Data Analysis Portfolio

Jeevan Kishore

Professional Background

I am Jeevan Kishore, I have Completed B.E. Electronics and Communication Engineering in 2022 from K.Ramakrishnan College of Technology, Trichy. I have secured 8.01 CGPA and have several skills including Data Analysis, Python, MySQL, Excel, Tableau,

R Programming.

I'm a quick learner and passionate about staying up-to-date with the latest trends and technologies. With my diverse skill set, I am capable of working on a wide range of projects. I have excellent communication and collaboration skills and enjoy working in a team environment to drive project success.

Table of Contents

Data Analytics Process

Project-1

Project Description:

We use Data Analytics in everyday life without even knowing it. Your task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.

Data Analytics Process

1. Scenario: Joining data analytics course in the best learning platform

• PLAN : Planned to shift my career to data analytics domain.

• PREPARE : Then decided to join a course based on my budget and

time that I can devote for learning.

• PROCESS : Then I searched and shortlisted few renowned learning

platforms like Trainity, Udemy, Upgrad, etc..

• ANALYZE : By analyzing these platforms through various criteria like

How it works?, What features they offers?, Placement Track, Fees

Structure, etc.. and finalized the best platform.

• SHARE : I registered for a webinar with trainity where I shared my

willingness to enroll in course.

• ACT : I paid the fees and enrolled in course and started learning.

2. Scenario: Cooking a dish

• PLAN : Planning to make non veg for lunch.

• PREPARE : Decided to cook chicken biryani based on the

preferences and time constrain to cook.

• PROCESS : Went to shop to purchase chicken, rice, curry leaves,

garlic, masalas, etc..

• ANALYZE : Ratio of rice, water quantity to added, ratio of meat, etc..

and the appropriate cooking time for chicken to be cooked.

• SHARE : Conveyed to my friends that I going to cook chicken biryani.

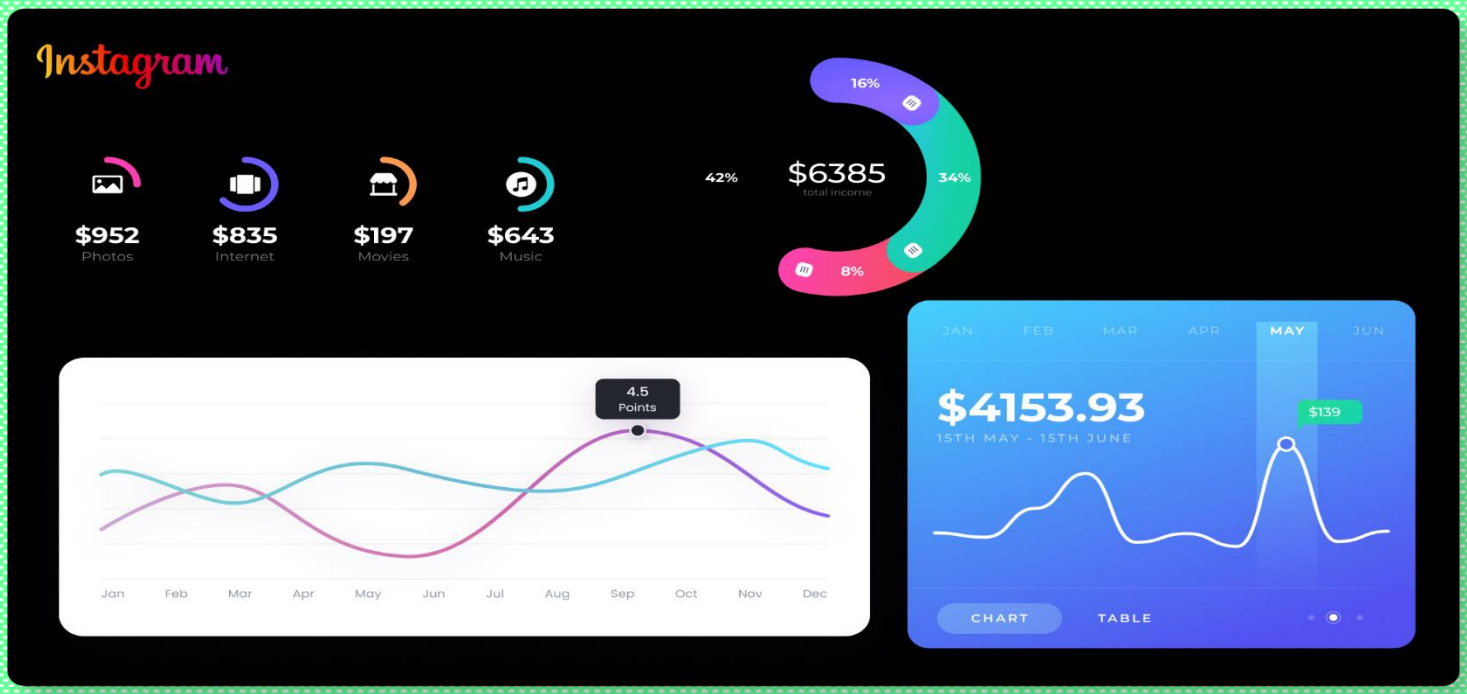
• ACT : The ingredients were organized and started cooking.

