Data Analysis Portfolio

Prepared By:- Aishwarya Desai

Professional Background

I did my post-graduation in statistics subject with first class from Sadguru Gadage Maharaj college Karad. And I am good at Python, R programming languages. Also I have learnt and used Machine Learning Algorithms, Data Cleaning and Analysis techniques in my projects.

I have worked with several companies as an intern like Technocolabs Inc Indore, Yoshops.com and Trainity as a Machine Learning Engineer and Data analyst. My main role was to analyze data and finding required insights also to representing it using PowerBi tools and deploying ML models.

I have also published a research paper titled- "A Statistical Study of Analysis of India’s Gross Domestic Product " in an ISBN journal and have worked on several projects related to data analysis and machine learning.

As a recent graduate, I'm eager to immerse myself in the complexities of the corporate world and gain insights into its inner workings. I'm highly adaptable and open to learning, given my limited practical experience. While my academic background has equipped me with theoretical knowledge, I'm enthusiastic about applying this knowledge in a tangible, real-world context. I'm confident that through dedicated efforts, I'll acquire valuable practical skills.

Table Of Contents

- Professional Background
- Table of Content
- Projects :

1. Instagram User Analytics
2. Operation Analytics and Investigating Metric Spike Analysis.
3. Hiring Process Analytics
4. IMDB Movie Analysis
5. Bank Loan Case Study
6. Impact Of Car Features
7. ABC Call Volume Trend Analysis

- Findings
- Conclusion

1. Instagram User Analytics

Project Overview

Exploring Instagram User Interactions for Informed Decision-Making

In this project, my role as a data analyst is like collaborating closely with Instagram's product team.

The core goal of this project revolves around deriving meaningful insights from user interactions and engagement data within the Instagram application.

Leveraging MySQL Workbench, the project strives to reveal underlying patterns, emerging trends, and user behaviors, ultimately furnishing actionable intelligence to inform and steer the company's strategies for sustainable growth.

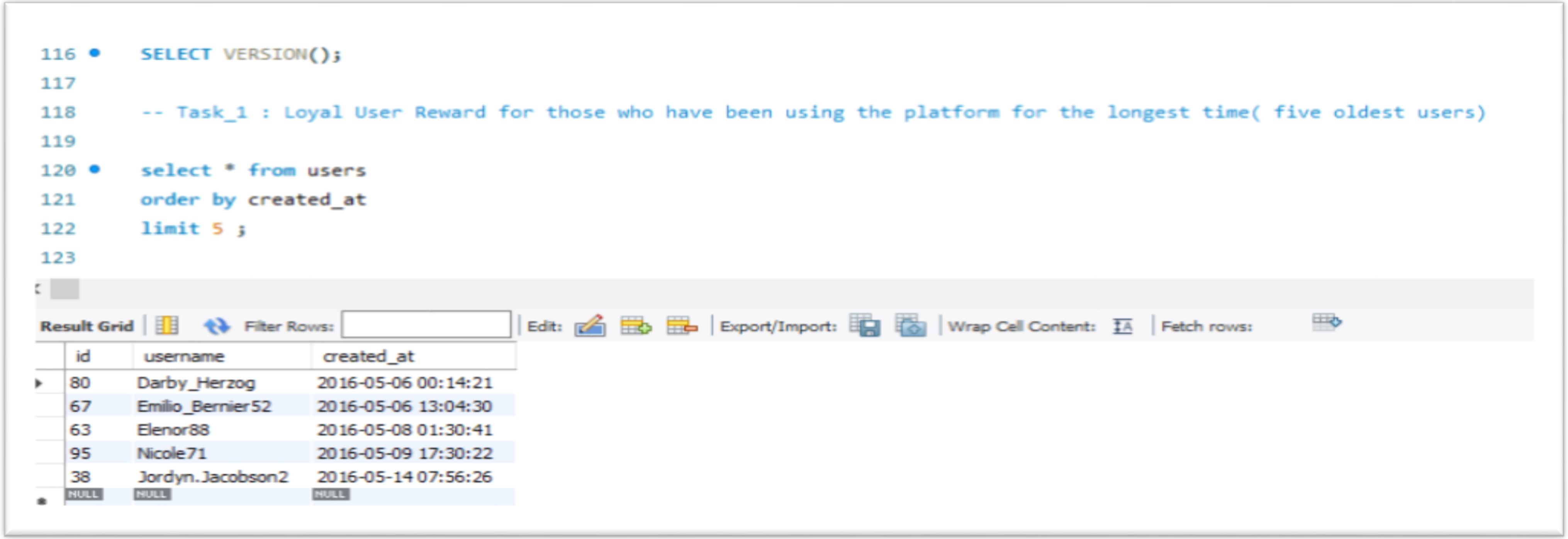
Project Description

The central aim of this project is to harness the power of data analysis techniques and SQL expertise to delve into the intricacies of user interactions and engagement with the Instagram application.

Our overarching goal is to cultivate a profound understanding of user behavior, thereby furnishing invaluable insights that will serve as a compass for decision-making across various organizational departments, encompassing marketing, product development, and the enhancement of user experience. Through meticulous scrutiny of the data, our endeavor is to unearth underlying patterns, emergent trends, and user preferences that will play an instrumental role in steering the strategic growth and continual refinement of the Instagram app.

A) Marketing Analysis: Insights

1. Loyal User Reward for those who have been using the platform for the longest time. That is to identify the five oldest users on Instagram from the database.

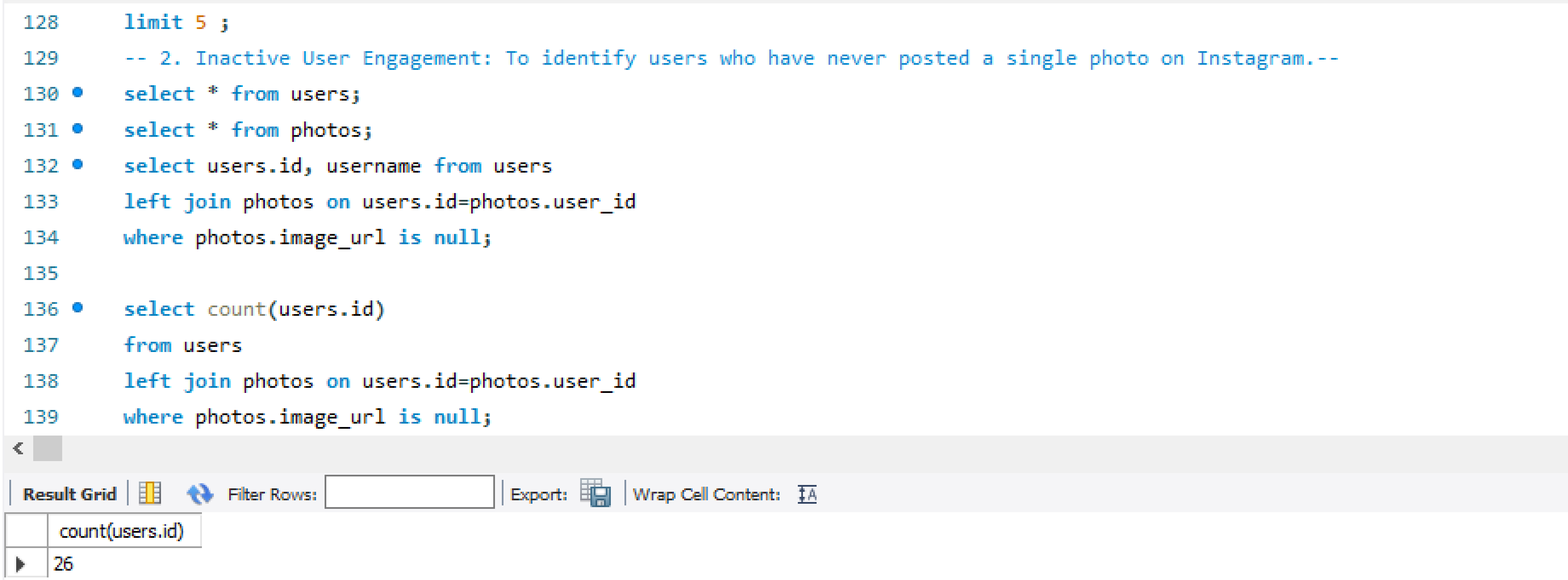


```
SELECT VERSION();

-- Task_1 : Loyal User Reward for those who have been using the platform for the longest time( five oldest users)

select * from users
order by created_at
limit 5 ;

```

id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26
NULL	NULL	NULL

Here we have got top five oldest user information, And the names of users are: Darby_Herzog, Emilio_Bernier52, Elenor88, Nicole71 and Jordyn.Jacobson2.

2. Inactive User Engagement: To identify users who have never posted a single photo on Instagram.

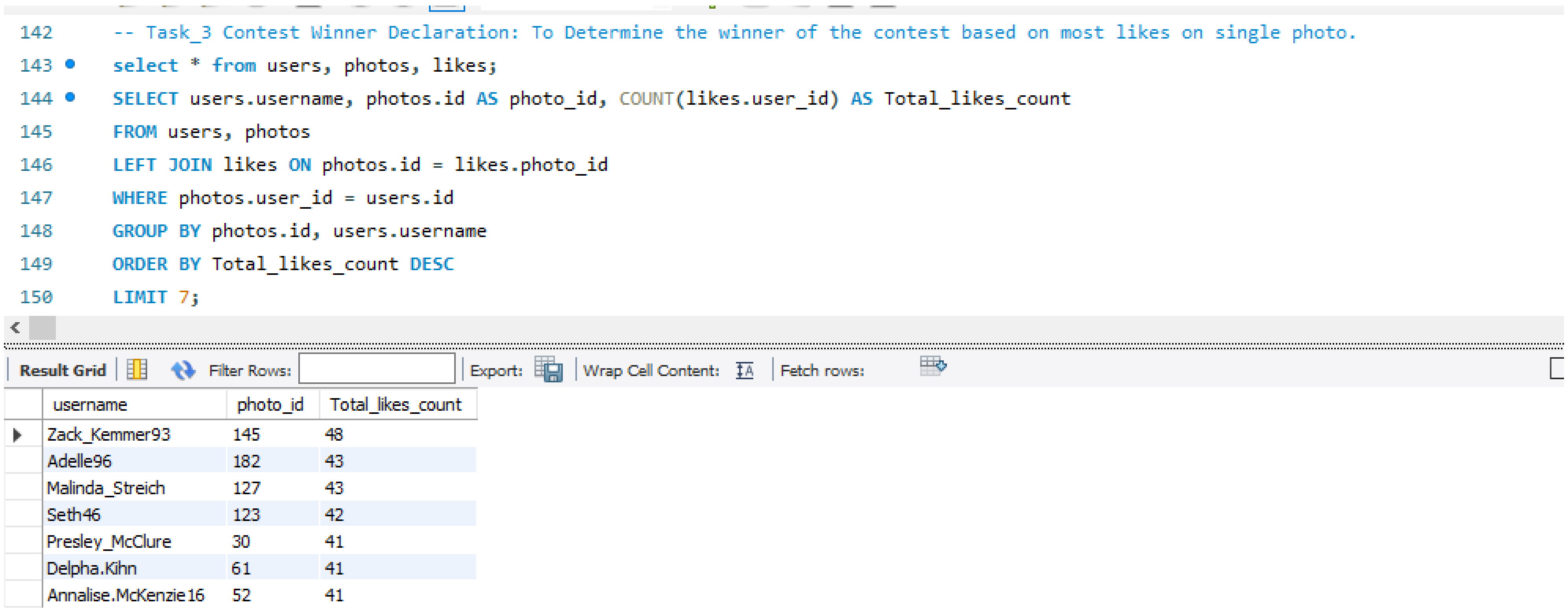


```
limit 5 ;
-- 2. Inactive User Engagement: To identify users who have never posted a single photo on Instagram.--
select * from users;
select * from photos;
select users.id, username from users
left join photos on users.id=photos.user_id
where photos.image_url is null;

select count(users.id)
from users
left join photos on users.id=photos.user_id
where photos.image_url is null;
```

```
count(users.id)
26
```

Here we have total 26 users and their names who have never posted a single photo on Instagram. So using this information, the team can encourage these inactive users to start posting by sending them promotional emails very easily.

3. Contest Winner Declaration: To Determine the winner of the contest based on most likes on single photo.



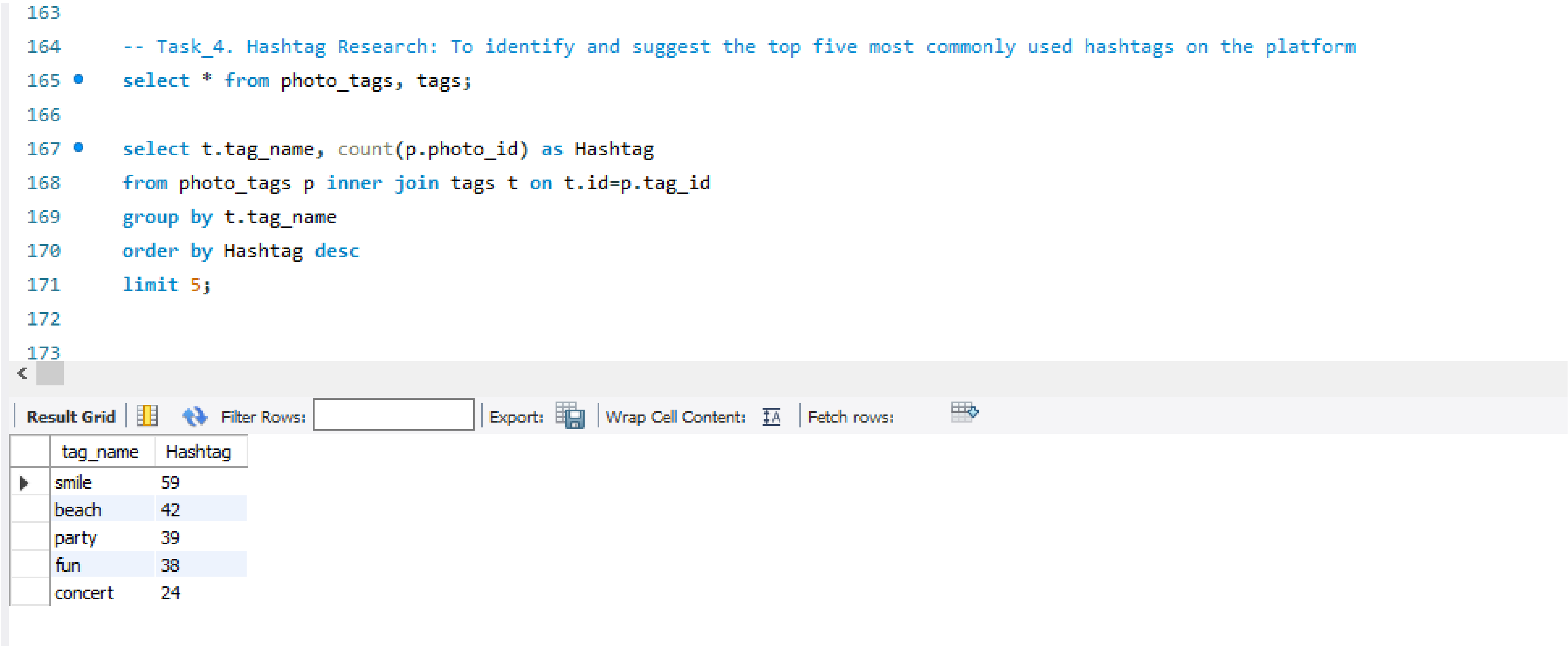
```sql
-- Task_3 Contest Winner Declaration: To Determine the winner of the contest based on most likes on single photo.
select * from users, photos, likes;
SELECT users.username, photos.id AS photo_id, COUNT(likes.user_id) AS Total_likes_count
FROM users, photos
LEFT JOIN likes ON photos.id = likes.photo_id
WHERE photos.user_id = users.id
GROUP BY photos.id, users.username
ORDER BY Total_likes_count DESC
LIMIT 7;
```

username	photo_id	Total_likes_count
Zack_Kemmer93	145	48
Adelle96	182	43
Malinda_Streich	127	43
Seth46	123	42
Presley_McClure	30	41
Delpha.Kihn	61	41
Annalise.McKenzie16	52	41

Here we have the top 7 users who has most likes on their single photo with respective count of their likes.

Amoung these 7 users The Zack_Kemmer93 has highest likes. So we can say that he must be the contest winner for having highest likes.

4. Hashtag Research: To identify and suggest the top five most commonly used hashtags on the platform.

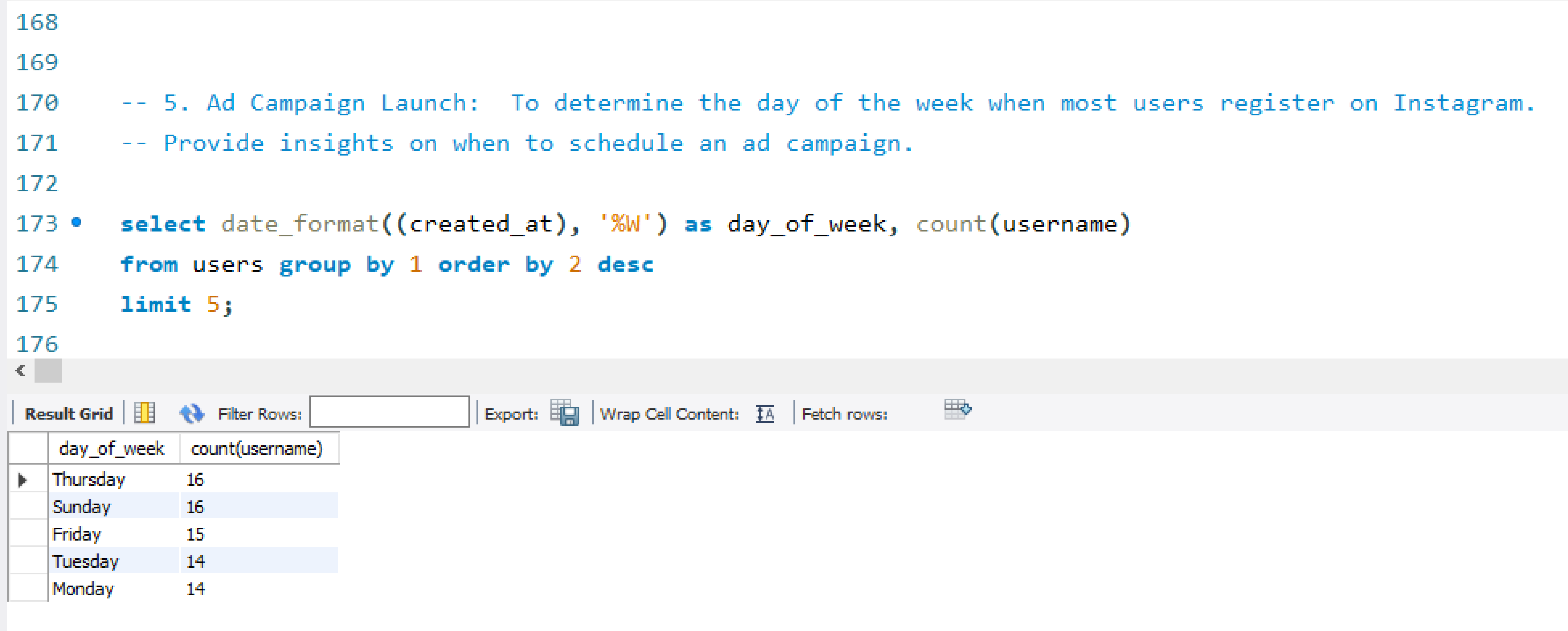


```sql
-- Task_4. Hashtag Research: To identify and suggest the top five most commonly used hashtags on the platform
select * from photo_tags, tags;

select t.tag_name, count(p.photo_id) as Hashtag
from photo_tags p inner join tags t on t.id=p.tag_id
group by t.tag_name
order by Hashtag desc
limit 5;
```

tag_name	Hashtag
smile	59
beach	42
party	39
fun	38
concert	24

The above result gives the top five mostly used hashtags, which are smile, beach, party, fun, and concert and their respective counts.

5. Ad Campaign Launch: To determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.



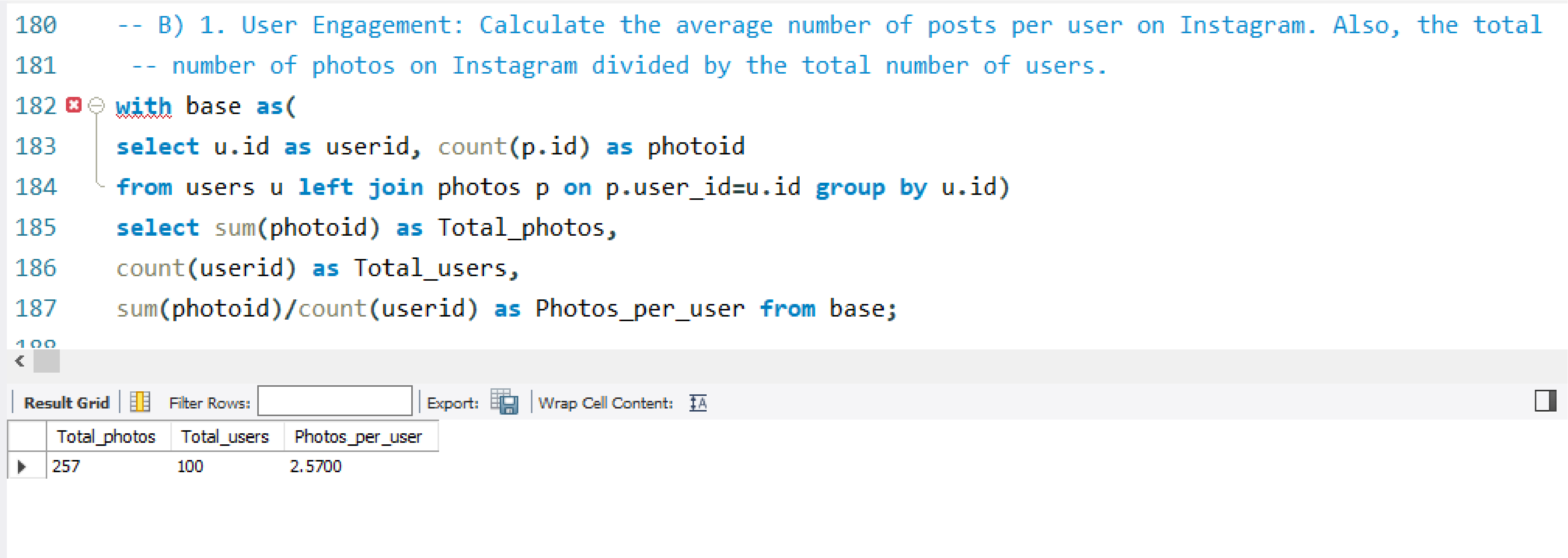
```sql


-- 5. Ad Campaign Launch:  To determine the day of the week when most users register on Instagram.
-- Provide insights on when to schedule an ad campaign.

select date_format((created_at), '%W') as day_of_week, count(username)
from users group by 1 order by 2 desc
limit 5;

```

day_of_week	count(username)
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14

From the above result we can say that among the all days Thursday and Sunday are the days on which the registration is higher. And hence we can schedule an ad campaign on both these days.

B) Investor Metrics: Insights

1. User Engagement: Calculate the average number of posts per user on Instagram. Also, the total number of photos on Instagram divided by the total number of users.

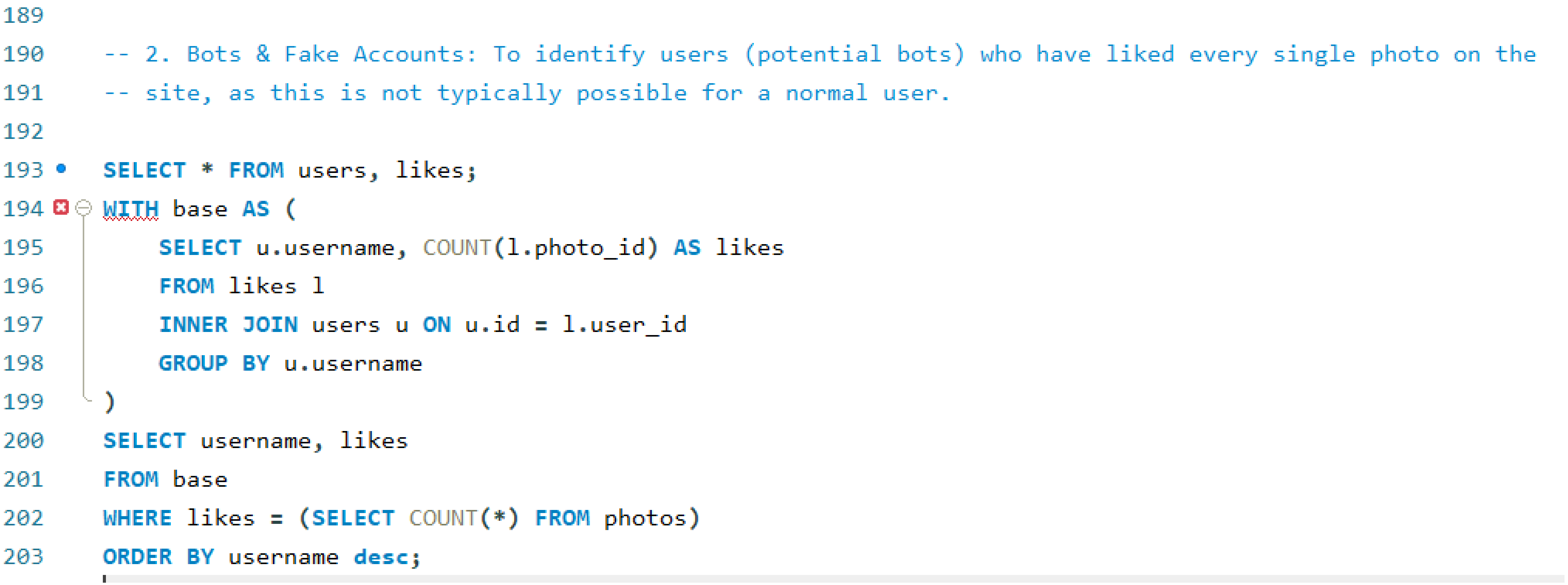


```
-- B) 1. User Engagement: Calculate the average number of posts per user on Instagram. Also, the total
 -- number of photos on Instagram divided by the total number of users.
with base as(
select u.id as userid, count(p.id) as photoid
from users u left join photos p on p.user_id=u.id group by u.id)
select sum(photoid) as Total_photos,
count(userid) as Total_users,
sum(photoid)/count(userid) as Photos_per_user from base;
```

Total_photos	Total_users	Photos_per_user
257	100	2.5700

From above result we can see that there are total 257 photos have been uploaded by100 users. So the photos per user would be 2.57

2. Bots & Fake Accounts: To identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.



```sql

-- 2. Bots & Fake Accounts: To identify users (potential bots) who have liked every single photo on the
-- site, as this is not typically possible for a normal user.

SELECT * FROM users, likes;
WITH base AS (
    SELECT u.username, COUNT(l.photo_id) AS likes
    FROM likes l
    INNER JOIN users u ON u.id = l.user_id
    GROUP BY u.username
)
SELECT username, likes
FROM base
WHERE likes = (SELECT COUNT(*) FROM photos)
ORDER BY username desc;
```

From above result we can say, there are total 13 users, who are fake based on their number of likes on photos.

Results

- I found this project very helpful to better understand, how to work on data and get useful insights from it using MY SQL.
- All the requirements got satisfied and the insights which drawn from it are as follows:

A) Marketing Anlysis:

1. We identified the top five oldest users to consider for the Loyal User Reward program.
2. There are 26 users who have never posted a single photo on the platform, and we plan to encourage them to start posting by sending promotional emails.
3. The user "Zack_Kemmer93" has the highest number of likes among the top 7 users, making him the contest winner for the highest likes.
4. The top five mostly used hashtags are "smile," "beach," "party," "fun," and "concert," along with their respective counts.
5. Thursday and Sunday have the highest registration rates, suggesting that scheduling ad campaigns on these days may lead to higher user acquisition.

A) Investor Metrics:

1. A total of 257 photos have been uploaded by 100 users.
2. We have identified 13 users who are potentially fake based on their patterns of liking photos.

Conclusion

1. Based on the insights and results we have got in analysis, we have a clear understanding of user behavior and platform metrics.
2. We can use this information to shape our marketing strategies, content creation, and ad campaign scheduling.
3. Additionally, we have taken steps to address fake users, ensuring the integrity of our user community. These efforts are aimed at driving the growth and success of our platform.

2. Operation Analytics and Investigating Metric Spike Analysis.

- Project Description

- This project mainly focuses on analyzing job data provided by a specific company with the aim of assisting the company in identifying areas for improvement. The objective is to uncover solutions to specific questions that can provide valuable insights to the operations, support, and marketing teams based on the gathered data.

