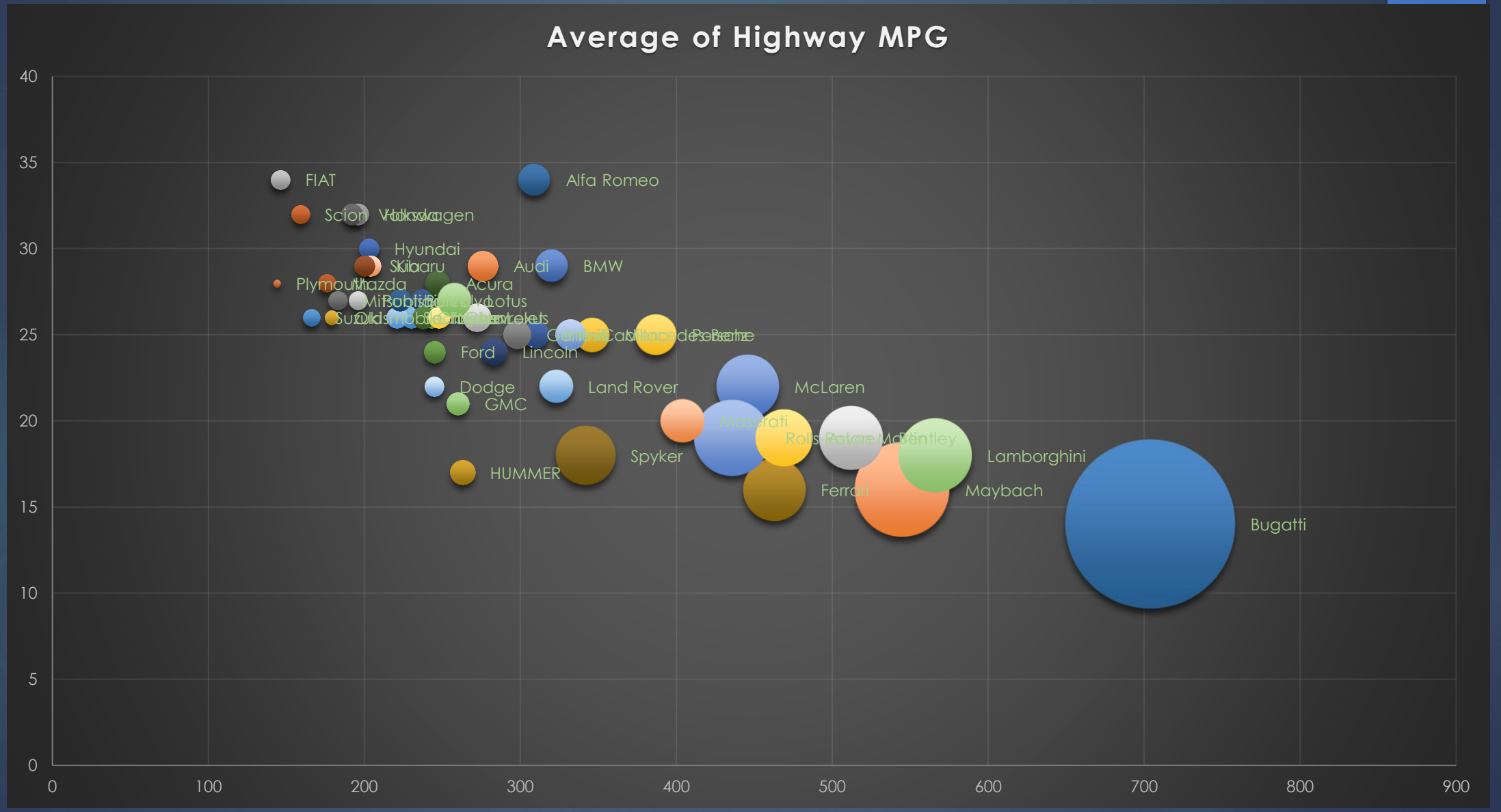
# Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?



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



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



# Result

- I have highlighted various texts in the visualisation with different colour so that it is easily visible
- I have created graphs wherever necessary I got to use the data labels built in function in which I can plot various car types with different colour
- But excel lags with such use data so in some queries I had to wait for a minute to get the changes affected
- The car with more engine cylinders has the most demand as compare to other features

# 3. Bank Loan Case Study

## Project Description

- First we should read the column description and determine whether to keep the column or not
- In this project I have cleaned the data i.e removing unnecessary columns from  the table
- I have also highlighted the column which I have removed in a separate sheet called column description
- We have to study the data to ensure that if we should approve the loan to people by reviewing their past application or by considering the factors such as their occupation type, housing type, members in the family and so on
- We also have a column where it is mention if the client has difficulty in payment issues or not

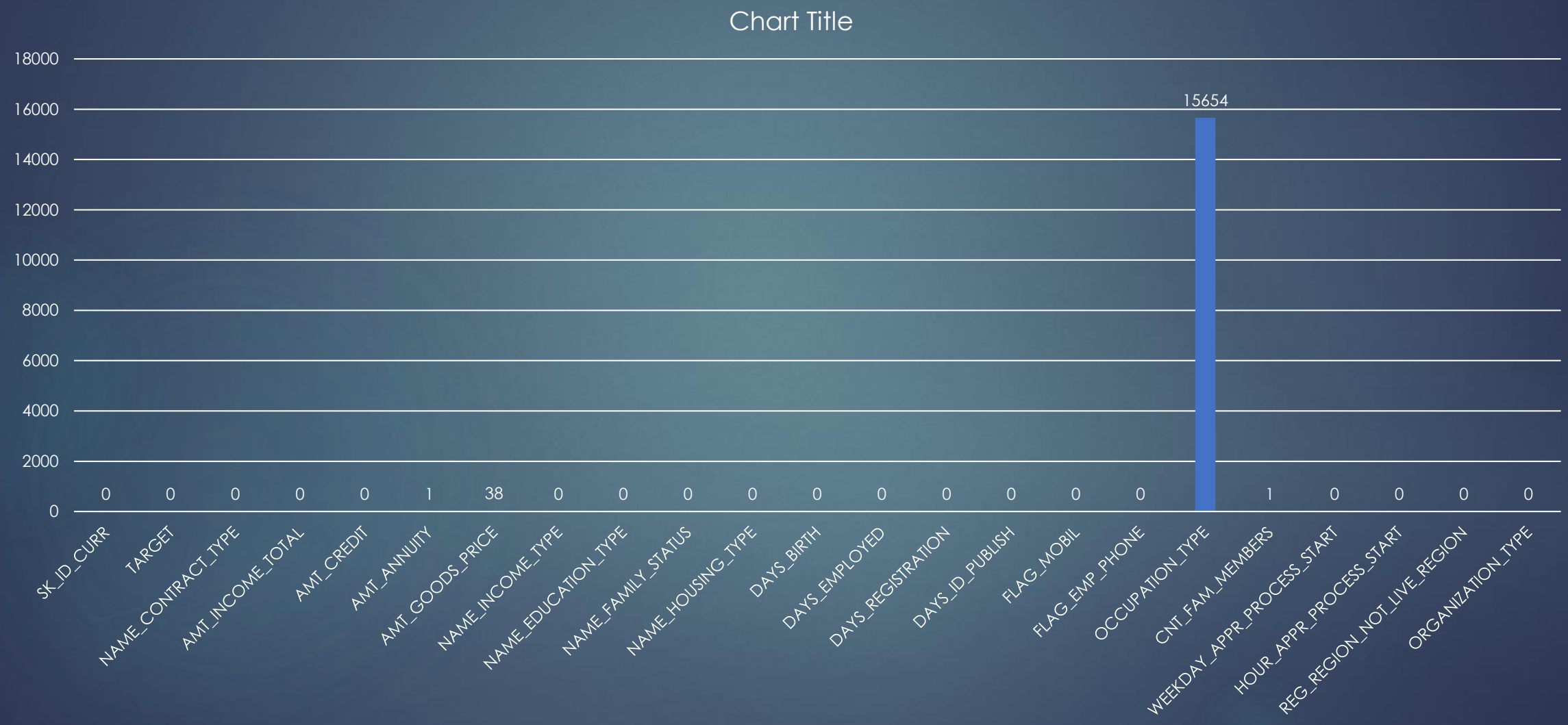
# Approach

- I have removed all the documents column since they had not details and they were irrelevant
- But in real life the documents column will be filled with actual documents provided by the costumer but here they were of no use, also the query didn't stated anything about the documents
- I have also cleaned the data and then used it to solve query
- After studying the data I have mostly used pivot table to solve most of the query since it have various built-in statistics function
- Plotting graph for most the result was must, so mainly I have used pie chart, bar graph, box and whiskers

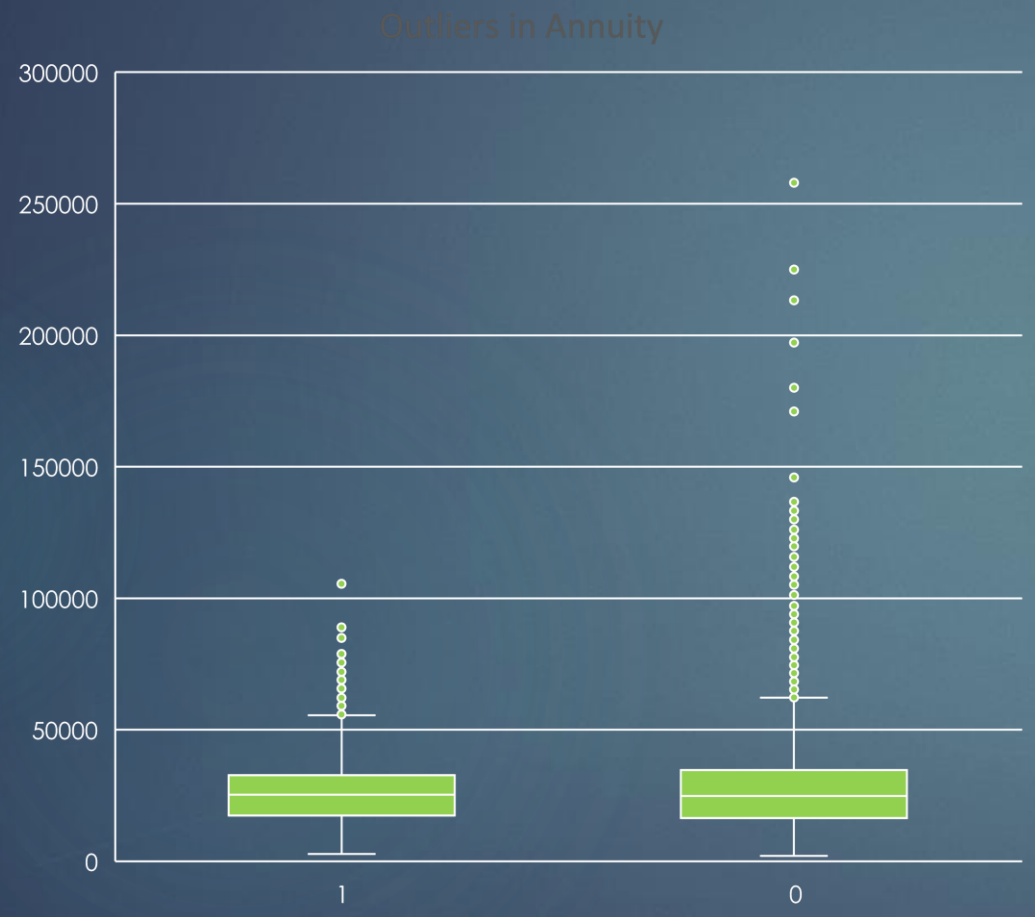
# Tech-Stack Used

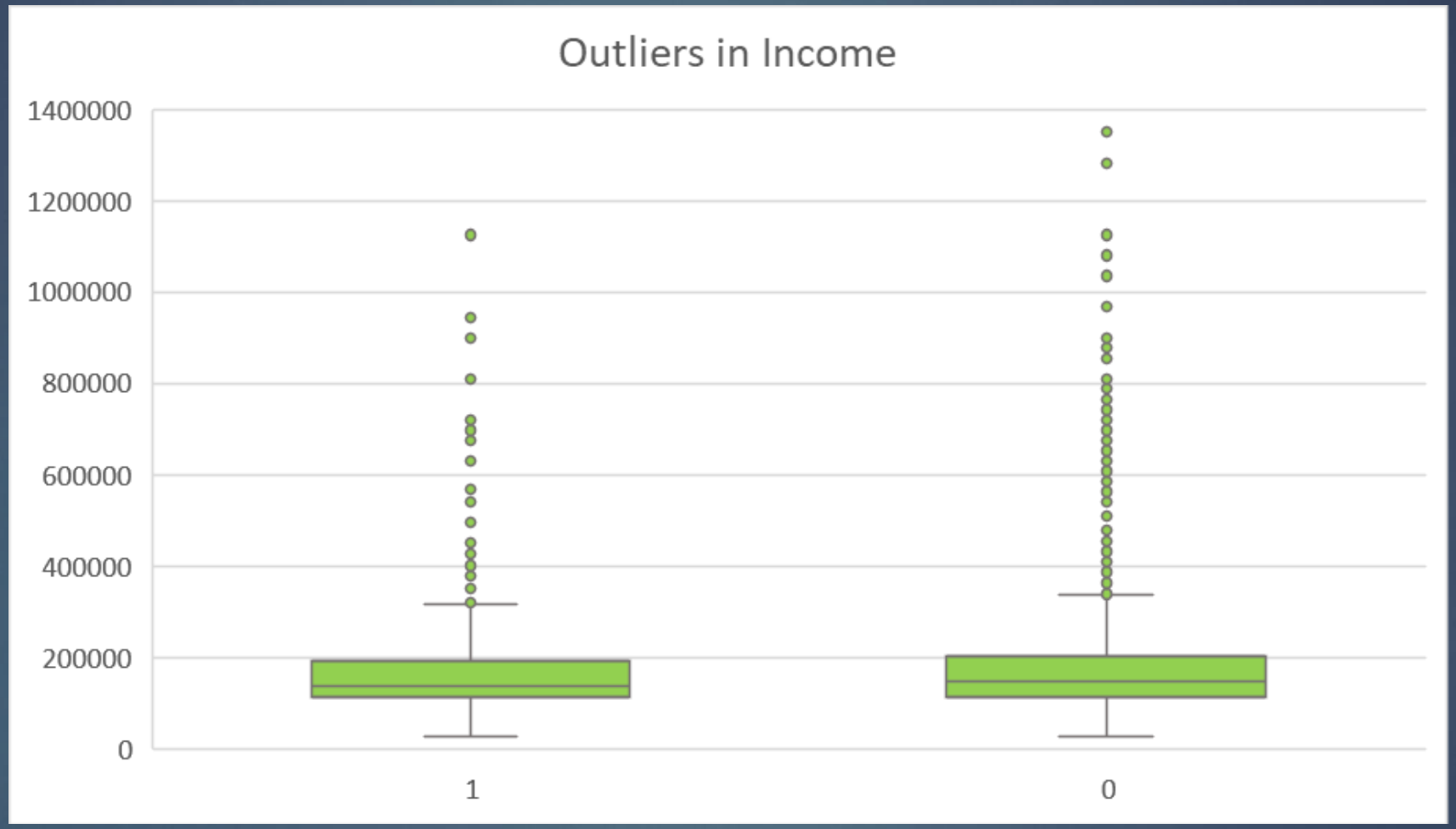
- Excel 2021 
- PowerPoint 2021 

# Task 1: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features



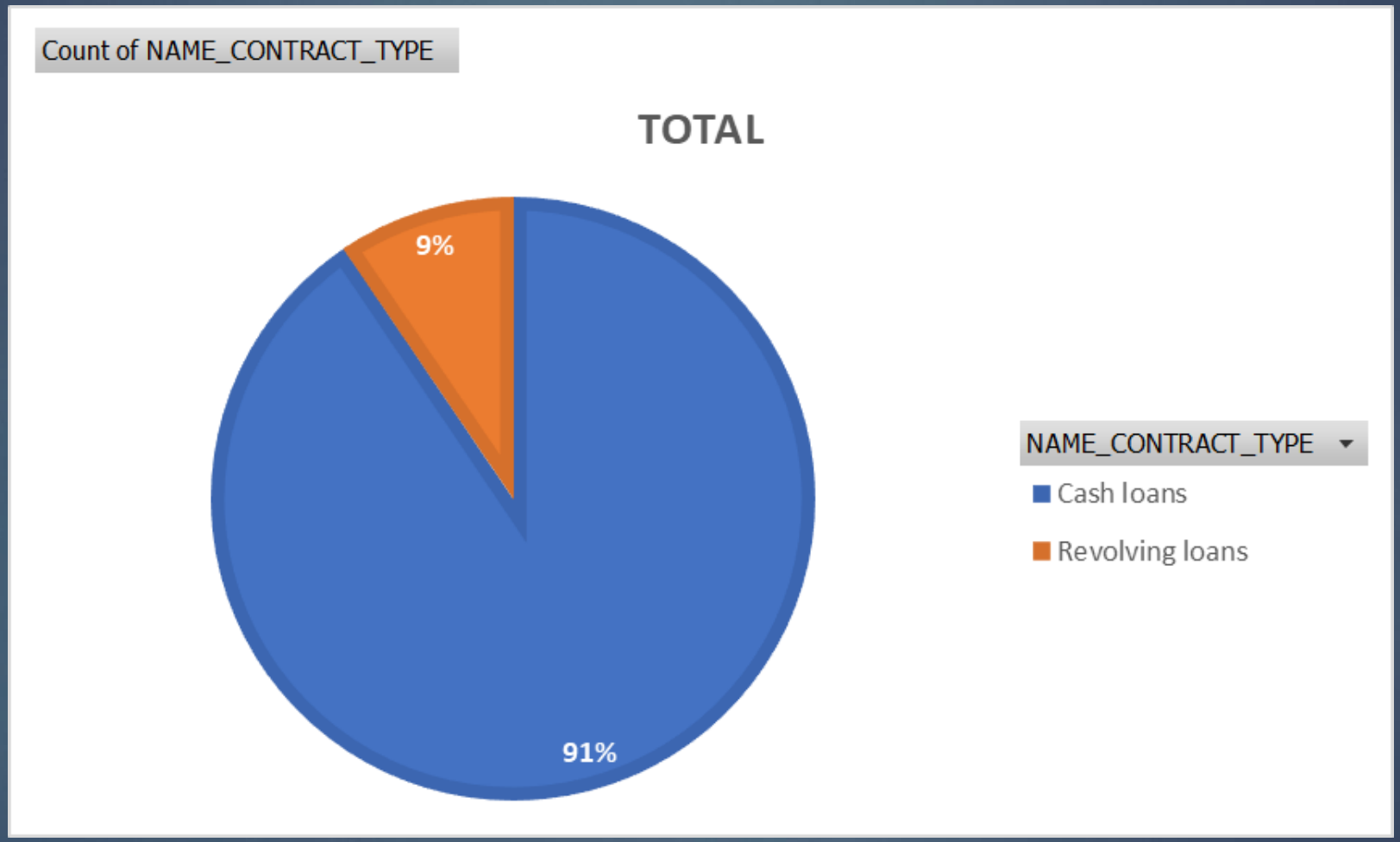
Task 2: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.