Conclusion

Hence, we have seen how we can use the 6 steps of Data Analytics Process while making any decision in real life scenarios.

Instagram User Analytics

Project-2



Project Description :

The aim of this project is to perform analysis on Instagram user data and gain insights into the behavior of Instagram users. SQL was used as a primary tool for data analysis and retrieval. The findings from the analysis provide valuable insights into user activity, preferences and demographics on the popular social media platform. This report presents the methodology and results of the analysis and discusses the implications of the findings for businesses and marketers.

Approach:

Created a database using MySQL that can store the data and support the queries that needs to be processed for analyzing and loaded the data into the database and used MySQL queries to perform the analysis and made insights about the queries that were asked by the management team.

Tech-Stack Used:

MySQL - To run the Query and extract data from database.

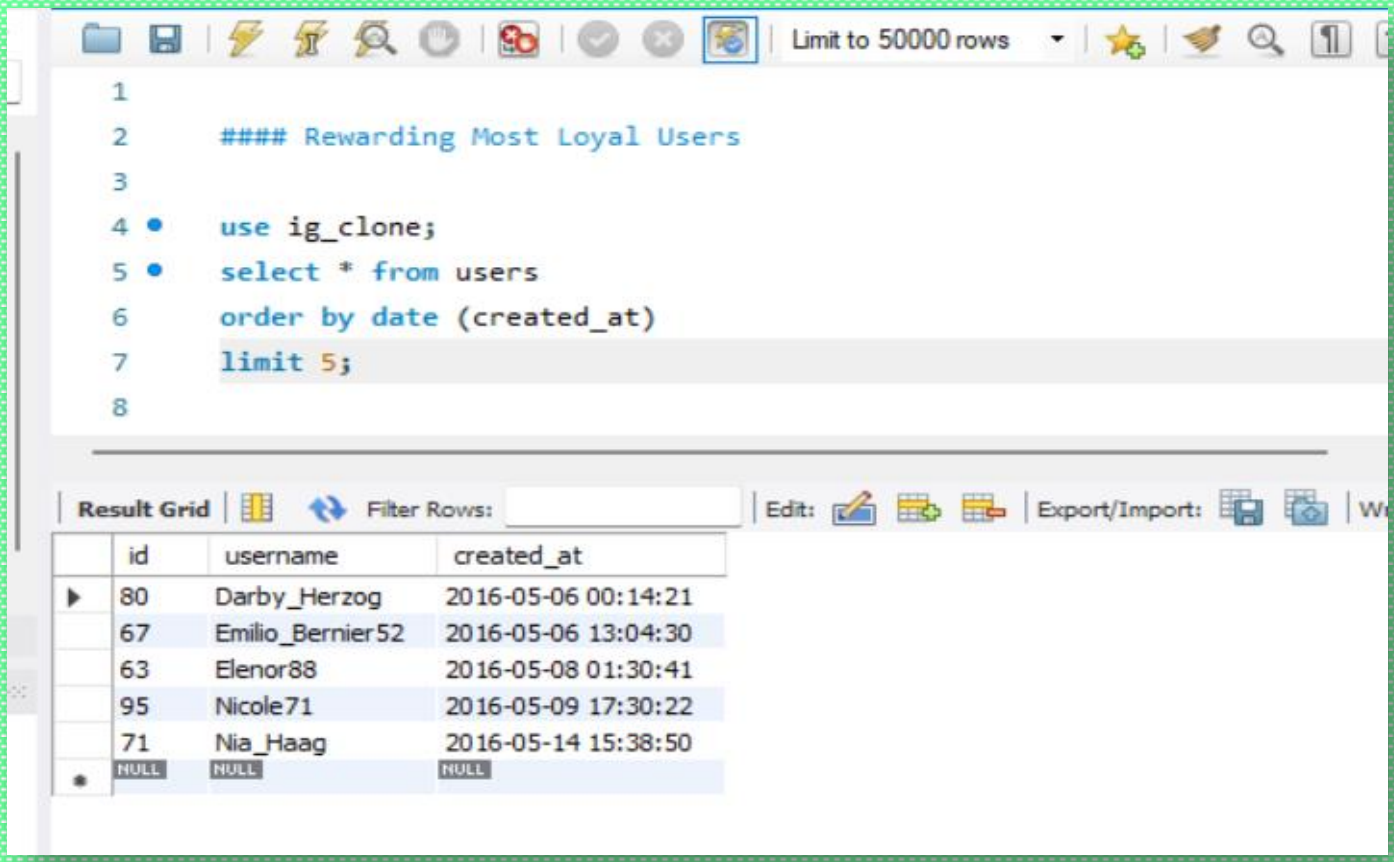
MS Excel – To Store the Extracted tables from the database.