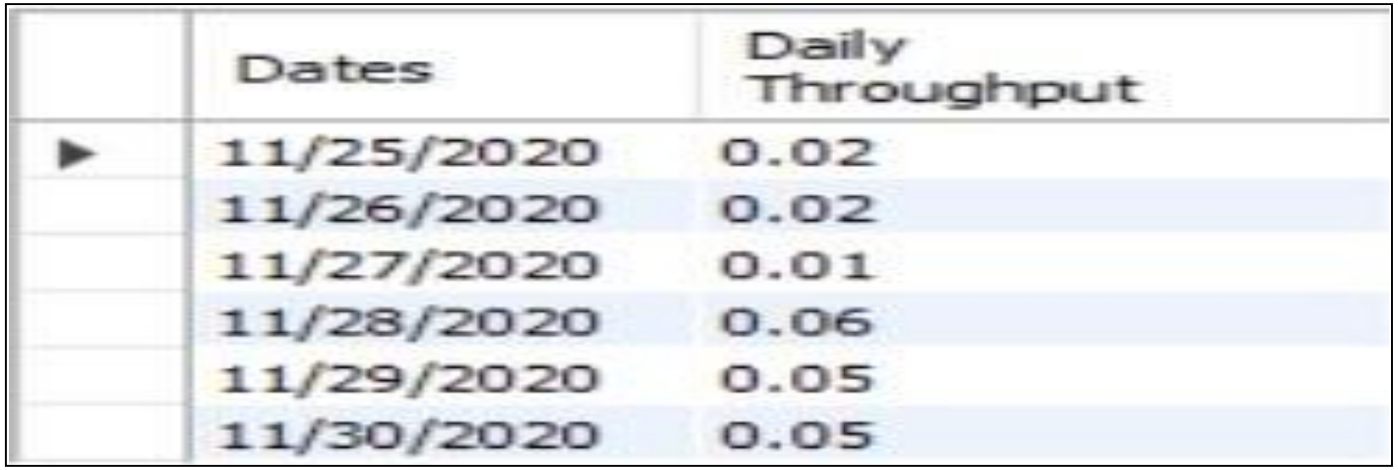
- Operational Analysis: Operational analysis involves the utilization of real-time data to inform the company's daily decision-making processes. This entails comprehending and elucidating abrupt shifts in critical performance indicators, such as drop in daily user engagement or a reduction in sales.

Case Study 1: Job Data Analysis

- 1. To determine the job review rate per hour for each day in November 2020,
- Ans: It was noticed that the highest number of jobs reviewed, reaching a peak of 218, occurred on November 28, 2020.

2. Througput Analysis:

- To Calculate the 7-day rolling average of throughput (number of events per second). To check whether prefer using the daily metric or the 7-day rolling average for throughput, and why?
    - Ans: The 7-day rolling average for throughput currently stands at 0.03.
    - When employing the daily metric, the highest throughput, at 0.06, occurred on November 28, 2020.
- .

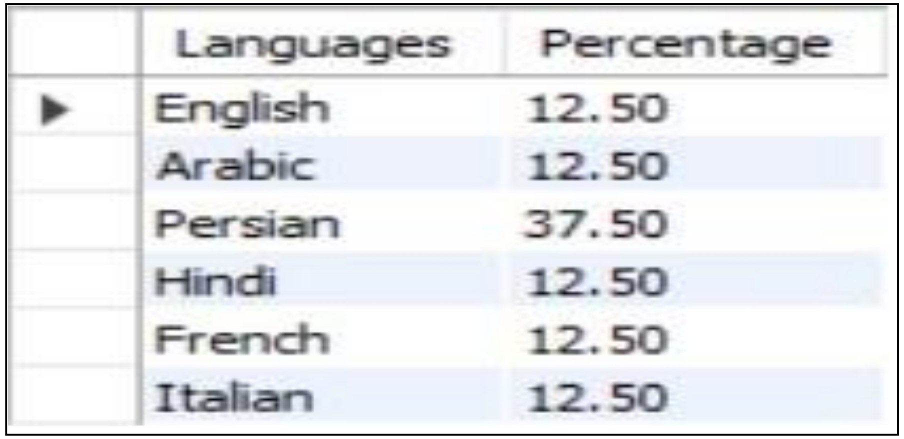


	Dates	Daily Throughput
▶	11/25/2020	0.02
	11/26/2020	0.02
	11/27/2020	0.01
	11/28/2020	0.06
	11/29/2020	0.05
	11/30/2020	0.05

3. Language Share Analysis:

To Calculate the percentage share of each language in the last 30 days.

The percentage share for each language in last 30 days is as below table:

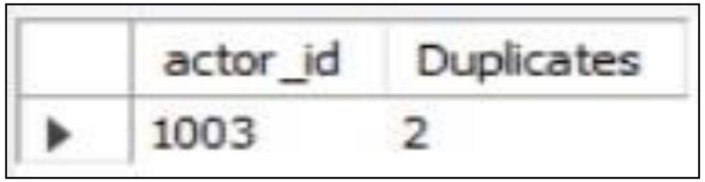


	Languages	Percentage
▶	English	12.50
	Arabic	12.50
	Persian	37.50
	Hindi	12.50
	French	12.50
	Italian	12.50

4. Duplicate Rows Detection:

To identify duplicate rows in the data.

It is observed that only single row is duplicated in the whole data.

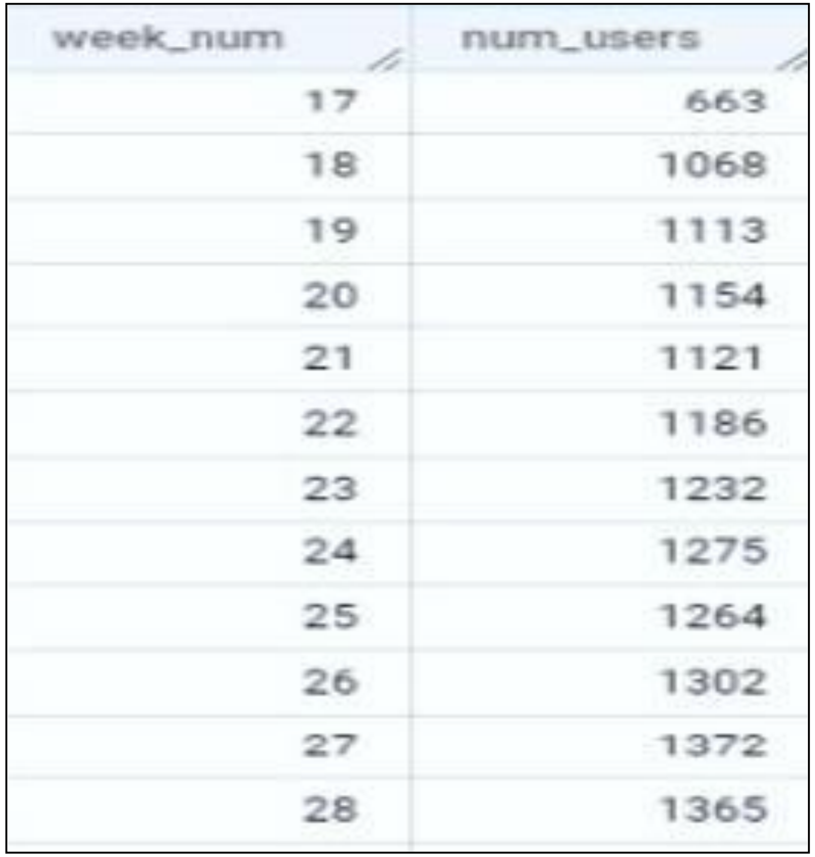


	actor_id	Duplicates
▶	1003	2

Case Study 2: Investigating Metric Spike

- **1. Weekly User Engagement**:
  - To measure the activeness of users on a weekly basis.

➢ The activeness of the users on weekly basis is as follows:

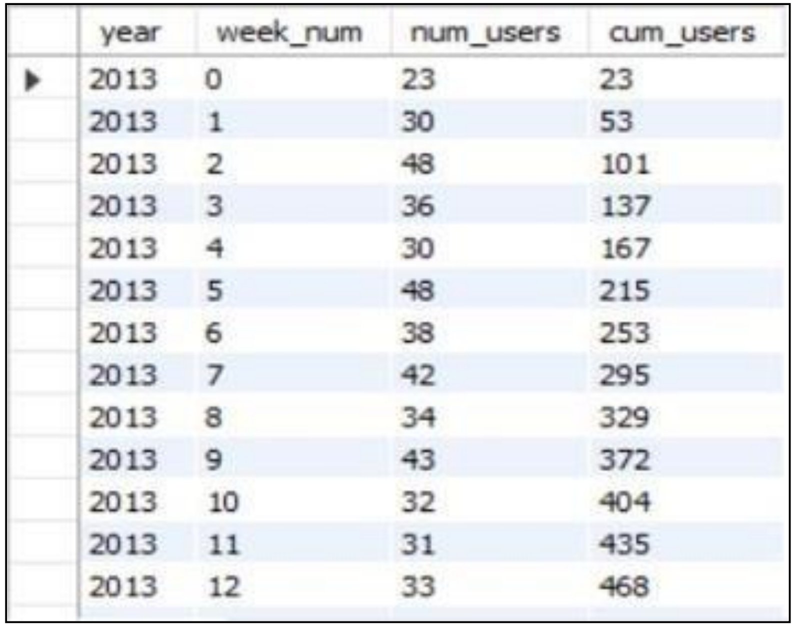


week_num	num_users
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365

2. User Growth Analysis:

Analyze the growth of users over time for a product.

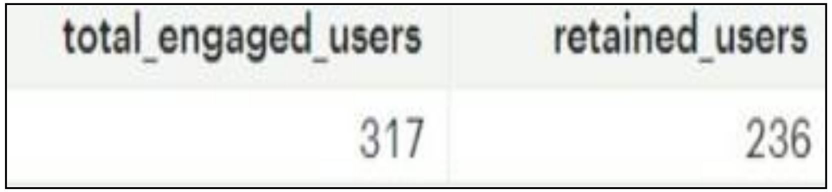
The observed growth of users over time for a product is:



	year	week_num	num_users	cum_users
▶	2013	0	23	23
	2013	1	30	53
	2013	2	48	101
	2013	3	36	137
	2013	4	30	167
	2013	5	48	215
	2013	6	38	253
	2013	7	42	295
	2013	8	34	329
	2013	9	43	372
	2013	10	32	404
	2013	11	31	435
	2013	12	33	468

3. Weekly Retention Analysis:

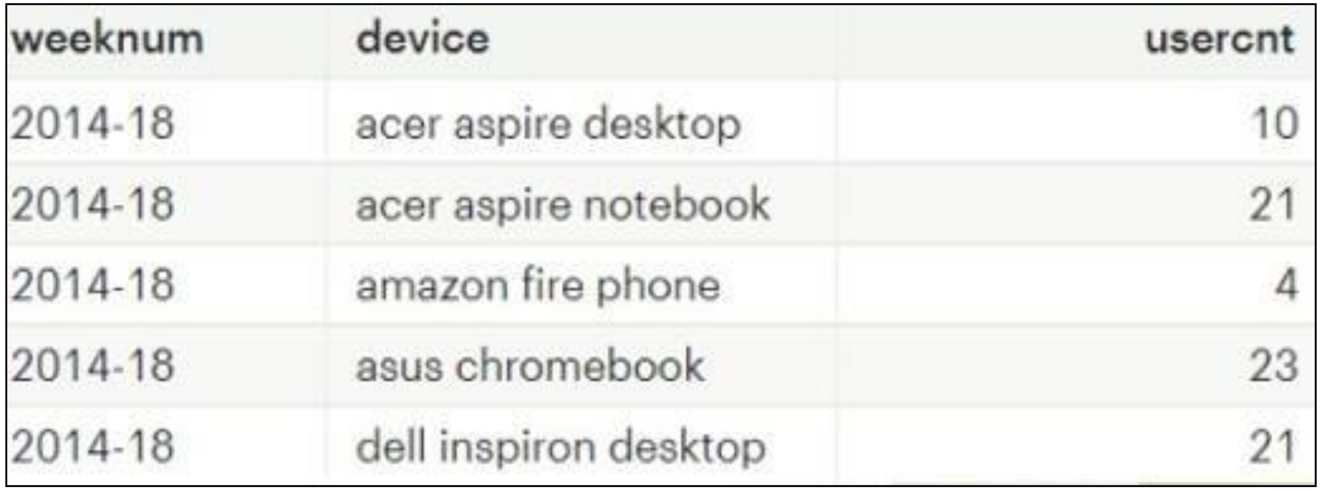
Analyze the retention of users on a weekly basis after signing up for a product.



total_engaged_users	retained_users
317	236

4. Weekly Engagement per Device:

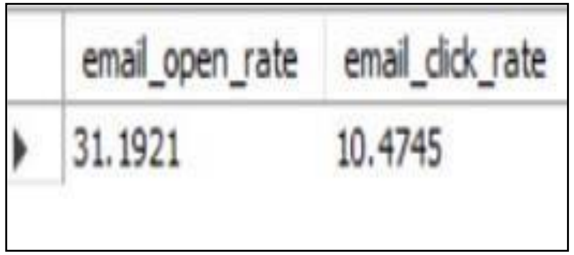
To Measure the activeness of users on a weekly basis per device.



weeknum	device	usercnt
2014-18	acer aspire desktop	10
2014-18	acer aspire notebook	21
2014-18	amazon fire phone	4
2014-18	asus chromebook	23
2014-18	dell inspiron desktop	21

5. Email Engagement Analysis:

To Analyze how users are engaging with the email service.



	email_open_rate	email_click_rate
▶	31.1921	10.4745

Result

i. On November 28, 2020, the highest number of job reviews recorded was 218.

ii. When evaluating throughput, a 7-day rolling average of 0.03 is preferred.

iii. It provides a more comprehensive representation, as opposed to the daily metric, which reached its peak at 0.06 on November 28, 2020. Persian language stands out with the largest share, accounting for 37.5% of all languages used.

iv. The 31st week stands as the pinnacle of weekly user engagement.

v. Notably, the 33rd week of 2014 witnessed the highest user engagement, while the lowest engagement was observed during the 35th week of the same year.

vi. The month of August consistently records the highest number of weekly digest emails received by users.

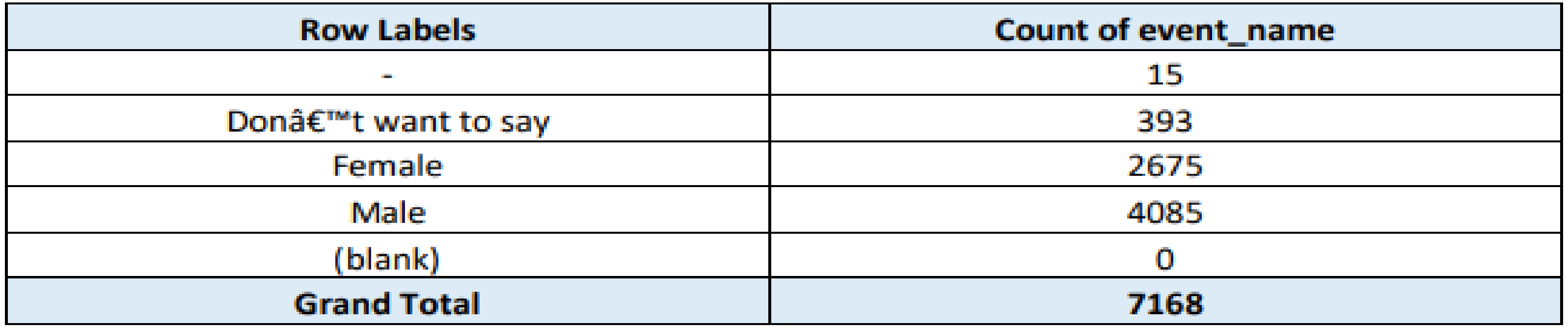
3. Hiring Process Analytics

Project Description

- The Hiring Process is process of in taking of people into an organization for various kinds of positions. The hiring process is a crucial function of any company. The primary objective revolves around grasping patterns such as the quantity of declined candidates, interviews conducted, job categories, and available job openings, all of which can furnish valuable discernments for the recruitment department.

Insights

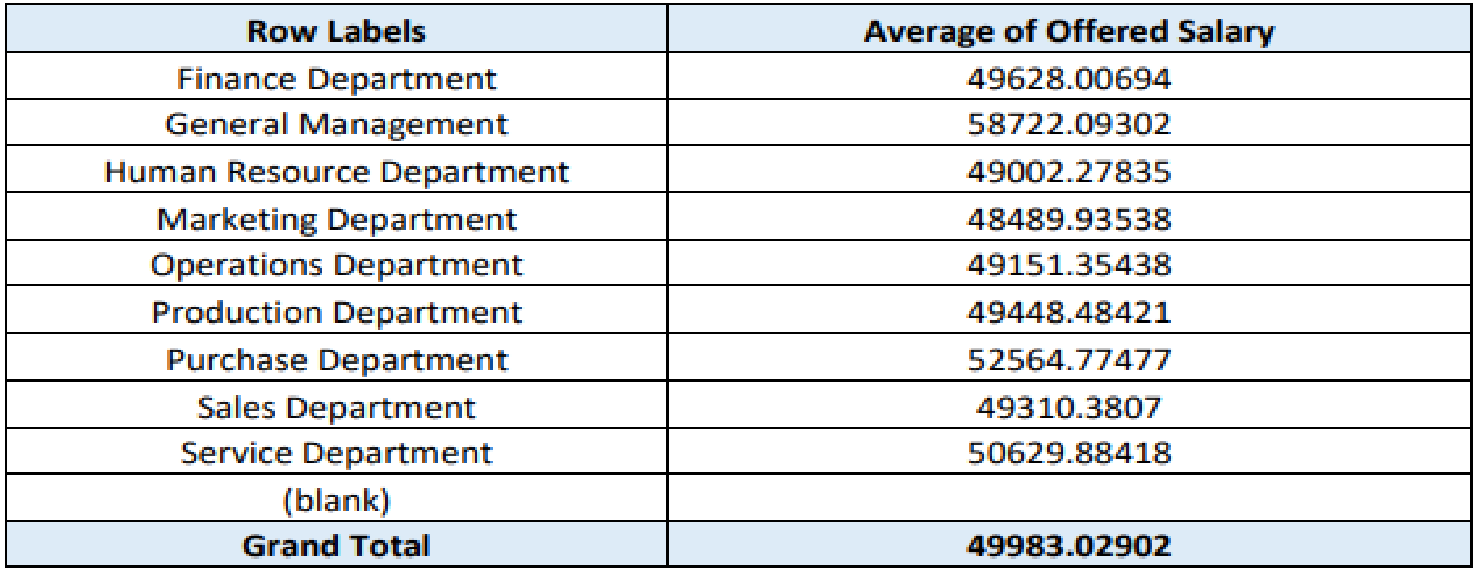
- A. Hiring Analysis: To determine the gender distribution of hires. To determine how many males and females have been hired by the company.



Row Labels	Count of event_name
-	15
Donâ€™t want to say	393
Female	2675
Male	4085
(blank)	0
Grand Total	**7168**

It suggest that out of total candidates 2675 females while 4085 males were hired by the company.

B. Salary Analysis: To determine the average salary offered by this company



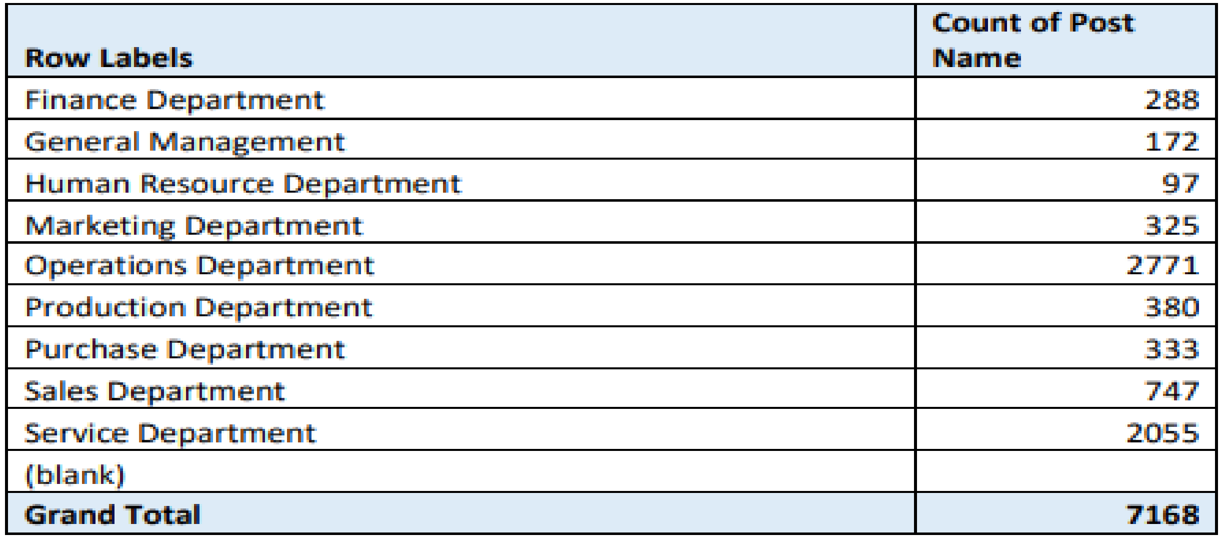
Row Labels	Average of Offered Salary
Finance Department	49628.00694
General Management	58722.09302
Human Resource Department	49002.27835
Marketing Department	48489.93538
Operations Department	49151.35438
Production Department	49448.48421
Purchase Department	52564.77477
Sales Department	49310.3807
Service Department	50629.88418
(blank)	
Grand Total	**49983.02902**

The overall average salary is **49983.02902**

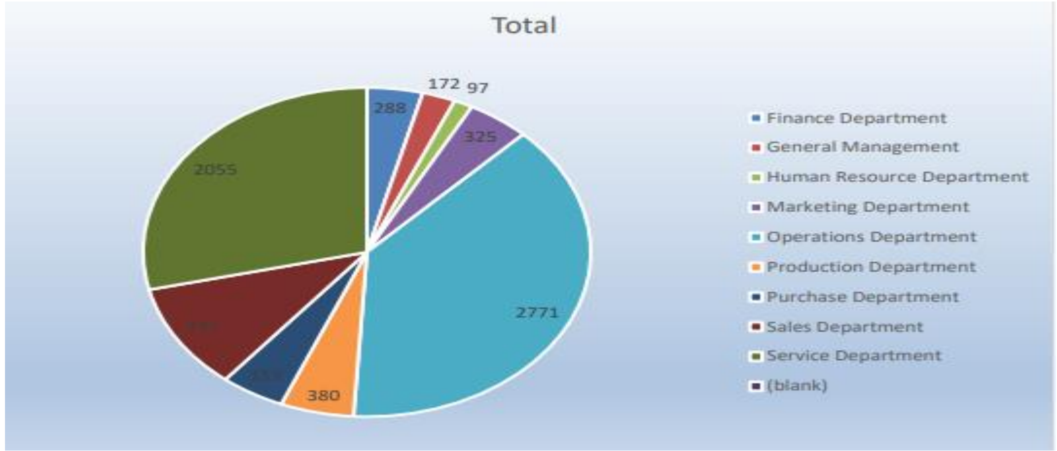
C. Salary Distribution: To Create class intervals for the salaries in the company. This will help you understand the salary distribution.

Class Interval for Salary	No. of Employees
1-50000	3613
50001-100000	3552
100001-150000	0
150001-200000	1
200001-250000	0
250001-300000	1
300001-350000	0
350001-400000	1

D. Departmental Analysis: To use the charts to show the proportion of people working in different departments.

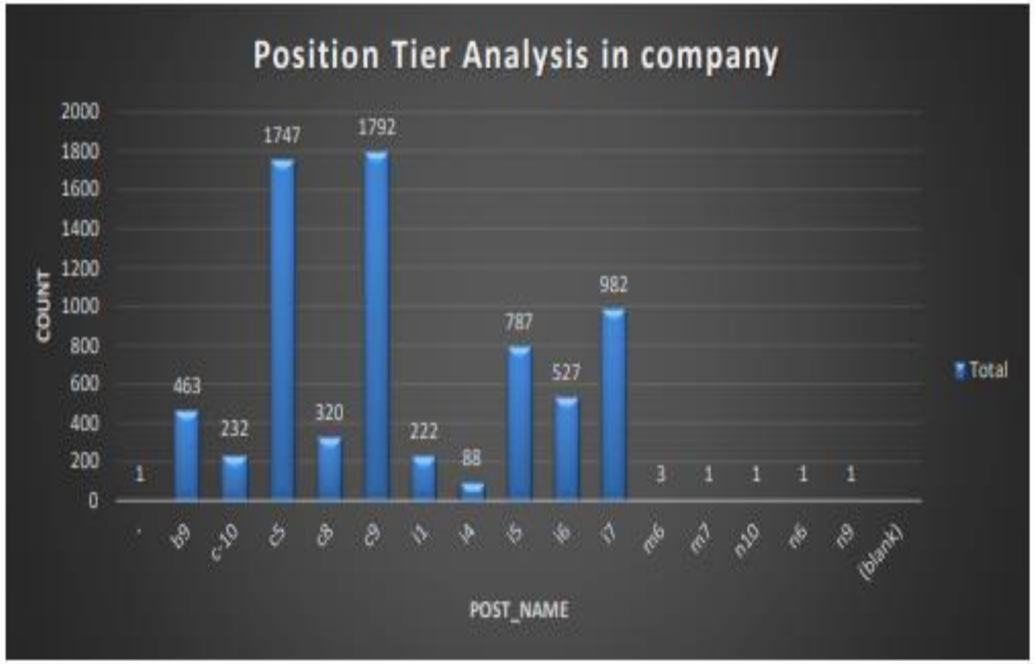


Row Labels	Count of Post Name
Finance Department	288
General Management	172
Human Resource Department	97
Marketing Department	325
Operations Department	2771
Production Department	380
Purchase Department	333
Sales Department	747
Service Department	2055
(blank)	
Grand Total	**7168**



E. Position Tier Analysis: To use suitable visualizations to represent the different position tiers within the company.

Row Labels	Count of Post Name
-	1
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1
(blank)	
Grand Total	**7168**



Results

- 1. From the total employees, 2675 females while 4085 males are hired.
- 2. The General Management Department has the highest salary 58722.09302 and the Marketingdepartment has lowest salary.
- 3. Most of the employees are in the salary interval 1-50000. 4. Most of the Employees are in the post tier of c9 i.e. 1792.

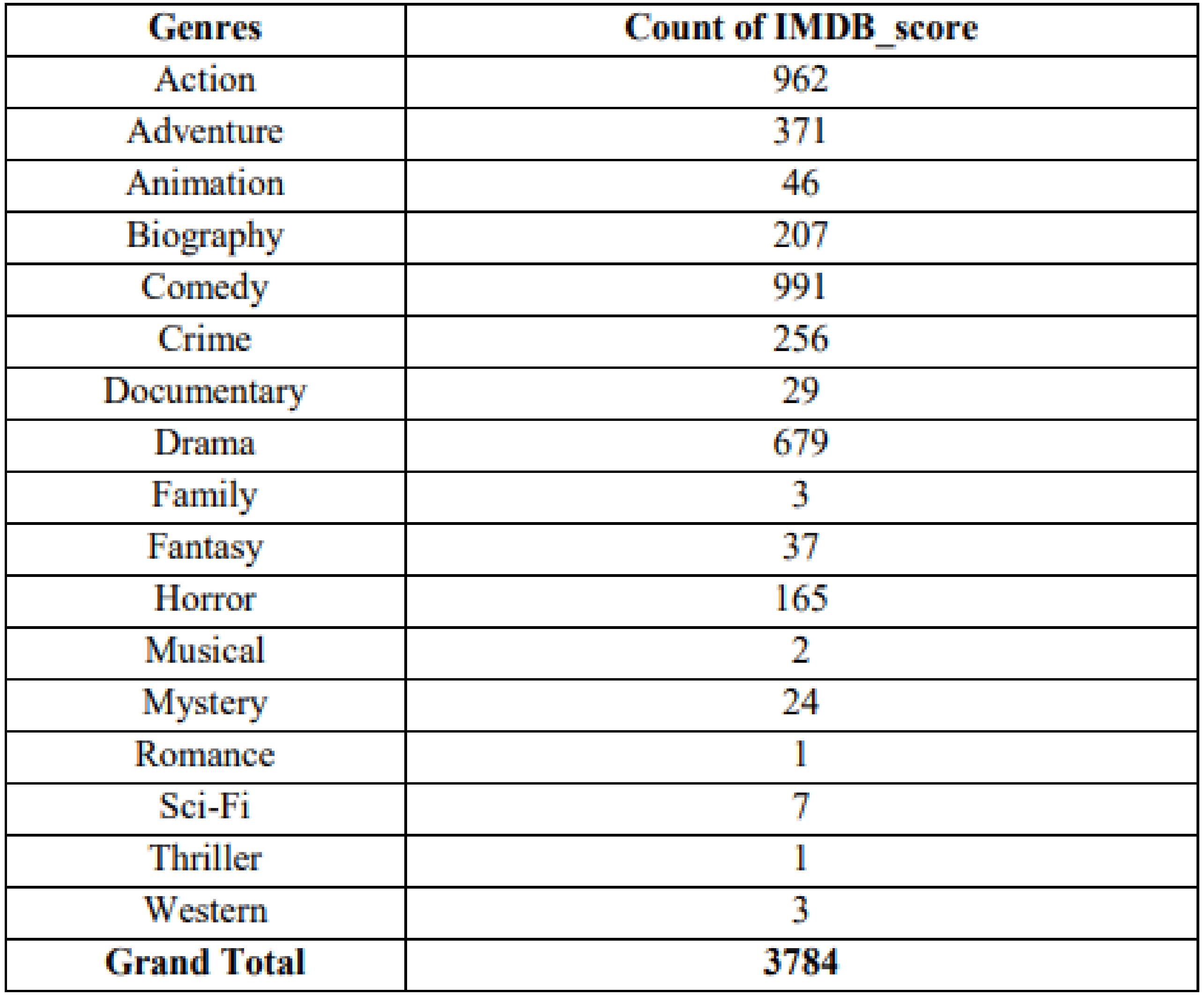
4. IMDB Movie Analysis

- Project Description

The Dataset provided for analysis contains information related to IMDB Movies. The success of any movie is determined by its IMDB ratings. In this project my role is to analyze what factors are affecting the success of a movie on IMDB which will help Movie Producers, Directors and Investors to take their decision for future Projects.