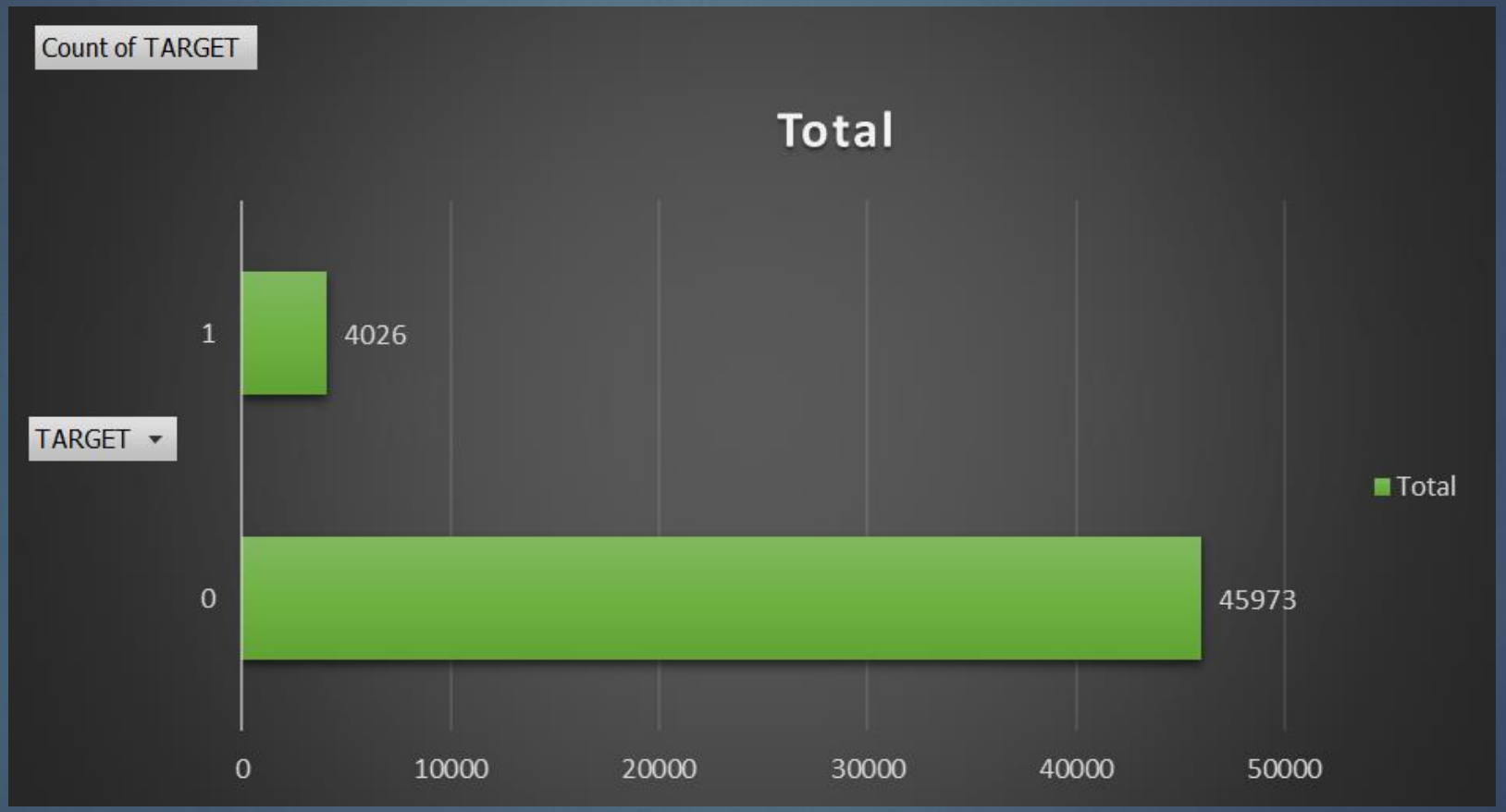


Outliers in Income
1400000
1200000
1000000
800000
600000
400000
200000
0
1
0


Outliers in Goods Price
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1
0

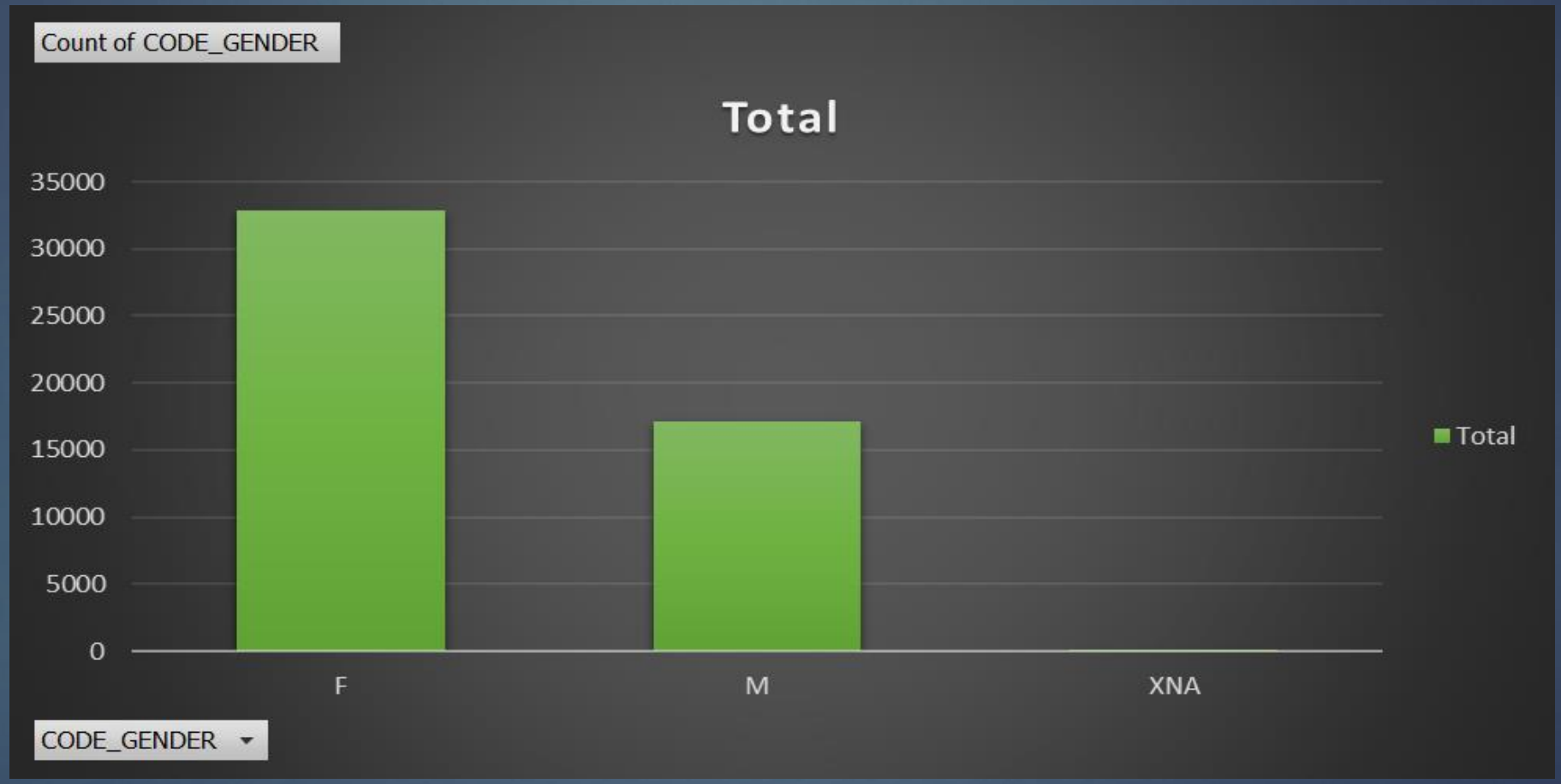
Task 3: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.



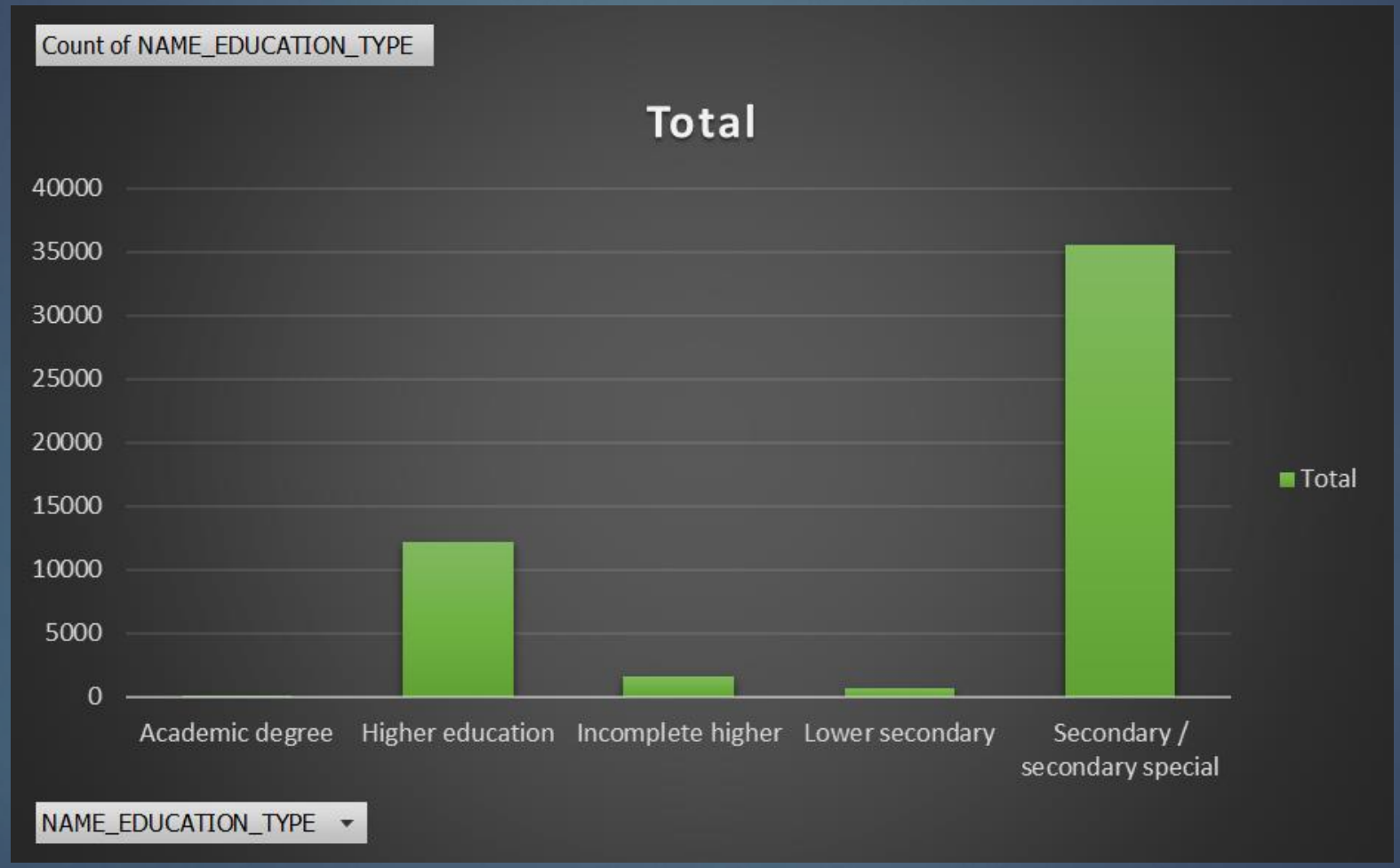
| Row Labels | Count of TARGET |
|---|---|
| 0 | 45973 |
| 1 | 4026 |
| **Grand Total** | **49999** |

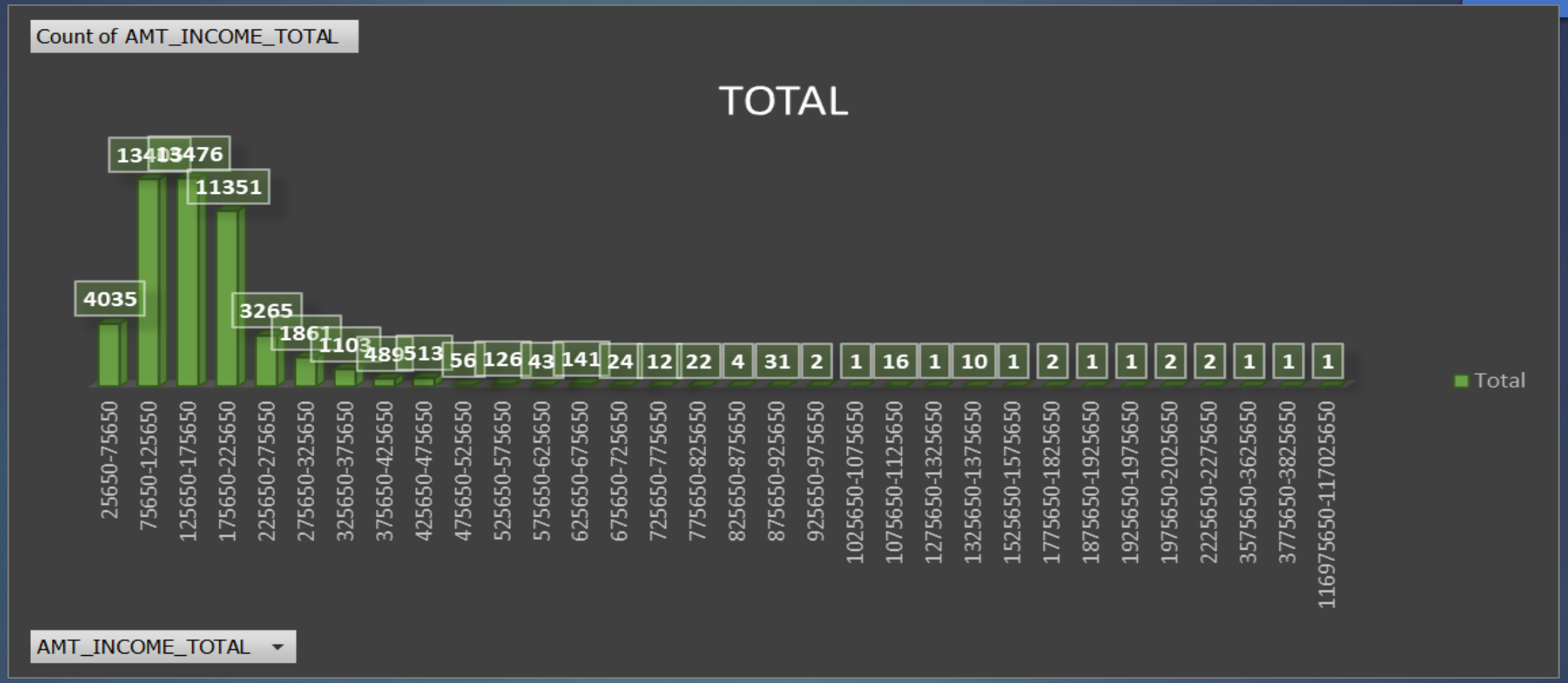


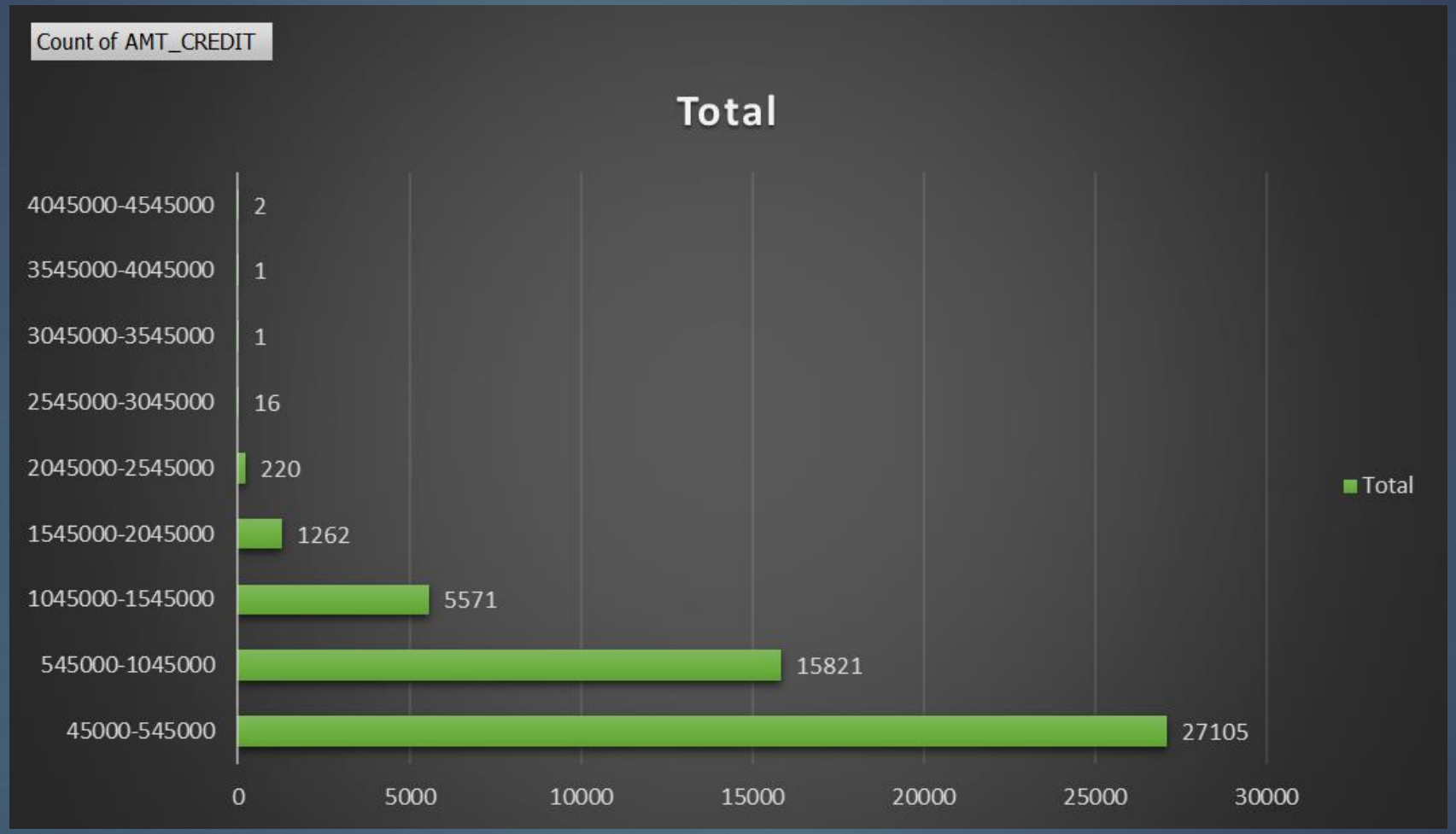
| Row Labels | Count of CODE_GENDER |
|---|---|
| F | 32823 |
| M | 17174 |
| XNA | 2 |
| **Grand Total** | **49999** |