MS PPT – To prepare the report based the insights from the database.

Findings

A) Marketing:

1. Rewarding Most Loyal Users:



They are the 5 oldest users of the Instagram from the database provided.

2. Remind Inactive Users to Start Posting:

```
8
9   ### Remind Inactive Users to Start Posting
10  select * from users
11  left join photos
12  on users.id = photos.user_id
13  where user_id is null;
14
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	username	created_at	id	image_url	user_id	created_dat
5	Aniya_Hackett	2016-12-07 01:04:39	NULL	NULL	NULL	NULL
7	Kasandra_Homenick	2016-12-12 06:50:08	NULL	NULL	NULL	NULL
14	Jadyn81	2017-02-06 23:29:16	NULL	NULL	NULL	NULL
21	Rocio33	2017-01-23 11:51:15	NULL	NULL	NULL	NULL
24	Maxwell.Halvorson	2017-04-18 02:32:44	NULL	NULL	NULL	NULL
25	Tierra.Trantow	2016-10-03 12:49:21	NULL	NULL	NULL	NULL
34	Pearl7	2016-07-08 21:42:01	NULL	NULL	NULL	NULL
36	Ollie_Ledner37	2016-08-04 15:42:20	NULL	NULL	NULL	NULL
41	Mckenna17	2016-07-17 17:25:45	NULL	NULL	NULL	NULL
45	David.Osinski47	2017-02-05 21:23:37	NULL	NULL	NULL	NULL
49	Morgan.Kassulke	2016-10-30 12:42:31	NULL	NULL	NULL	NULL
53	Linnea59	2017-02-07 07:49:34	NULL	NULL	NULL	NULL
54	Duane60	2016-12-21 04:43:38	NULL	NULL	NULL	NULL