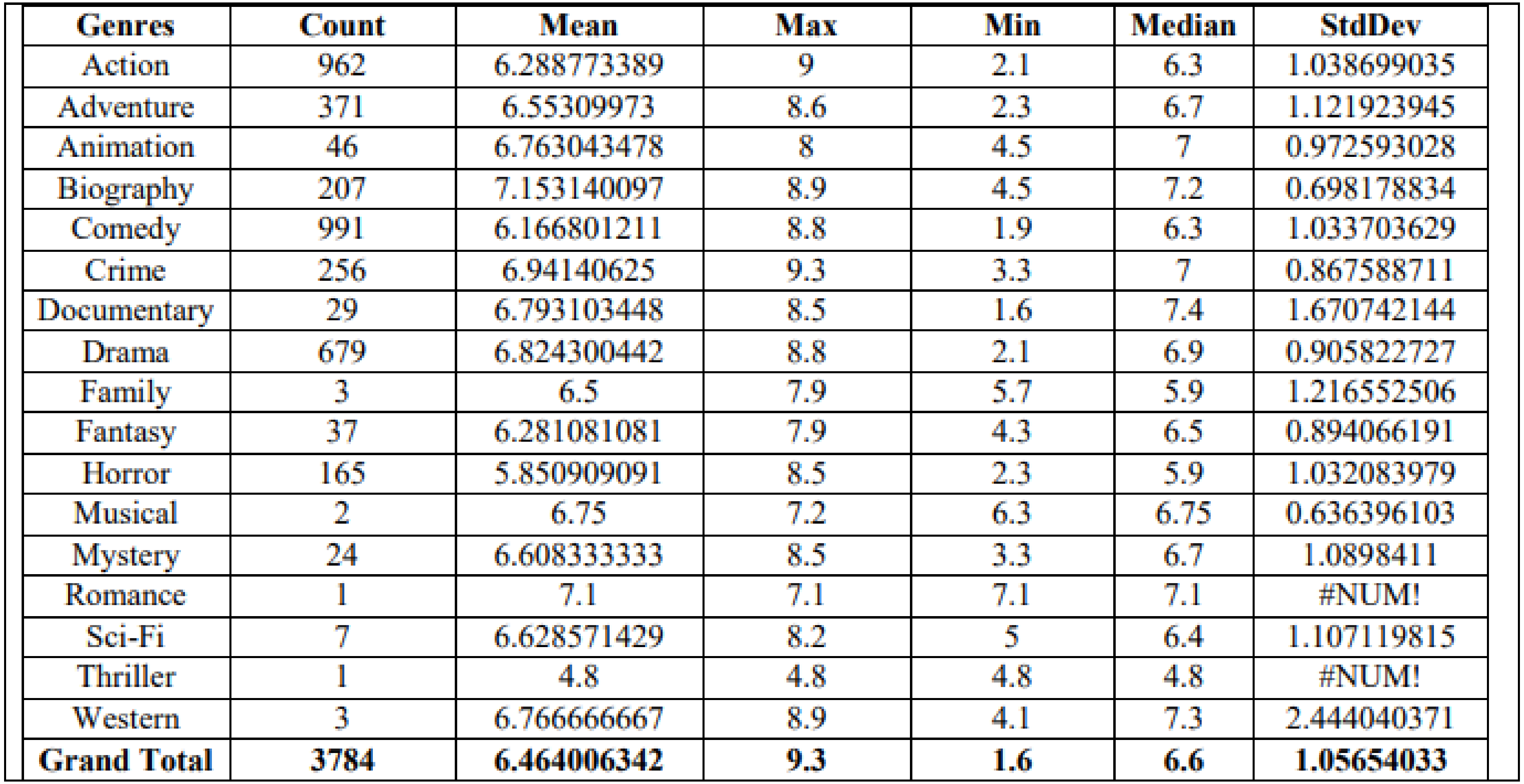
Insights

TASK 1: Movie Genre Analysis To determine most common genres of movie in the dataset and calculate descriptive statistics (mean, median, range, variance, standard deviation)



Genres	Count of IMDB_score
Action	962
Adventure	371
Animation	46
Biography	207
Comedy	991
Crime	256
Documentary	29
Drama	679
Family	3
Fantasy	37
Horror	165
Musical	2
Mystery	24
Romance	1
Sci-Fi	7
Thriller	1
Western	3
Grand Total	**3784**

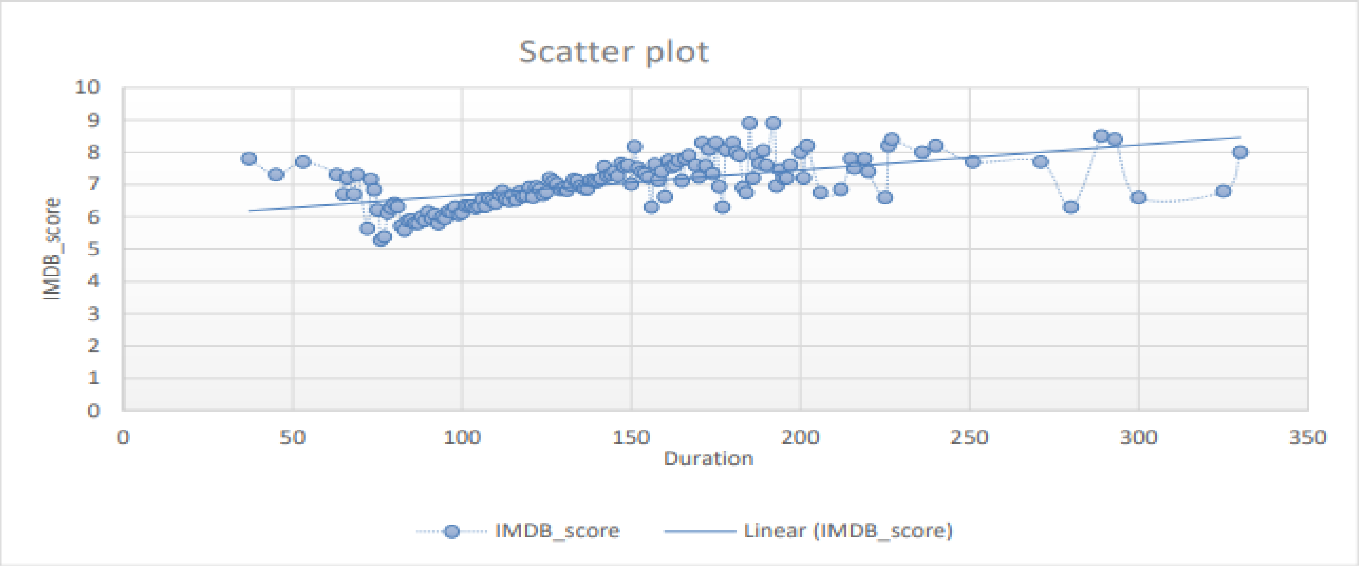
Genre wise Descriptive Statistics of IMDB Scores



Genres	Count	Mean	Max	Min	Median	StdDev
Action	962	6.288773389	9	2.1	6.3	1.038699035
Adventure	371	6.55309973	8.6	2.3	6.7	1.121923945
Animation	46	6.763043478	8	4.5	7	0.972593028
Biography	207	7.153140097	8.9	4.5	7.2	0.698178834
Comedy	991	6.166801211	8.8	1.9	6.3	1.033703629
Crime	256	6.94140625	9.3	3.3	7	0.867588711
Documentary	29	6.793103448	8.5	1.6	7.4	1.670742144
Drama	679	6.824300442	8.8	2.1	6.9	0.905822727
Family	3	6.5	7.9	5.7	5.9	1.216552506
Fantasy	37	6.281081081	7.9	4.3	6.5	0.894066191
Horror	165	5.850909091	8.5	2.3	5.9	1.032083979
Musical	2	6.75	7.2	6.3	6.75	0.636396103
Mystery	24	6.608333333	8.5	3.3	6.7	1.0898411
Romance	1	7.1	7.1	7.1	7.1	#NUM!
Sci-Fi	7	6.628571429	8.2	5	6.4	1.107119815
Thriller	1	4.8	4.8	4.8	4.8	#NUM!
Western	3	6.766666667	8.9	4.1	7.3	2.444040371
Grand Total	**3784**	**6.464006342**	**9.3**	**1.6**	**6.6**	**1.05654033**

Task 2: Movie Duration Analysis

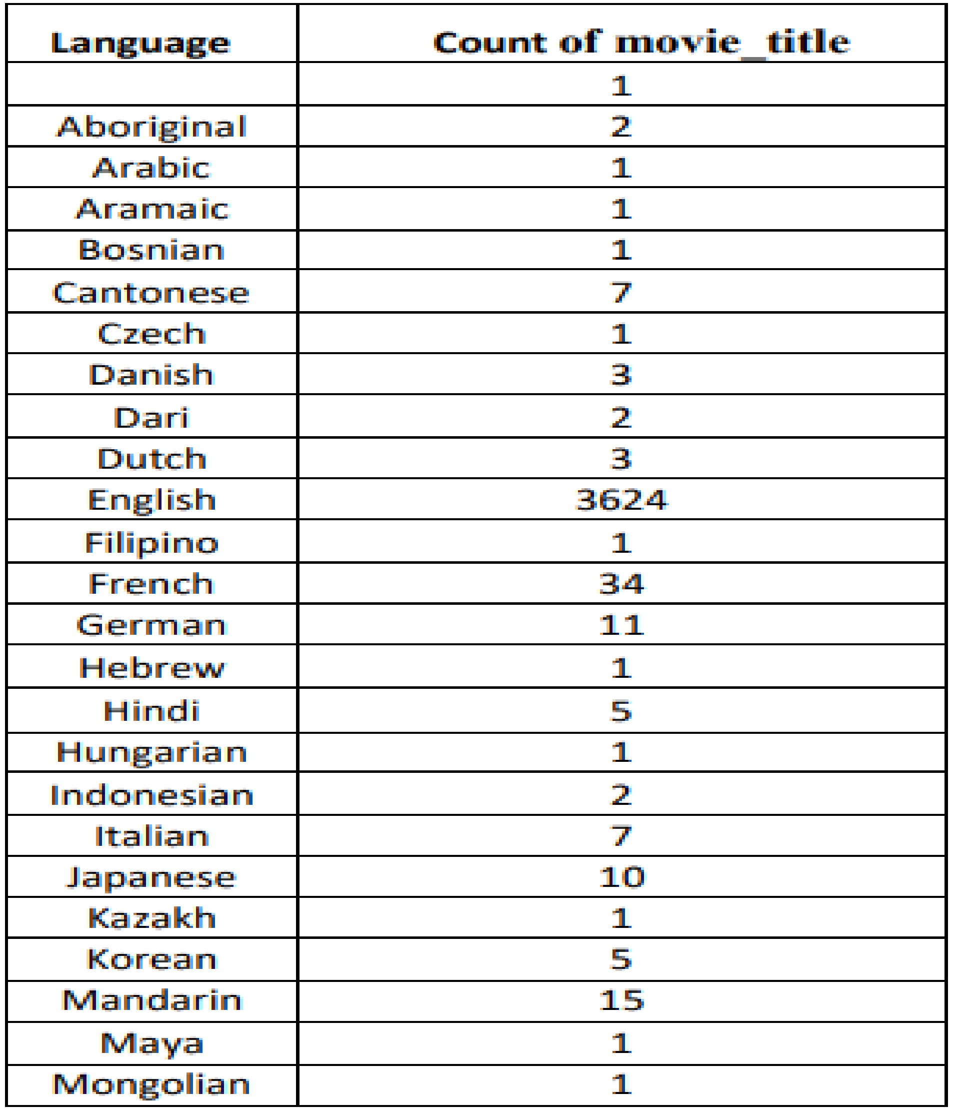
To Analyze distribution of movie durations and identify relationship between movie duration and IMDB score.



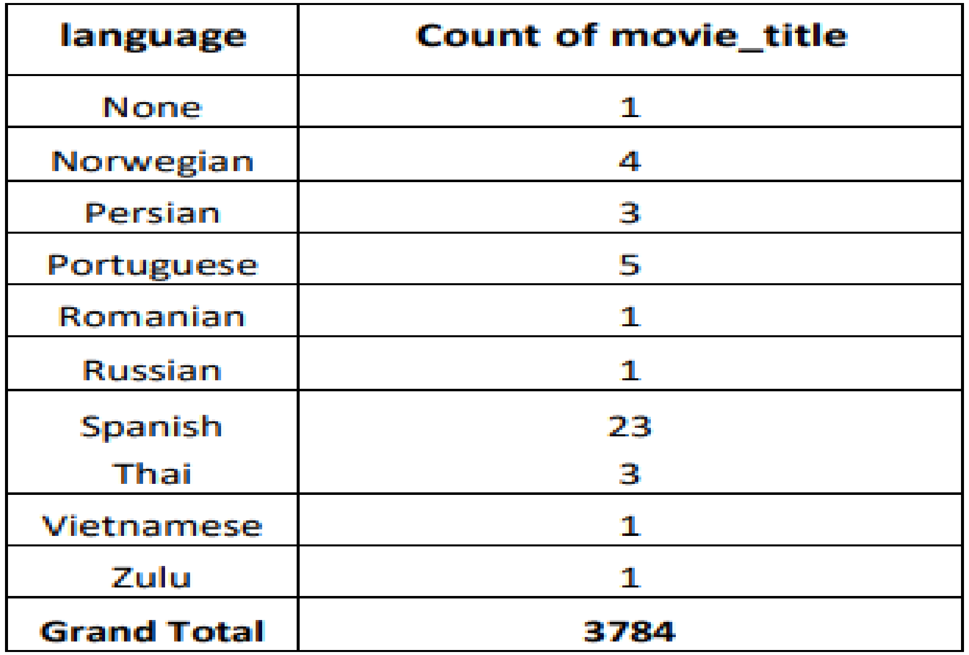
Mean	145.4837
Median	140
Standard Deviation	56.90598

Task 3: Language Analysis

To determine the most common languages used in movies and to analyze their impact on the IMDB score using descriptive statistics.

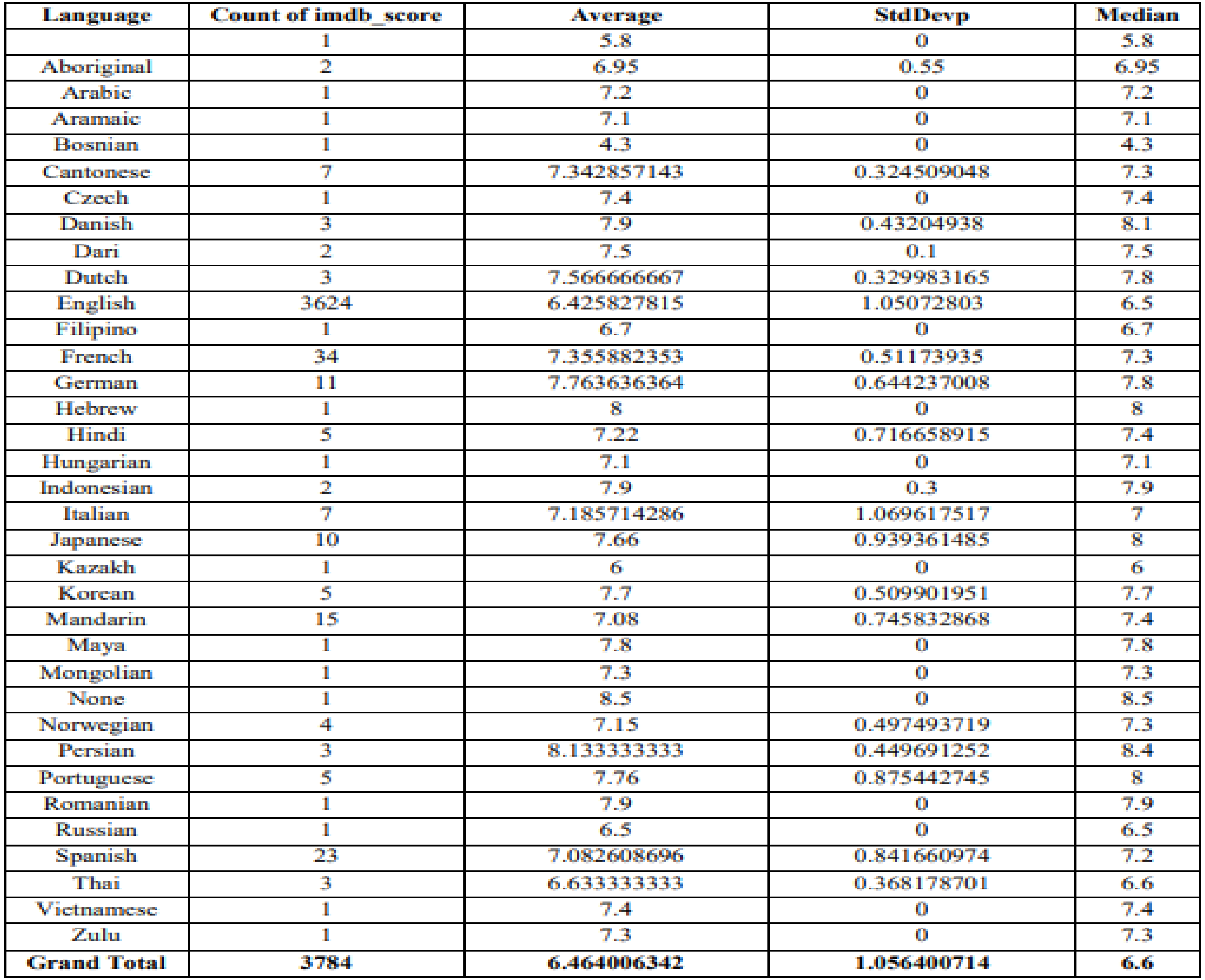


Language	Count of movie_title
	1
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	7
Czech	1
Danish	3
Dari	2
Dutch	3
English	3624
Filipino	1
French	34
German	11
Hebrew	1
Hindi	5
Hungarian	1
Indonesian	2
Italian	7
Japanese	10
Kazakh	1
Korean	5
Mandarin	15
Maya	1
Mongolian	1



language	Count of movie_title
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	23
Thai	3
Vietnamese	1
Zulu	1
Grand Total	**3784**

Language wise Descriptive Statistics of IMDB Score

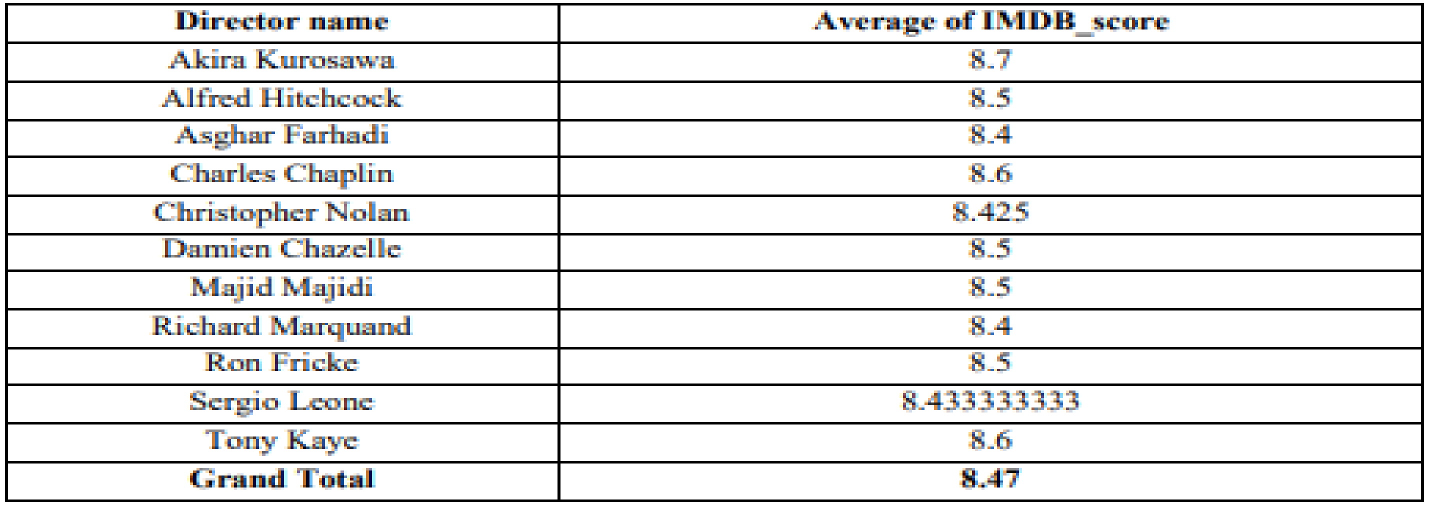


Language	Count of imdb_score	Average	StdDevp	Median
	1	5.8	0	5.8
Aboriginal	2	6.95	0.55	6.95
Arabic	1	7.2	0	7.2
Aramaic	1	7.1	0	7.1
Bosnian	1	4.3	0	4.3
Cantonese	7	7.342857143	0.324509048	7.3
Czech	1	7.4	0	7.4
Danish	3	7.9	0.43204938	8.1
Dari	2	7.5	0.1	7.5
Dutch	3	7.566666667	0.329983165	7.8
English	3624	6.425827815	1.05072803	6.5
Filipino	1	6.7	0	6.7
French	34	7.355882353	0.51173935	7.3
German	11	7.763636364	0.644237008	7.8
Hebrew	1	8	0	8
Hindi	5	7.22	0.716658915	7.4
Hungarian	1	7.1	0	7.1
Indonesian	2	7.9	0.3	7.9
Italian	7	7.185714286	1.069617517	7
Japanese	10	7.66	0.939361485	8
Kazakh	1	6	0	6
Korean	5	7.7	0.509901951	7.7
Mandarin	15	7.08	0.745832868	7.4
Maya	1	7.8	0	7.8
Mongolian	1	7.3	0	7.3
None	1	8.5	0	8.5
Norwegian	4	7.15	0.497493719	7.3
Persian	3	8.133333333	0.449691252	8.4
Portuguese	5	7.76	0.875442745	8
Romanian	1	7.9	0	7.9
Russian	1	6.5	0	6.5
Spanish	23	7.082608696	0.841660974	7.2
Thai	3	6.633333333	0.368178701	6.6
Vietnamese	1	7.4	0	7.4
Zulu	1	7.3	0	7.3
Grand Total	**3784**	**6.464006342**	**1.056400714**	**6.6**

Task 4: Director Analysis

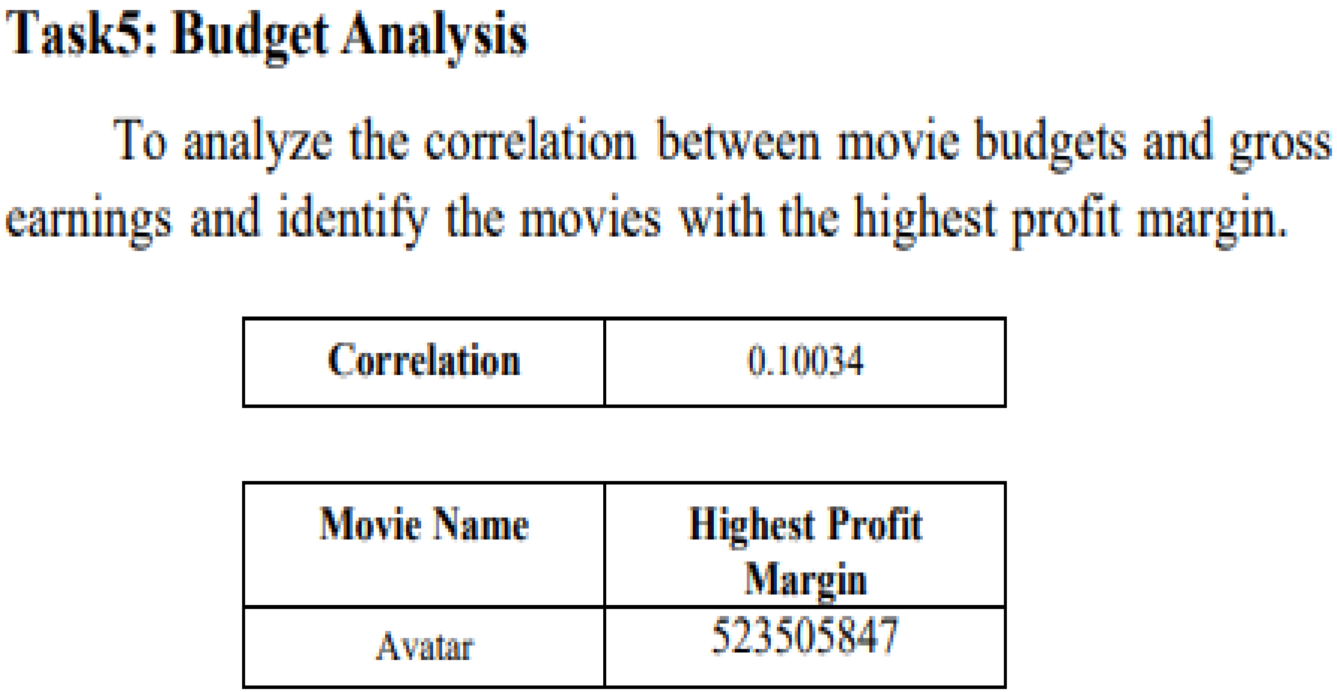
To identify top directors on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

10 Directors with highest average of IMDB Score



Director name	Average of IMDB_score
Akira Kurosawa	8.7
Alfred Hitchcock	8.5
Asghar Farhadi	8.4
Charles Chaplin	8.6
Christopher Nolan	8.425
Damien Chazelle	8.5
Majid Majidi	8.5
Richard Marquand	8.4
Ron Fricke	8.5
Sergio Leone	8.433333333
Tony Kaye	8.6
Grand Total	**8.47**

Percentile	1



Task5: Budget Analysis

To analyze the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

Correlation	0.10034

Movie Name	Highest Profit Margin
Avatar	523505847

Results

- Most common genre in the movie is Comedy having highest count of movies i.e. 991
- Film-noir genre has highest average IMDB score ie.7.6 whereas Comedy genre has highest maximum value for IMDB score is 9.5 and Documentary has lowest min value for IMDB Score ie.1.6
- Average duration for the movie is 145.48 with median 140 and std deviation is 56.90
- Scatter plot shows that duration 185 has highest IMDB score ie. 9 while duration 76 has lowest ie.5.28 & it kept fluctuating between them.
- Most common language used in the movie is English with highest count 3624
- Director Akira Kurosawa has highest average IMDB score which is 8.7
- Percentile is 8.7 and Percentilerank is 1 Directors have 100% success rate.
- Correlation value 0.10034 between movie budgets and ross earning shows that there is relatively week positive linear relation between them.
- Movie Avatar has highest profit margin ie. 523505847.

5. Bank Loan Case Study

- Project Description

- This case study attempts to analyze patterns in the data using Exploratory Data Analysis(EDA) and ensure that capable applicants are not rejected. The main Aim of study is to identify patterns that indicates if a customer will have difficulty paying their installments which can be used to make decision such as denying the loan, reducing the amount of loan or lending at a higher interest rate to risky applicants

- Insights

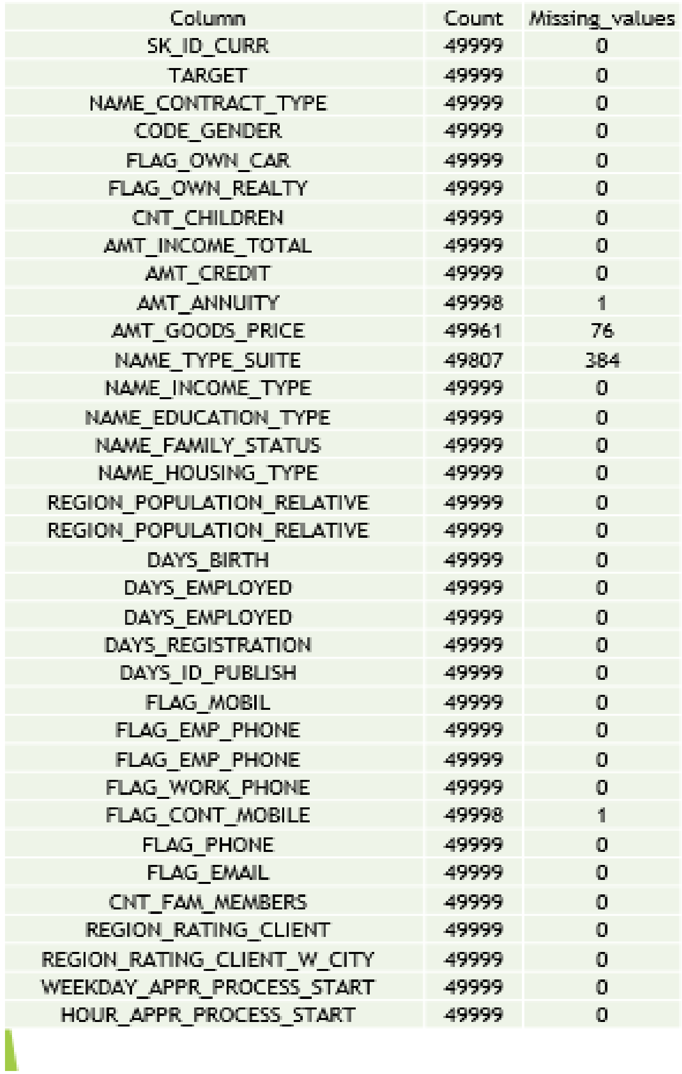
Task : A

To Identify Missing Data & Deal with it appropriately.