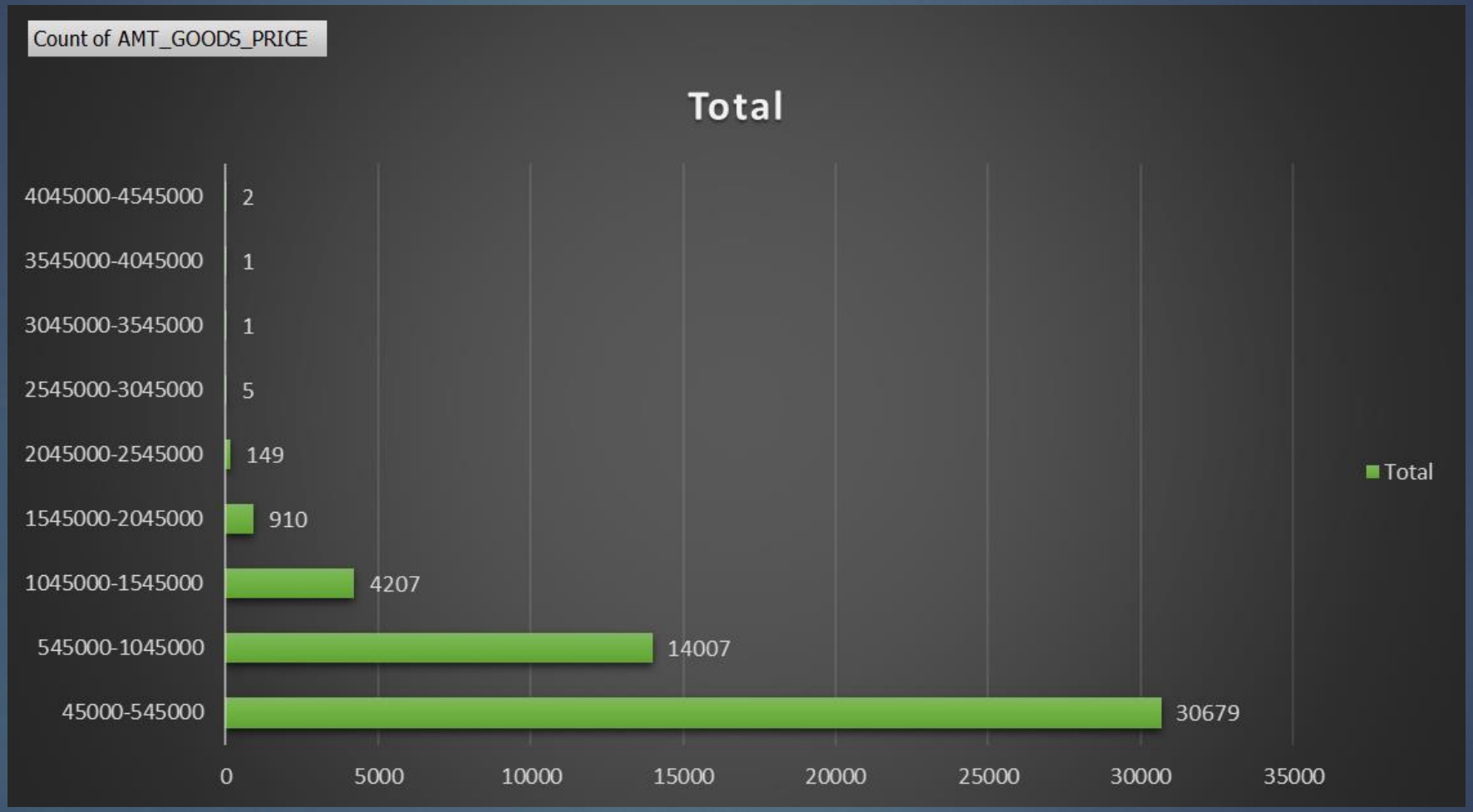


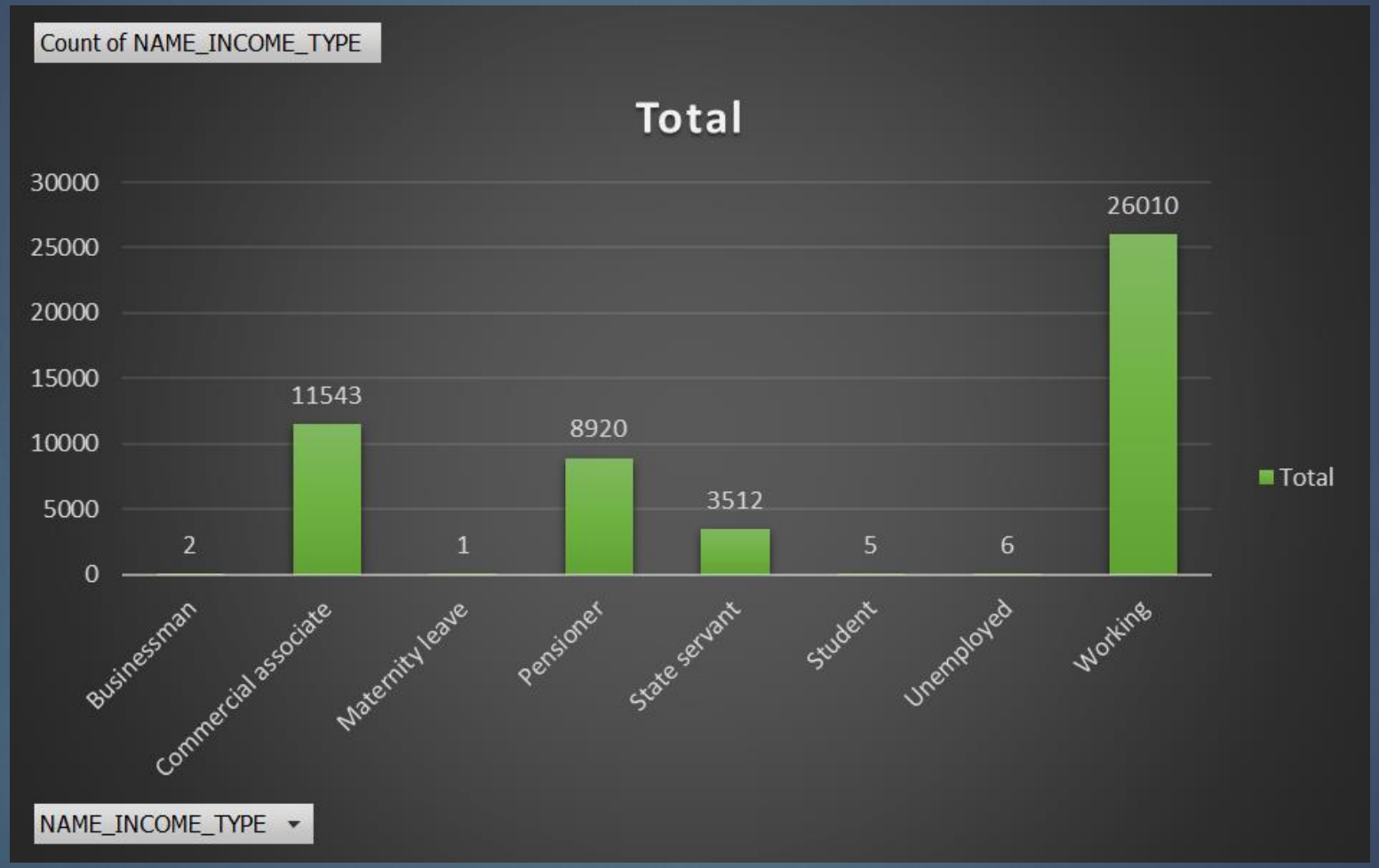
Task 4: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

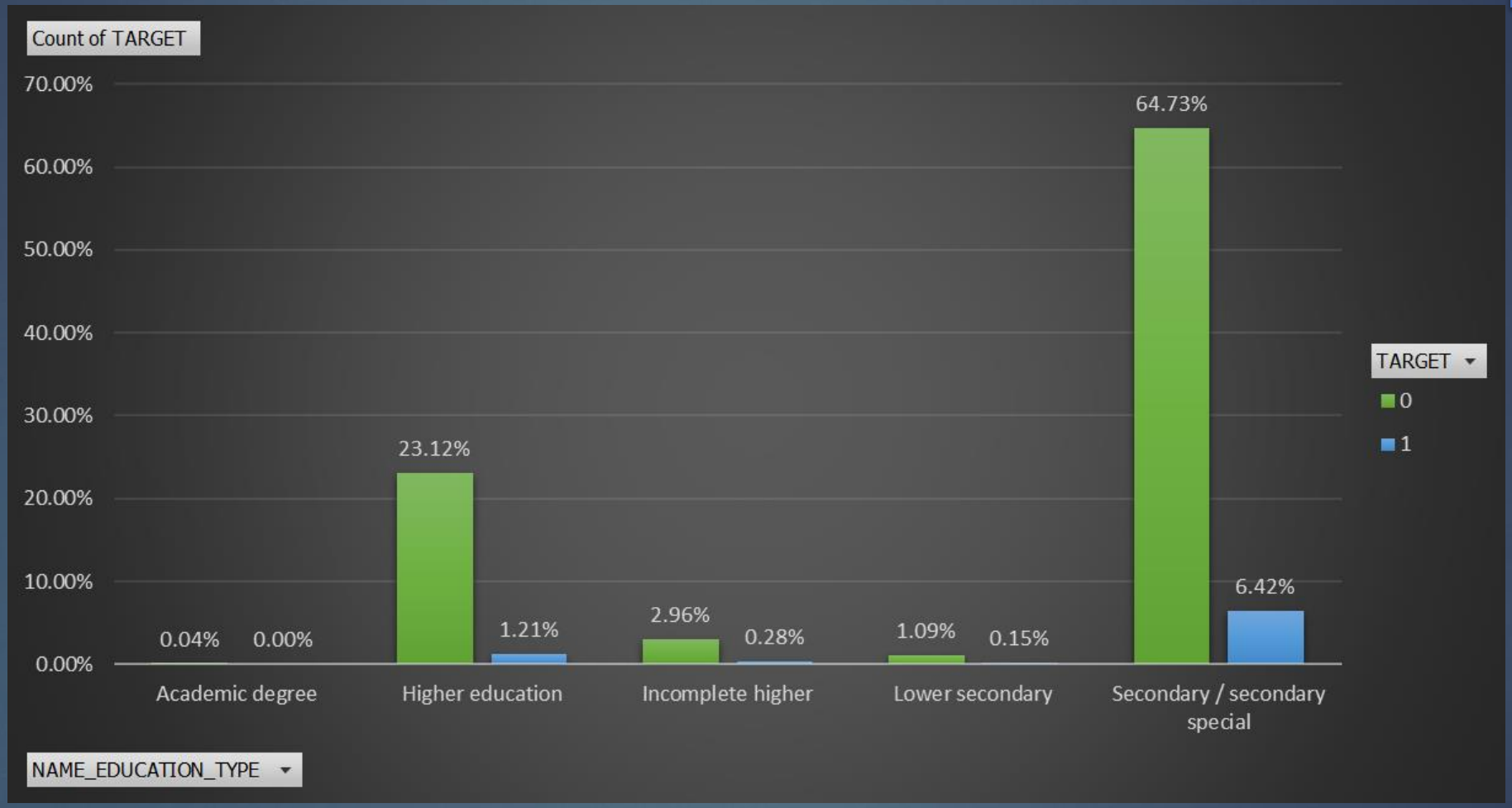


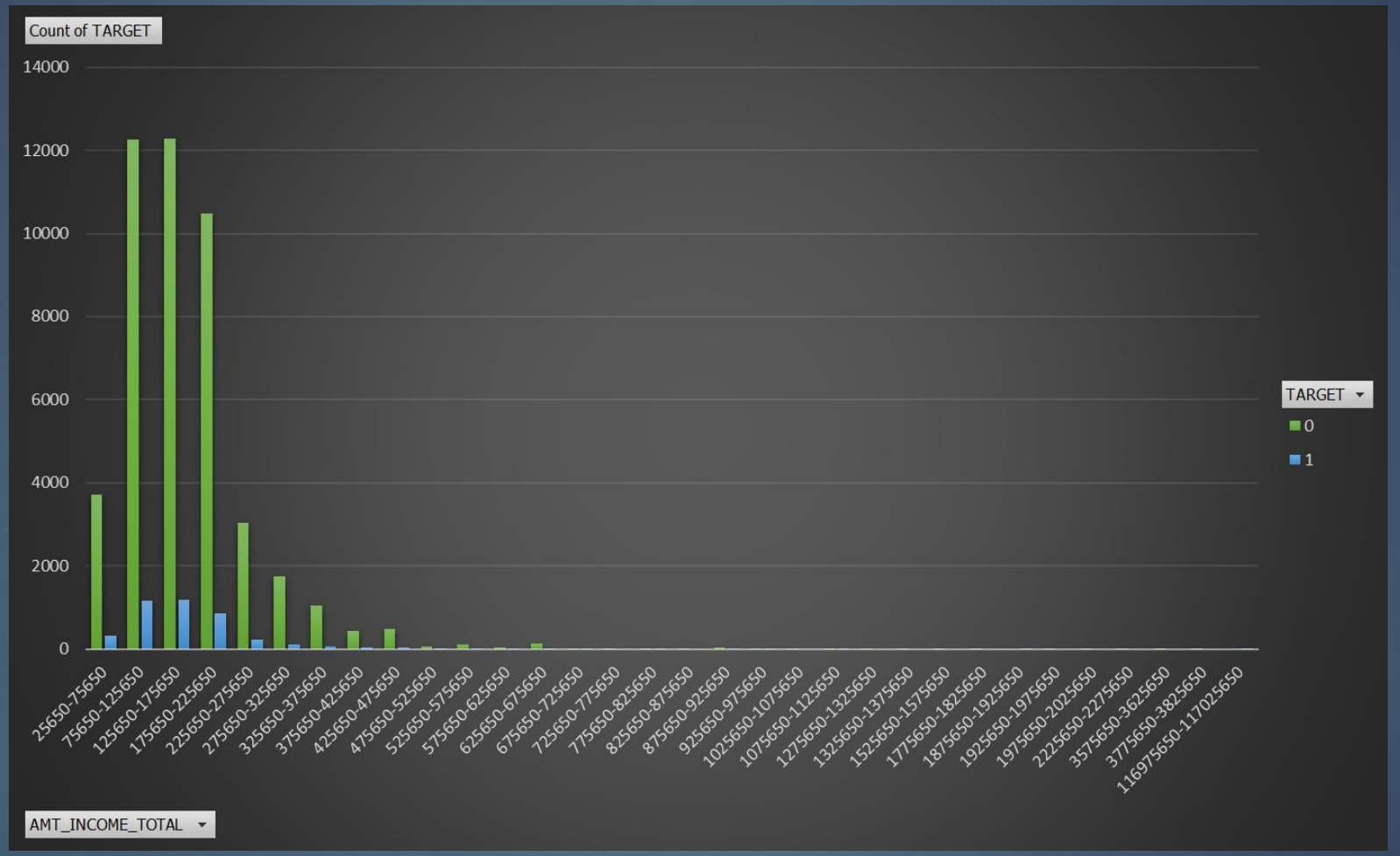

Count of AMT_INCOME_TOTAL
TOTAL
13476
11351
4035
3265
1861
1103
489
513
56
126
43
141
24
12
22
4
31
2
1
16
1
10
1
2
1
1
2
2
1
1
1
25650-75650
75650-125650
125650-175650
175650-225650
225650-275650
275650-325650
325650-375650
375650-425650
425650-475650
475650-525650
525650-575650
575650-625650
625650-675650
675650-725650
725650-775650
775650-825650
825650-875650
875650-925650
925650-975650
1025650-1075650
1075650-1125650
1275650-1325650
1325650-1375650
1525650-1575650
1775650-1825650
1875650-1925650
1925650-1975650
1975650-2025650
2225650-2275650
3575650-3625650
3775650-3825650
11697565O-11702565O
Total
AMT_INCOME_TOTAL