57	Julien_Schmidt	2017-02-02 23:12:48	NULL	NULL	NULL	NULL
66	Mike.Auer39	2016-07-01 17:36:15	NULL	NULL	NULL	NULL

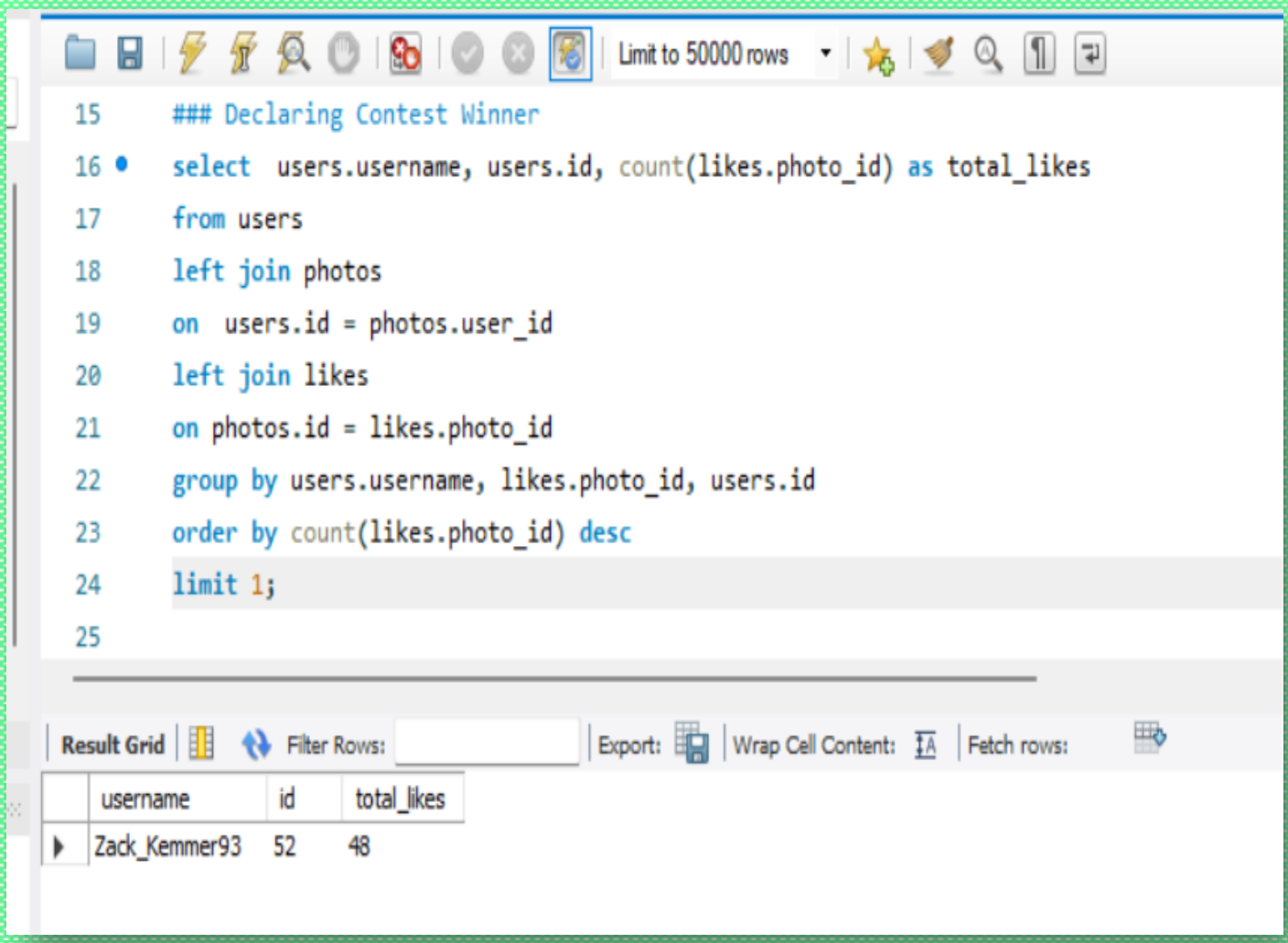
Result 5

```
8
9   ### Remind Inactive Users to Start Posting
10  select * from users
11  left join photos
12  on users.id = photos.user_id
13  where user_id is null;
14
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	username	created_at	id	image_url	user_id	created_dat
54	Duane60	2016-12-21 04:43:38	NULL	NULL	NULL	NULL
57	Julien_Schmidt	2017-02-02 23:12:48	NULL	NULL	NULL	NULL
66	Mike.Auer39	2016-07-01 17:36:15	NULL	NULL	NULL	NULL
68	Franco_Keebler64	2016-11-13 20:09:27	NULL	NULL	NULL	NULL
71	Nia_Haag	2016-05-14 15:38:50	NULL	NULL	NULL	NULL
74	Hulda.Macejkovic	2017-01-25 17:17:28	NULL	NULL	NULL	NULL
75	Leslie67	2016-09-21 05:14:01	NULL	NULL	NULL	NULL
76	Janelle.Nikolaus81	2016-07-21 09:26:09	NULL	NULL	NULL	NULL
80	Darby_Herzog	2016-05-06 00:14:21	NULL	NULL	NULL	NULL
81	Esther.Zulauf61	2017-01-14 17:02:34	NULL	NULL	NULL	NULL
83	Bartholome.Bernhard	2016-11-06 02:31:23	NULL	NULL	NULL	NULL
89	Jessyca_West	2016-09-14 23:47:05	NULL	NULL	NULL	NULL
90	Esmeralda.Mraz57	2017-03-03 11:52:27	NULL	NULL	NULL	NULL
91	Bethany20	2016-06-03 23:31:53	NULL	NULL	NULL	NULL