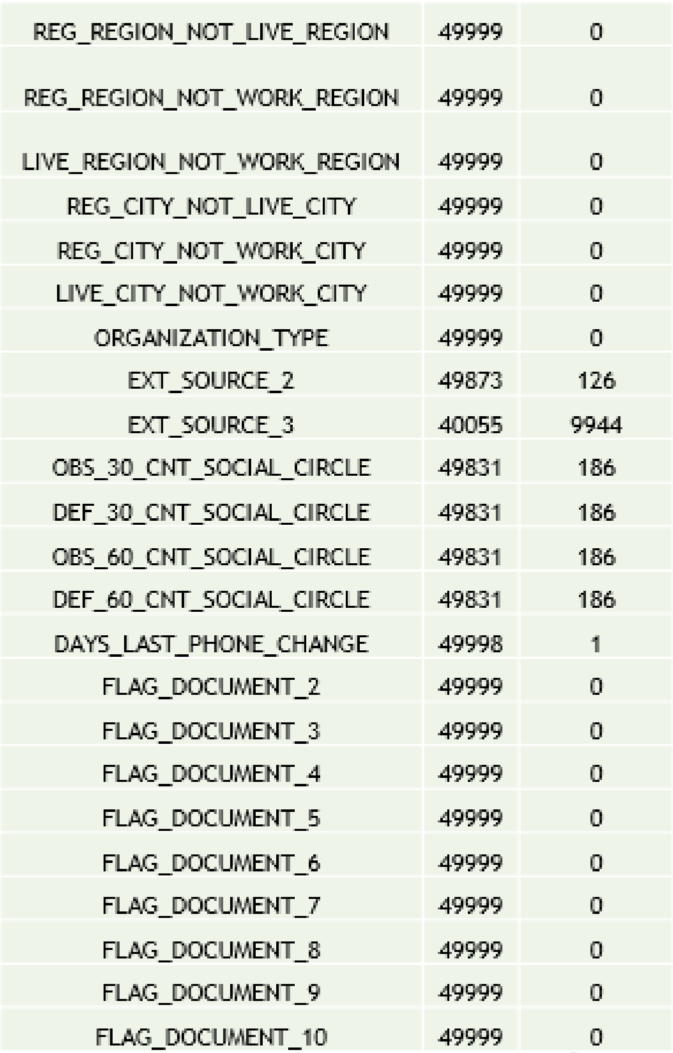
To Identify missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Ans:

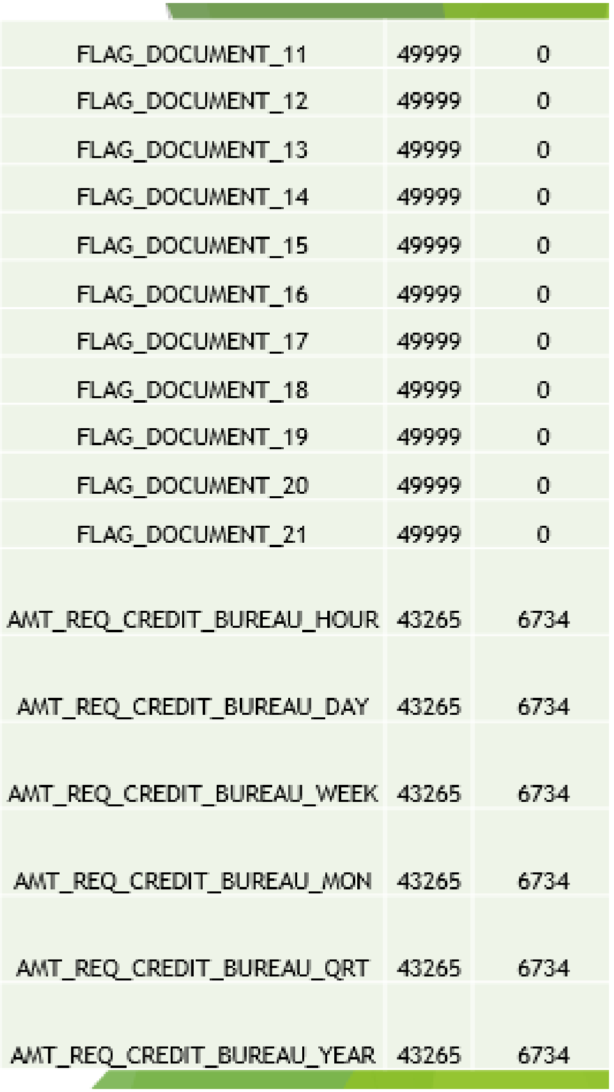
Initially there were 123 columns and 49999 rows in the data. To deal with missing data I firstly calculate null values percentage on the basis of that I removed the column having null value percentage greater than 30%, also there were some undesirable columns so I delete those columns too. Then for the remaining columns I used median method to fill missing values for numerical columns. Also for better analysis I converted DAYS_BIRTH as DAYS_BIRTH(YEAR), DAYS_EMPLOYED as DAYS_EMPLOYED(YEAR), DAYS_REGISTRATION as DAYS_REGISTRATION (YEAR), DAYS_ID_PUBLISH as DAYS_ID_PUBLISH(YEAR).



Column	Count	Missing_values
SK_ID_CURR	49999	0
TARGET	49999	0
NAME_CONTRACT_TYPE	49999	0
CODE_GENDER	49999	0
FLAG_OWN_CAR	49999	0
FLAG_OWN_REALTY	49999	0
CNT_CHILDREN	49999	0
AMT_INCOME_TOTAL	49999	0
AMT_CREDIT	49999	0
AMT_ANNUITY	49998	1
AMT_GOODS_PRICE	49961	76
NAME_TYPE_SUITE	49807	384
NAME_INCOME_TYPE	49999	0
NAME_EDUCATION_TYPE	49999	0
NAME_FAMILY_STATUS	49999	0
NAME_HOUSING_TYPE	49999	0
REGION_POPULATION_RELATIVE	49999	0
REGION_POPULATION_RELATIVE	49999	0
DAYS_BIRTH	49999	0
DAYS_EMPLOYED	49999	0
DAYS_EMPLOYED	49999	0
DAYS_REGISTRATION	49999	0
DAYS_ID_PUBLISH	49999	0
FLAG_MOBIL	49999	0
FLAG_EMP_PHONE	49999	0
FLAG_EMP_PHONE	49999	0
FLAG_WORK_PHONE	49999	0
FLAG_CONT_MOBILE	49998	1
FLAG_PHONE	49999	0
FLAG_EMAIL	49999	0
CNT_FAM_MEMBERS	49999	0
REGION_RATING_CLIENT	49999	0
REGION_RATING_CLIENT_W_CITY	49999	0
WEEKDAY_APPR_PROCESS_START	49999	0
HOUR_APPR_PROCESS_START	49999	0



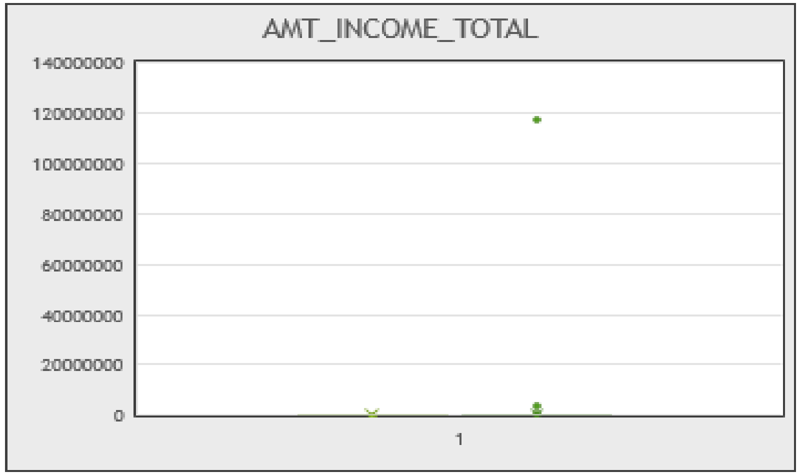
REG_REGION_NOT_LIVE_REGION	49999	0
REG_REGION_NOT_WORK_REGION	49999	0
LIVE_REGION_NOT_WORK_REGION	49999	0
REG_CITY_NOT_LIVE_CITY	49999	0
REG_CITY_NOT_WORK_CITY	49999	0
LIVE_CITY_NOT_WORK_CITY	49999	0
ORGANIZATION_TYPE	49999	0
EXT_SOURCE_2	49873	126
EXT_SOURCE_3	40055	9944
OBS_30_CNT_SOCIAL_CIRCLE	49831	186
DEF_30_CNT_SOCIAL_CIRCLE	49831	186
OBS_60_CNT_SOCIAL_CIRCLE	49831	186
DEF_60_CNT_SOCIAL_CIRCLE	49831	186
DAYS_LAST_PHONE_CHANGE	49998	1
FLAG_DOCUMENT_2	49999	0
FLAG_DOCUMENT_3	49999	0
FLAG_DOCUMENT_4	49999	0
FLAG_DOCUMENT_5	49999	0
FLAG_DOCUMENT_6	49999	0
FLAG_DOCUMENT_7	49999	0
FLAG_DOCUMENT_8	49999	0
FLAG_DOCUMENT_9	49999	0
FLAG_DOCUMENT_10	49999	0

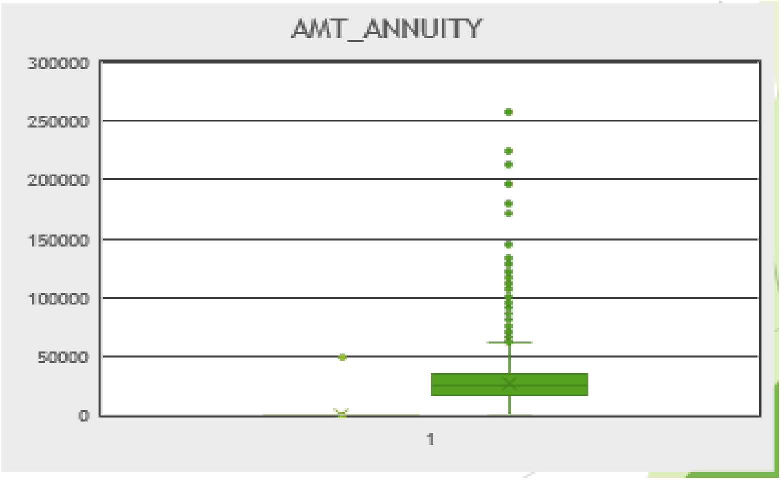


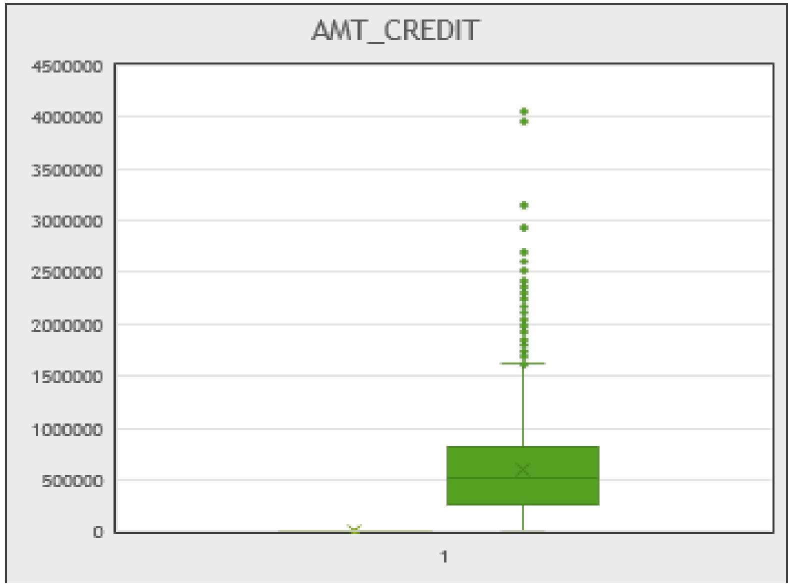
FLAG_DOCUMENT_11	49999	0
FLAG_DOCUMENT_12	49999	0
FLAG_DOCUMENT_13	49999	0
FLAG_DOCUMENT_14	49999	0
FLAG_DOCUMENT_15	49999	0
FLAG_DOCUMENT_16	49999	0
FLAG_DOCUMENT_17	49999	0
FLAG_DOCUMENT_18	49999	0
FLAG_DOCUMENT_19	49999	0
FLAG_DOCUMENT_20	49999	0
FLAG_DOCUMENT_21	49999	0
AMT_REQ_CREDIT_BUREAU_HOUR	43265	6734
AMT_REQ_CREDIT_BUREAU_DAY	43265	6734
AMT_REQ_CREDIT_BUREAU_WEEK	43265	6734
AMT_REQ_CREDIT_BUREAU_MON	43265	6734
AMT_REQ_CREDIT_BUREAU_QRT	43265	6734
AMT_REQ_CREDIT_BUREAU_YEAR	43265	6734

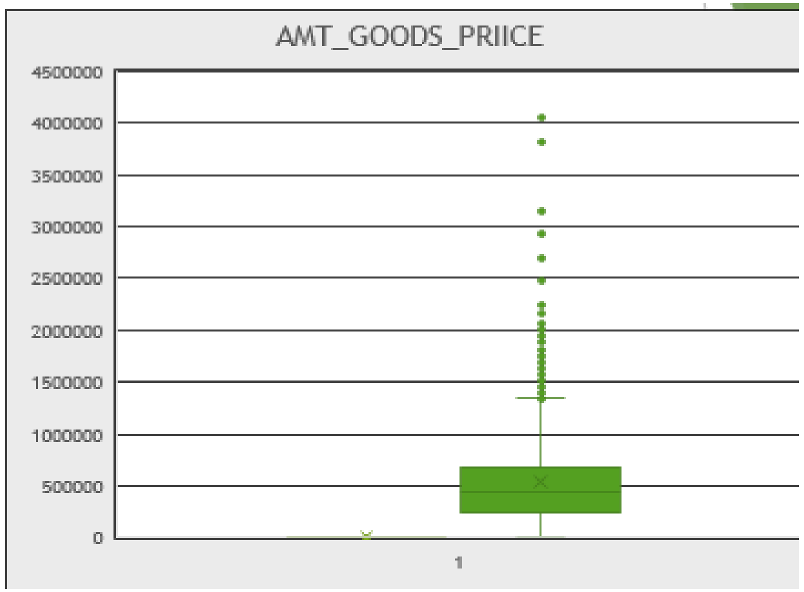
Task 2: Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- Box plots of Target column vs
- 1. AMT_INCOME_TOTAL 2. AMT_ANNUITY 3. AMT_CREDIT 4. AMT_GOODS_PRICE 5. CNT_FAM_MEMBERS



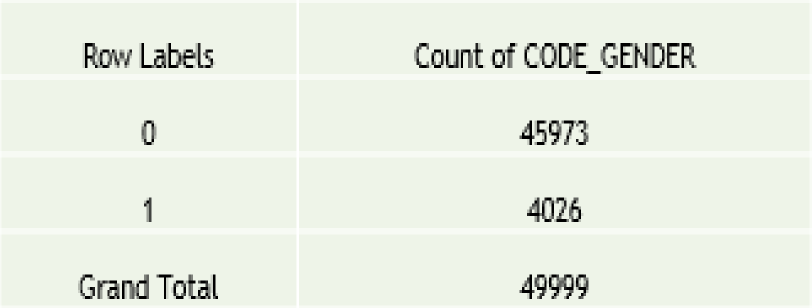




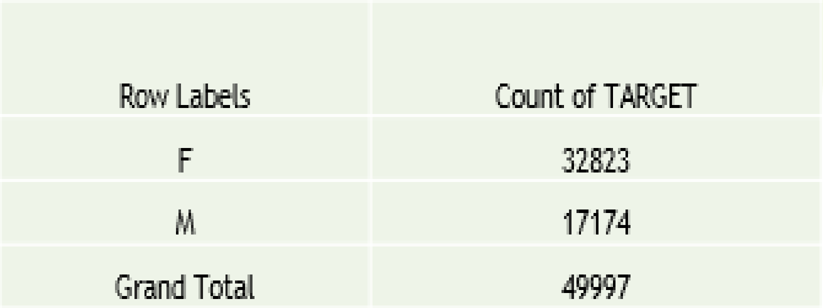


C. Analyse Data Imbalance Task: Determine if there is data imbalance in the loan application and calculate ratio of data imbalance using excel function

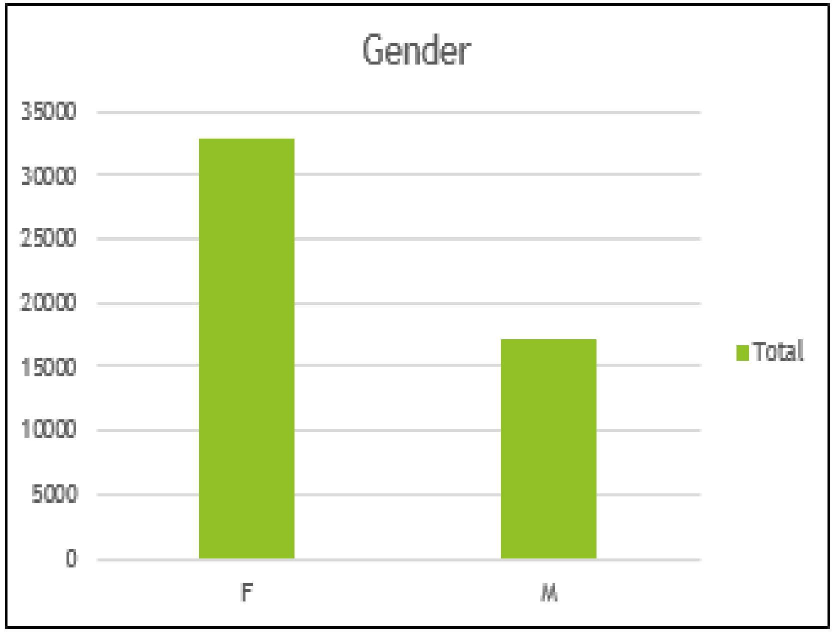
We can see the data imbalance using following Pivot tables and charts:

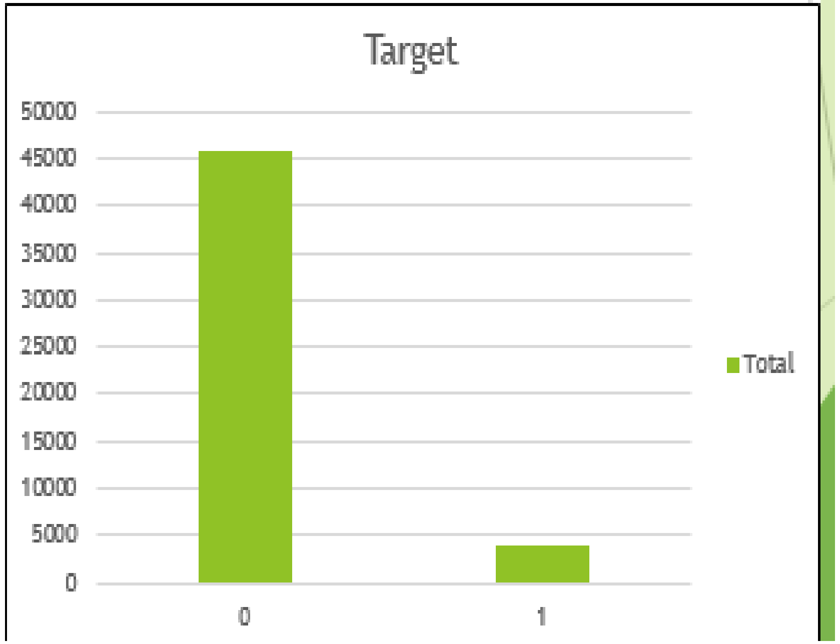


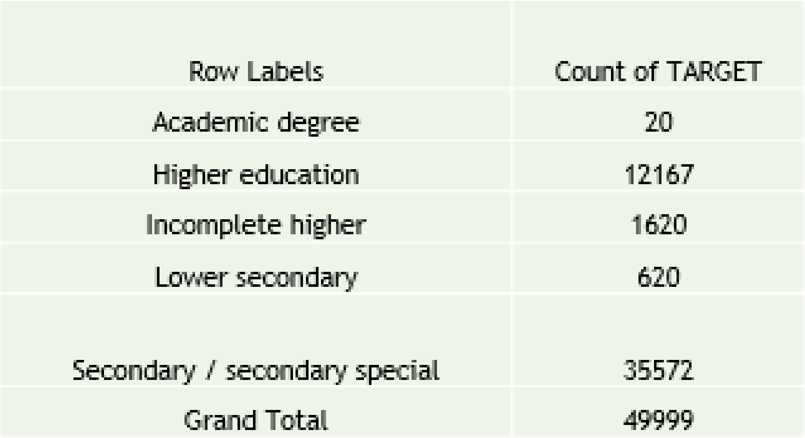
Row Labels	Count of CODE_GENDER
0	45973
1	4026
Grand Total	49999



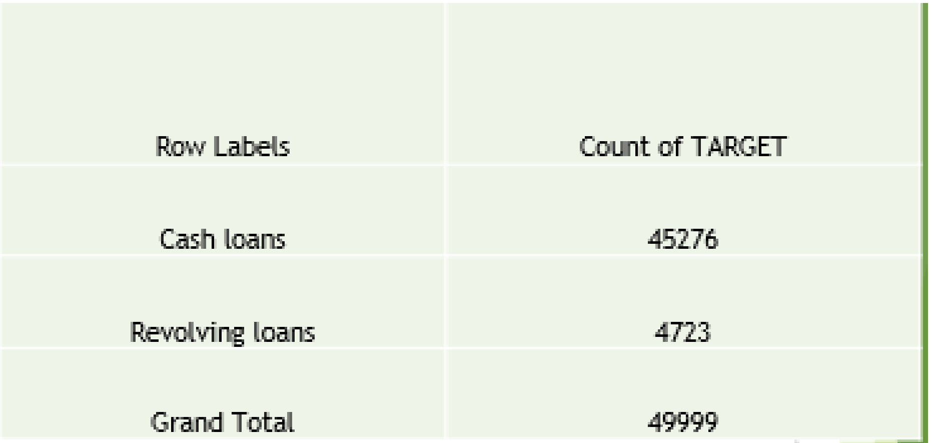
Row Labels	Count of TARGET
F	32823
M	17174
Grand Total	49997



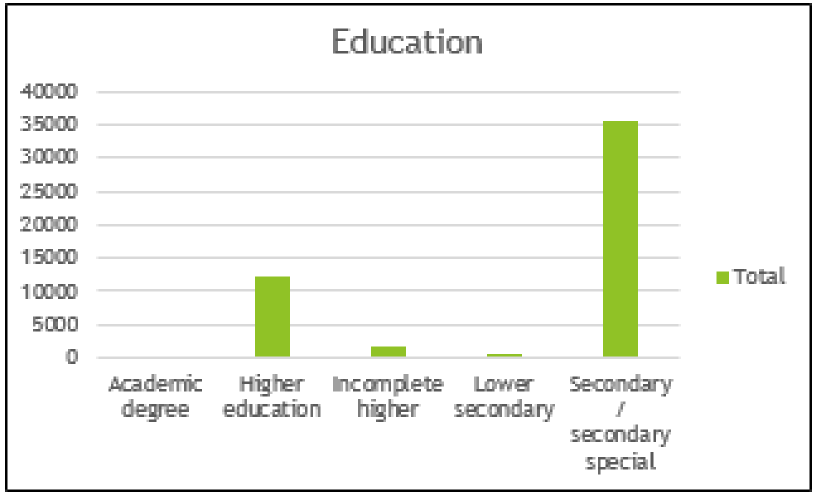


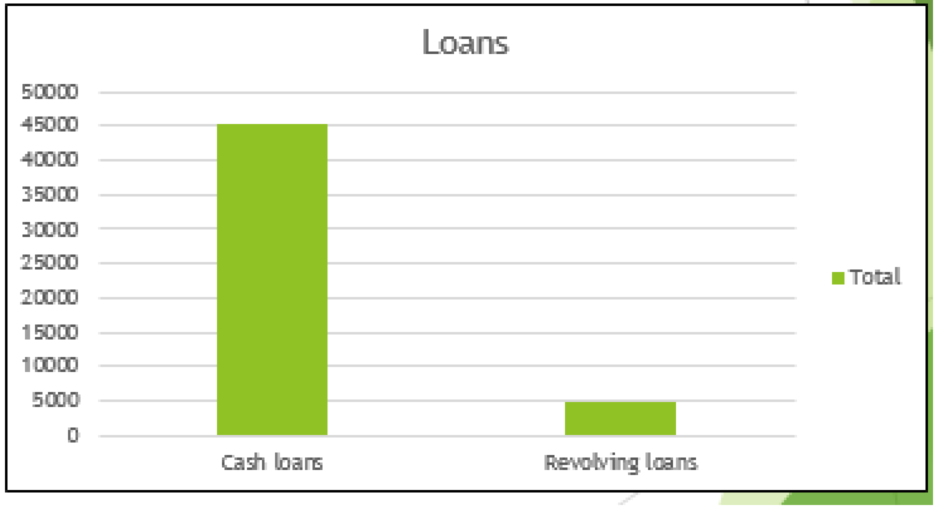


Row Labels	Count of TARGET
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35572
Grand Total	49999



Row Labels	Count of TARGET
Cash loans	45276
Revolving loans	4723
Grand Total	49999

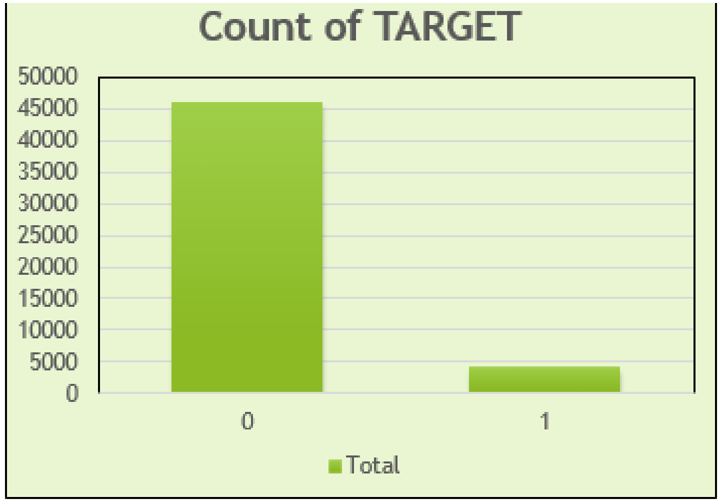


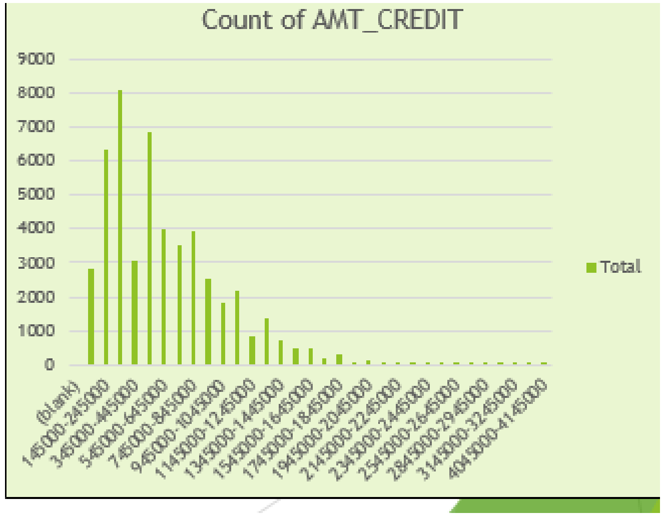


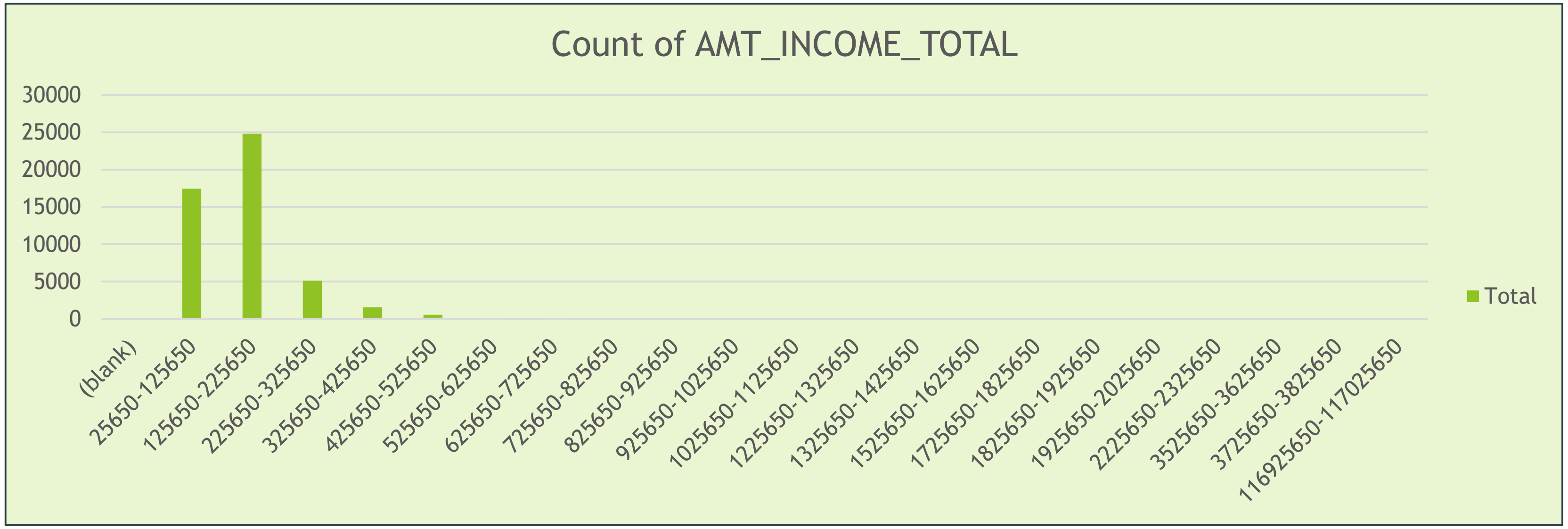
D. Perform Univariate, Segmented Univariate and Bivariate Analysis Task: Perform univariate analysis to understand distribution of individual variables. Segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationship between variables and target variable

Univariate Analysis: The Univariate Analysis focuses on examining and describing the individual variables in isolation. The summary and analysis for the single variable.