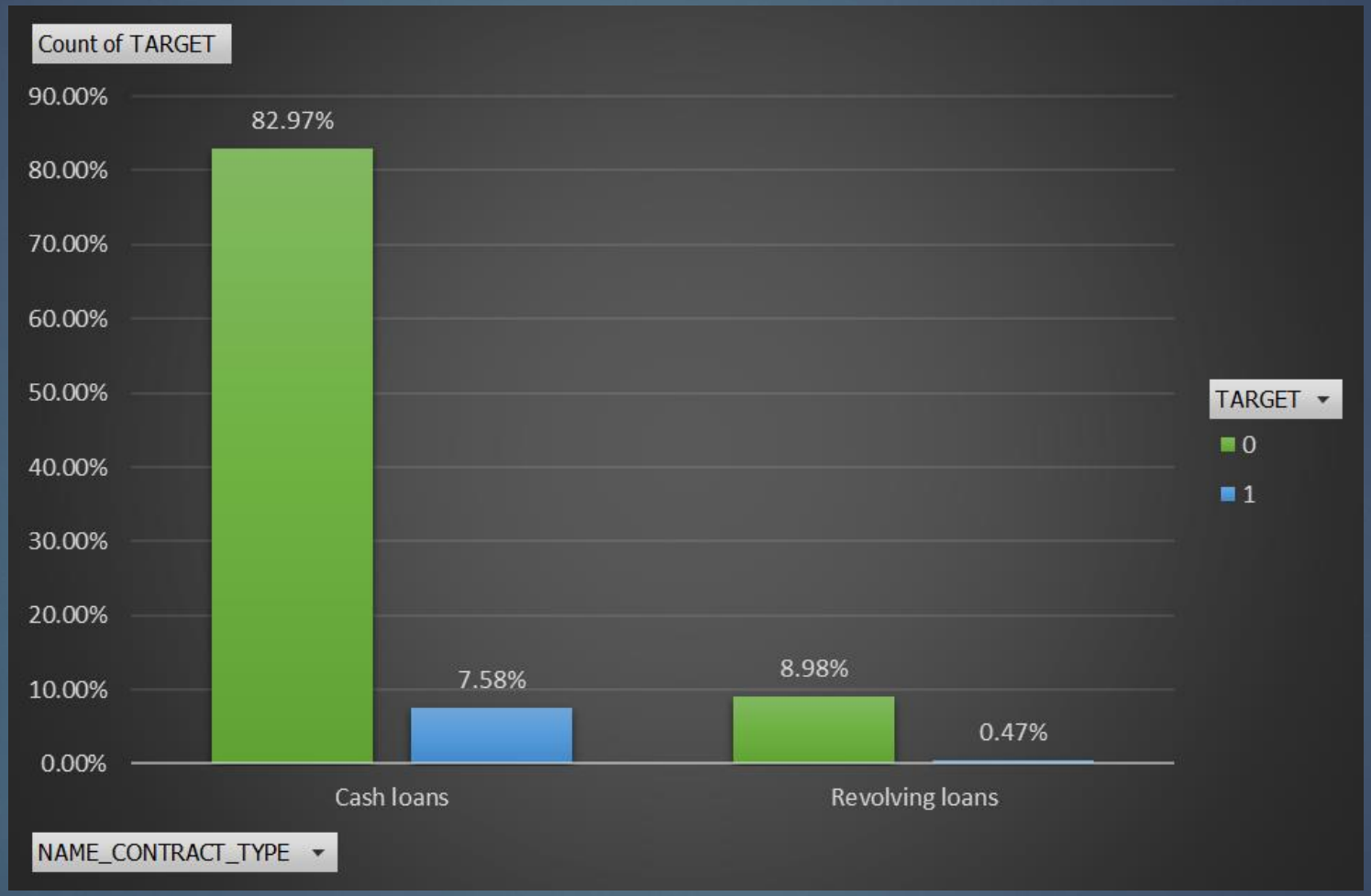

Count of AMT_CREDIT
Total
4045000-4545000
3545000-4045000
3045000-3545000
2545000-3045000
2045000-2545000
1545000-2045000
1045000-1545000
545000-1045000
45000-545000
2
1
1
16
220
1262
5571
15821
27105
0
5000
10000
15000
20000
25000
30000
Total

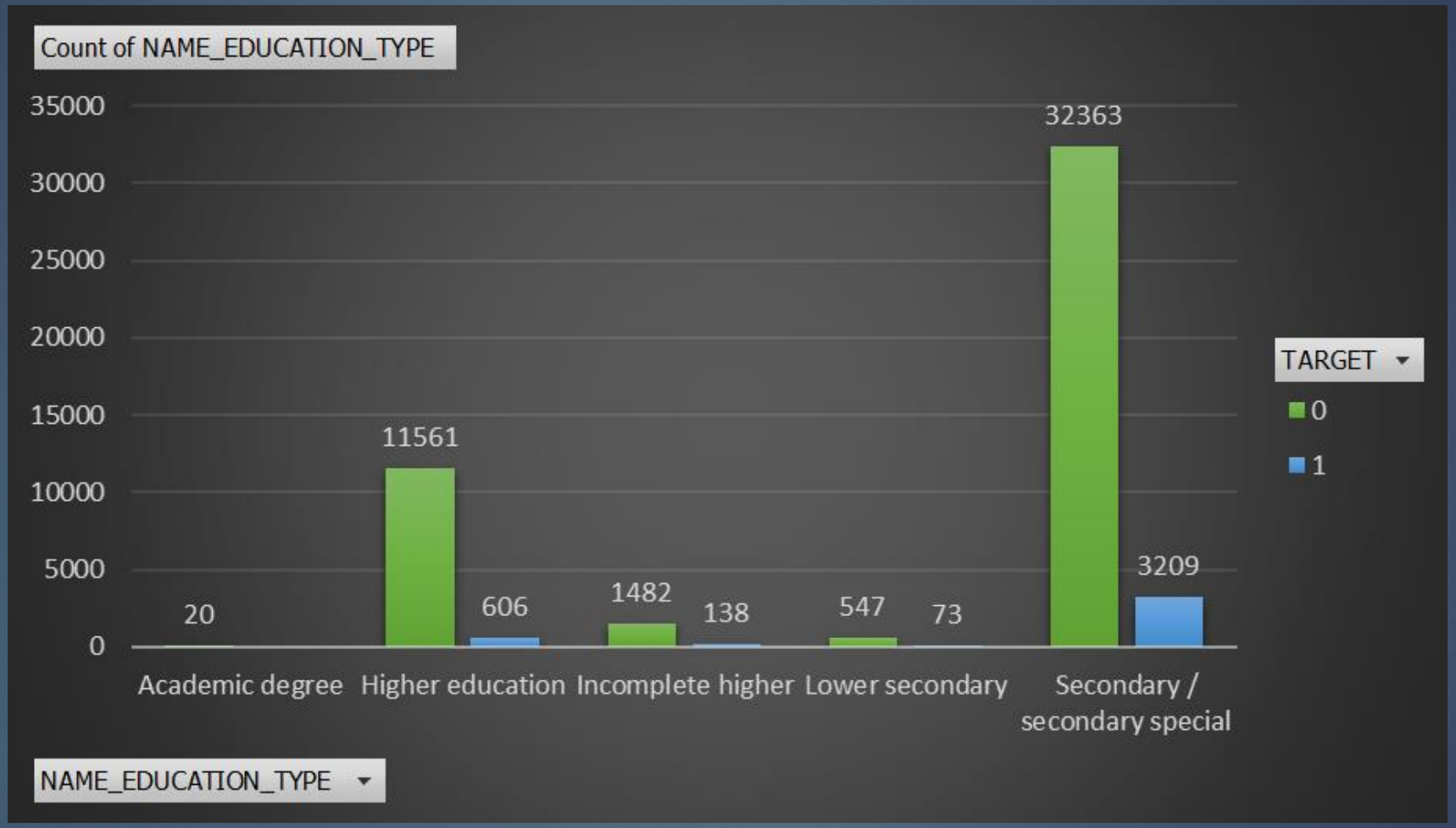

Count of AMT_GOODS_PRICE
Total
4045000-4545000
3545000-4045000
3045000-3545000
2545000-3045000
2045000-2545000
1545000-2045000
1045000-1545000
545000-1045000
45000-545000
2
1
1
5
149
910
4207
14007
30679
0
5000
10000
15000
20000
25000
30000
35000
Total

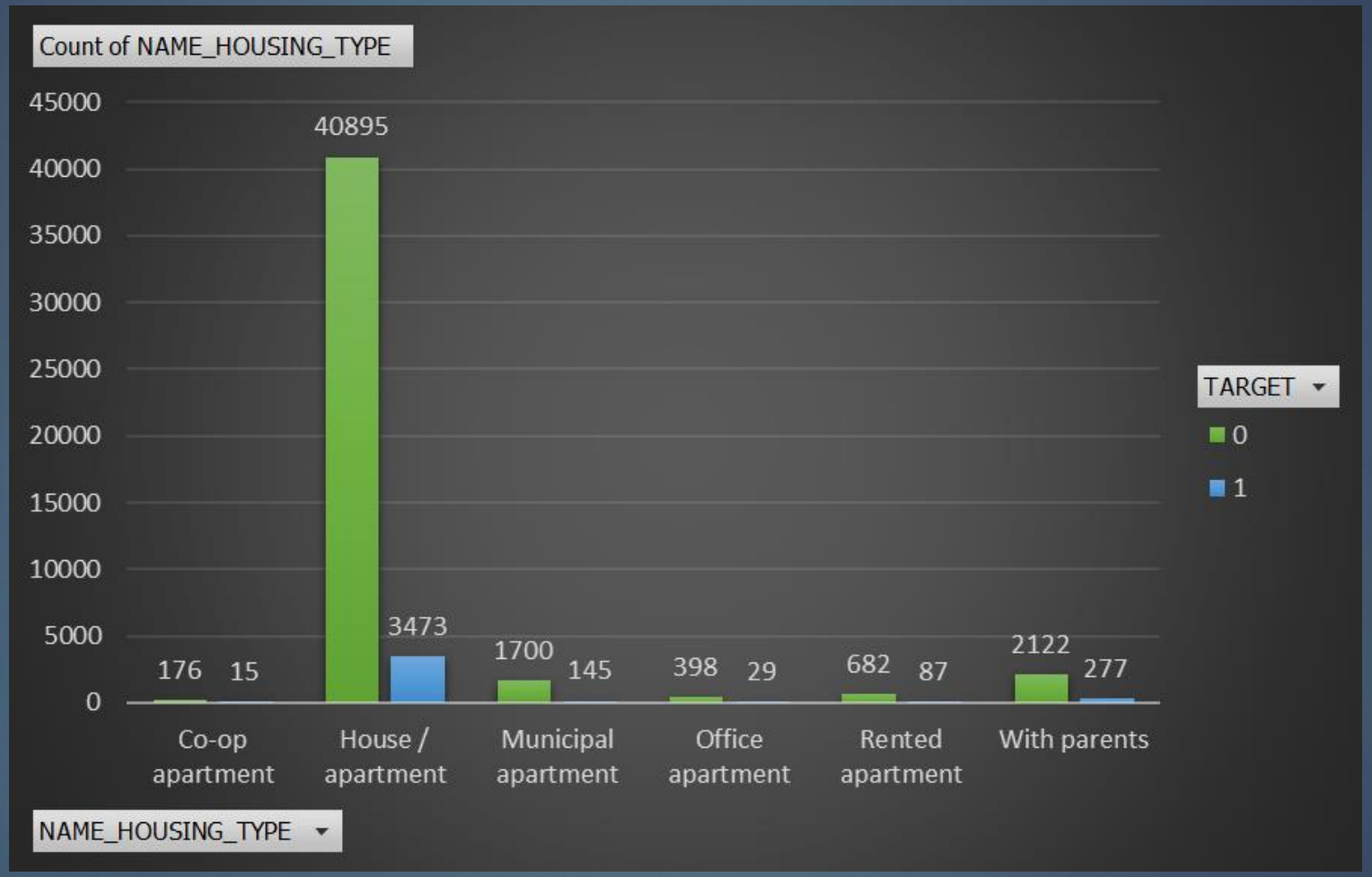

Count of NAME_INCOME_TYPE
Total
30000
25000
20000
15000
10000
5000
0
2
11543
1
8920
3512
5
6
26010
Total
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
NAME_INCOME_TYPE

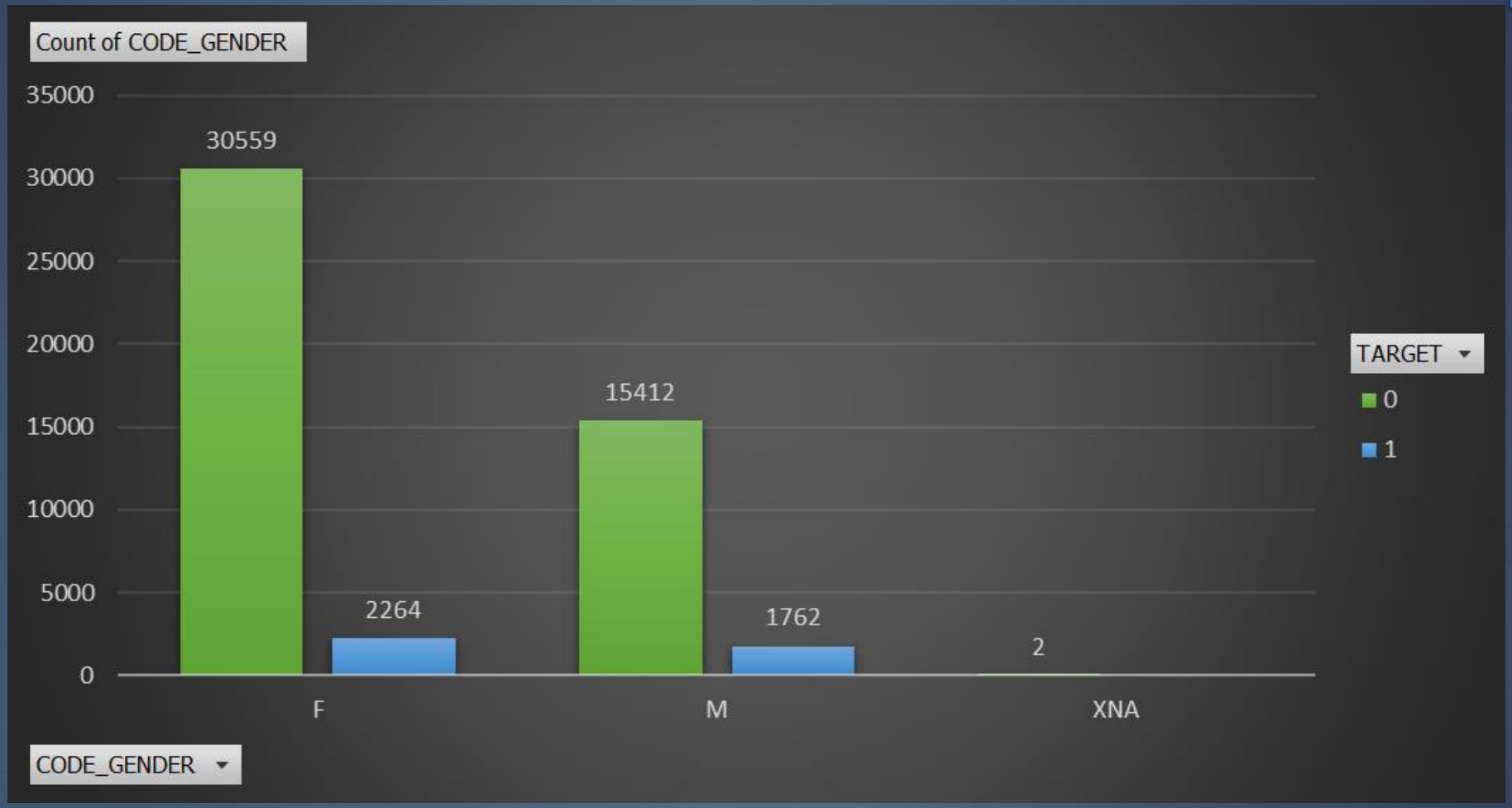

Count of TARGET
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0.04%
0.00%
23.12%
1.21%
2.96%
0.28%
1.09%
0.15%
64.73%
6.42%
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE
TARGET
0
1


Count of TARGET
14000
12000
10000
8000
6000
4000
2000
0
25650-75650
75650-125650
125650-175650
175650-225650
225650-275650
275650-325650
325650-375650
375650-425650
425650-475650
475650-525650
525650-575650
575650-625650
625650-675650
675650-725650
725650-775650
775650-825650
825650-875650
875650-925650
925650-975650
1025650-1075650
1075650-1125650
1275650-1325650
1325650-1375650
1525650-1575650
1775650-1825650
1875650-1925650
1925650-1975650
1975650-2025650
2225650-2275650
3575650-3625650
3775650-3825650
116975650-117025650
TARGET
0
1
AMT_INCOME_TOTAL


Count of TARGET
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
82.97%
7.58%
8.98%
0.47%
Cash loans
Revolving loans
TARGET
0
1
NAME_CONTRACT_TYPE


Count of NAME_EDUCATION_TYPE
35000
30000
25000
20000
15000
10000
5000
0
20
11561
606
1482
138
547
73
32363
3209
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
TARGET
0
1
NAME_EDUCATION_TYPE


Count of NAME_HOUSING_TYPE
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
40895
3473
1700
145
398
29
682
87
2122
277
176
15
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
TARGET
0
1
NAME_HOUSING_TYPE


Count of CODE_GENDER
35000
30000
25000
20000
15000
10000
5000
0
30559
2264
15412
1762
2
F
M
XNA
TARGET
0
1
CODE_GENDER

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

| CORELATION | AMT_INCOME | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | DAYS_EMPLOYED | CNT_FAM_MEMBERS |
|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | | | | | |
| AMT_CREDIT | 0.069315897 | 1 | | | | |
| AMT_ANNUITY | 0.083008508 | 0.769498914 | 1 | | | |
| AMT_GOODS_PRICE | 0.069885575 | 0.98694373 | 0.774433947 | 1 | | |
| DAYS_EMPLOYED | -0.03161555 | -0.070471393 | -0.110449038 | -0.067845372 | 1 | |
| CNT_FAM_MEMBERS | 0.011227377 | 0.06399785 | 0.077378847 | 0.061624351 | -0.229818862 | 1 |

Correlation for values with Target 0

| CORRELATION | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | DAYS_EMPLOYED | CNT_FAM_MEMBERS |
|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | | | | | |
| AMT_CREDIT | 0.015271444 | 1 | | | | |
| AMT_ANNUITY | 0.018004594 | 0.749665201 | 1 | | | |
| AMT_GOODS_PRICE | 0.013266279 | 0.982432318 | 0.749705184 | 1 | | |
| DAYS_EMPLOYED | -0.011555963 | 0.016039571 | -0.079556008 | 0.020213912 | 1 | |
| CNT_FAM_MEMBERS | 0.013121678 | 0.06124869 | 0.075838463 | 0.055103609 | -0.183560113 | 1 |

Correlation for values with Target 1

# Result

- By determining the outliers we can blacklist the costumers who fall in it since they might not be paying the loan on time despite for their income
- We can also see that there are more female costumers who has applied for loan then male customers
- We can also consider correlation to check if they have a strong relationship or not so we  can approve the loan to them by considering these factor as well as where they live and is the area well developed or not
- Most people who require loan are working
- I got to use pivot table a lot since it has lot of built in statistics function which are really helpful in such condition
- I got to use excel at its full potential and also explore various function which were not known to me

# 4. IMDB Movie Analysis

## Project Description

- In this project we have been given various movies data from IMDB website
- We have various column such as movie title, director name, imdb_score for the movie and so on
- We can use all of this data to study the success of movie and also likes and dislikes of the people
- Studying the likes and dislikes the director can make various changes to his/her new movie so that people will have a positive review
- We also have budget and gross income of the movie through which we can easily calculate the profit gained

# Approach

- First step was to clean data i.e removing empty cells, duplicate cells, unnecessary column such as actor likes, color and so on
- Then using pivot table to solve query as it help very effectively
- Using various excel function and even combining them to solve certain query
- Using average, median, mode function for descriptive statistics
- Analysing the relation between movie budgets and gross earnings but subtracting them to get profit and then choosing the highest profit movie