The Plots obtained while performing the Univariate Analysis are as follows.

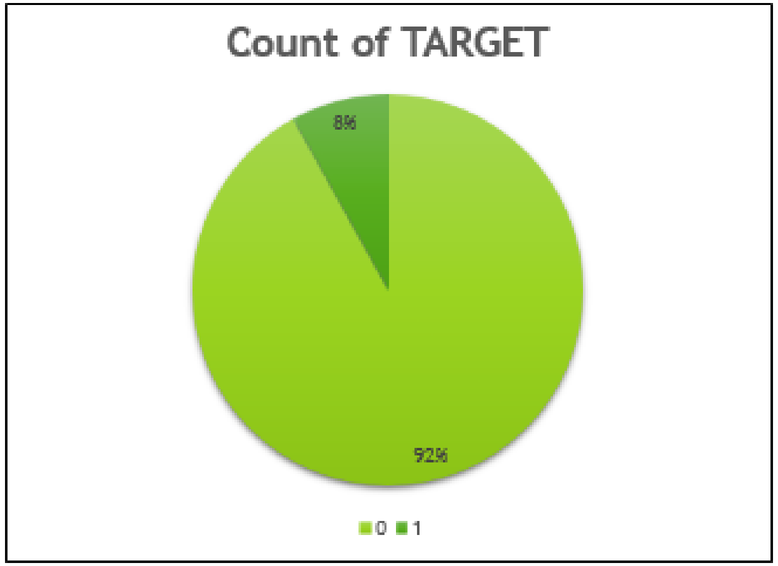


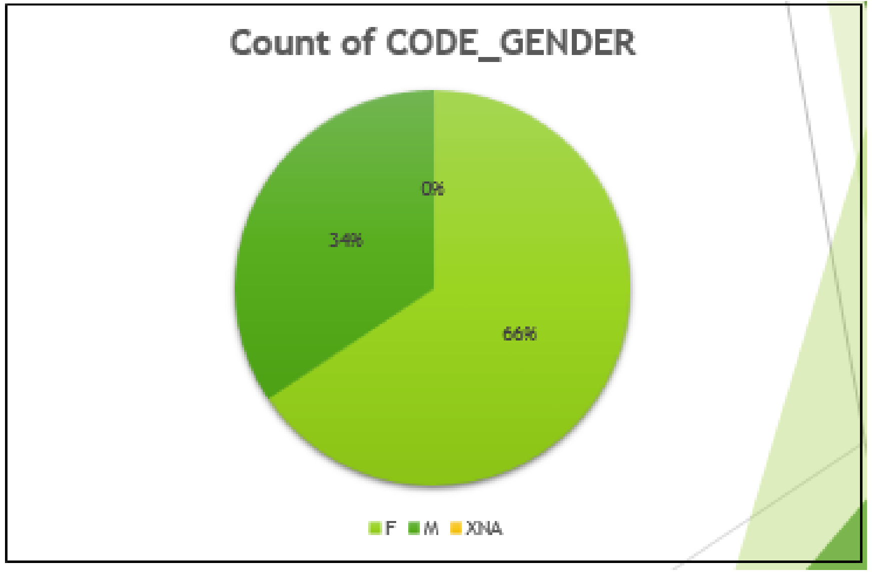


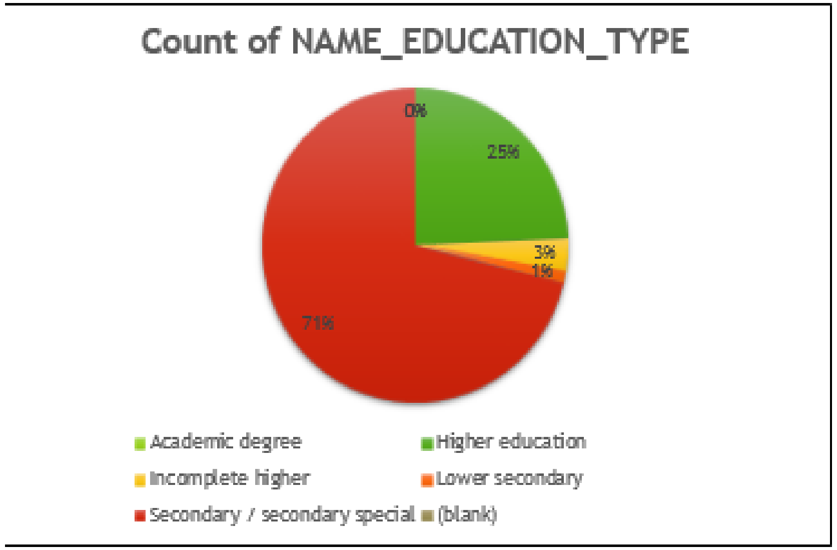


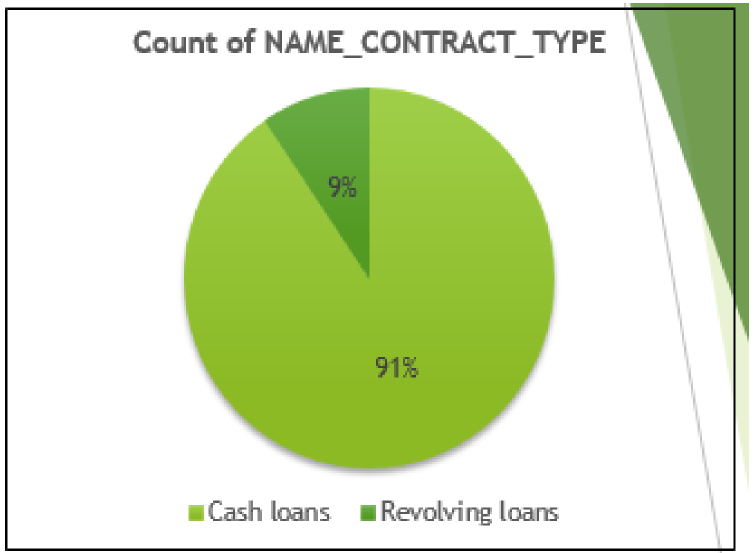
Segmented Univariate Analysis:

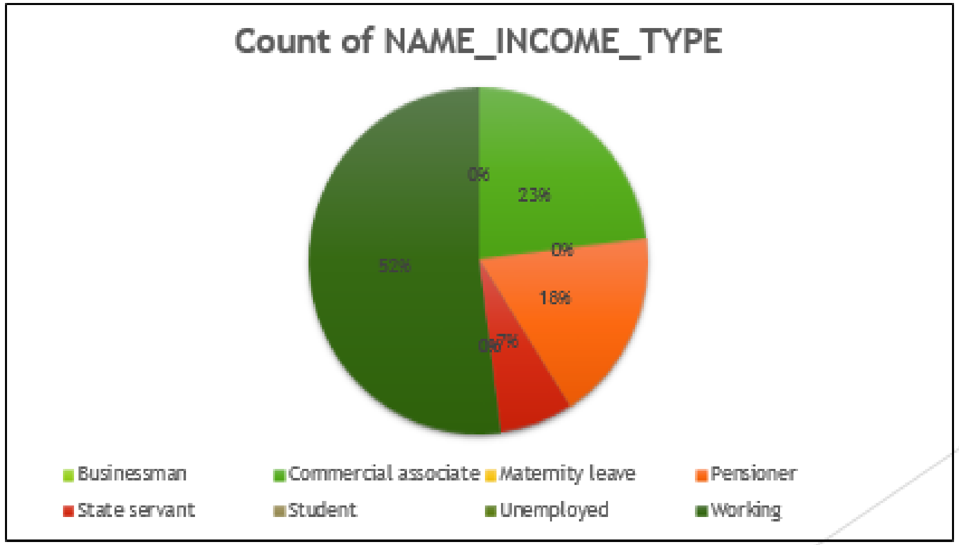
- It is the extension of Univariate Analysis which involves the splitting of data into specific segments or the groups contained.



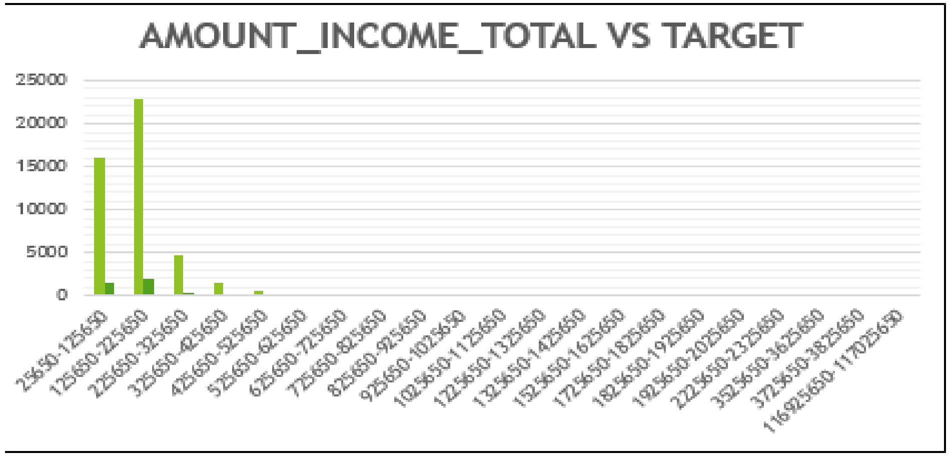


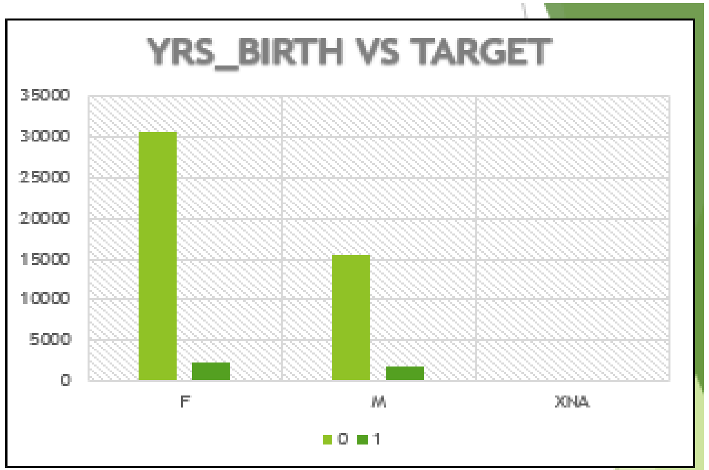


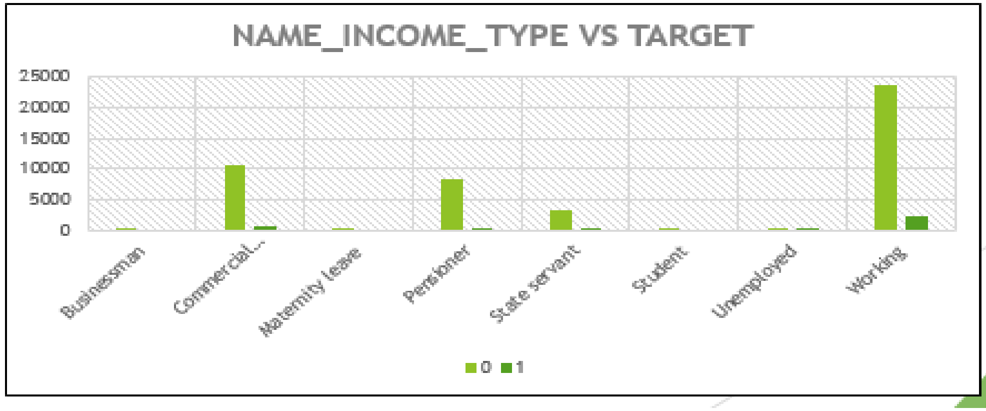




Bivariate Analysis: These examines the relationship between the two variables. Following are the plots obtained while performing the Bivariate Analysis:

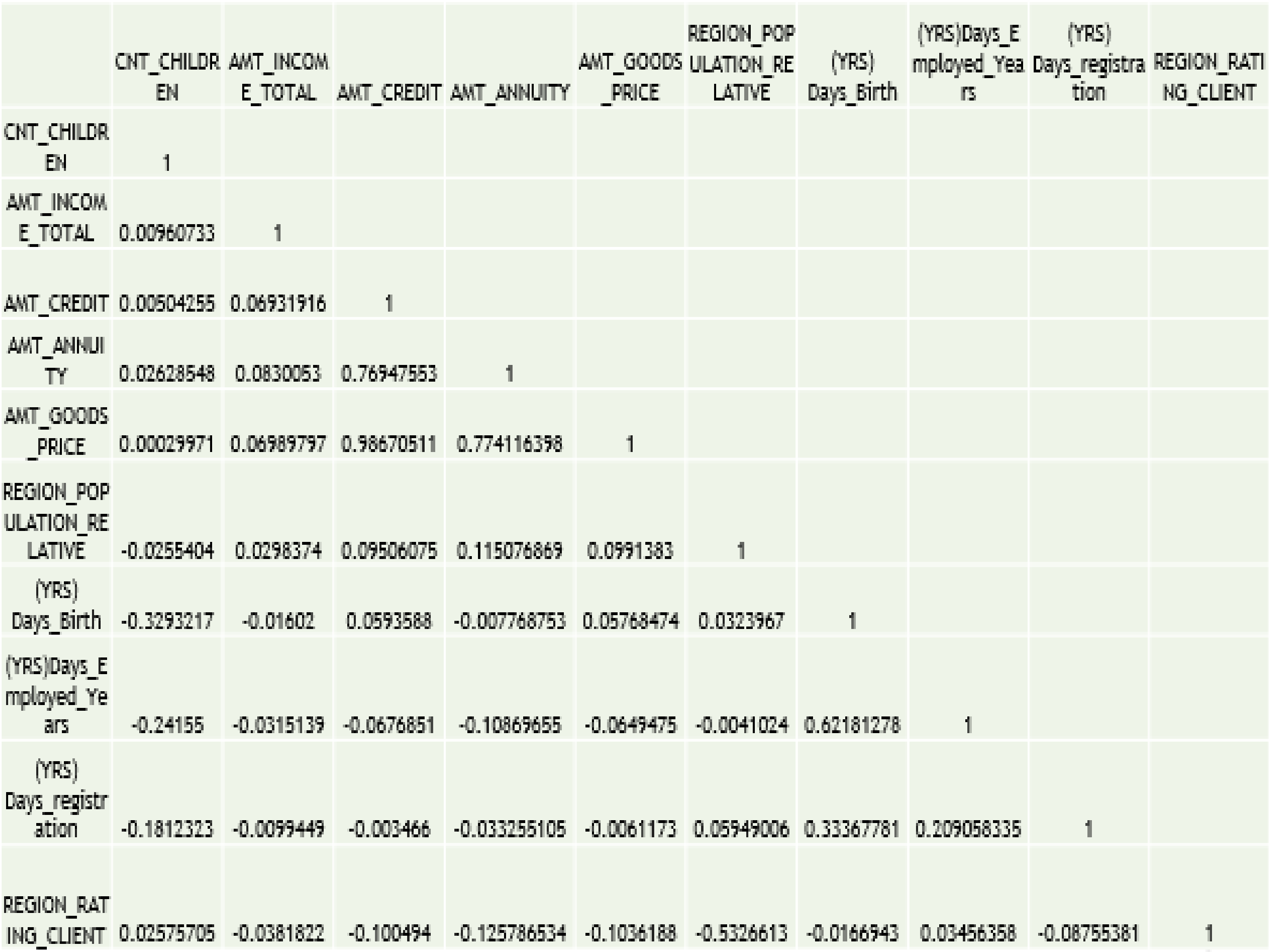






E. Identify Top Correlation for Different scenarios.
Task: To determine top correlation for each segmented data(client with payment difficulties and all other cases.

The correlation between the variables with TARGET 0



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	(YRS) Days_Birth	(YRS)Days_Employed_Years	(YRS) Days_registration	REGION_RATING_CLIENT
CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.00960733	1								
AMT_CREDIT	0.00504255	0.06931916	1							
AMT_ANNUITY	0.02628548	0.0830053	0.76947553	1						
AMT_GOODS_PRICE	0.00029971	0.06989797	0.98670511	0.774116398	1					
REGION_POPULATION_RELATIVE	-0.0255404	0.0298374	0.09506075	0.115076869	0.0991383	1				
(YRS) Days_Birth	-0.3293217	-0.01602	0.0593588	-0.007768753	0.05768474	0.0323967	1			
(YRS)Days_Employed_Years	-0.24155	-0.0315139	-0.0676851	-0.10869655	-0.0649475	-0.0041024	0.62181278	1		
(YRS) Days_registration	-0.1812323	-0.0099449	-0.003466	-0.033255105	-0.0061173	0.05949006	0.33367781	0.209058335	1	
REGION_RATING_CLIENT	0.02575705	-0.0381822	-0.100494	-0.125786534	-0.1036188	-0.5326613	-0.0166943	0.03456358	-0.08755381	1

Highly correlated variables are:

Var 1	Var2	Corr coeff
AMT_ANNUITY	AMT_CREDIT	0.76949879
AMT_GOODS_PRICE	AMT_CREDIT	0.98670439
AMT_GOODS_PRICE	AMT_ANNUITY	0.77413414
(YRS) DAYS_EMPLOYED	(YRS)DAYS_BIRTH	0.62148914

The correlation between the variables with TARGET 1

CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.00960733	1								
AMT_CREDIT	0.005042552	0.069319162	1							
AMT_ANNUITY	0.026285477	0.083005301	0.76947553	1						
AMT_GOODS_PRICE	0.000299712	0.069897973	0.98670511	0.774116398	1					
REGION_POPULATION_RELATIVE	-0.025540408	0.029837396	0.09506075	0.115076869	0.099138301	1				
(YRS) Days_Birth	-0.329321666	-0.016019977	0.0593588	-0.007768753	0.057684741	0.0323967	1			
(YRS)Days_Employed_Years	-0.24154996	-0.031513894	-0.0676851	-0.10869655	-0.064947528	-0.0041024	0.621812778	1		
(YRS) Days_registration	-0.181232261	-0.00994491	-0.003466	-0.033255105	-0.006117338	0.05949006	0.333677806	0.209058335	1	
REGION_RATING_CLIENT	0.02575705	-0.038182249	-0.100494	-0.125786534	-0.103618752	-0.5326613	-0.016694261	-0.03456358	-0.08755381	1

The high Correlation coefficients are obtained as:

VAR1	VAR2	Correlation Coeff
AMT_CREDIT	AMT_GOODS_PRICE	0.982268
AMT_ANNUITY	AMT_CREDIT	0.749665
AMT_GOODS_PRICE	AMT_ANNUITY	0.749504
(YRS)DAYS_EMPLOYED	(YRS)DAYS_BIRTH	0.587858

RESULTS

- It is observed that most of the clients become default due to other cases.
- Cash Loans have much higher defaults than Revolving Loans.
- The Education Type Academic Degree has a less number of defaults.
- The Clients with Income ranging between 25000-1025000 has the highest defaults.
- Credit amount of the bank loan is generally falling in the range of 45000-1,45000.
- There are no defaults for clients who are Businessman and students.
- Males are less inclined towards the defaults than the females.

6. Analyzing The Impact Of Car Features On Price And Profitability

- PROJECT DESCRIPTION

- The automotive industry has witnessed rapid transformations in recent decades, characterized by an increasing emphasis on fuel efficiency, environmental sustainability, and technological advancements. In this context, comprehending the diverse factors that steer consumer preferences for automobiles has become more critical than ever.
- • As a Data Analyst, the primary objective is to ascertain how a car manufacturer can enhance pricing strategies and product development decisions to achieve the twin goals of maximizing profitability and aligning with consumer demand.
- • Addressing this business challenge will necessitate the application of advanced Excel proficiency and a deep understanding of data analysis methodologies, such as regression analysis and the utilization of pivot tables.

PROJECT OVERVIEW

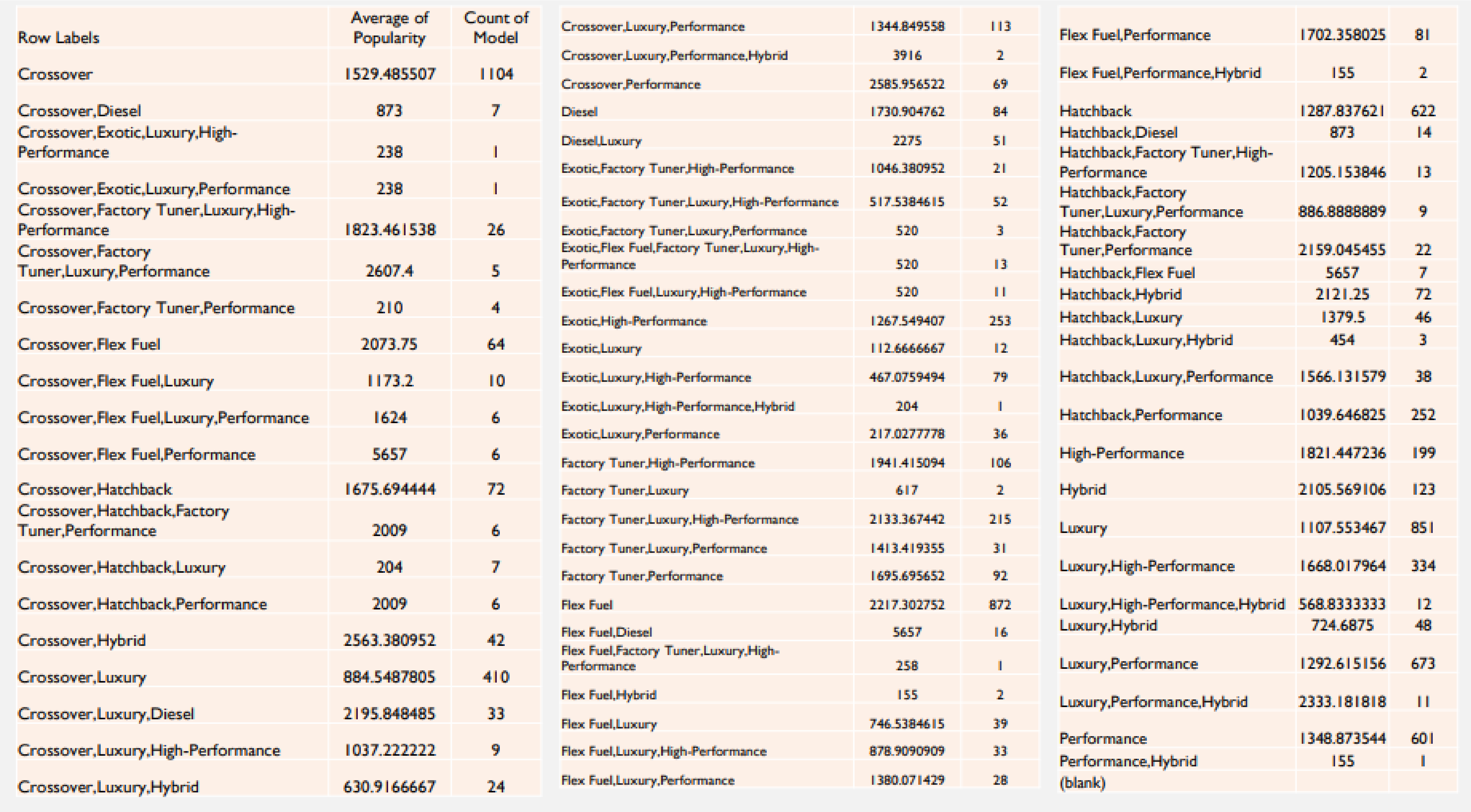
- The analysis of the projects is going to answer the following questions:
- 1. How does the popularity of a car model vary across different market categories?
- 2. What is the relationship between a car's engine power and its price?
- 3. Which car features are most important in determining a car's price? • 4. How does the average price of a car vary across different manufacturers?
- 5. What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

INSIGHTS

- As first step of analysis is data cleaning. Initially there were 11915 rows and 16 columns were there and after removal of blank values from data those rows got reduced to 8114. But we can’t remove column as it will lead to loss of information about cars, for better understanding I have renamed ‘Make’ column to ‘Brands’.

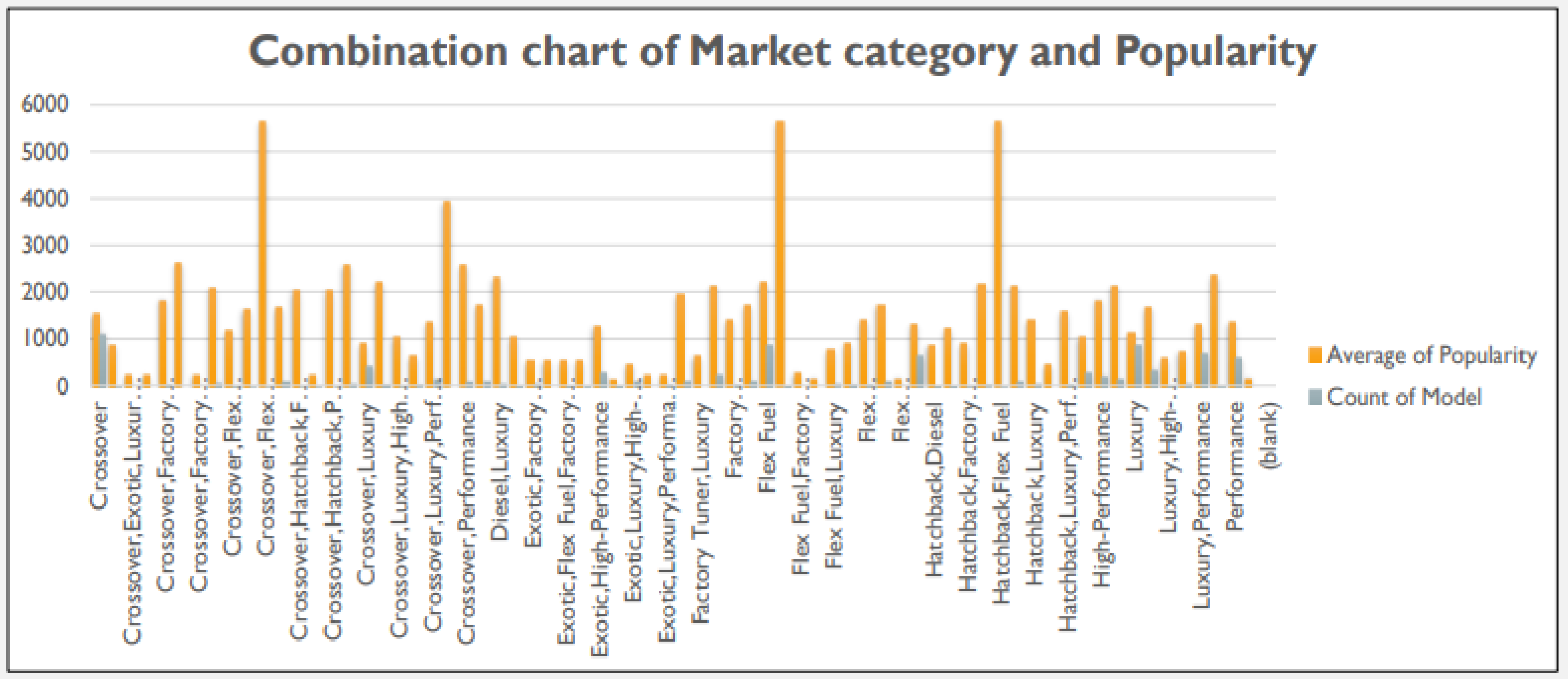
• Insight Required: How does the popularity of a car model vary across different market categories?

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



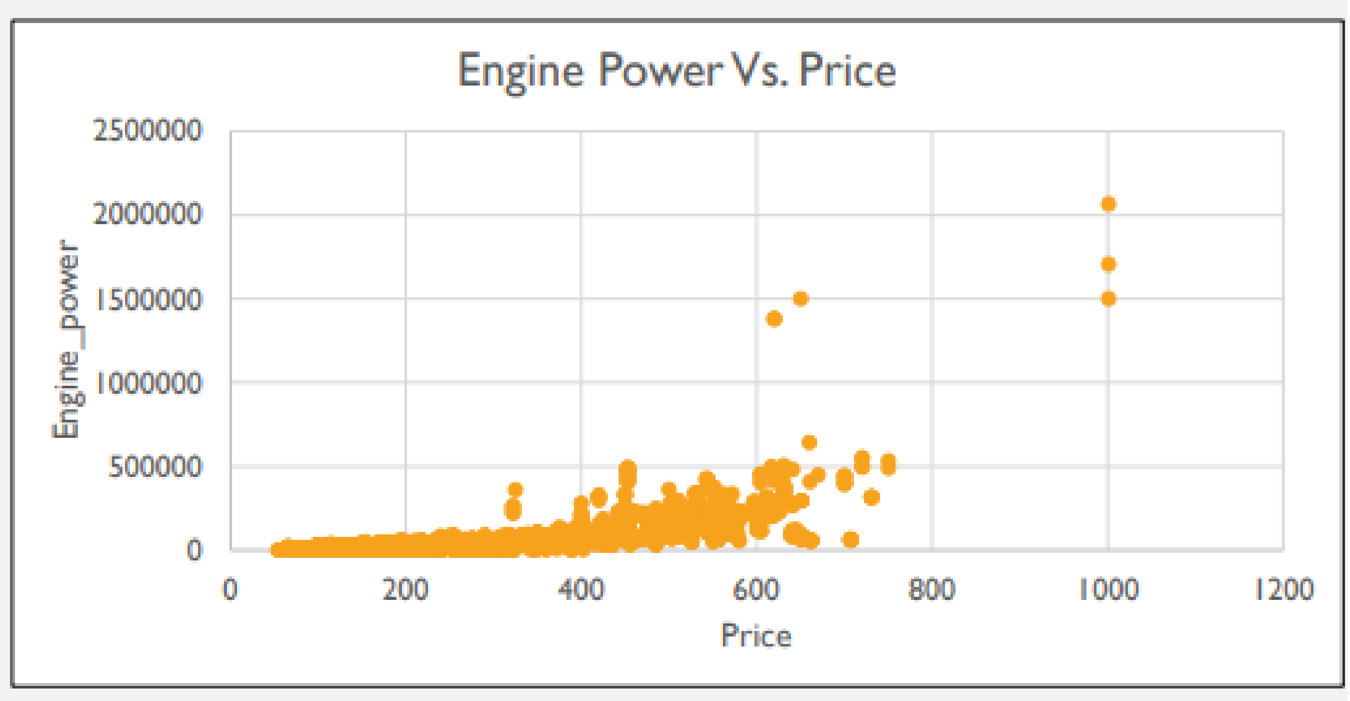
Row Labels	Average of Popularity	Count of Model
Crossover	1529.485507	1104
Crossover,Diesel	873	7
Crossover,Exotic,Luxury,High-Performance	238	1
Crossover,Exotic,Luxury,Performance	238	1
Crossover,Factory Tuner,Luxury,High-Performance	1823.461538	26
Crossover,Factory Tuner,Luxury,Performance	2607.4	5
Crossover,Factory Tuner,Performance	210	4
Crossover,Flex Fuel	2073.75	64
Crossover,Flex Fuel,Luxury	1173.2	10
Crossover,Flex Fuel,Luxury,Performance	1624	6
Crossover,Flex Fuel,Performance	5657	6
Crossover,Hatchback	1675.694444	72
Crossover,Hatchback,Factory Tuner,Performance	2009	6
Crossover,Hatchback,Luxury	204	7
Crossover,Hatchback,Performance	2009	6
Crossover,Hybrid	2563.380952	42
Crossover,Luxury	884.5487805	410
Crossover,Luxury,Diesel	2195.848485	33
Crossover,Luxury,High-Performance	1037.222222	9
Crossover,Luxury,Hybrid	630.9166667	24
Crossover,Luxury,Performance	1344.849558	113
Crossover,Luxury,Performance,Hybrid	3916	2
Crossover,Performance	2585.956522	69
Diesel	1730.904762	84
Diesel,Luxury	2275	51
Exotic,Factory Tuner,High-Performance	1046.380952	21
Exotic,Factory Tuner,Luxury,High-Performance	517.5384615	52
Exotic,Factory Tuner,Luxury,Performance	520	3
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	520	13
Exotic,Flex Fuel,Luxury,High-Performance	520	11
Exotic,High-Performance	1267.549407	253
Exotic,Luxury	112.6666667	12
Exotic,Luxury,High-Performance	467.0759494	79
Exotic,Luxury,High-Performance,Hybrid	204	1
Exotic,Luxury,Performance	217.0277778	36
Factory Tuner,High-Performance	1941.415094	106
Factory Tuner,Luxury	617	2
Factory Tuner,Luxury,High-Performance	2133.367442	215
Factory Tuner,Luxury,Performance	1413.419355	31
Factory Tuner,Performance	1695.695652	92
Flex Fuel	2217.302752	872
Flex Fuel,Diesel	5657	16
Flex Fuel,Factory Tuner,Luxury,High-Performance	258	1
Flex Fuel,Hybrid	155	2
Flex Fuel,Luxury	746.5384615	39
Flex Fuel,Luxury,High-Performance	878.9090909	33
Flex Fuel,Luxury,Performance	1380.071429	28
Flex Fuel,Performance	1702.358025	81
Flex Fuel,Performance,Hybrid	155	2
Hatchback	1287.837621	622
Hatchback,Diesel	873	14
Hatchback,Factory Tuner,High-Performance	1205.153846	13
Hatchback,Factory Tuner,Luxury,Performance	886.8888889	9
Hatchback,Factory Tuner,Performance	2159.045455	22
Hatchback,Flex Fuel	5657	7
Hatchback,Hybrid	2121.25	72
Hatchback,Luxury	1379.5	46
Hatchback,Luxury,Hybrid	454	3
Hatchback,Luxury,Performance	1566.131579	38
Hatchback,Performance	1039.646825	252
High-Performance	1821.447236	199
Hybrid	2105.569106	123
Luxury	1107.553467	851
Luxury,High-Performance	1668.017964	334
Luxury,High-Performance,Hybrid	568.8333333	12
Luxury,Hybrid	724.6875	48
Luxury,Performance	1292.615156	673
Luxury,Performance,Hybrid	2333.181818	11
Performance	1348.873544	601
Performance,Hybrid	155	1
(blank)		

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



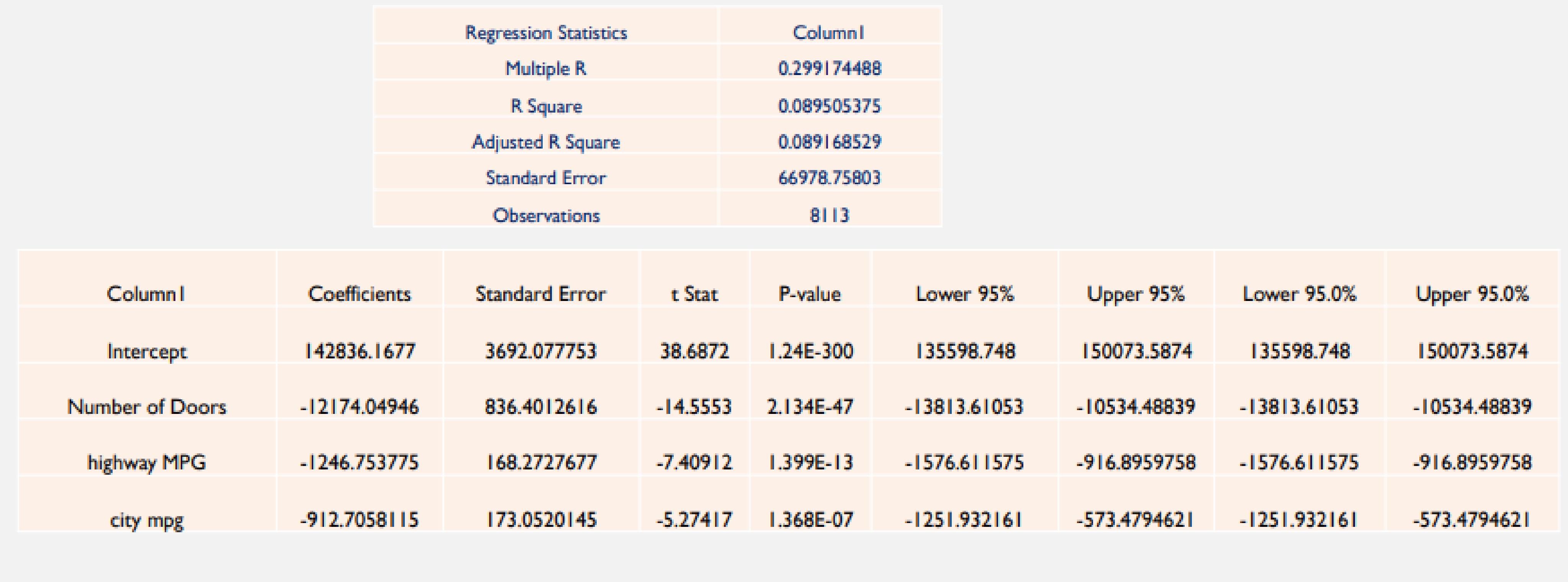
Results: 1. It is observed that the Market Category with Flex Fuel and Diesel has the highest Average Popularity, so it can be concluded that most of the people popularly use the cars with flex fuel and diesels. 2. It is also observed that the count for the model of market category Crossover is the highest.

Task 2: To Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.



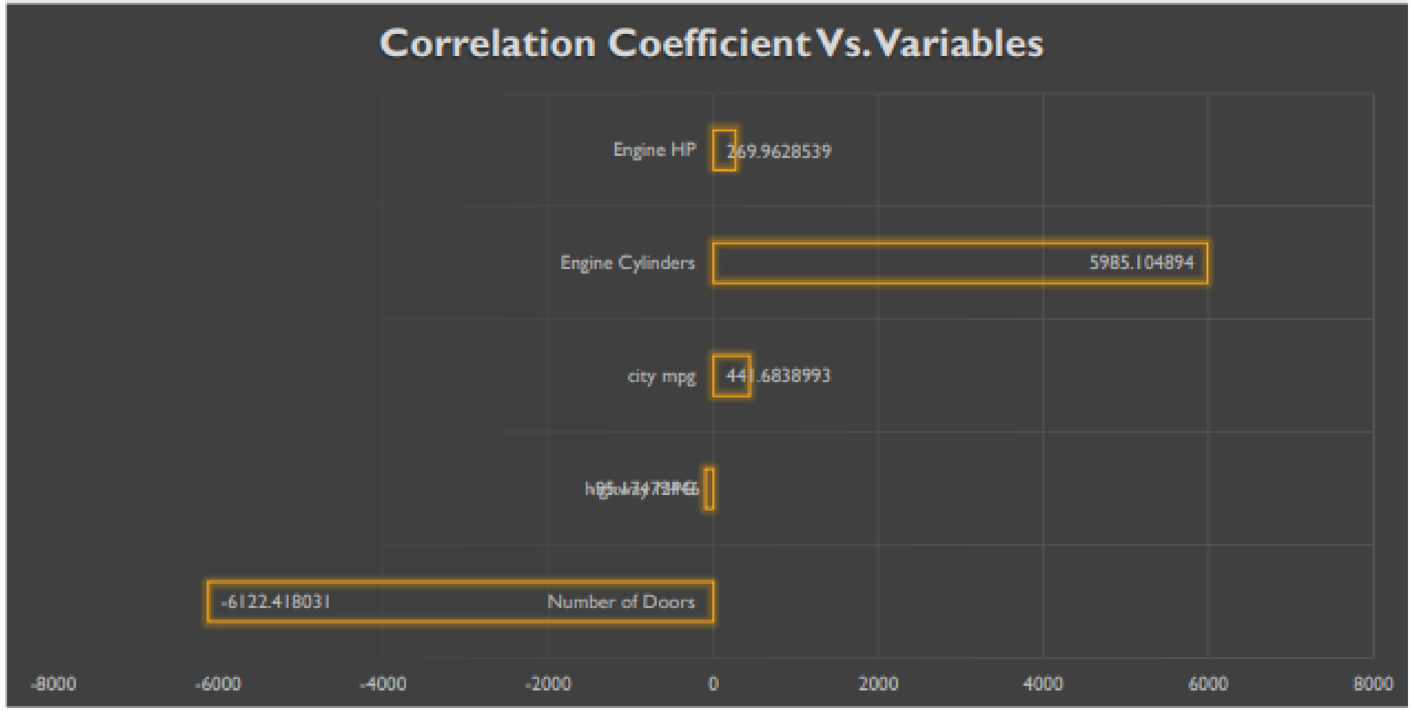
Result: It is observed that the Engine Power and Price are dependent on each other. As the Engine Power increases the Price of the model increases.

Task 3: To 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. With the help of Regression given in the Statistical Analysis tab the following tables have been obtained:



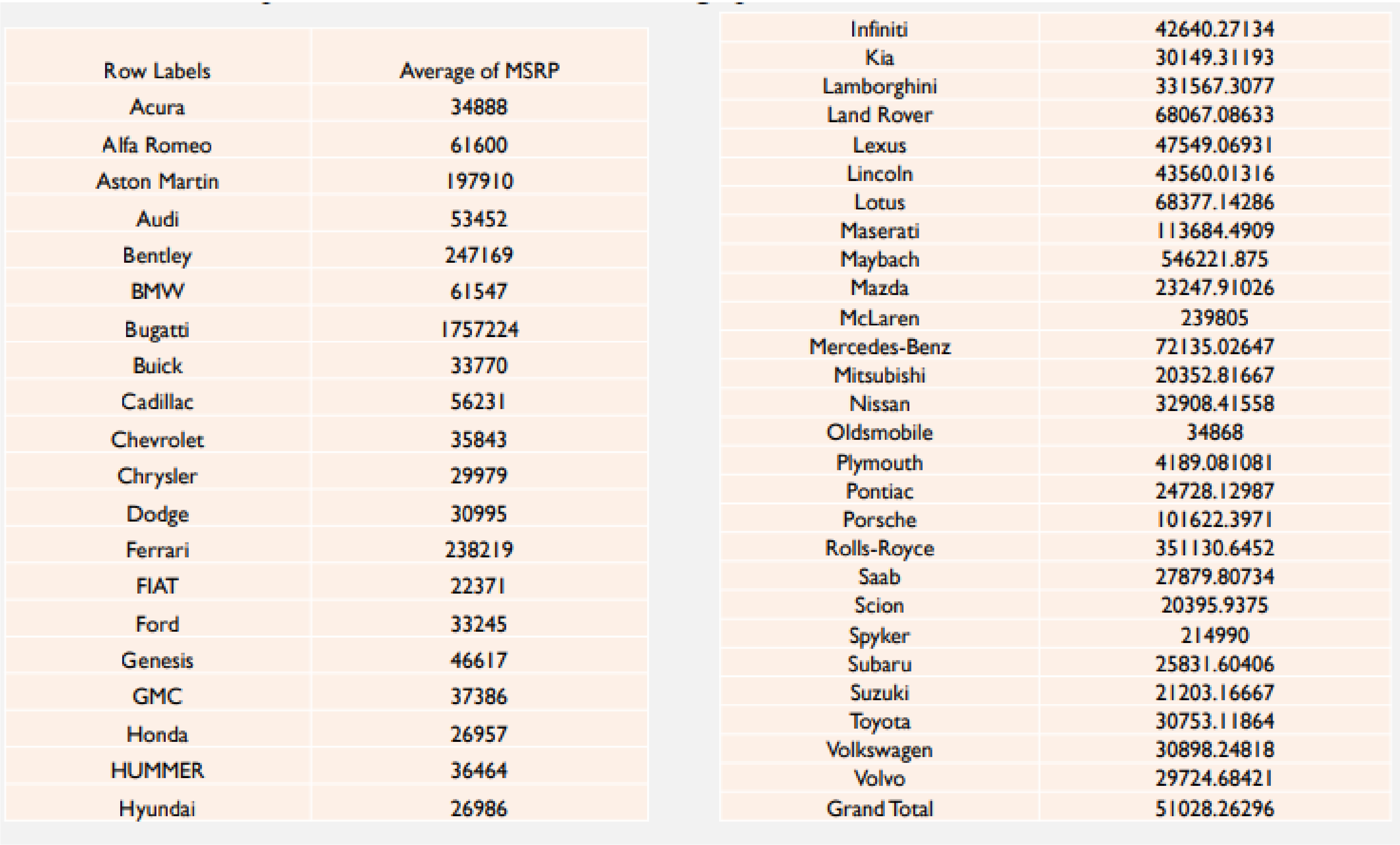
Regression Statistics	Column1
Multiple R	0.299174488
R Square	0.089505375
Adjusted R Square	0.089168529
Standard Error	66978.75803
Observations	8113

Column1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	142836.1677	3692.077753	38.6872	1.24E-300	135598.748	150073.5874	135598.748	150073.5874
Number of Doors	-12174.04946	836.4012616	-14.5553	2.134E-47	-13813.61053	-10534.48839	-13813.61053	-10534.48839
highway MPG	-1246.753775	168.2727677	-7.40912	1.399E-13	-1576.611575	-916.8959758	-1576.611575	-916.8959758
city mpg	-912.7058115	173.0520145	-5.27417	1.368E-07	-1251.932161	-573.4794621	-1251.932161	-573.4794621



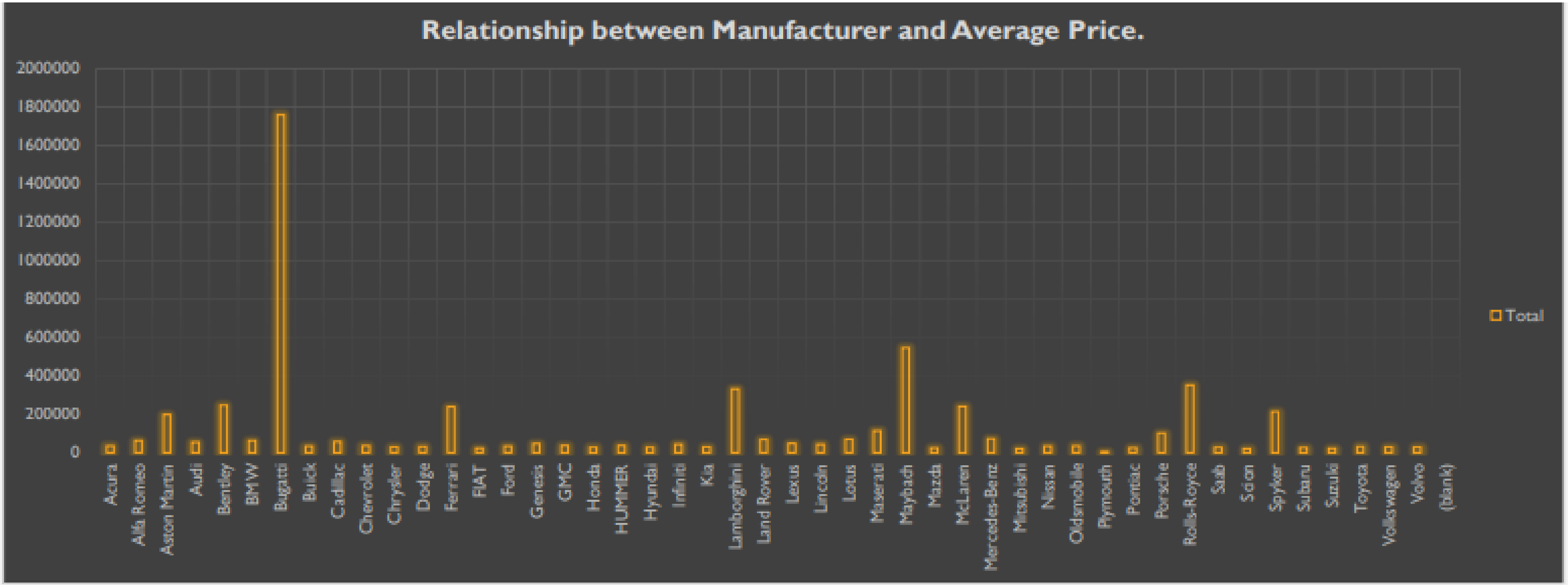
Results: It is observed that the Engine Cylinders has the highest correlation coefficient with the Car Price.

Task 4.A: To Create a pivot table that shows the average price of cars for each manufacturer.



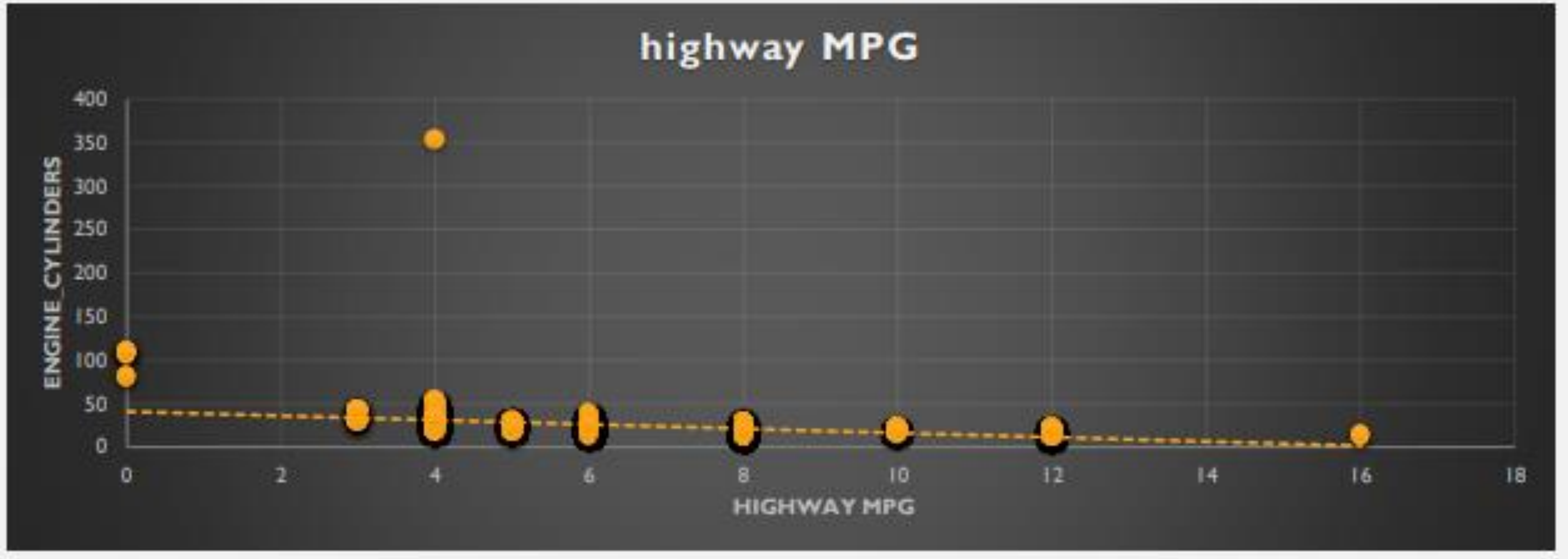
Row Labels	Average of MSRP
Acura	34888
Alfa Romeo	61600
Aston Martin	197910
Audi	53452
Bentley	247169
BMW	61547
Bugatti	1757224
Buick	33770
Cadillac	56231
Chevrolet	35843
Chrysler	29979
Dodge	30995
Ferrari	238219
FIAT	22371
Ford	33245
Genesis	46617
GMC	37386
Honda	26957
HUMMER	36464
Hyundai	26986
Infiniti	42640.27134
Kia	30149.31193
Lamborghini	331567.3077
Land Rover	68067.08633
Lexus	47549.06931
Lincoln	43560.01316
Lotus	68377.14286
Maserati	113684.4909
Maybach	546221.875
Mazda	23247.91026
McLaren	239805
Mercedes-Benz	72135.02647
Mitsubishi	20352.81667
Nissan	32908.41558
Oldsmobile	34868
Plymouth	4189.081081
Pontiac	24728.12987
Porsche	101622.3971
Rolls-Royce	351130.6452
Saab	27879.80734
Scion	20395.9375
Spyker	214990
Subaru	25831.60406
Suzuki	21203.16667
Toyota	30753.11864
Volkswagen	30898.24818
Volvo	29724.68421
Grand Total	51028.26296

Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



Results: It is observed from the plot that the Manufacturer named Bugati has the Highest Average Price.

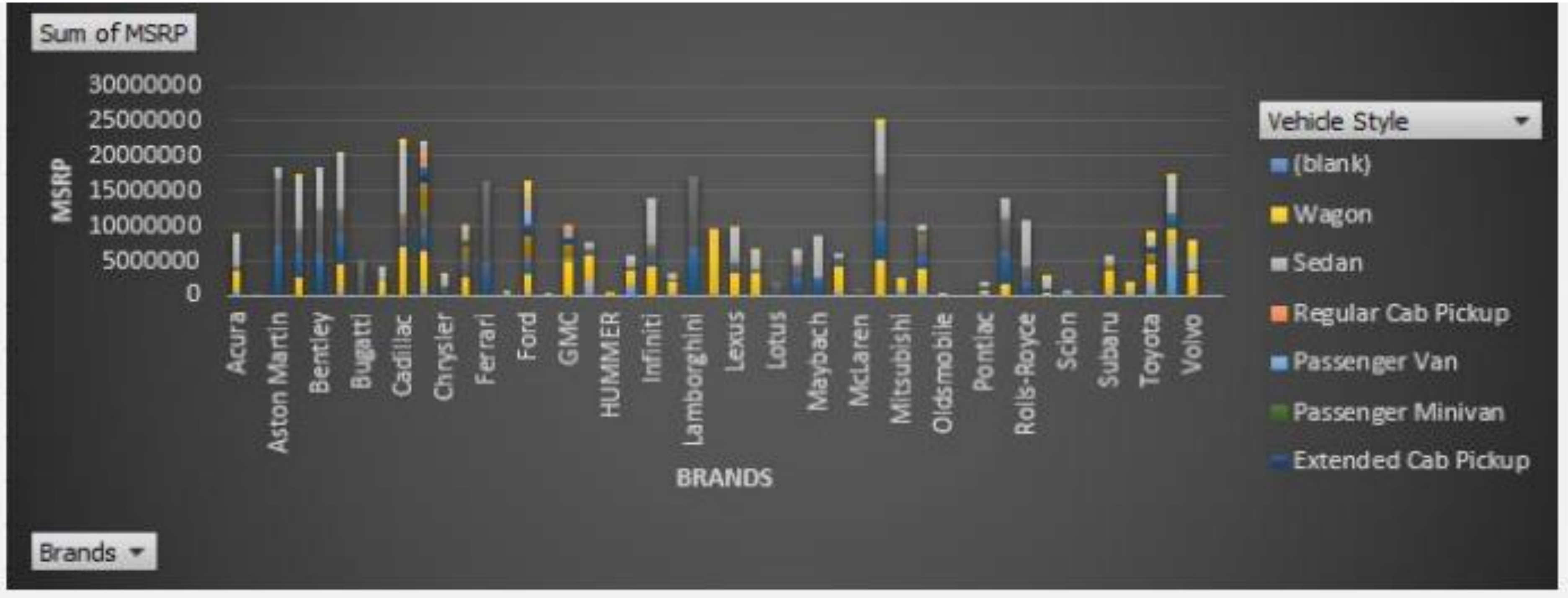
Task 5.A: To Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance. Task 5.B: To Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.



Results: It is observed from the trendline that as the number of Engine Cylinders goes on increasing the estimated miles per gallon the car gets on the highway goes on decreasing. The correlation between the number of cylinders and Highway MPG is -0.6033, which indicates there is a negative correlation between them.