# Tech-Stack Used

- Excel 2021
- PowerPoint 2021

Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Solution: The most common genre are comedy, drama, romance, crime, action, thriller, adventure, science fiction. Some of these genre are together for many movies such as for comedy, drama, romance we have 145 movies which is highest of all

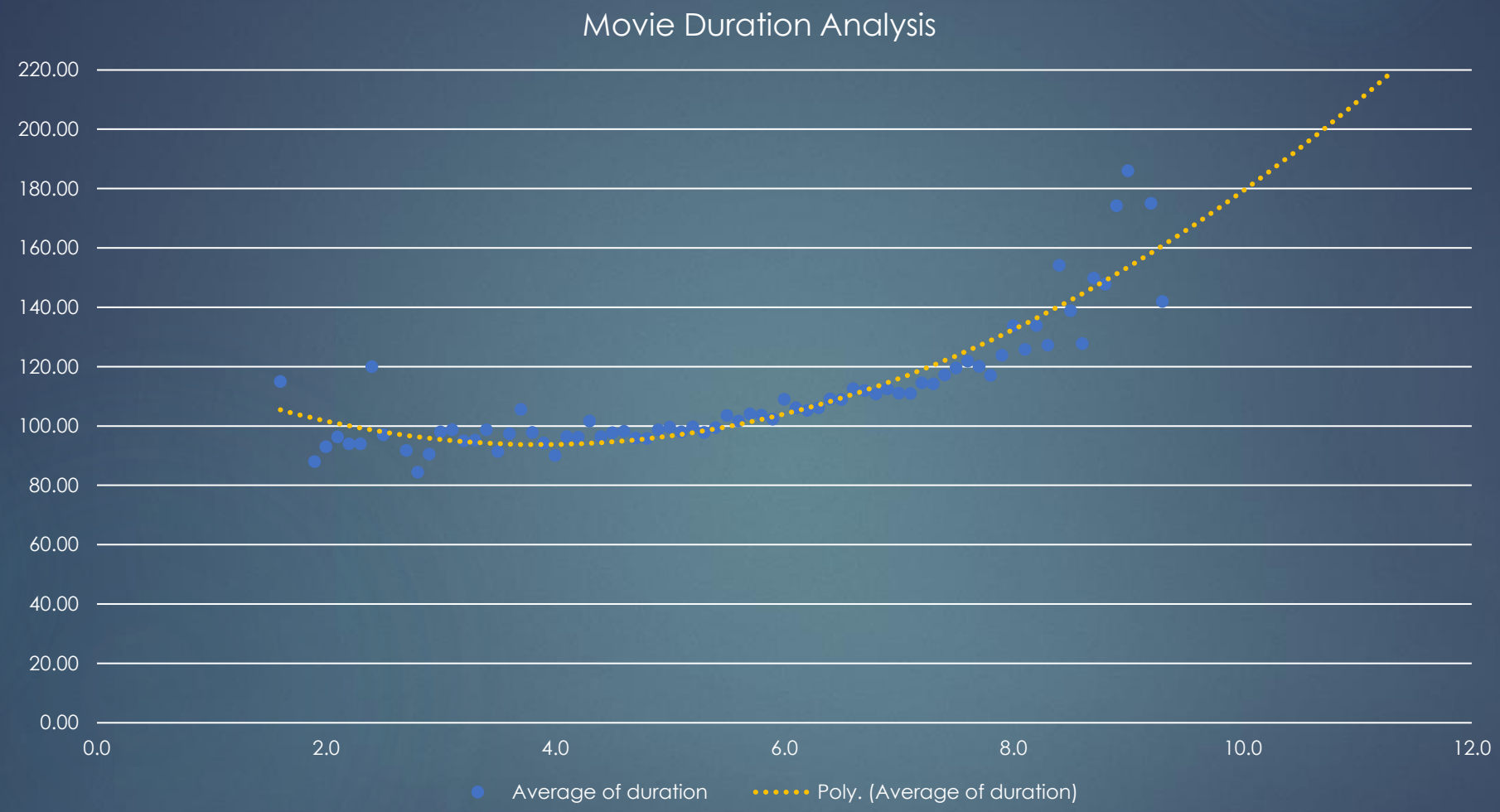
| Row Labels | Count of movie_title | Average of imdb_score | Max of imdb_score | Min of imdb_score |
|---|---|---|---|---|
| Comedy\|Drama\|Romance | 145 | 6.49 | 8 | 4.3 |
| Drama | 139 | 7.08 | 8.8 | 3.4 |
| Comedy\|Drama | 137 | 6.59 | 8.8 | 3.3 |
| Comedy | 137 | 5.84 | 8 | 1.9 |
| Comedy\|Romance | 130 | 5.94 | 8.4 | 2.7 |
| Drama\|Romance | 113 | 6.96 | 8.1 | 4.1 |
| Crime\|Drama\|Thriller | 79 | 6.85 | 8.5 | 5.1 |
| Action\|Crime\|Thriller | 54 | 6.40 | 7.6 | 4.4 |
| Action\|Crime\|Drama\|Thriller | 48 | 6.52 | 9 | 5.1 |
| Comedy\|Crime | 45 | 6.04 | 8.3 | 3.1 |
| Action\|Adventure\|Sci-Fi | 45 | 6.67 | 8.4 | 2.4 |
| Action\|Adventure\|Thriller | 43 | 6.70 | 8 | 4.4 |
| Horror | 41 | 5.93 | 8 | 4.2 |
| Crime\|Drama\|Mystery\|Thriller | 40 | 6.92 | 8.6 | 5.6 |
| Drama\|Thriller | 40 | 6.63 | 8.5 | 3.9 |

| Var of imdb_score | StdDev of imdb_score | Range of imdb_score | Median of imdb_score | Mode of imdb_score |
|---|---|---|---|---|
| 0.57 | 0.75 | 3.7 | 6.5 | 6.5 |
| 0.69 | 0.83 | 5.4 | 7.2 | 7.3 |
| 0.69 | 0.83 | 5.5 | 6.7 | 6.7 |
| 1.50 | 1.22 | 6.1 | 6 | 6.2 |
| 0.70 | 0.84 | 5.7 | 6 | 6.1 |
| 0.55 | 0.74 | 4 | 7.1 | 7.2 |
| 0.61 | 0.78 | 3.4 | 7 | 6.1 |
| 0.41 | 0.64 | 3.2 | 6.5 | 6.5 |
| 0.52 | 0.72 | 3.9 | 6.5 | 6.5 |
| 1.39 | 1.18 | 5.2 | 6.1 | 6.7 |
| 1.47 | 1.21 | 6 | 6.8 | 6.6 |
| 0.56 | 0.75 | 3.6 | 6.7 | 6.8 |
| 0.64 | 0.80 | 3.8 | 5.9 | 5.9 |
| 0.61 | 0.78 | 3 | 6.65 | 6.6 |
| 0.78 | 0.88 | 4.6 | 6.65 | 6.8 |

Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Solution:

1. In this analysis I have calculated mean/average, standard deviation, median duration of all movies on the basis of imdb_score and created a scattered chart on this basis.
2. In standard deviation we get #DIV/0! Error as there are less than 1 movie which have imdb_score as 1.6, 2.0, 2.5, 3.2, 9.2, 9.3 due to which we get this error
3. In this chart I have only taken average duration which help us to study how imdb_score is affected by the duration of movies.
4. The average duration for imdb score 1.6 starts from 115 and it goes down a bit as we go ahead in the imdb_score
5. Then it rises gradually as we keep going further as the duration stays ranging from 90 minutes to 120 minutes which has most of the movies having an imdb_score of 4.0 to 8.0
6. As score rises from 8.0 the duration also rises from 120 minutes to 190 minutes on an average


Movie Duration Analysis
220.00
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
0.0
2.0
4.0
6.0
8.0
10.0
12.0
Average of duration
Poly. (Average of duration)

Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Solution:

| Languages | Number of Movies | Average of Movies | Median of Movies | StdDev ofMovies |
|---|---|---|---|---|
| English | 3499 | 6.42 | 6.5 | 1.05 |
| Mandarin | 14 | 7.02 | 7.25 | 0.74 |
| Aboriginal | 2 | 6.95 | 6.95 | 0.55 |
| Spanish | 23 | 7.08 | 7.2 | 0.84 |
| French | 34 | 7.36 | 7.3 | 0.51 |
| Filipino | 1 | 6.7 | 6.7 | 0 |
| Maya | 1 | 7.8 | 7.8 | 0 |
| Kazakh | 1 | 6 | 6 | 0 |
| Cantonese | 7 | 7.34 | 7.3 | 0.32 |
| Japanese | 10 | 7.66 | 8 | 0.94 |
| Aramaic | 1 | 7.1 | 7.1 | 0 |
| Italian | 7 | 7.19 | 7 | 1.07 |
| Dutch | 3 | 7.57 | 7.8 | 0.33 |
| Dari | 2 | 7.5 | 7.5 | 0.1 |
| German | 10 | 7.77 | 7.8 | 0.68 |
| Mongolian | 1 | 7.3 | 7.3 | 0 |
| Thai | 3 | 6.63 | 6.6 | 0.37 |
| Bosnian | 1 | 4.3 | 4.3 | 0 |
| Korean | 5 | 7.7 | 7.7 | 0.51 |
| Hungarian | 1 | 7.1 | 7.1 | 0 |
| Hindi | 5 | 7.22 | 7.4 | 0.72 |
| Danish | 3 | 7.9 | 8.1 | 0.43 |
| Portuguese | 5 | 7.76 | 8 | 0.88 |
| Norwegian | 4 | 7.15 | 7.3 | 0.5 |
| Czech | 1 | 7.4 | 7.4 | 0 |
| Russian | 1 | 6.5 | 6.5 | 0 |
| None | 1 | 8.5 | 8.5 | 0 |
| Zulu | 1 | 7.3 | 7.3 | 0 |
| Hebrew | 1 | 8 | 8 | 0 |
| Arabic | 1 | 7.2 | 7.2 | 0 |
| Vietnamese | 1 | 7.4 | 7.4 | 0 |
| Indonesian | 2 | 7.9 | 7.9 | 0.3 |
| Romanian | 1 | 7.9 | 7.9 | 0 |
| Persian | 3 | 8.13 | 8.4 | 0.45 |

1. Since we have counted the mean, median, and standard deviation for each language we can analyze the imdb_score
2. The most common languages for the movies is English with a total of 3499 movies having an average score of 6.42 but few movies also have imdb_score  of 5
3. Then we have French language having 34 movies in total with an average imdb_score of 7.36 its rating differs from 0.51 from its average rating in some movies
4. The languages which have movies ranging from 1 to 40 have an average rating of 7 or 8 and also the differ from 0 to 0.80 from the average rating
5. As the number of movies is less for various languages its average rating is high and on the other hand english language having 3499 movies has average rating less then the others

Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Solution:

| Director name | Average imdb_score | Rank | Top 10% Directors | Difference from average score | 90th Percentile |
|---|---|---|---|---|---|
| James Cameron | 7.91 | 38 | James Cameron | 0.41 | 7.5 |
| Gore Verbinski | 6.99 | | | | |
| Sam Mendes | 7.46 | | | | |
| Christopher Nolan | 8.43 | 8 | Christopher Nolan | 0.93 | |
| Andrew Stanton | 7.73 | 77 | Andrew Stanton | 0.23 | |
| Sam Raimi | 6.96 | | | | |
| Nathan Greno | 7.8 | 53 | Nathan Greno | 0.3 | |
| Joss Whedon | 7.87 | 50 | Joss Whedon | 0.37 | |
| David Yates | 7.2 | | | | |
| Zack Snyder | 7.14 | | | | |
| Bryan Singer | 7.29 | | | | |
| Marc Forster | 7.23 | | | | |
| Andrew Adamson | 7.15 | | | | |
| Rob Marshall | 6.6 | | | | |
| Barry Sonnenfeld | 6.46 | | | | |
| Peter Jackson | 7.89 | 49 | Peter Jackson | 0.39 | |
| Marc Webb | 7.13 | | | | |
| Ridley Scott | 7.13 | | | | |
| Chris Weitz | 6.08 | | | | |
| Anthony Russo | 7 | | | | |
| Peter Berg | 6.67 | | | | |
| Colin Trevorrow | 7 | | | | |
| Shane Black | 7.4 | | | | |
| Tim Burton | 7.05 | | | | |
| Brett Ratner | 6.46 | | | | |
| Dan Scanlon | 7.3 | | | | |
| Michael Bay | 6.62 | | | | |
| Joseph Kosinski | 6.9 | | | | |
| John Lasseter | 7.38 | | | | |
| Martin Campbell | 6.55 | | | | |
| Lee Unkrich | 8.3 | 12 | Lee Unkrich | 0.8 | |
| McG | 6.08 | | | | |
| James Wan | 7.2 | | | | |
| J.J. Abrams | 7.45 | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Baz Luhrmann | 7.08 | | | | |
| Mike Newell | 6.84 | | | | |
| Guillermo del Toro | 7.14 | | | | |
| Steven Spielberg | 7.54 | 150 | Steven Spielberg | | 0.04 |
| Mark Andrews | 7.2 | | | | |
| Justin Lin | 6.84 | | | | |
| Roland Emmerich | 6.19 | | | | |
| Robert Zemeckis | 7.31 | | | | |
| Lana Wachowski | 6.92 | | | | |
| Pete Docter | 8.23 | 16 | Pete Docter | | 0.73 |
| Rob Letterman | 5.93 | | | | |
| Jon Favreau | 7.02 | | | | |
| Martin Scorsese | 7.68 | 103 | Martin Scorsese | | 0.18 |
| Rob Cohen | 5.58 | | | | |
| David Ayer | 6.95 | | | | |
| Tom Shadyac | 6.31 | | | | |
| Doug Liman | 7.13 | | | | |
| Kevin Reynolds | 6.68 | | | | |
| Stephen Sommers | 6.08 | | | | |
| Rupert Sanders | 6.1 | | | | |
| Robert Stromberg | 7 | | | | |
| Matt Reeves | 7.27 | | | | |
| Carl Rinsch | 6.3 | | | | |
| Mike Mitchell | 5.68 | | | | |
| Brad Bird | 7.58 | 143 | Brad Bird | | 0.08 |
| Don Hall | 7.9 | 39 | Don Hall | | 0.4 |
| Rich Moore | 7.8 | 53 | Rich Moore | | 0.3 |
| Dean DeBlois | 7.77 | 71 | Dean DeBlois | | 0.27 |
| Jonathan Mostow | 6.55 | | | | |
| James Gunn | 7.13 | | | | |
| David Fincher | 7.75 | 73 | David Fincher | | 0.25 |
| Matthew Vaughn | 7.65 | 106 | Matthew Vaughn | | 0.15 |
| Francis Lawrence | 7 | | | | |
| Jon Turteltaub | 6.5 | | | | |
| Wolfgang Petersen | 6.77 | | | | |

| | | | | | |
|---|---|---|---|---|---|
| James Bobin | 6.63 | | | | |
| Chris Miller | 6.4 | | | | |
| Duncan Jones | 7.57 | 145 | Duncan Jones | | 0.07 |
| Alan Taylor | 6.85 | | | | |
| Michael Apted | 6.32 | | | | |
| Oliver Stone | 6.91 | | | | |
| Eric Darnell | 6.76 | | | | |
| Shawn Levy | 6.09 | | | | |
| Gavin Hood | 6.88 | | | | |
| Chris Buck | 7.6 | 116 | Chris Buck | | 0.1 |
| George Miller | 6.68 | | | | |
| Ron Howard | 6.93 | | | | |
| Kenneth Branagh | 7.18 | | | | |
| Byron Howard | 6.9 | | | | |
| Hoyt Yeatman | 5.1 | | | | |
| Jonathan Liebesman | 5.66 | | | | |
| Christopher McQuarrie | 7.03 | | | | |
| Joe Johnston | 6.33 | | | | |
| Steve Hickner | 6.2 | | | | |
| Jennifer Yuh Nelson | 7.3 | | | | |
| M. Night Shyamalan | 6.03 | | | | |
| Simon Wells | 5.65 | | | | |
| David Bowers | 6.5 | | | | |
| Joe Wright | 6.98 | | | | |
| Rob Minkoff | 5.94 | | | | |
| Lee Tamahori | 5.95 | | | | |
| Paul Feig | 6 | | | | |
| Alessandro Carloni | 7.2 | | | | |
| Peter Ramsey | 7.3 | | | | |
| Dean Parisot | 6.28 | | | | |
| Edward Zwick | 7.24 | | | | |
| Alex Proyas | 6.82 | | | | |
| Richard Donner | 6.63 | | | | |
| Ang Lee | 7.25 | | | | |
| Jon M. Chu | 5.34 | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Breck Eisner | 6.17 | | | | |
| Hironobu Sakaguchi | 6.4 | | | | |
| Peter Weir | 7.73 | 77 | Peter Weir | | 0.23 |
| Bill Condon | 6.56 | | | | |
| Louis Leterrier | 6.67 | | | | |
| Alejandro G. Iaarritu | 7.84 | 52 | Alejandro G. Iaarritu | | 0.34 |
| David Soren | 6.5 | | | | |
| Paul Greengrass | 7.59 | 142 | Paul Greengrass | | 0.09 |
| Mark Osborne | 7.7 | 80 | Mark Osborne | | 0.2 |
| Peyton Reed | 6.48 | | | | |
| Tim Johnson | 6.75 | | | | |
| Phillip Noyce | 6.77 | | | | |
| Darren Aronofsky | 7.48 | | | | |
| Alfonso Cuaran | 7.8 | 53 | Alfonso Cuaran | | 0.3 |
| Eric Leighton | 6.5 | | | | |
| Tom McGrath | 7.3 | | | | |
| Chris Columbus | 6.65 | | | | |
| Robert Schwentke | 6.35 | | | | |
| Carlos Saldanha | 6.83 | | | | |
| Guy Ritchie | 7.11 | | | | |
| Paul Verhoeven | 6.56 | | | | |
| John McTiernan | 6.63 | | | | |
| Tony Gilroy | 6.73 | | | | |
| Joel Schumacher | 6.34 | | | | |
| John Woo | 6.52 | | | | |
| Tim Story | 6.05 | | | | |
| Mark Steven Johnson | 6.05 | | | | |
| Neill Blomkamp | 7.17 | | | | |
| David Twohy | 6.68 | | | | |
| Josa Padilha | 7.15 | | | | |
| James L. Brooks | 6.5 | | | | |
| James Mangold | 7.08 | | | | |
| George Lucas | 7.4 | | | | |
| Kirk De Micco | 5.9 | | | | |
| Cedric Nicolas-Troyan | 6.1 | | | | |

1. There are 5 column as follows:-
    1. Director name:- It has all unique director names
    2. Average imdb_score:- This column has average of imdb_score for each director
    3. Rank:- This column has ranking with respect to director who fall in top 10% category
    4. Top 10% Directors:- This column has directors who are above 90$^{th}$ percent of average imdb_score
    5. Difference from average score:- Directors having difference from average imdb_score and 90$^{th}$ percentile
    6. 90$^{th}$ Percentile:- The 90$^{th}$ percentile of average imdb_score

2. Their difference from the average in listed in column 5 having data bars for visualization
3. The top 3 directors with most contribution to the success of movies is Akira Kurosava, Tony Kaye, Charles Chaplin and Ron Fricke as Tony and Charles both share same rank

Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Solution:

| Movie Name | Correlation coefficient | Profit |
|---|---|---|
| Avatar | 0.094 | 523505847 |
| Pirates of the Caribbean: At World's End | | 9404152 |
| Spectre | | -44925825 |
| The Dark Knight Rises | | 198130642 |
| John Carter | | -190641321 |
| Spider-Man 3 | | 78530303 |
| Tangled | | -59192738 |
| Avengers: Age of Ultron | | 208991599 |
| Harry Potter and the Half-Blood Prince | | 51956980 |
| Batman v Superman: Dawn of Justice | | 80249062 |
| Superman Returns | | -8930592 |
| Quantum of Solace | | -31631573 |
| Pirates of the Caribbean: Dead Man's Chest | | 198032628 |
| The Lone Ranger | | -125710090 |
| Man of Steel | | 66021565 |
| The Chronicles of Narnia: Prince Caspian | | -83385977 |
| The Avengers | | 403279547 |
| Pirates of the Caribbean: On Stranger Tides | | -8936125 |
| Men in Black 3 | | -45979146 |
| The Hobbit: The Battle of the Five Armies | | 5108370 |
| The Amazing Spider-Man | | 32030663 |
| Robin Hood | | -94780265 |
| The Hobbit: The Desolation of Smaug | | 33355354 |
| The Golden Compass | | -109916481 |
| King Kong | | 11051260 |
| Titanic | | 458672302 |
| Captain America: Civil War | | 157197282 |
| Battleship | | -143826840 |

| Rank | Highest profit | Movie name |
|---|---|---|
| 1 | 523505847 | Avatar |
| 2 | 502177271 | Jurassic World |
| 3 | 458672302 | Titanic |
| 4 | 449935665 | Star Wars: Episode IV - A New Hope |
| 5 | 424449459 | E.T. the Extra-Terrestrial |
| 6 | 403279547 | The Avengers |
| 7 | 377783777 | The Lion King |
| 8 | 359544677 | Star Wars: Episode I - The Phantom Menace |
| 9 | 348316061 | The Dark Knight |
| 10 | 329999255 | The Hunger Games |

These are the top 10 movies having highest profit

1. When the correlation is close to 1, the movie budget increases and the gross earning also increases
2. When the correlation is close to 0, the movie budget increases and the gross earning decreases
3. When the correlation is close to -1, the movie budget and the gross earnings are very low
4. I have also used large function along with sequence to get top 10 highest profit movies

# Result

- Using Excel to its full potential
- I used pivot table the most along with average, sum, countif, correl, large and sequence in various cases
- I have also used data visualization to show various relationship between two or more entities
- In various cases such as the standard deviation of movies gave error since there were less number of movies filtered by the genre

# 5. Hiring Process Analytics

## Project Description

- We have a company's data in which they have showed us how many people they have hired at various post in various departments
- We also have offered salary to each person they have interviewed
- We have to use formulas and pivot table to show the data accordingly
- We get to test pivot table and its function in various manner which helps building our creative thinking

# Approach

- We will be using average, sum, count functions mainly to solve the queries
- For some queries we have to use Pivot table which even helps us in making charts/graph
- We will also use pie diagram to show the number of males and females hired
- Using Bar graph wherever necessary for better understanding
- We will use sheets to for clean and precise view of the result we will get

# Tech-Stack Used

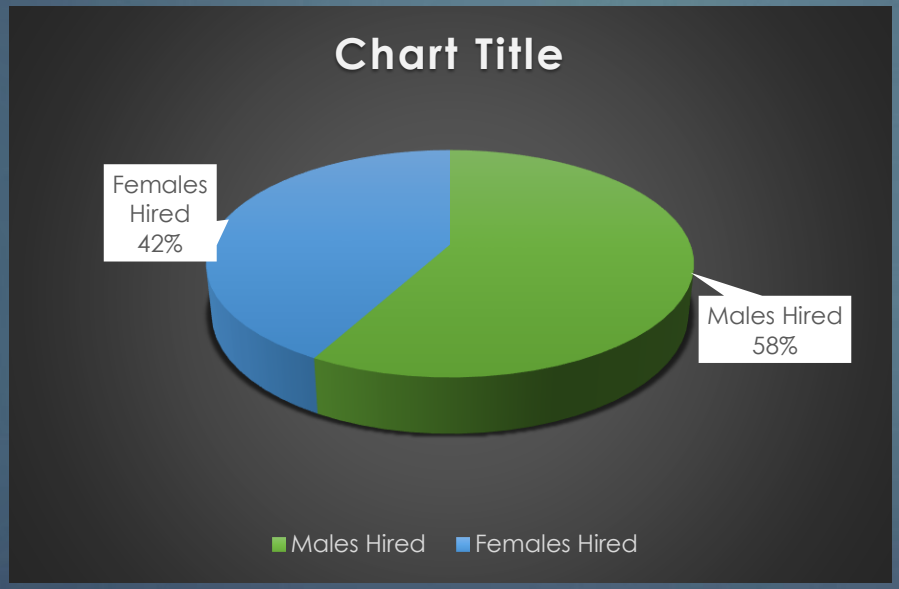
- Excel 2021
- Powerpoint 2021

# Your task: How many males and females are Hired ?

Solution: Male:- =COUNTIFS(D:D,D2,C:C,C2)

Female:- =COUNTIFS(D:D,D3,C:C,C2)

| Males Hired | Females Hired |
|---|---|
| 2563 | 1856 |

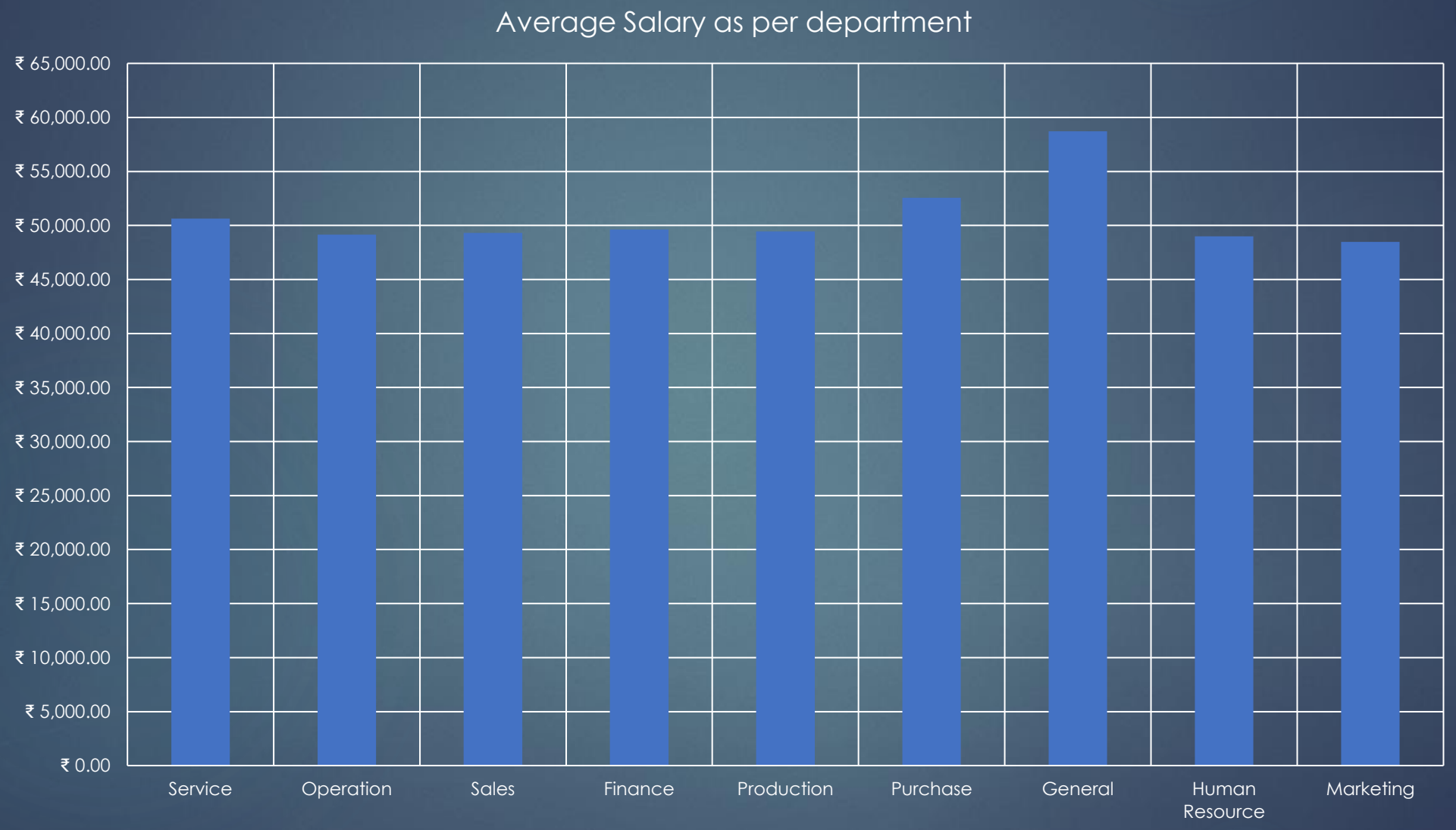


Your task: What is the average salary offered in this company ?

Solution: =ROUND(AVERAGE(G:G),2)

| Average Salary |
| --- |
| ₹ 49,983.03 |

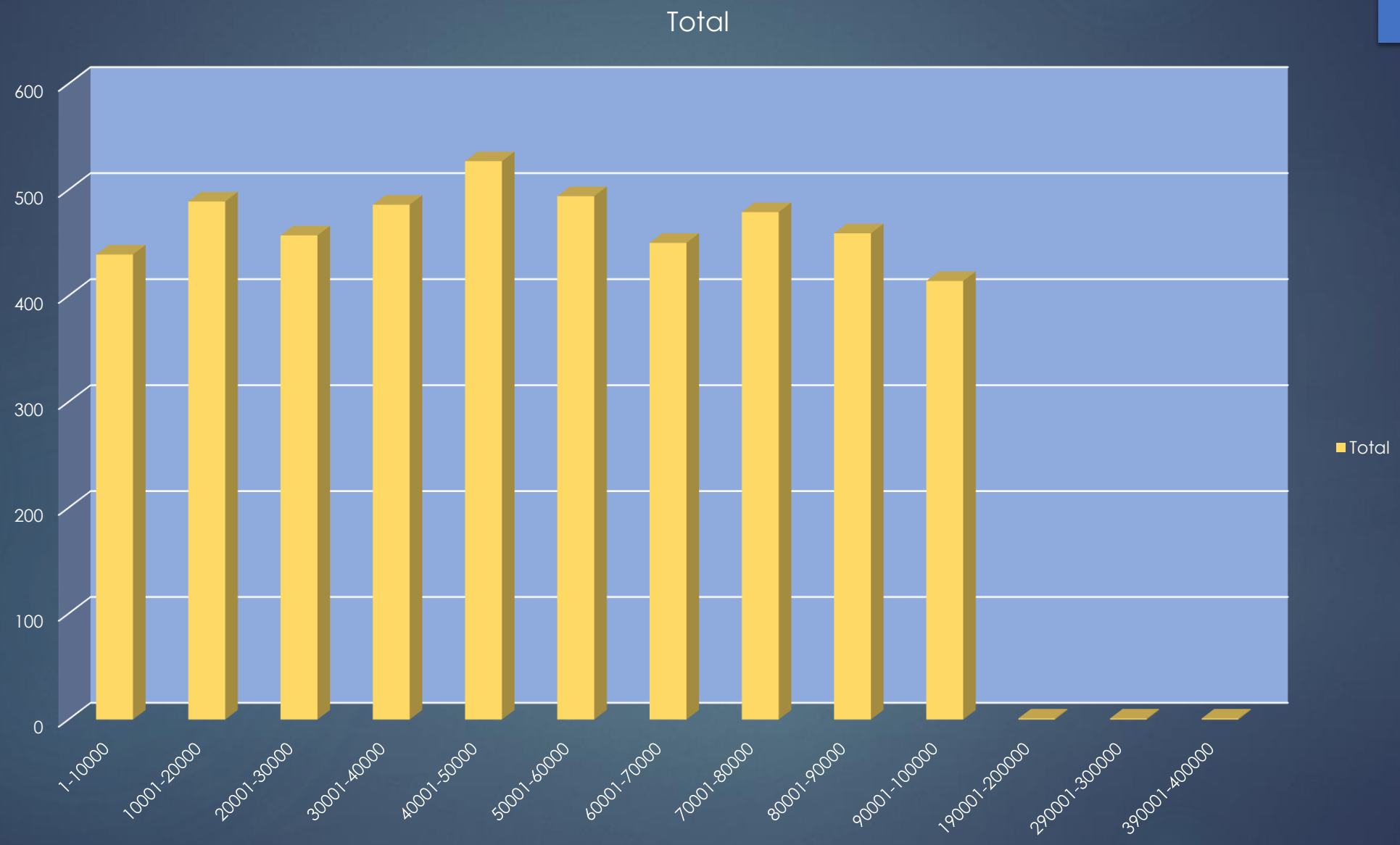
| Department | Average |
| --- | --- |
| Service | ₹ 50,629.88 |
| Operation | ₹ 49,151.35 |
| Sales | ₹ 49,310.38 |
| Finance | ₹ 49,628.01 |
| Production | ₹ 49,448.48 |
| Purchase | ₹ 52,564.77 |
| General | ₹ 58,722.09 |
| Human Resource | ₹ 49,002.28 |
| Marketing | ₹ 48,489.94 |


Average Salary as per department
₹ 65,000.00
₹ 60,000.00
₹ 55,000.00
₹ 50,000.00
₹ 45,000.00
₹ 40,000.00
₹ 35,000.00
₹ 30,000.00
₹ 25,000.00
₹ 20,000.00
₹ 15,000.00
₹ 10,000.00
₹ 5,000.00
₹ 0.00
Service
Operation
Sales
Finance
Production
Purchase
General
Human Resource
Marketing

# Your task: Draw the class intervals for salary in the company ?

Solution: This class interval is just for the people who got hired

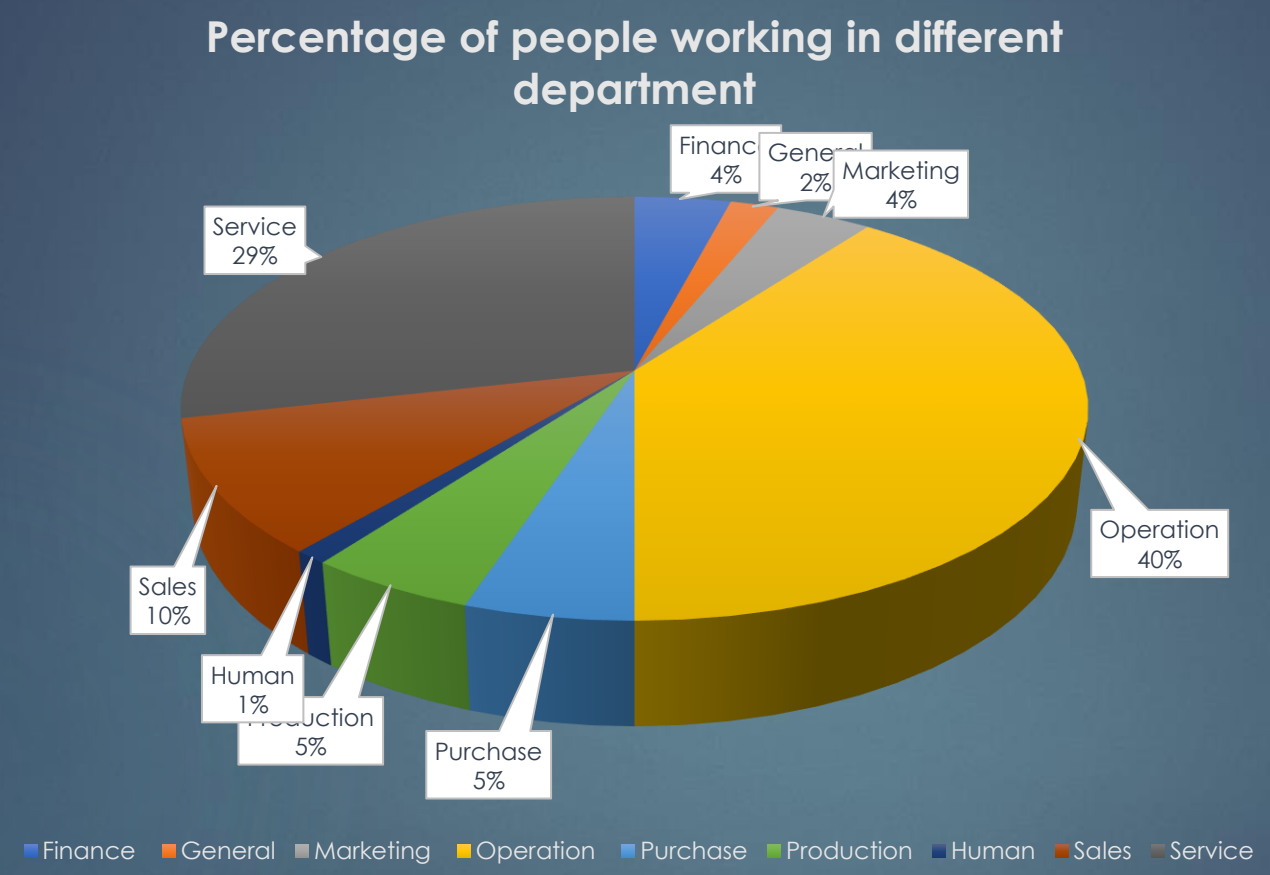
| Status | Hired |
|---|---|
| | |
| **Row Labels** | **Count of event_name** |
| 1-10000 | 439 |
| 10001-20000 | 489 |
| 20001-30000 | 457 |
| 30001-40000 | 486 |
| 40001-50000 | 527 |
| 50001-60000 | 494 |
| 60001-70000 | 450 |
| 70001-80000 | 479 |
| 80001-90000 | 459 |
| 90001-100000 | 414 |
| 190001-200000 | 1 |
| 290001-300000 | 1 |
| 390001-400000 | 1 |
| **Grand Total** | **4697** |


Total
600
500
400
300
200
100
0
1-10000
10001-20000
20001-30000
30001-40000
40001-50000
50001-60000
60001-70000
70001-80000
80001-90000
90001-100000
190001-200000
290001-300000
390001-400000
Total

Your task: Draw Pie Chart / Bar Graph ( or any other graph ) to show proportion of people working different department ?