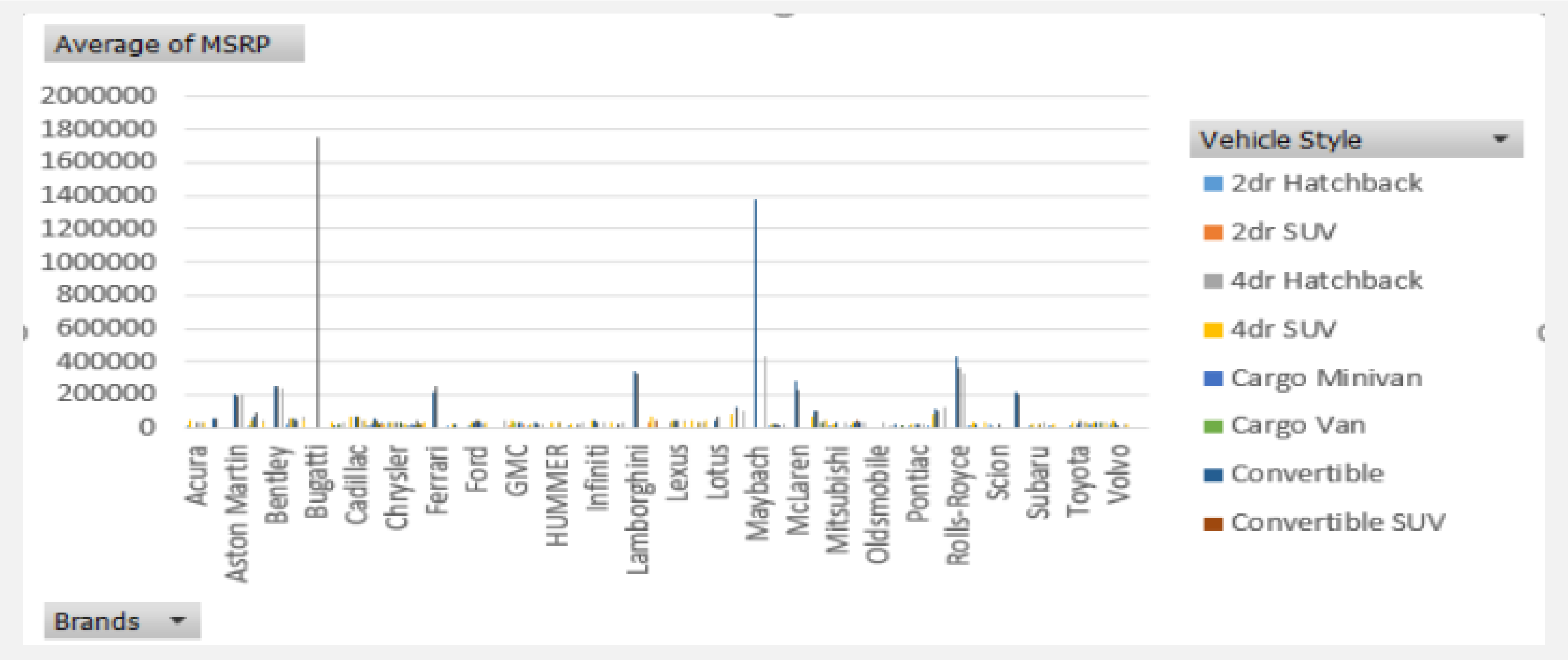
BUILDING THE DASHBOARDS

- Task 1: How does the distribution of car prices vary by brand and body style? Answer: From the cleaned data, the variables named Brands, Vehicle Style and MSRP were copied to the new sheet. The Pivot Table is created by taking the Car brand to the rows, Vehicle Style to column and sum of MSRP to the Values.



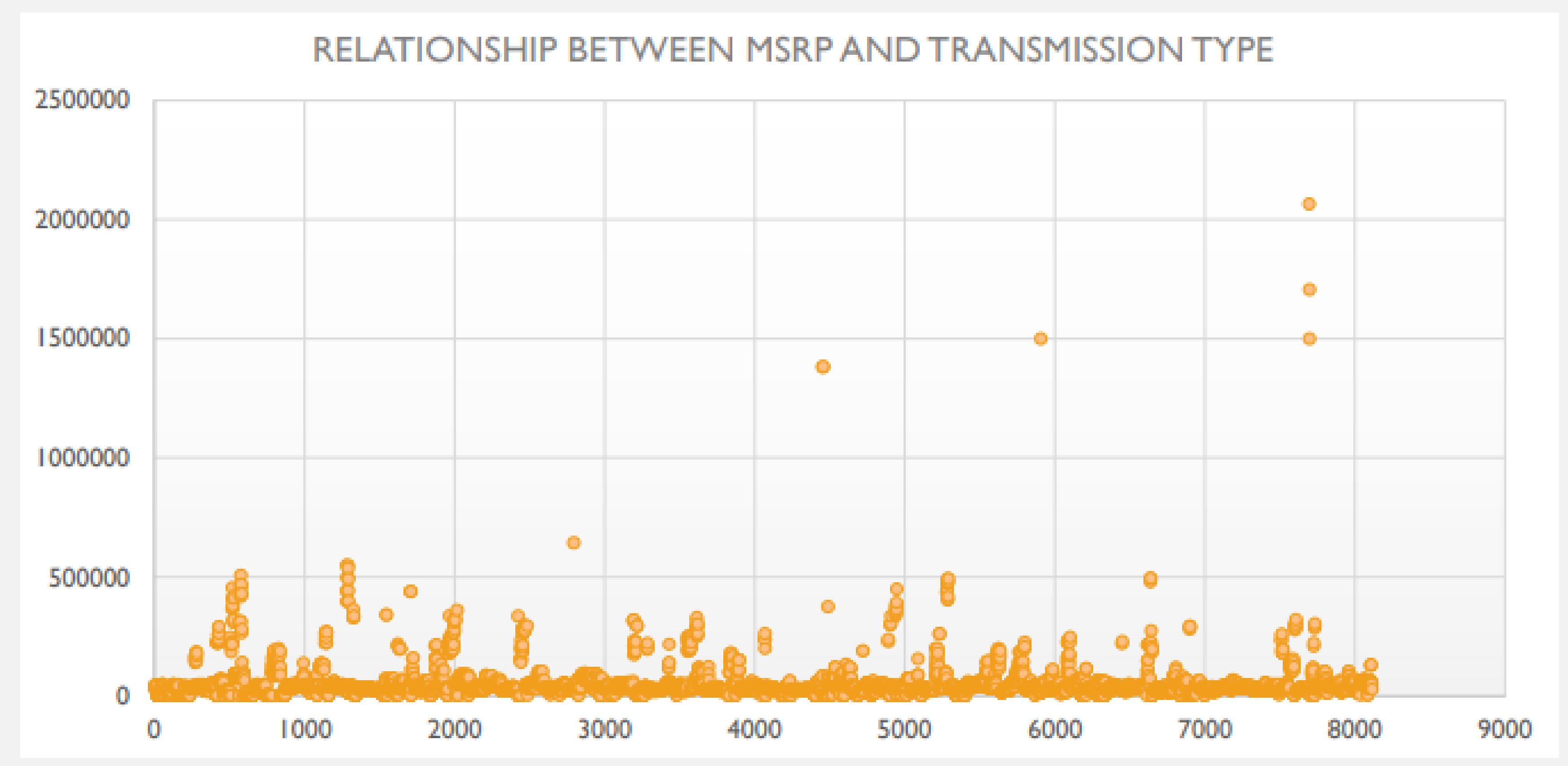
Results: It is observed that the Car brand Mercedes-Benz has the highest Price.

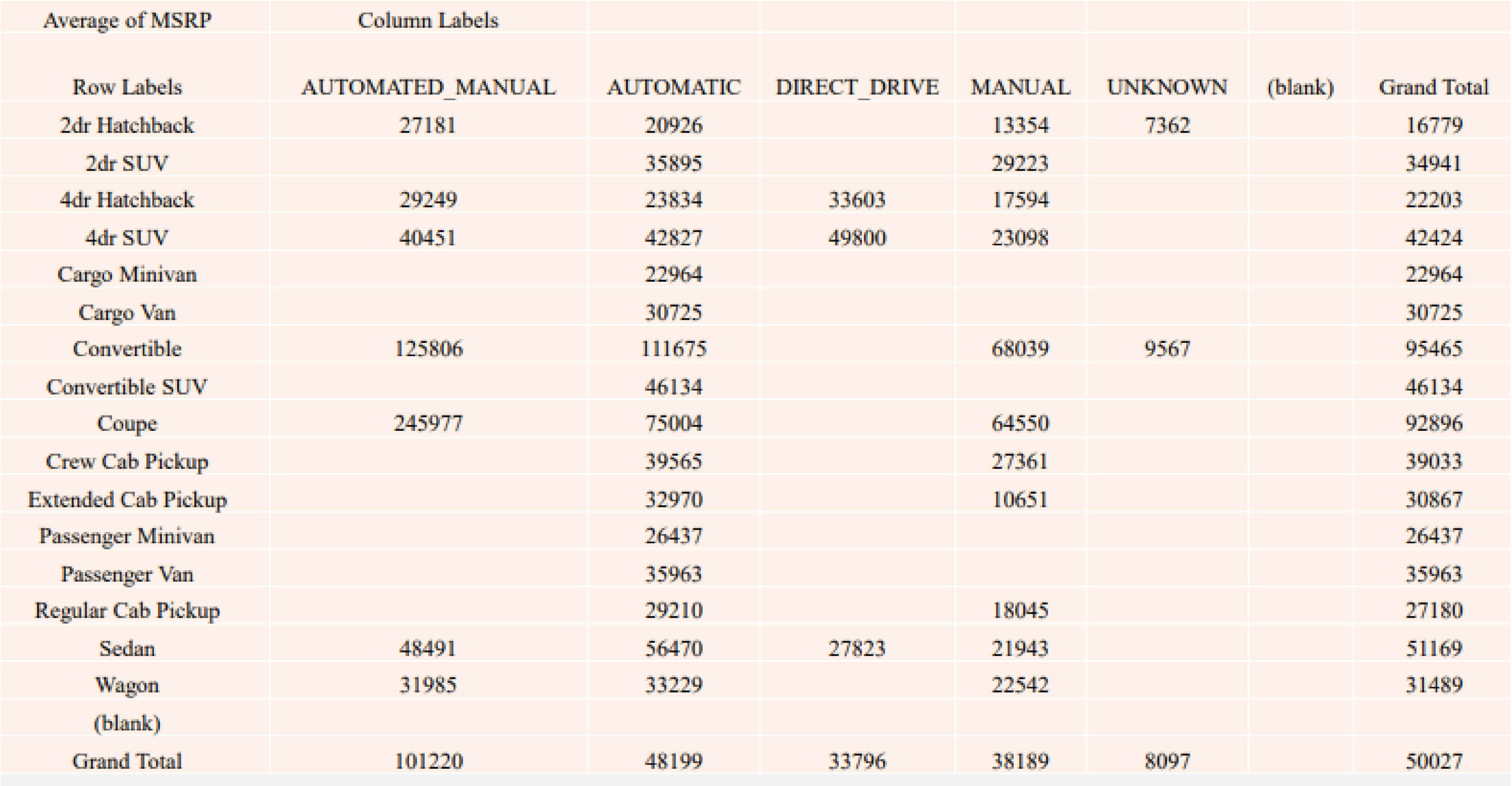
Task 2: To identify which car brands have the highest and lowest average MSRPs, and how does this vary by body style.



Results: It is observed from the plot that the Manufacturer named Bugati has the Highest Average Price.

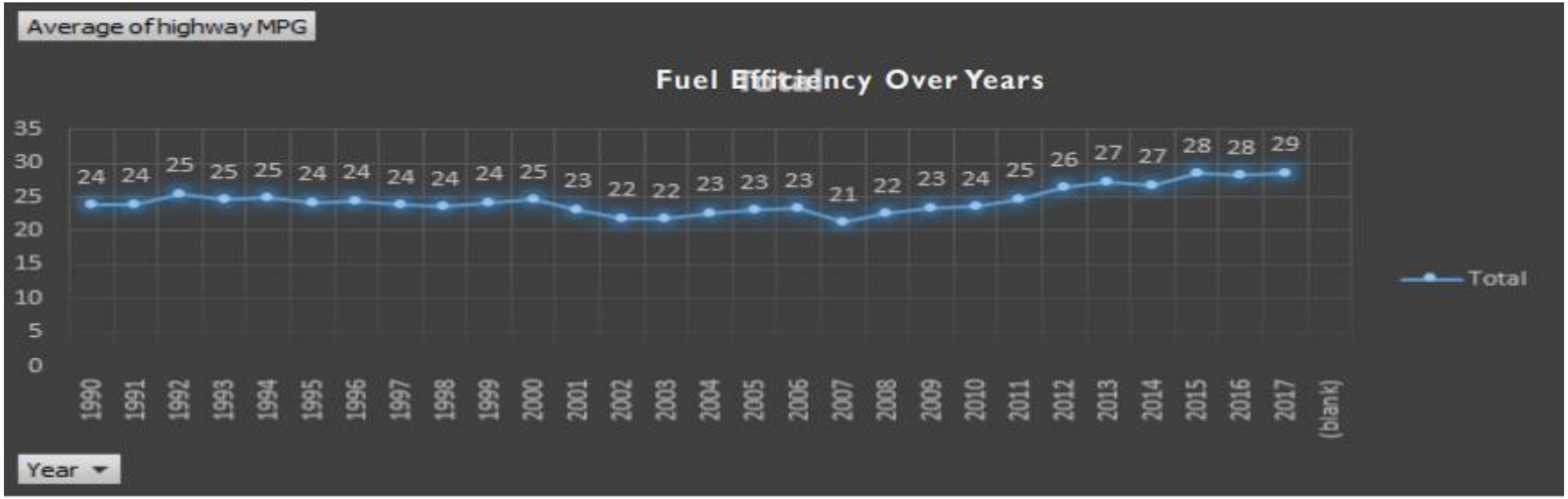
Task 3: To visualize how the different feature such as transmission type affect the MSRP, and how does this vary by body style. Answer: To create the Pivot table Vehicle Style is taken as the rows and Transmission Type to the column with the Average of MSRP as the values.



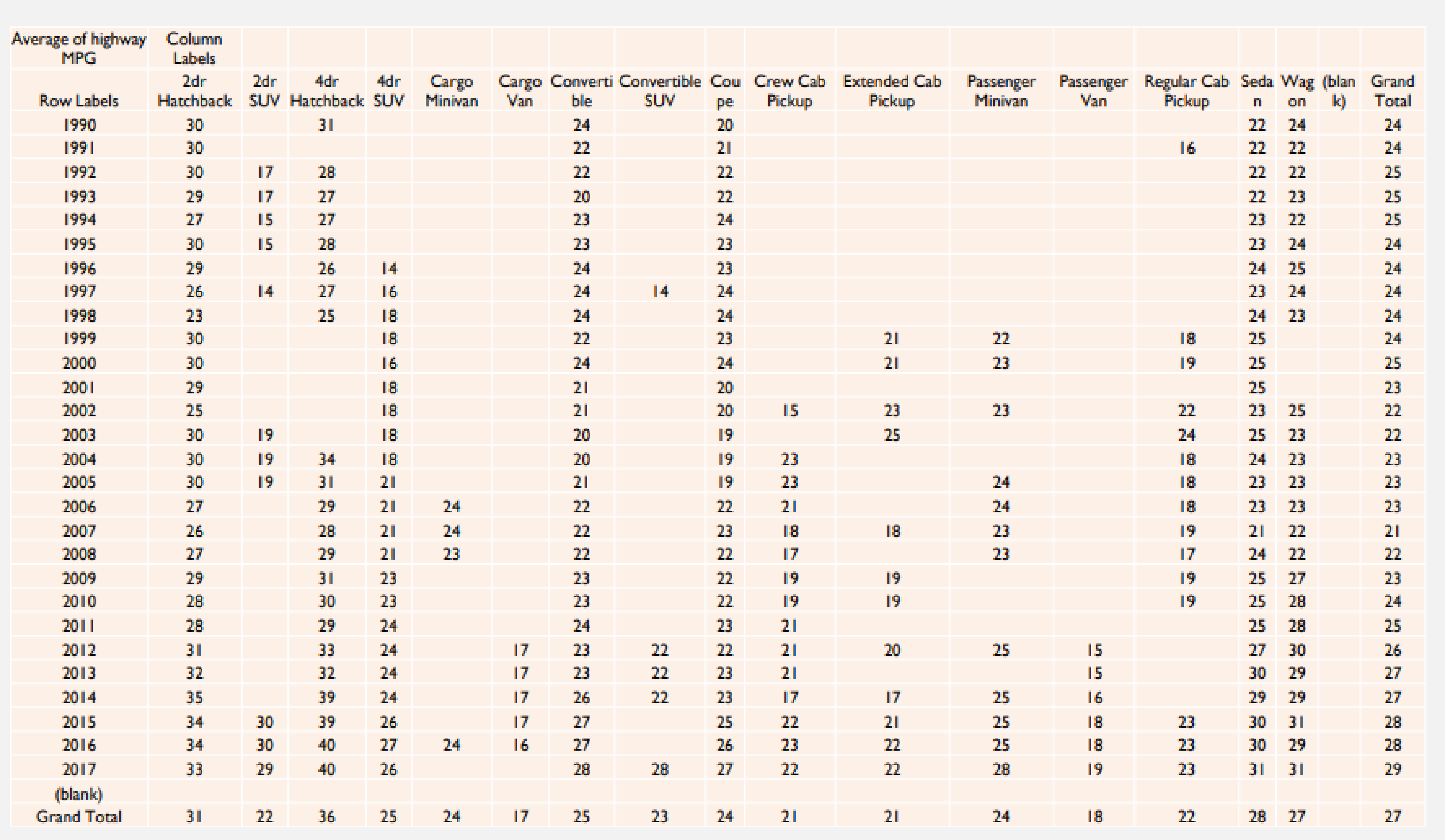


Average of MSRP	Column Labels						
Row Labels	AUTOMATED_MANUAL	AUTOMATIC	DIRECT_DRIVE	MANUAL	UNKNOWN	(blank)	Grand Total
2dr Hatchback	27181	20926		13354	7362		16779
2dr SUV		35895		29223			34941
4dr Hatchback	29249	23834	33603	17594			22203
4dr SUV	40451	42827	49800	23098			42424
Cargo Minivan		22964					22964
Cargo Van		30725					30725
Convertible	125806	111675		68039	9567		95465
Convertible SUV		46134					46134
Coupe	245977	75004		64550			92896
Crew Cab Pickup		39565		27361			39033
Extended Cab Pickup		32970		10651			30867
Passenger Minivan		26437					26437
Passenger Van		35963					35963
Regular Cab Pickup		29210		18045			27180
Sedan	48491	56470	27823	21943			51169
Wagon	31985	33229		22542			31489
(blank)							
Grand Total	101220	48199	33796	38189	8097		50027

Task 4: To visualize how the fuel efficiency of cars vary across different body styles and model years.
Answer: To create the Pivot Table, Year and vehicle Style from the cleaned data are taken in the rows and the average of Highway MPG to the values.

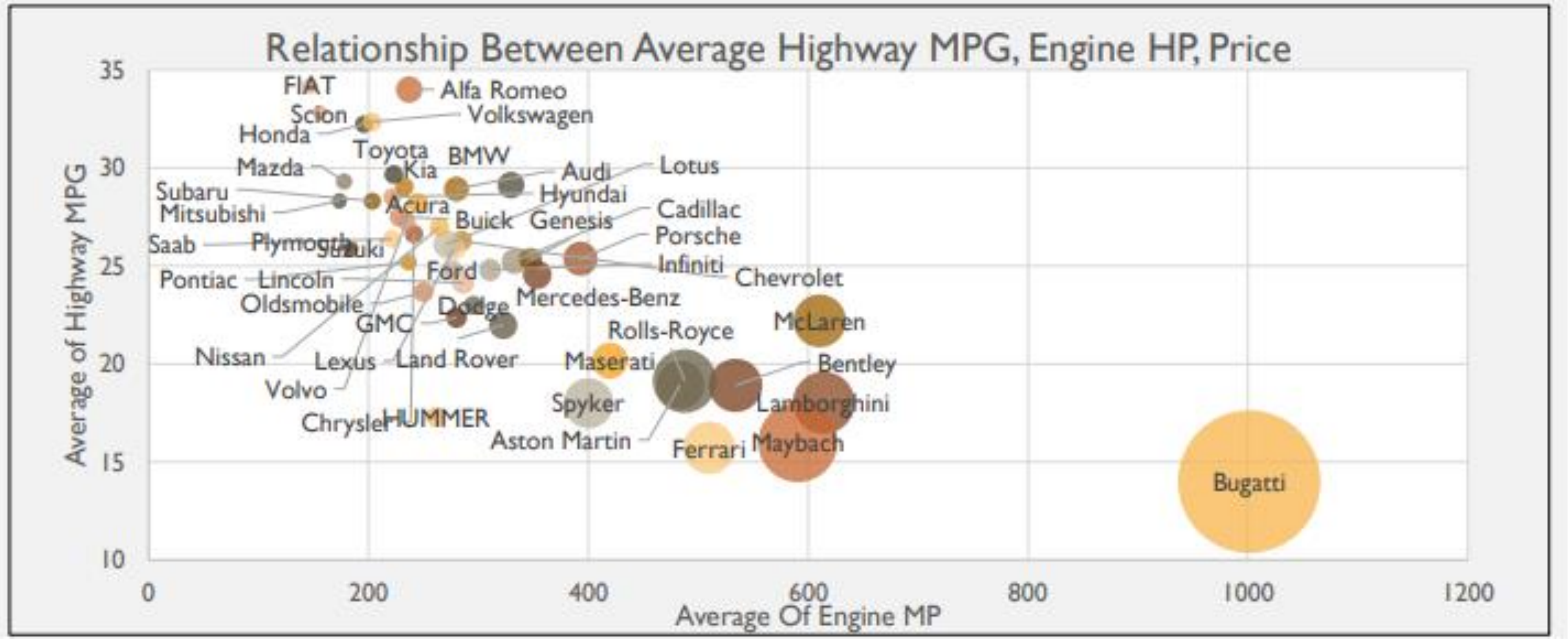


Result: This line chart shows that, the trend having increasing nature after the year 2007 and before that it was in constantly changeable nature.



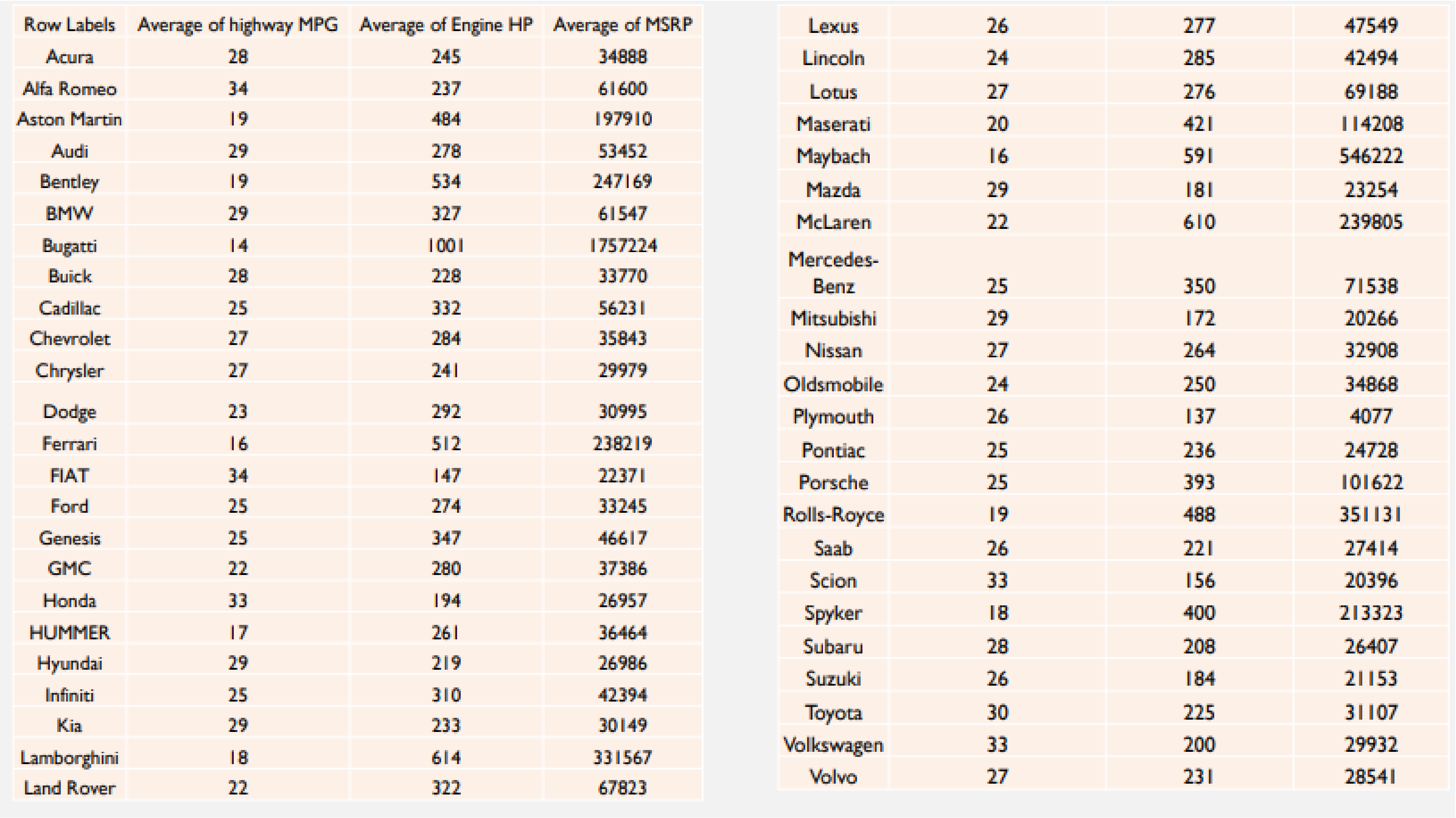
Average of highway MPG	Column Labels																	
Row Labels	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	(blank)	Grand Total
1990	30		31				24		20						22	24		24
1991	30						22		21					16	22	22		24
1992	30	17	28				22		22						22	22		25
1993	29	17	27				20		22						22	23		25
1994	27	15	27				23		24						23	22		25
1995	30	15	28				23		23						23	24		24
1996	29		26	14			24		23						24	25		24
1997	26	14	27	16			24	14	24						23	24		24
1998	23		25	18			24		24						24	23		24
1999	30			18			22		23		21	22		18	25			24
2000	30			16			24		24		21	23		19	25			25
2001	29			18			21		20						25			23
2002	25			18			21		20	15	23	23		22	23	25		22
2003	30	19		18			20		19		25			24	25	23		22
2004	30	19	34	18			20		19	23				18	24	23		23
2005	30	19	31	21			21		19	23		24		18	23	23		23
2006	27		29	21	24		22		22	21		24		18	23	23		23
2007	26		28	21	24		22		23	18	18	23		19	21	22		21
2008	27		29	21	23		22		22	17		23		17	24	22		22
2009	29		31	23			23		22	19	19			19	25	27		23
2010	28		30	23			23		22	19	19			19	25	28		24
2011	28		29	24			24		23	21					25	28		25
2012	31		33	24		17	23	22	22	21	20	25	15		27	30		26
2013	32		32	24		17	23	22	23	21			15		30	29		27
2014	35		39	24		17	26	22	23	17	17	25	16		29	29		27
2015	34	30	39	26		17	27		25	22	21	25	18	23	30	31		28
2016	34	30	40	27	24	16	27		26	23	22	25	18	23	30	29		28
2017	33	29	40	26			28	28	27	22	22	28	19	23	31	31		29
(blank)																		
Grand Total	31	22	36	25	24	17	25	23	24	21	21	24	18	22	28	27		27

Task 5: To visualize how the car's horsepower, MPG, and price vary across different Brands. Answer: The Pivot Table is created by taking the Car Brand from the cleaned data to the row and Average of Engine HP, Highway MPG and MSRP to the values.



Result: It is observed that as the Engine HP increases, the Highway MPG decreases and the price also increases.

The pivot table is given below



Row Labels	Average of highway MPG	Average of Engine HP	Average of MSRP
Acura	28	245	34888
Alfa Romeo	34	237	61600
Aston Martin	19	484	197910
Audi	29	278	53452
Bentley	19	534	247169
BMW	29	327	61547
Bugatti	14	1001	1757224
Buick	28	228	33770
Cadillac	25	332	56231
Chevrolet	27	284	35843
Chrysler	27	241	29979
Dodge	23	292	30995
Ferrari	16	512	238219
FIAT	34	147	22371
Ford	25	274	33245
Genesis	25	347	46617
GMC	22	280	37386
Honda	33	194	26957
HUMMER	17	261	36464
Hyundai	29	219	26986
Infiniti	25	310	42394
Kia	29	233	30149
Lamborghini	18	614	331567
Land Rover	22	322	67823
Lexus	26	277	47549
Lincoln	24	285	42494
Lotus	27	276	69188
Maserati	20	421	114208
Maybach	16	591	546222
Mazda	29	181	23254
McLaren	22	610	239805
Mercedes-Benz	25	350	71538
Mitsubishi	29	172	20266
Nissan	27	264	32908
Oldsmobile	24	250	34868
Plymouth	26	137	4077
Pontiac	25	236	24728
Porsche	25	393	101622
Rolls-Royce	19	488	351131
Saab	26	221	27414
Scion	33	156	20396
Spyker	18	400	213323
Subaru	28	208	26407
Suzuki	26	184	21153
Toyota	30	225	31107
Volkswagen	33	200	29932
Volvo	27	231	28541

RESULTS

- It is observed that the count for the model of market category Crossover is the highest.
- It is observed that the Engine Power and Price are dependent on each other. As the Engine Power increases the Price of the model increases.
- It is observed that the Engine Cylinders has the highest correlation coefficient with the Car Price. i.e, the price of car increases as the Number of Engine cylinders increases.
- The Manufacturer named Bugati has the Highest Average Price.
- As the number of Engine Cylinders goes on increasing the estimated miles per gallon the car gets on the highway goes on decreasing.
- The correlation between the number of cylinders and Highway MPG is -0.6033.
- It is observed that the Car brand Mercedes-Benz has the highest Price.
- Car Bugatti has the highest average price and Plymouth has the lowest average price.