Solution: As I have put all the formulas in different sheets it will be confusing to put it in her so I have just pasted the tables and graph

| Departments | Total People | Pecent |
|---|---|---|
| Finance | 176 | 4 |
| General | 113 | 2 |
| Marketing | 202 | 4 |
| Operation | 1843 | 39 |
| Purchase | 230 | 5 |
| Production | 246 | 5 |
| Human | 70 | 1 |
| Sales | 485 | 10 |
| Service | 1332 | 28 |
| **Total** | **4697** | |

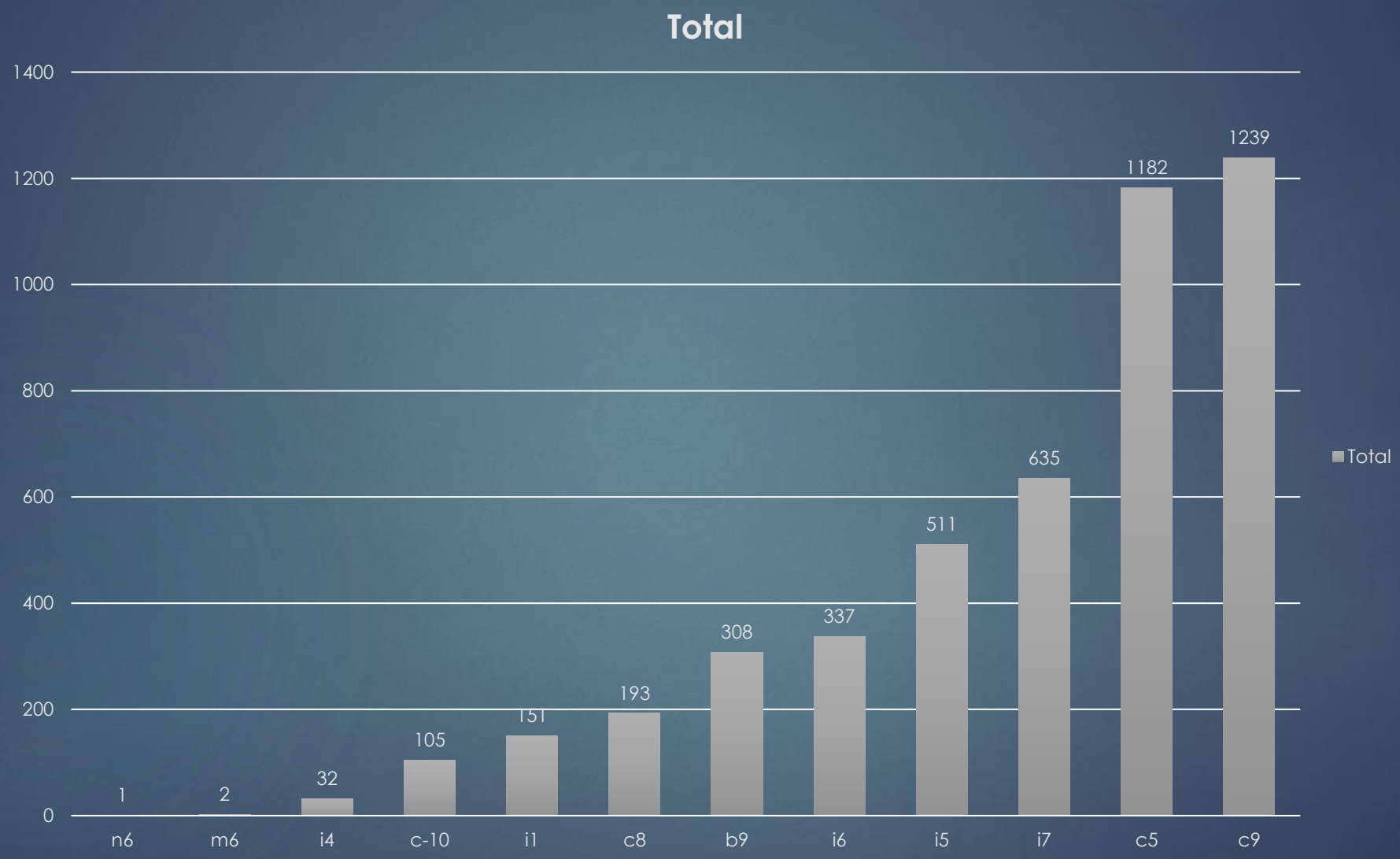



By Akshay Gosavi

# Your task: Represent different post tiers using chart/graph?

Solution:

| Status | Hired |
|---|---|
| | |
| **Row Labels** | **Count of application_id** |
| n6 | 1 |
| m6 | 2 |
| i4 | 32 |
| c-10 | 105 |
| i1 | 151 |
| c8 | 193 |
| b9 | 308 |
| i6 | 337 |
| i5 | 511 |
| i7 | 635 |
| c5 | 1182 |
| c9 | 1239 |
| **Grand Total** | **4696** |


Total
1400
1200
1000
800
600
400
200
0
1
2
32
105
151
193
308
337
511
635
1182
1239
n6
m6
i4
c-10
i1
c8
b9
i6
i5
i7
c5
c9
Total

# Result

- I have learnt various uses of pivot table and its chart
- Many formulas were copied and passed in different sheets which made them complex to understand in this pdf
- So instead I have put the tables from excel to make understanding better
- Using various formulas to get desired result

# 6. Operation Analytics and Investigating Metric Spike

## Project Description

- In this project we are provided with job data and application data of a certain company and we have to solve queries such as average jobs reviewed, number of jobs, language percent, detect duplicate rows and display them
- Some queries are not clear yet we need to dive deep to understand them and solve them
- There are two types of queries
  1. Case Study 1 (Job Data)
  2. Case Study 2 (Investigating metric spike)
- We have to use over, joins, where from structure query language to solve all this query
- The database itself is little complicated as it as datetime datatype which have null values and these values throw error while importing it in SQL database

# Approach

- Using Excel to open dataset and understand the queries there first
- Then using SQL to load data and do operations there to find the actual expected output
- Dividing problems into two or more parts depending upon the query for better understanding
- Helping company with understanding their growth/downfall over time with the help of user data
- Then after solving all the queries analyzing the results and queries and make a final report

# Software Used

- Excel 2022 
- SQL workbench 8.0 
- PowerPoint 2022 version 

## A.Number of jobs reviewed: Amount of jobs reviewed over time.

Your task: Calculate the number of jobs reviewed per hour per day for November 2020?

Solution:

```
select ds as date, round((count(job_id)/sum(time_spent))*(60*60)) as Jobs_review_per_hr
from job_data
where ds between '2020-11-1' and '2020-11-30'
group by ds
order by Jobs_review_per_hr desc;
```

Output:

| date | Jobs_review_per_hr |
|---|---|
| 2020-11-28 | 218 |
| 2020-11-30 | 180 |
| 2020-11-29 | 180 |
| 2020-11-25 | 80 |
| 2020-11-26 | 64 |
| 2020-11-27 | 35 |

Throughput: It is the no. of events happening per second.

Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?

Solution:

```
select ds as dates,avg(Jobs_review_per_hr)
        over(order by ds rows between 6 preceding and current row) as
Seven_day_rolling_avg
        from(select ds, round((count(job_id)/sum(time_spent))*(60*60)) as
Jobs_review_per_hr
        from job_data
        where ds between '2020-11-1' and '2020-11-30'
        group by ds) as derived_job_data
        order by derived_job_data.ds;
```

Output:

| | dates | Seven_day_rolling_avg |
|---|---|---|
| ▶ | 2020-11-25 | 80.0000 |
| | 2020-11-26 | 72.0000 |
| | 2020-11-27 | 59.6667 |
| | 2020-11-28 | 99.2500 |
| | 2020-11-29 | 115.4000 |
| | 2020-11-30 | 126.1667 |

## Percentage share of each language: Share of each language for different contents.

Your task: Calculate the percentage share of each language in the last 30 days?

Solution:

```
select language, round((count(*)/total)*100) as language_share
from job_data
join(select count(*) as total
from job_data) total_cal
group by language
order by language_share asc;
```

Output:

| language | language_share |
|---|---|
| English | 13 |
| Arabic | 13 |
| Hindi | 13 |
| French | 13 |
| Italian | 13 |
| Persian | 38 |

Duplicate rows: Rows that have the same value present in them.

Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

Solution:

```
select ds, count(*) as duplicate_value
from job_data
group by ds
having duplicate_value>1;
```

Output:

| ds | duplicate_value |
|---|---|
| 2020-11-30 | 2 |
| 2020-11-28 | 2 |

User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.

Your task: Calculate the weekly user engagement?

Solution:

```
select extract(week from occured_at)-16 as 'No_of_weeks',
count(distinct user_id) as 'Users_count_per_week'
from events
where event_type = 'engagement'
group by No_of_weeks;
```

Output:

| No_of_weeks | Users_count_per_week |
|---|---|
| 1 | 663 |
| 2 | 1068 |
| 3 | 1113 |
| 4 | 1154 |
| 5 | 1121 |
| 6 | 1186 |
| 7 | 1232 |
| 8 | 1275 |
| 9 | 1264 |
| 10 | 1302 |
| 11 | 1372 |
| 12 | 1365 |
| 13 | 1376 |
| 14 | 1467 |
| 15 | 1299 |
| 16 | 1225 |
| 17 | 1225 |
| 18 | 1204 |
| 19 | 104 |

## User Growth: Amount of users growing over time for a product.

Your task: Calculate the user growth for product?

Solution:

```
select sq.user_count, date, sum(user_count) over (order by date) as user_growth
from( select date(activated_at) as date, count(user_id) as user_count
From users
where state = 'active'
group by date) as sq
order by date;
```

Output:

| user_count | date | user_growth |
|---|---|---|
| 7 | 2013-01-01 | 7 |
| 7 | 2013-01-02 | 14 |
| 6 | 2013-01-03 | 20 |
| 1 | 2013-01-04 | 21 |
| 2 | 2013-01-05 | 23 |
| 3 | 2013-01-06 | 26 |
| 4 | 2013-01-07 | 30 |
| 2 | 2013-01-08 | 32 |
| 6 | 2013-01-09 | 38 |
| 6 | 2013-01-10 | 44 |
| 6 | 2013-01-11 | 50 |
| 3 | 2013-01-12 | 53 |
| 2 | 2013-01-13 | 55 |
| 8 | 2013-01-14 | 63 |
| 11 | 2013-01-15 | 74 |
| 7 | 2013-01-16 | 81 |
| 9 | 2013-01-17 | 90 |

| user_count | date | user_growth |
|---|---|---|
| 1 | 2013-01-19 | 101 |
| 1 | 2013-01-20 | 102 |
| 7 | 2013-01-21 | 109 |
| 5 | 2013-01-22 | 114 |
| 7 | 2013-01-23 | 121 |
| 5 | 2013-01-24 | 126 |
| 8 | 2013-01-25 | 134 |
| 3 | 2013-01-26 | 137 |
| 1 | 2013-01-27 | 138 |
| 7 | 2013-01-28 | 145 |
| 3 | 2013-01-29 | 148 |
| 6 | 2013-01-30 | 154 |
| 6 | 2013-01-31 | 160 |
| 4 | 2013-02-01 | 164 |
| 3 | 2013-02-02 | 167 |
| 1 | 2013-02-03 | 168 |
| 10 | 2013-02-04 | 178 |
| 12 | 2013-02-05 | 190 |
| 6 | 2013-02-06 | 196 |
| 10 | 2013-02-07 | 206 |
| 7 | 2013-02-08 | 213 |
| 2 | 2013-02-09 | 215 |
| 1 | 2013-02-10 | 216 |

| user_count | date | user_growth |
|---|---|---|
| 9 | 2013-02-12 | 229 |
| 7 | 2013-02-13 | 236 |
| 11 | 2013-02-14 | 247 |
| 4 | 2013-02-15 | 251 |
| 2 | 2013-02-16 | 253 |
| 4 | 2013-02-17 | 257 |
| 7 | 2013-02-18 | 264 |
| 5 | 2013-02-19 | 269 |
| 9 | 2013-02-20 | 278 |
| 8 | 2013-02-21 | 286 |
| 7 | 2013-02-22 | 293 |
| 2 | 2013-02-23 | 295 |
| 1 | 2013-02-24 | 296 |
| 6 | 2013-02-25 | 302 |
| 5 | 2013-02-26 | 307 |
| 7 | 2013-02-27 | 314 |
| 6 | 2013-02-28 | 320 |
| 8 | 2013-03-01 | 328 |
| 1 | 2013-03-02 | 329 |
| 9 | 2013-03-04 | 338 |
| 7 | 2013-03-05 | 345 |
| 7 | 2013-03-06 | 352 |
| 9 | 2013-03-07 | 361 |

| user_count | date | user_growth |
|---|---|---|
| 3 | 2013-03-09 | 372 |
| 8 | 2013-03-11 | 380 |
| 5 | 2013-03-12 | 385 |
| 6 | 2013-03-13 | 391 |
| 8 | 2013-03-14 | 399 |
| 4 | 2013-03-15 | 403 |
| 1 | 2013-03-16 | 404 |
| 1 | 2013-03-17 | 405 |
| 6 | 2013-03-18 | 411 |
| 4 | 2013-03-19 | 415 |
| 5 | 2013-03-20 | 420 |
| 7 | 2013-03-21 | 427 |
| 8 | 2013-03-22 | 435 |
| 2 | 2013-03-24 | 437 |
| 7 | 2013-03-25 | 444 |
| 7 | 2013-03-26 | 451 |
| 5 | 2013-03-27 | 456 |
| 3 | 2013-03-28 | 459 |
| 6 | 2013-03-29 | 465 |
| 3 | 2013-03-30 | 468 |
| 2 | 2013-03-31 | 470 |
| 7 | 2013-04-01 | 477 |
| 10 | 2013-04-02 | 487 |

## Weekly Retention: Users getting retained weekly after signing-up for a product.

Your task: Calculate the weekly retention of users-sign up cohort?

Solution:

```
select week, first_value(No_of_users_retained)
over (order by week) as total_users,No_of_users_retained
From(select timestampdiff(week, u.activated_at, e.occurred_at) as week,
count(distinct u.user_id) as No_of_users_retained
From (select user_id, activated_at from users
where state='active'
group by 1) u
left join(select user_id,occurred_at from events
where event_name='login') e
on u.user_id=e.user_id
group by 1) c
order by week;
```

Output:

| week | total_users | No_of_users_retained |
|---|---|---|
| NULL | 3239 | 3239 |
| 0 | 3239 | 3772 |
| 1 | 3239 | 1709 |
| 2 | 3239 | 1226 |
| 3 | 3239 | 842 |
| 4 | 3239 | 654 |
| 5 | 3239 | 501 |
| 6 | 3239 | 431 |
| 7 | 3239 | 369 |
| 8 | 3239 | 360 |
| 9 | 3239 | 331 |
| 10 | 3239 | 292 |
| 11 | 3239 | 295 |
| 12 | 3239 | 266 |
| 13 | 3239 | 282 |
| 14 | 3239 | 289 |
| 15 | 3239 | 258 |
| 16 | 3239 | 276 |
| 17 | 3239 | 283 |
| 18 | 3239 | 269 |
| 19 | 3239 | 276 |
| 20 | 3239 | 277 |
| 21 | 3239 | 232 |
| 22 | 3239 | 221 |
| 23 | 3239 | 219 |
| 24 | 3239 | 207 |
| 25 | 3239 | 239 |
| 26 | 3239 | 225 |
| 27 | 3239 | 196 |
| 28 | 3239 | 197 |
| 29 | 3239 | 191 |
| 30 | 3239 | 194 |
| 31 | 3239 | 179 |
| 32 | 3239 | 192 |
| 33 | 3239 | 153 |
| 34 | 3239 | 159 |
| 35 | 3239 | 155 |
| 36 | 3239 | 153 |
| 37 | 3239 | 161 |
| 38 | 3239 | 165 |
| 39 | 3239 | 169 |
| 40 | 3239 | 168 |
| 41 | 3239 | 158 |
| 42 | 3239 | 152 |
| 43 | 3239 | 160 |
| 44 | 3239 | 162 |
| 45 | 3239 | 144 |
| 46 | 3239 | 137 |
| 47 | 3239 | 124 |
| 48 | 3239 | 106 |
| 49 | 3239 | 112 |
| 50 | 3239 | 112 |
| 51 | 3239 | 100 |
| 52 | 3239 | 109 |
| 53 | 3239 | 114 |
| 54 | 3239 | 109 |
| 55 | 3239 | 100 |
| 56 | 3239 | 106 |
| 57 | 3239 | 101 |
| 58 | 3239 | 97 |
| 59 | 3239 | 90 |
| 60 | 3239 | 100 |
| 61 | 3239 | 95 |
| 62 | 3239 | 86 |
| 63 | 3239 | 98 |
| 64 | 3239 | 96 |
| 65 | 3239 | 98 |
| 66 | 3239 | 101 |
| 67 | 3239 | 99 |
| 68 | 3239 | 105 |
| 69 | 3239 | 101 |
| 70 | 3239 | 85 |
| 71 | 3239 | 82 |
| 72 | 3239 | 77 |
| 73 | 3239 | 72 |
| 74 | 3239 | 56 |
| 75 | 3239 | 50 |
| 76 | 3239 | 49 |
| 77 | 3239 | 41 |
| 78 | 3239 | 26 |
| 79 | 3239 | 29 |
| 80 | 3239 | 22 |
| 81 | 3239 | 9 |
| 82 | 3239 | 15 |
| 83 | 3239 | 15 |
| 84 | 3239 | 9 |
| 85 | 3239 | 3 |
| 86 | 3239 | 2 |

Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.

Your task: Calculate the weekly engagement per device?

Solution:

```
select date(occurred_at) as date, device, count(distinct user_id) as user_count ,
extract(week from occurred_at)-16 as week
from events
group by device, week
order by week;
```

Output:

| date | device | user_count | week |
|---|---|---|---|
| 2014-05-01 | acer aspire desktop | 12 | 1 |
| 2014-05-02 | acer aspire notebook | 23 | 1 |
| 2014-05-02 | amazon fire phone | 4 | 1 |
| 2014-05-02 | asus chromebook | 23 | 1 |
| 2014-05-02 | dell inspiron desktop | 20 | 1 |
| 2014-05-01 | dell inspiron notebook | 48 | 1 |
| 2014-05-01 | hp pavilion desktop | 17 | 1 |
| 2014-05-02 | htc one | 19 | 1 |
| 2014-05-02 | ipad air | 29 | 1 |
| 2014-05-02 | ipad mini | 19 | 1 |
| 2014-05-02 | iphone 4s | 27 | 1 |
| 2014-05-02 | iphone 5 | 69 | 1 |
| 2014-05-01 | iphone 5s | 47 | 1 |
| 2014-05-02 | kindle fire | 6 | 1 |
| 2014-05-02 | lenovo thinkpad | 94 | 1 |
| 2014-05-01 | mac mini | 7 | 1 |
| 2014-05-02 | macbook air | 61 | 1 |
| 2014-05-02 | macbook pro | 154 | 1 |
| 2014-05-01 | nexus 10 | 16 | 1 |
| 2014-05-02 | nexus 5 | 45 | 1 |
| 2014-05-02 | nexus 7 | 19 | 1 |
| 2014-05-01 | nokia lumia 635 | 18 | 1 |
| 2014-05-02 | samsumg galaxy tablet | 8 | 1 |
| 2014-05-02 | samsung galaxy note | 8 | 1 |
| 2014-05-01 | samsung galaxy s4 | 58 | 1 |

| | | | | |
|---|---|---|---|---|
| | 2014-05-07 | acer aspire desktop | 33 | 2 |
| | 2014-05-08 | acer aspire notebook | 41 | 2 |
| | 2014-05-06 | amazon fire phone | 9 | 2 |
| | 2014-05-09 | asus chromebook | 49 | 2 |
| | 2014-05-07 | dell inspiron desktop | 63 | 2 |
| | 2014-05-06 | dell inspiron notebook | 84 | 2 |
| | 2014-05-09 | hp pavilion desktop | 39 | 2 |
| | 2014-05-05 | htc one | 20 | 2 |
| | 2014-05-04 | ipad air | 62 | 2 |
| | 2014-05-10 | ipad mini | 37 | 2 |
| | 2014-05-07 | iphone 4s | 51 | 2 |
| | 2014-05-05 | iphone 5 | 130 | 2 |
| | 2014-05-06 | iphone 5s | 76 | 2 |
| | 2014-05-08 | kindle fire | 29 | 2 |
| | 2014-05-06 | lenovo thinkpad | 174 | 2 |
| | 2014-05-08 | mac mini | 15 | 2 |
| | 2014-05-07 | macbook air | 136 | 2 |
| | 2014-05-10 | macbook pro | 280 | 2 |
| | 2014-05-09 | nexus 10 | 34 | 2 |
| | 2014-05-10 | nexus 5 | 89 | 2 |
| | 2014-05-08 | nexus 7 | 35 | 2 |
| | 2014-05-07 | nokia lumia 635 | 35 | 2 |
| | 2014-05-09 | samsumg galaxy tablet | 15 | 2 |
| | 2014-05-08 | samsung galaxy note | 18 | 2 |
| | 2014-05-06 | samsung galaxy s4 | 90 | 2 |
| | 2014-05-05 | windows surface | 13 | 2 |
| | 2014-05-13 | acer aspire desktop | 26 | 3 |
| | 2014-05-16 | acer aspire notebook | 48 | 3 |
| | 2014-05-14 | amazon fire phone | 14 | 3 |
| | 2014-05-11 | asus chromebook | 32 | 3 |
| | 2014-05-13 | dell inspiron desktop | 38 | 3 |
| | 2014-05-12 | dell inspiron notebook | 93 | 3 |
| | 2014-05-16 | hp pavilion desktop | 47 | 3 |
| | 2014-05-17 | htc one | 34 | 3 |
| | 2014-05-15 | ipad air | 57 | 3 |
| | 2014-05-13 | ipad mini | 40 | 3 |
| | 2014-05-13 | iphone 4s | 48 | 3 |
| | 2014-05-17 | iphone 5 | 123 | 3 |
| | 2014-05-13 | iphone 5s | 88 | 3 |
| | 2014-05-15 | kindle fire | 22 | 3 |
| | 2014-05-14 | lenovo thinkpad | 204 | 3 |
| | 2014-05-13 | mac mini | 20 | 3 |
| | 2014-05-14 | macbook air | 123 | 3 |

| | | | | |
|---|---|---|---|---|
| | 2014-05-14 | nexus 7 | 47 | 3 |
| | 2014-05-13 | nokia lumia 635 | 25 | 3 |
| | 2014-05-15 | samsumg galaxy tablet | 6 | 3 |
| | 2014-05-14 | samsung galaxy note | 12 | 3 |
| | 2014-05-16 | samsung galaxy s4 | 103 | 3 |
| | 2014-05-13 | windows surface | 17 | 3 |
| | 2014-05-21 | acer aspire desktop | 25 | 4 |
| | 2014-05-22 | acer aspire notebook | 49 | 4 |
| | 2014-05-21 | amazon fire phone | 12 | 4 |
| | 2014-05-22 | asus chromebook | 45 | 4 |
| | 2014-05-18 | dell inspiron desktop | 55 | 4 |
| | 2014-05-20 | dell inspiron notebook | 93 | 4 |
| | 2014-05-20 | hp pavilion desktop | 32 | 4 |
| | 2014-05-21 | htc one | 32 | 4 |
| | 2014-05-23 | ipad air | 63 | 4 |
| | 2014-05-23 | ipad mini | 39 | 4 |
| | 2014-05-21 | iphone 4s | 63 | 4 |
| | 2014-05-19 | iphone 5 | 142 | 4 |
| | 2014-05-23 | iphone 5s | 86 | 4 |
| | 2014-05-21 | kindle fire | 24 | 4 |
| | 2014-05-24 | lenovo thinkpad | 201 | 4 |
| | 2014-05-21 | mac mini | 30 | 4 |
| | 2014-05-22 | macbook air | 134 | 4 |
| | 2014-05-19 | macbook pro | 289 | 4 |
| | 2014-05-19 | nexus 10 | 27 | 4 |
| | 2014-05-22 | nexus 5 | 110 | 4 |
| | 2014-05-20 | nexus 7 | 36 | 4 |
| | 2014-05-20 | nokia lumia 635 | 27 | 4 |
| | 2014-05-21 | samsumg galaxy tablet | 11 | 4 |
| | 2014-05-21 | samsung galaxy note | 19 | 4 |
| | 2014-05-21 | samsung galaxy s4 | 106 | 4 |
| | 2014-05-22 | windows surface | 24 | 4 |
| | 2014-05-27 | acer aspire desktop | 31 | 5 |
| | 2014-05-28 | acer aspire notebook | 49 | 5 |
| | 2014-05-26 | amazon fire phone | 6 | 5 |
| | 2014-05-26 | asus chromebook | 39 | 5 |
| | 2014-05-27 | dell inspiron desktop | 46 | 5 |
| | 2014-05-29 | dell inspiron notebook | 88 | 5 |
| | 2014-05-29 | hp pavilion desktop | 49 | 5 |
| | 2014-05-30 | htc one | 29 | 5 |
| | 2014-05-28 | ipad air | 55 | 5 |
| | 2014-05-26 | ipad mini | 29 | 5 |
| | 2014-05-29 | iphone 4s | 49 | 5 |
| | 2014-05-30 | iphone 5 | 147 | 5 |
| | 2014-05-30 | iphone 5s | 86 | 5 |
| | 2014-05-27 | kindle fire | 31 | 5 |

## Email Engagement: Users engaging with the email service.

Your task: Calculate the email engagement metrics?

Solution:

```
select Week, (email_click/email_open)*100 as click_to_open_rate,
        (email_open/total_user)*100 as Email_Open_Rate,
        (weekly_digest/total_user)*100 as Weekly_digest_rate,
        (sent_emails/total_user)*100 as Sent_eamils_rate
        from(select count(user_id)as total_user, extract(week from occurred_at)-16 as Week,
        count(case when action ='email_clickthrough' then user_id else null end) as email_click,
        count(case when action ='email_open' then user_id else null end) as email_open,
        count(case when action='sent_reengagement_email' then user_id else null end)as sent_emails,
        count(case when action='sent_weekly_digest' then user_id else null end) as weekly_digest
        from email_events
        group by week) sq
        group by week
        order by week;
```

Output:

| Week | click_to_open_rate | Email_Open_Rate | Weekly_digest_rate | Sent_eamils_rate |
|---|---|---|---|---|
| 1 | 53.3762 | 21.2722 | 62.3803 | 4.9932 |
| 2 | 47.0652 | 22.2169 | 63.5354 | 3.7914 |
| 3 | 49.2323 | 22.5896 | 62.2890 | 4.0000 |
| 4 | 50.1972 | 22.6542 | 61.7069 | 4.2672 |
| 5 | 43.6098 | 22.8438 | 63.5391 | 3.6550 |
| 6 | 49.4980 | 21.5538 | 63.6226 | 4.1549 |
| 7 | 49.8152 | 22.3093 | 62.5155 | 4.0619 |
| 8 | 48.0274 | 22.9257 | 61.6201 | 4.4436 |
| 9 | 48.2759 | 21.7529 | 63.8768 | 3.8689 |
| 10 | 47.6109 | 22.1970 | 63.0871 | 4.1477 |
| 11 | 50.2419 | 22.5291 | 62.2820 | 3.8699 |
| 12 | 47.6911 | 22.4446 | 63.0450 | 3.8063 |
| 13 | 48.3687 | 21.6876 | 64.0545 | 3.7679 |
| 14 | 45.4220 | 23.3039 | 62.2604 | 3.8506 |
| 15 | 33.1374 | 23.2411 | 65.2664 | 3.7910 |
| 16 | 31.1573 | 22.8707 | 66.6101 | 3.3933 |
| 17 | 34.1869 | 23.1422 | 64.7181 | 4.2281 |
| 18 | 31.9481 | 23.9391 | 64.3557 | 4.0572 |
| 19 | 92.6829 | 32.2835 | 0.0000 | 37.7953 |

# Result

- Case 1 was much easier then the case 2
- In case 2 we were given big data and we had to handle it and solve the query accordingly
- I got to use sub queries with joins which was too complicated
- But it gave me a good real life experience of critical thinking and also thinking out of the box
- Using various sql function and applying them in the queries which helped me in better understanding

# 7. Instagram User Analytics

## Project Description

### Marketing point of view

In this given project we were given several queries such as:-

1. In the first query we need to identify the five most oldest users on Instagram using the dates on which their accounts were created
2. In second query we had to find a user who never posted even a single photo on Instagram and promote him to post by sending a email
3. In third query we had to identify the winner of the contest hosted by the team in which we had to select a user who gets the most likes on a single photo and provide the following user details to the team
4. In this fourth query we had to identify and suggest the top 5 most commonly used hashtags on the platform for a partner brand so that they can reach most of the people on Instagram platform
5. In this last marketing query we had to identify on which day did most of the users registered on this platform. In this query we also needed to provide an insights on when can we schedule an ad campaign so that they can promote their products to users

## Investor Point of view

In this second part we only had two queries given by the investor:-

1. In the first query we have to provide the average time of users posts on instagram and the total number of users and also total number of posts on instagram
2. As there are companies who try to increase the users count on their platform by using bots and creating fake accounts the user wanted to know if there are any such fake and bot accounts before investing in the product. This can be done by checking the data for users who like every single post on instagram since bot can like all the posts

# Approach

1. Using SQL Query Language to solve all the queries
2. Using SQL commands to extract necessary information from the database
3. Dividing problems into parts for better understanding
4. Helping investors by giving them the average posts per user
5. Providing investors with total number of posts and users on Instagram

# Softwares used

Software used from filtering and sorting the data according to the query to making the detailed project report:-

1.SQL workbench 8.0

2.PowerPoint 2022 version

# Marketing Queries

**Task 1**: Find the 5 oldest users of the Instagram from the database provided

**Solution**: select username, created_at from users

ORDER BY created_at asc

limit 5;

**Output**:

| username | created_at |
|---|---|
| Darby_Herzog | 2016-05-06 00:14:21 |
| Emilio_Bernier52 | 2016-05-06 13:04:30 |
| Elenor88 | 2016-05-08 01:30:41 |
| Nicole71 | 2016-05-09 17:30:22 |
| Jordyn.Jacobson2 | 2016-05-14 07:56:26 |

**Task 2**: Find the users who have never posted a single photo on Instagram

**Solution**: select * from users

where id not in (select user_id from photos);

**Output**:

| id | username | created_at |
|---|---|---|
| 5 | Aniya_Hackett | 2016-12-07 01:04:39 |
| 7 | Kasandra_Homenick | 2016-12-12 06:50:08 |
| 14 | Jadyn81 | 2017-02-06 23:29:16 |
| 21 | Rocio33 | 2017-01-23 11:51:15 |
| 24 | Maxwell.Halvorson | 2017-04-18 02:32:44 |
| 25 | Tierra.Trantow | 2016-10-03 12:49:21 |
| 34 | Pearl7 | 2016-07-08 21:42:01 |
| 36 | Ollie_Ledner37 | 2016-08-04 15:42:20 |
| 41 | Mckenna17 | 2016-07-17 17:25:45 |
| 45 | David.Osinski47 | 2017-02-05 21:23:37 |
| 49 | Morgan.Kassulke | 2016-10-30 12:42:31 |
| 53 | Linnea59 | 2017-02-07 07:49:34 |
| 54 | Duane60 | 2016-12-21 04:43:38 |
| 57 | Julien_Schmidt | 2017-02-02 23:12:48 |
| 66 | Mike.Auer39 | 2016-07-01 17:36:15 |
| 68 | Franco_Keebler64 | 2016-11-13 20:09:27 |
| 71 | Nia_Haag | 2016-05-14 15:38:50 |
| 74 | Hulda.Macejkovic | 2017-01-25 17:17:28 |
| 75 | Leslie67 | 2016-09-21 05:14:01 |
| 76 | Janelle.Nikolaus81 | 2016-07-21 09:26:09 |
| 80 | Darby_Herzog | 2016-05-06 00:14:21 |
| 81 | Esther.Zulauf61 | 2017-01-14 17:02:34 |
| 83 | Bartholome.Bernhard | 2016-11-06 02:31:23 |

**Task 3**: Identify the winner of the contest and provide their details to the team

**Solution**: select users.id, users.username, count(photo_id) as No_of_likes,photos.image_url from likes

```
join photos
on photos.id = likes.photo_id
join users
on users.id=likes.photo_id
group by photo_id
order by No_of_likes desc
limit 1;
```

**Output**:

| | id | username | No_of_likes | image_url |
|---|---|---|---|---|
| ▶ | 30 | Kaley9 | 41 | http://kenny.com |

**Task 4:** Identify and suggest the top 5 most commonly used hashtags on the platform

**Solution:**
```
select tags.id, tag_name, count(tag_id) as hastag_count from tags
    join photo_tags
    on photo_tags.tag_id=tags.id
    group by tag_id
    order by hastag_count desc
    limit 5;
```

**Output**:

| id | tag_name | hastag_count |
|---|---|---|
| 21 | smile | 59 |
| 20 | beach | 42 |
| 17 | party | 39 |
| 13 | fun | 38 |
| 18 | concert | 24 |

**Task 5**: What day of the week do most users register on? Provide insights on when to schedule an ad campaign

**Solution**: select dayname(created_at) as dayofweek,

count(*) as most_users_registered from users

group by dayofweek

order by most_users_registered desc;

**Output**:

| | dayofweek | most_users_registered |
|---|---|---|
| ▶ | Thursday | 16 |
| | Sunday | 16 |

# Investor Queries

**Task 1**: Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users

**Solution**: select(select count(id) as total_photos from photos)/(select(select count(id) as total_users) from users) as Average_user_posting;
select count(id) as total_photos from photos;
select count(id) as total_users from users;

**Output**:

| Average_user_posting |
| --- |
| 2.5700 |

Average posts by a user is 2.5700 i.e 3 posts

| total_photos |
| --- |
| 257 |

| total_users |
| --- |
| 100 |

**Task 2**: Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).

**Solution**: select likes.user_id, users.username, count(likes.created_at) as NoOfLikes from likes

join users on users.id=likes.user_id

group by likes.user_id

having NoOfLikes = 257;

**Output**:

| user_id | username | NoOfLikes |
|---|---|---|
| 5 | Aniya_Hackett | 257 |
| 14 | Jadyn81 | 257 |
| 21 | Rocio33 | 257 |
| 24 | Maxwell.Halvorson | 257 |
| 36 | Ollie_Ledner37 | 257 |
| 41 | Mckenna17 | 257 |
| 54 | Duane60 | 257 |
| 57 | Julien_Schmidt | 257 |
| 66 | Mike.Auer39 | 257 |
| 71 | Nia_Haag | 257 |
| 75 | Leslie67 | 257 |
| 76 | Janelle.Nikolaus81 | 257 |
| 91 | Bethany20 | 257 |

# Insights

1. In this project Instagram user analytics we were given a SQL data with tables such as users, posts, tags, comments, likes, follows, photo_tags.
2. Using this tables we can easily get every detail for every user such as his name, userid, birthdate, account creation date, how many times he has posted and on which post he had commented.
3. We might sometimes get confused with the column name as it is same in various tables.
4. Using the joint the problem was countered easily in many questions.
5. The data was clean and precise which helped me to filter and extract the information easily.

# Result

1. Using this given database I learn data analysis process on Instagram user data which bought me 1 step closer to my goal of becoming a data analyst.
2. Real life example of data analyst and use my skills at right point.
3. It gave me a good experience to work with big company data and how to handle it.
4. Solving those various queries helped me think outside the box and put forth my knowledge and my SQL skills

# 8. Data Analytics In Everyday Life

Real life example where data analysis is used (Looking for a job)

Plan - Suppose I am looking for a job as a backend developer

Prepare - I will learn all the necessary skills required to become a backend developer such as Java, Python, etc.

Process - Then I will give various exams to get certified as backend developer. This certificate will really helpful to make me stand out from others to get a job.

Analyze - Now while applying for certain jobs I might need check its description such as they might want someone who is expert in C++. So I need to analyze such description according to my knowledge as I only know Java and python.

Share - Once I am shortlisted, I might need to talk to the HR regarding what kind of responsibilities they need in me, what kind of experience they want in this role. Then, if I am selected there will be salary negotiation.

Act - If I like their terms and the package, they offer me. Then I will accept the offer and get the job

# 9. Appendix: Links To All Projects

- Links to all projects :- https://drive.google.com/drive/folders/1jMLtnWgV2MoeJsladvxPhgedJYXOoYsT?usp=sharing


Thank
you!