7. ABC Call Volume Trend Analysis

- Project Description:

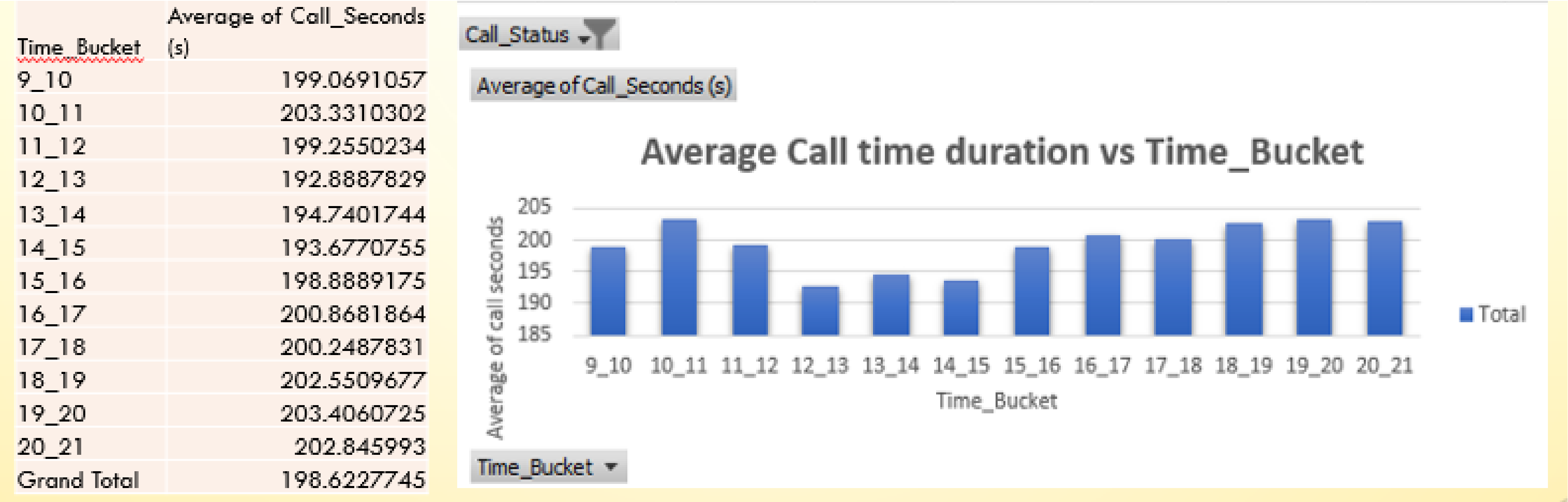
- As a data analyst in this project, I'm delving into customer experience (cx) analytics, with a specific focus on the inbound calling team of a company. The dataset I'm working with covers 23 days and contains details like agent information, queue time, call timing, call duration, and call status (abandoned, answered, or transferred). The cx team's primary role is to gather and analyze customer feedback and data, drawing valuable insights for the organization. They manage customer experience programs, communication, journey mapping, and data management. Tn the modern era, ai tools like ivr, rpa, predictive analytics, and intelligent routing are integral to enhancing customer experience.
- Within the cx team, customer service representatives, or call center agents, play a pivotal role in providing support to customers through various channels, including inbound calls. the ultimate goal is to engage and satisfy customers, nurturing their loyalty and advocacy for the business.
- Inbound customer support is the focus of this project, involves handling incoming calls from existing or prospective customers. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business.

Insights

- The first step of analysis is data cleaning. Initially the data contains total117988 rows and 13 columns and after that I replaces NA values with special character ‘-’.
- Also I proffered to remove unwanted columns, so I removed two columns from data which are: Agent_name and Agaent_ID. Then I started analysis with 70111 rows and 11 columns.
- Now lets see the tasks and insights gained from each task that asked in this project.

To determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.

Task: what is the average duration of calls for each time bucket?



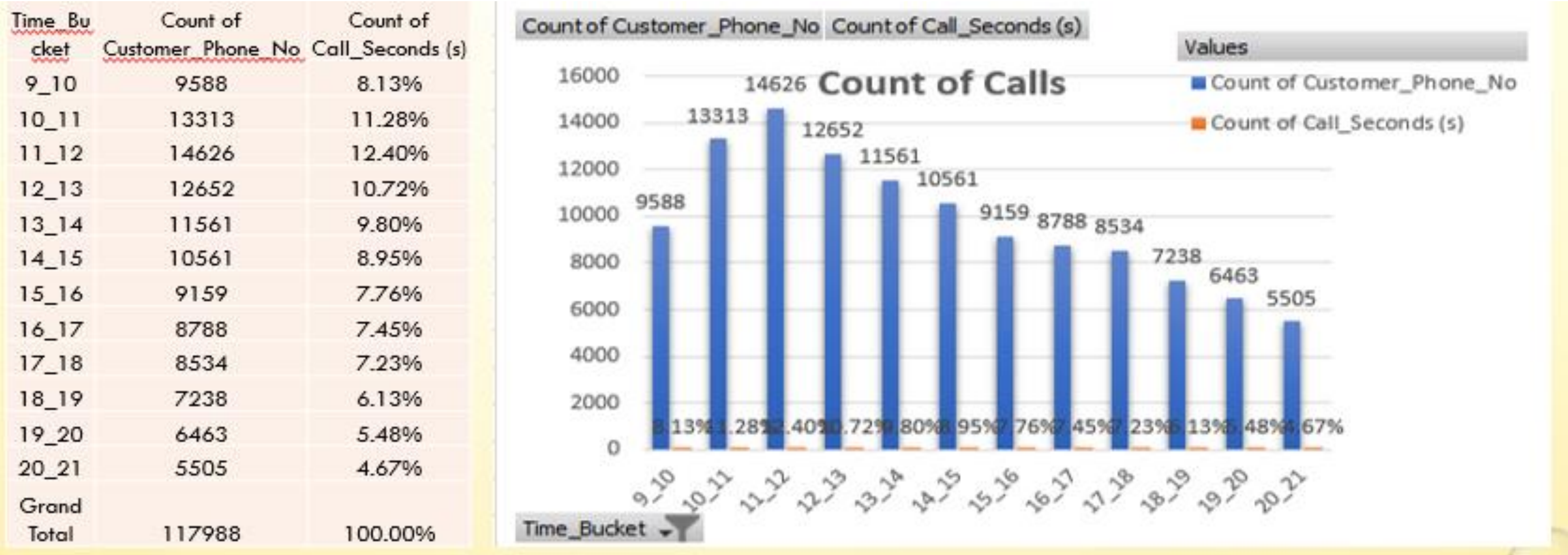
Time_Bucket	Average of Call_Seconds (s)
9_10	199.0691057
10_11	203.3310302
11_12	199.2550234
12_13	192.8887829
13_14	194.7401744
14_15	193.6770755
15_16	198.8889175
16_17	200.8681864
17_18	200.2487831
18_19	202.5509677
19_20	203.4060725
20_21	202.845993
Grand Total	198.6227745

Results: From above table and chart the average duration of calls for each time bucket can be easily seen. And we can see that at after 10_11 number of calls falls down till 13_14 and later again they increases till 20_21.

Task-2: Call volume analysis

To visualize the total number of calls received.

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



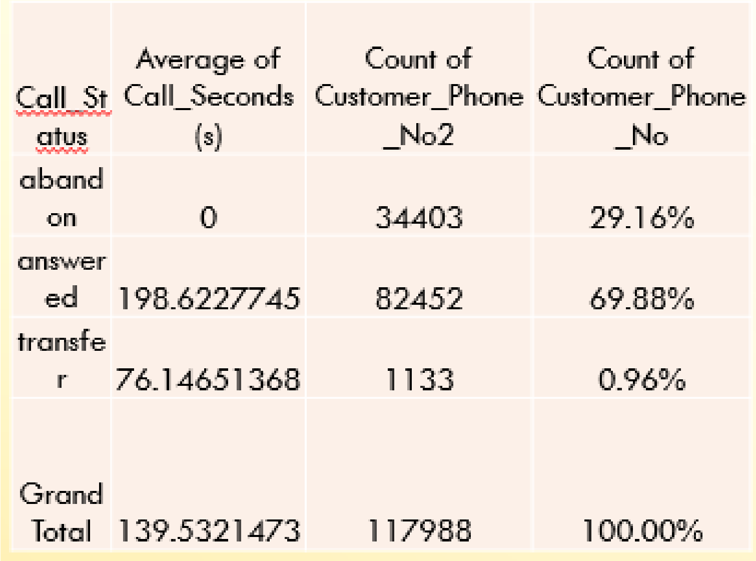
Time_Bucket	Count of Customer_Phone_No	Count of Call_Seconds (s)
9_10	9588	8.13%
10_11	13313	11.28%
11_12	14626	12.40%
12_13	12652	10.72%
13_14	11561	9.80%
14_15	10561	8.95%
15_16	9159	7.76%
16_17	8788	7.45%
17_18	8534	7.23%
18_19	7238	6.13%
19_20	6463	5.48%
20_21	5505	4.67%
Grand Total	117988	100.00%

Results: From above table and chart we can see the total number of calls received in each time bucket. Here many calls received in between 9_10 to 14_15 time bucket and then the least calls are observed in between 20_21 time bucket

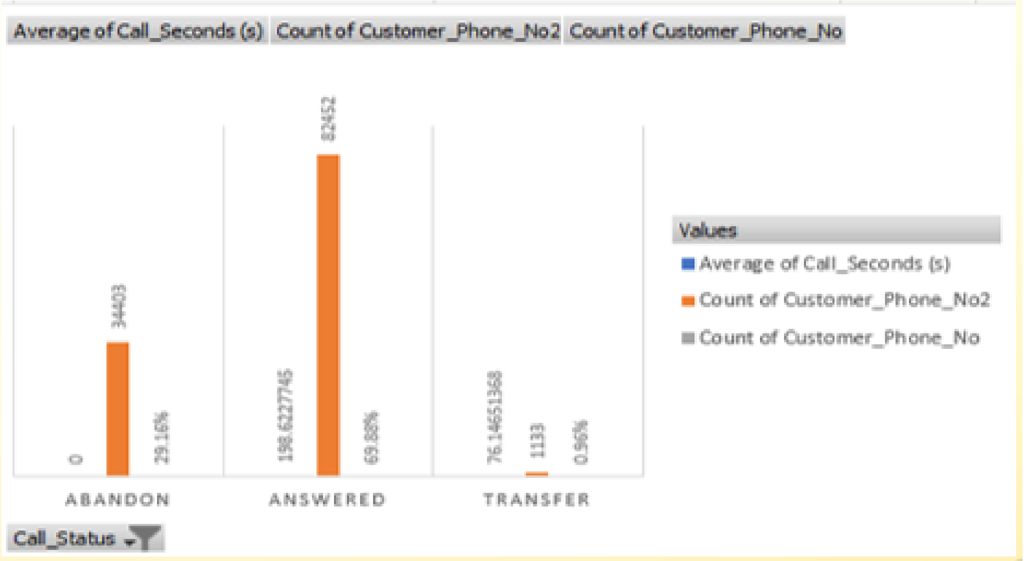
Task-3: Manpower planning

The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%.

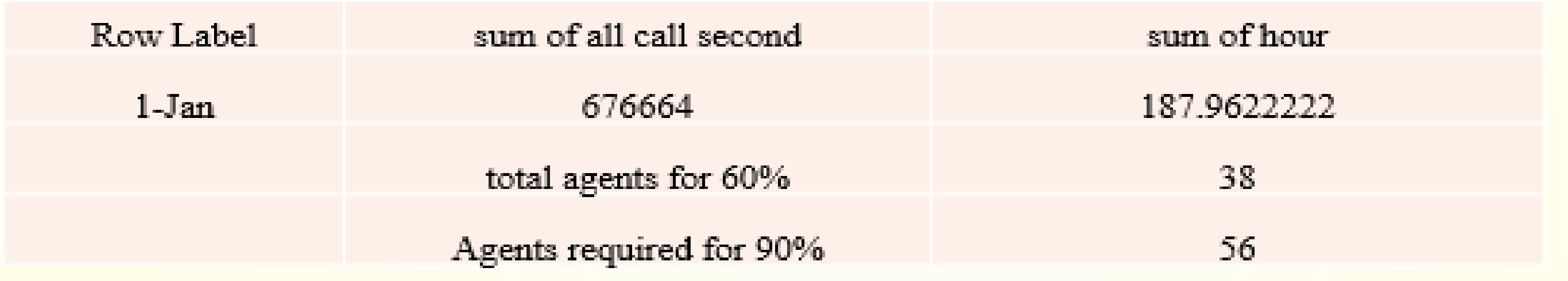
Task: what is the minimum number of agents required in each time bucket to reduce the abandon rate to 10



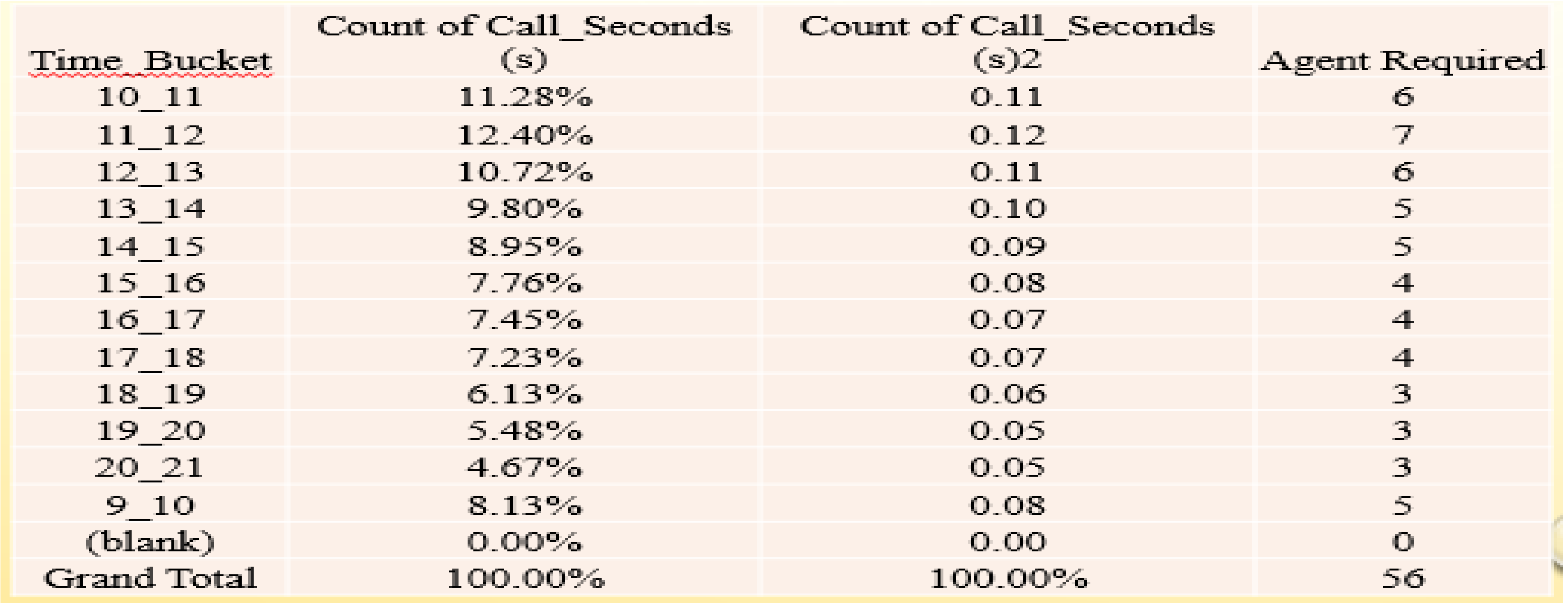
Call_Status	Average of Call_Seconds (s)	Count of Customer_Phone_No2	Count of Customer_Phone_No
abandon	0	34403	29.16%
answered	198.6227745	82452	69.88%
transfer	76.14651368	1133	0.96%
Grand Total	139.5321473	117988	100.00%



Results: From above table and chart we see that ,the abandoned call rate is 29% which is approximately 30%.



Row Label	sum of all call second	sum of hour
1-Jan	676664	187.9622222
	total agents for 60%	38
	Agents required for 90%	56



Time_Bucket	Count of Call_Seconds (s)	Count of Call_Seconds (s)2	Agent Required
10_11	11.28%	0.11	6
11_12	12.40%	0.12	7
12_13	10.72%	0.11	6
13_14	9.80%	0.10	5
14_15	8.95%	0.09	5
15_16	7.76%	0.08	4
16_17	7.45%	0.07	4
17_18	7.23%	0.07	4
18_19	6.13%	0.06	3
19_20	5.48%	0.05	3
20_21	4.67%	0.05	3
9_10	8.13%	0.08	5
(blank)	0.00%	0.00	0
Grand Total	100.00%	100.00%	56

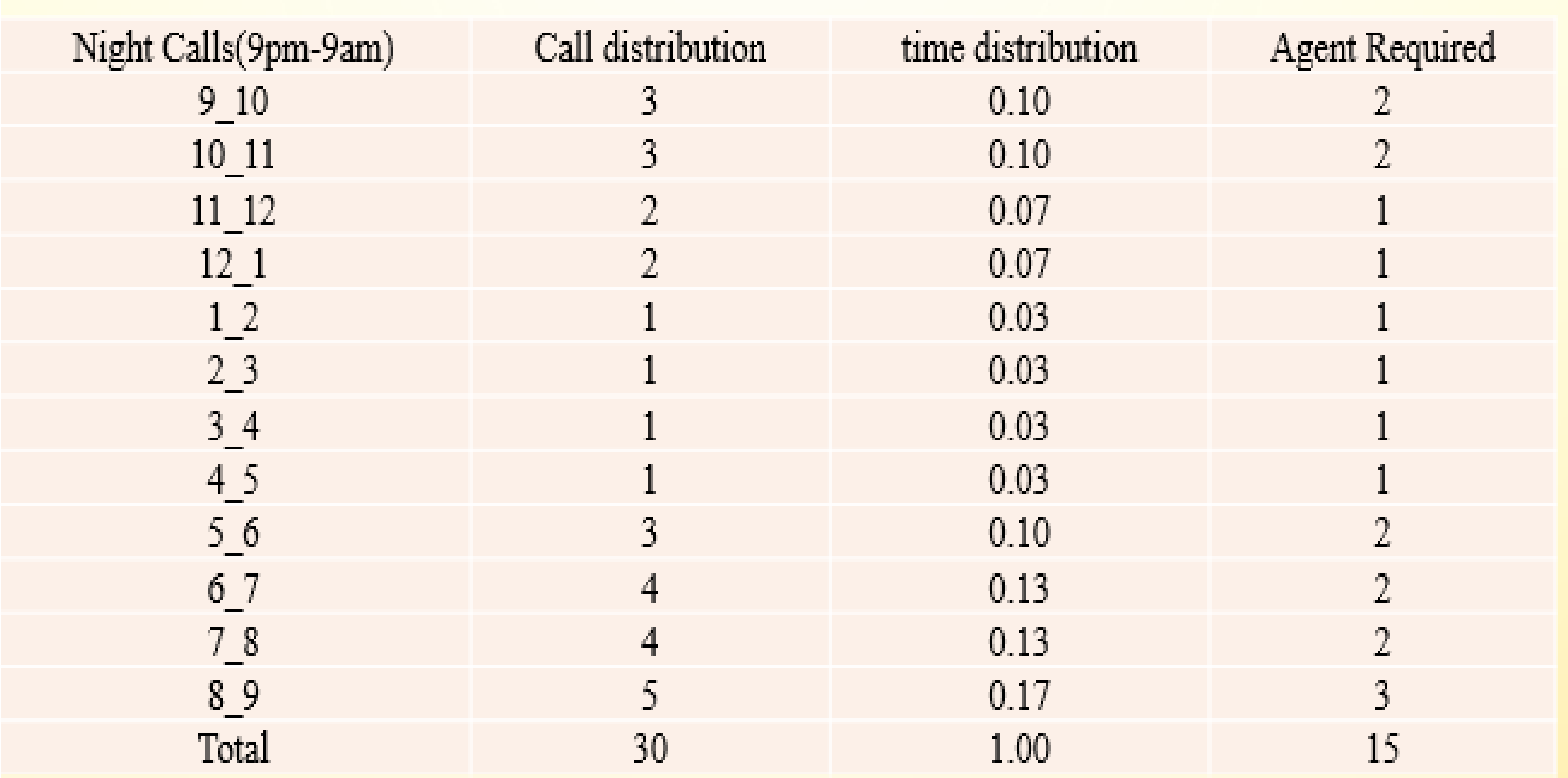
Results: From both thses tables we can easily see the agents required to receive 90% calls.

The top most requirement is of 7 agents to receive calls between 11_12.

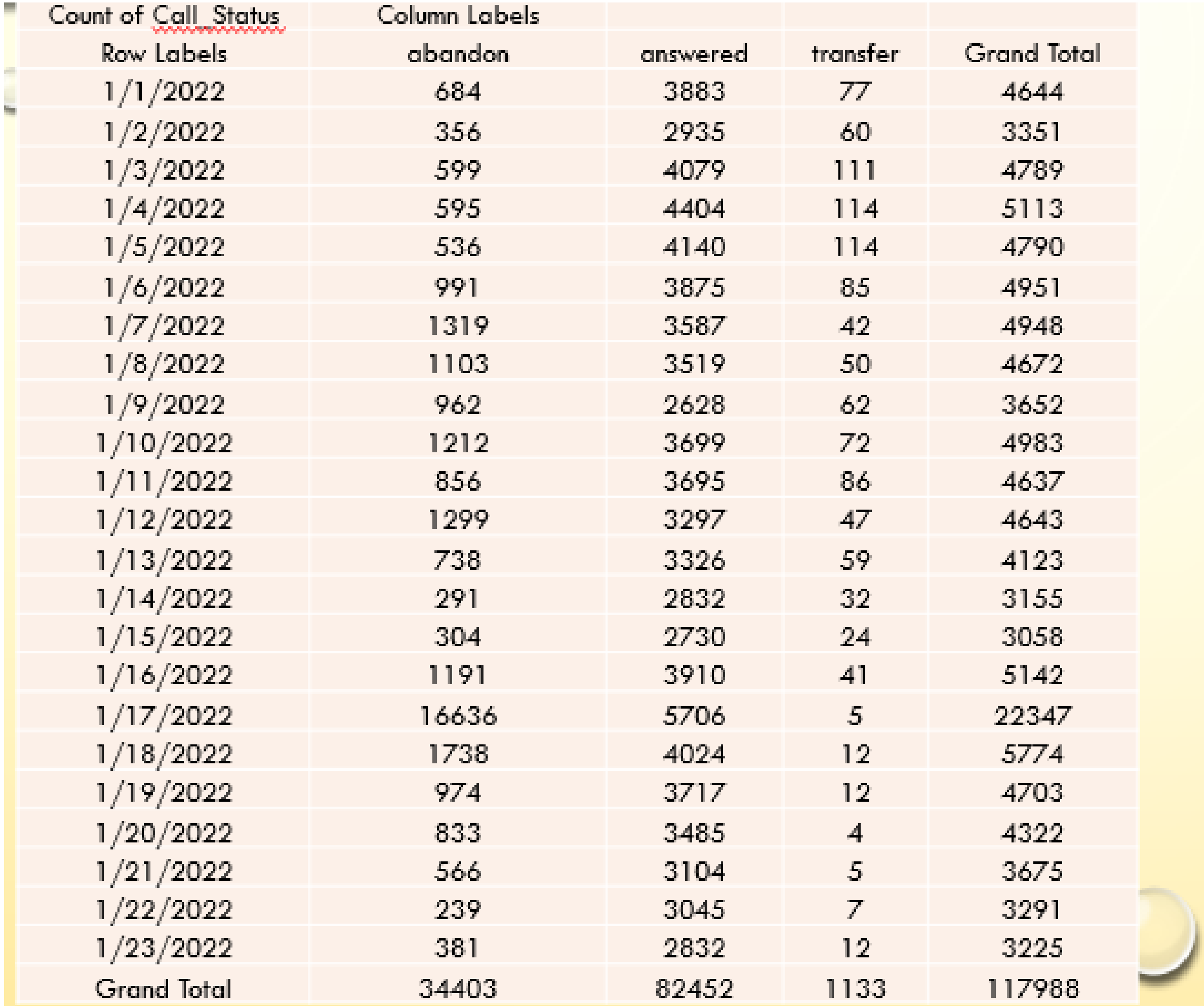
Task-4: Night shift manpower planning

Customers also call ABC insurance company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am.

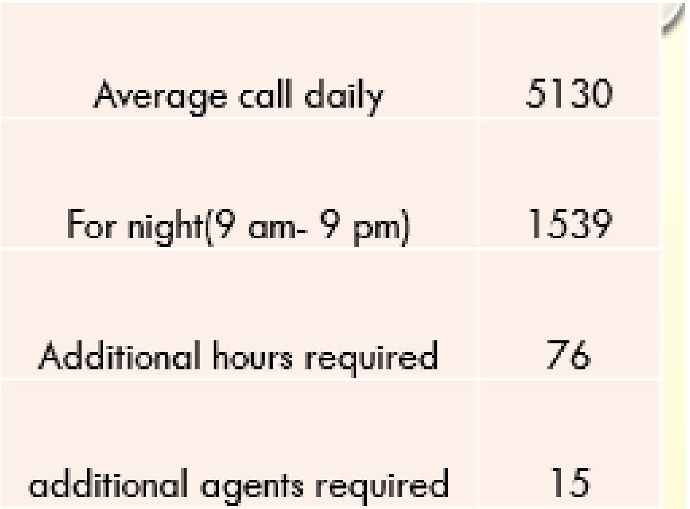
Task: propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.



Night Calls(9pm-9am)	Call distribution	time distribution	Agent Required
9_10	3	0.10	2
10_11	3	0.10	2
11_12	2	0.07	1
12_1	2	0.07	1
1_2	1	0.03	1
2_3	1	0.03	1
3_4	1	0.03	1
4_5	1	0.03	1
5_6	3	0.10	2
6_7	4	0.13	2
7_8	4	0.13	2
8_9	5	0.17	3
Total	30	1.00	15



Count of Call_Status	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
1/1/2022	684	3883	77	4644
1/2/2022	356	2935	60	3351
1/3/2022	599	4079	111	4789
1/4/2022	595	4404	114	5113
1/5/2022	536	4140	114	4790
1/6/2022	991	3875	85	4951
1/7/2022	1319	3587	42	4948
1/8/2022	1103	3519	50	4672
1/9/2022	962	2628	62	3652
1/10/2022	1212	3699	72	4983
1/11/2022	856	3695	86	4637
1/12/2022	1299	3297	47	4643
1/13/2022	738	3326	59	4123
1/14/2022	291	2832	32	3155
1/15/2022	304	2730	24	3058
1/16/2022	1191	3910	41	5142
1/17/2022	16636	5706	5	22347
1/18/2022	1738	4024	12	5774
1/19/2022	974	3717	12	4703
1/20/2022	833	3485	4	4322
1/21/2022	566	3104	5	3675
1/22/2022	239	3045	7	3291
1/23/2022	381	2832	12	3225
Grand Total	34403	82452	1133	117988



Average call daily	5130
For night(9 am- 9 pm)	1539
Additional hours required	76
additional agents required	15

Results: From both theses tables we can easily see that, From 9 pm to 9 am we need 15 agents to receive calls.

Findings

- Based on the project done, here are some major findings:

1. User Engagement Insights:

- - The 33rd week of 2014 had the highest user engagement, while the 35th week of the same year saw the lowest engagement, indicating a significant fluctuation in user activity during that period.

2. Marketing Analysis:

- - The most commonly used hashtags on the platform are "smile," "beach," "party," "fun," and "concert," which can be leveraged for marketing campaigns.
- - Consider running ad campaigns on Thursdays and Sundays, as these days have the highest registration rates, potentially leading to increased user acquisition.

3. Investor Metrics:

- - Among 100 users, 13 were identified as potentially fake based on their liking patterns, which may impact the platform's credibility.

4. Employee Demographics:

- - The company employs 2,675 females and 4,085 males, indicating a gender imbalance in the workforce.
- - The General Management Department has the highest average salary, while the Marketing Department has the lowest, suggesting potential areas for salary adjustment or improvement.

5. Movie Insights:

- - Comedy is the most common movie genre, with 991 movies, while Film-noir has the highest average IMDB score (7.6).
- - Director Akira Kurosawa stands out with the highest average IMDB score of 8.7 and a 100% success rate.

6. Financial Data Analysis:

- - The majority of clients become default due to reasons other than market category or loan type.
- - Cash Loans show a higher default rate compared to Revolving Loans, and clients with incomes ranging from 25,000 to 1,025,000 tend to have the highest default rates.
- - There are no defaults among Businessmen and students, suggesting they are more reliable borrowers.

7. Car Market Observations:

- - Market category "Crossover" has the highest count, indicating strong consumer interest in this type of vehicle.
- - There is a clear positive correlation between Engine Power and Car Price, and the number of Engine Cylinders has the highest correlation coefficient with Car Price, implying that more cylinders tend to increase the price.
- - Bugatti is the manufacturer with the highest average car price, and as the number of cylinders increases, the estimated highway MPG decreases.

conclusion:

- User engagement on the platform exhibits fluctuations, with peaks and troughs during different weeks. It's crucial to understand these patterns for content scheduling. Leveraging popular hashtags and running ad campaigns on Thursdays and Sundays can enhance user acquisition.

- Identifying potential fake users is important for maintaining platform credibility and integrity. Regularly monitoring user behavior is essential.

- The workforce shows a gender imbalance, and variations in departmental salaries suggest the need for further analysis and potential adjustments to ensure fair compensation.

- The prevalence of comedy films and the success of director Akira Kurosawa highlight areas for content creation and investment in future film projects.

- Understanding the causes of defaults, differentiating between loan types, and recognizing income ranges prone to defaults can inform risk assessment and lending strategies.

- The popularity of "Crossover" models, the correlation between car features and prices, and the brand preferences provide insights for the automotive industry and pricing strategies.

Thank You