3. Declaring Contest Winner:



```
15  ### Declaring Contest Winner
16  select  users.username, users.id, count(likes.photo_id) as total_likes
17  from users
18  left join photos
19  on  users.id = photos.user_id
20  left join likes
21  on photos.id = likes.photo_id
22  group by users.username, likes.photo_id, users.id
23  order by count(likes.photo_id) desc
24  limit 1;
25
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

username	id	total_likes
Zack_Kemmer93	52	48

Zack_Kemmer93 is the winner with most number of likes for a single photo.

4. Hashtag Researching :

```
25
26      ### Top 5 Hashtag Researching
27 •    select tag_name as "TOP 5 Tags", count(tags.id) as "No. of Times Used"
28      from photos
29      left join photo_tags
30      on photos.id = photo_tags.photo_id
31      left join tags
32      on photo_tags.tag_id = tags.id
33      Group by tags.id
34      order by count(id) desc
35      limit 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

TOP 5 Tags	No. of Times Used
smile	59
beach	42
party	39
fun	38
concert	24

Result 30

tag_name	count(t.id)
smile	59
beach	42
party	39
fun	38
concert	24

The list of all the hashtags were the most used in Instagram.

5. Launch AD Campaign:

```
36
37      ### Best Day for Launch of AD Campaign
38 •    select dayname(created_at) as Days_of_the_week, count(*) as Total
39      from users
40      group by Days_of_the_week
41      order by Total desc
42
43
44
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Days_of_the_week	Total
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Result 34

days_of_week	count(days_of_week)
Thursday	16
Sunday	16
Tuesday	14
Saturday	12
Wednesday	13
Monday	14
Friday	15

These days of the week were the most users registered on.

B) Investor Metrics

6. User Engagement:

```
43    ###Investor Metrics
44    ###User Engagement
45
46    #total no of users of instagram
47    select count(id) as total_no_of_users from users;
48
49
50
51
52
53
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_no_of_users
100

Total Number of Users in Instagram

```
43    ###Investor Metrics
44    ###User Engagement
45
46    #total no of users of instagram
47    select count(id) as total_no_of_users from users;
48
49    #total no of photos on instagram
50    select count(id) as total_no_of_photos from photos;
51
52
53
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_no_of_photos
257

Total Number of Photos in Instagram

```sql
#average post per user
with cte as
( select users.id, count(photos.id) as post_per_user
from users
left join photos
on users.id = photos.user_id
group by users.id
order by post_per_user desc )
select avg(post_per_user) as "Average Post per User"
from cte;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Average Post per User
2.5700

Average Post per User

7. Bots & Fake Accounts:

```sql
###Bots & Fake Accounts
with cte as
(select likes.user_id, count(likes.user_id) as no_of_likes
from photos
left join likes
on photos.id = likes.photo_id
left join users
on photos.user_id = users.id
group by likes.user_id)
select user_id, users.username,no_of_likes
from cte
left join users
on cte.user_id = users.id
where no_of_likes = 257;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

user_id	username	no_of_likes
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

Result 55

These are the accounts that have liked all the photos that are available, Hence we can conclude that these are bots.

Conclusion

Analyzed many useful insights that could help the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether would help the business grow and this would help in making data driven decisions for the business.

Operational and Metrics Analytics

Project-3

Project Description:

This project involved using SQL to analyse operational data and investigate metric spikes. The goal was to generate insights and answer complex questions related to the organization's operations. I used a range of SQL techniques to answer complex questions, such as subqueries, joins, and aggregations. For example, I used subqueries to compare the performance of different teams or regions, and joins to combine data from multiple

sources. I also calculated metrics such as rolling averages and percentiles to gain a deeper understanding of the data.

Approach:

Use SQL to create a database that can store the data and support the queries that we need to perform the analysis and load the data into the database. Use SQL queries to perform the analysis and retrieve the insights you want to extract.

Tech-Stack Used:

MySQL Workbench 8.0 CE

Case Study 1 (Job Data)

1.Number of jobs reviewed:

```
1  use cs1;
2
3  ###Number of jobs reviewed
4  select ds, count(job_id) as 'Jobs Reviewed per day', sum(time_spen
5  from job_data
6  where ds >='2020-11-01'  and ds <='2020-11-30'
7  group by ds ;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ds	Jobs Reviewed per day	Hours spent per Day
2020-11-30	2	0.0111
2020-11-29	1	0.0056
2020-11-28	2	0.0092
2020-11-27	1	0.0289
2020-11-26	1	0.0156
2020-11-25	1	0.0125

Result 3

2.Throughput:

```
9   ###Throughput
10  SELECT ds,
11  sum(jobs) over (order by ds rows between 6 preceding and current row) /
12  sum(total_time) as throughput_7day_rolling_avg
13  from
14  (SELECT ds, COUNT(job_id) as jobs, SUM(time_spent) as total_time
15  FROM job_data where ds >= '2020-11-01' and ds <='2020-11-30'
16  GROUP BY ds
17  ) d
18  group by ds;
19
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ds	throughput_7d_rolling_avg
2020-11-25	0.0222
2020-11-26	0.0357
2020-11-27	0.0288
2020-11-28	0.1515
2020-11-29	0.3000
2020-11-30	0.2000

0.1515

3.Percentage share of each language:

```
19
20  ###Percentage share of each language
21  select language, (count(*)/(select count(*) from job_data))*100 as percentage
22  from job_data
23  group by language;
24
25
26
27
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

language	percentage
English	12.5000
Arabic	12.5000
Persian	37.5000
Hindi	12.5000
French	12.5000
Italian	12.5000

4.Duplicate rows:

```
24
25      ###Duplicate rows
26      select job_id, count(job_id) as duplicate_rows
27      from job_data
28      group by job_id
29      having count(*) > 1;
30
31
32
33
34
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_id	duplicate_rows
23	3

Case Study 2 (Investigating metric spike)

1.User Engagement:

```
3   SELECT
4       extract(year from occurred_at) as year,
5       extract(week from occurred_at) as weeknum,
6       count(distinct user_id) as user_enagagement
7   FROM
8       events
9   GROUP BY
10      year, weeknum
11  order by
12      year, weeknum
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year	weeknum	user_enagagement
2014	17	663
2014	18	1068
2014	19	1113
2014	20	1154
2014	21	1121

2.User Growth:

```sql
select
year_num,
week_num,
num_active_users,
SUM(num_active_users)OVER(ORDER BY year_num, week_num ROWS BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW) AS cum_active_users
from
(
select
extract(year from activated_at) as year_num,
extract(week from activated_at) as week_num,
count(distinct user_id) as num_active_users
from
users
WHERE
state = 'active'
group by year_num,week_num
order by year_num,week_num
) a;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year_num	week_num	num_active_users	cum_active_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

3.Weekly Retention:

```sql
SELECT
distinct user_id,
COUNT(user_id),
SUM(CASE WHEN retention_week = 1 Then 1 Else 0 END) as per_week_retention
FROM
(
SELECT
a.user_id,
a.signup_week,
b.engagement_week,
b.engagement_week - a.signup_week as retention_week
FROM
(
(SELECT distinct user_id, extract(week from occurred_at) as signup_week from events_data
WHERE event_type = 'signup_flow'
and event_name = 'complete_signup'
and extract(week from occurred_at) = 18
)a
LEFT JOIN
(SELECT distinct user_id, extract(week from occurred_at) as engagement_week FROM events_data
where event_type = 'engagement'
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

user_id	COUNT(user_id)	per_week_retention
11926	1	0
11928	1	0
11929	1	0
11931	1	0
11933	1	0

```sql
SELECT
a.user_id,
a.signup_week,
b.engagement_week,
b.engagement_week - a.signup_week as retention_week
FROM
(
(SELECT distinct user_id, extract(week from occurred_at) as signup_week from events_data
WHERE event_type = 'signup_flow'
and event_name = 'complete_signup'
and extract(week from occurred_at) = 18
)a
LEFT JOIN
(SELECT distinct user_id, extract(week from occurred_at) as engagement_week FROM events_data
where event_type = 'engagement'
)b
on a.user_id = b.user_id
)
)d
group by user_id
order by user_id;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

user_id	COUNT(user_id)	per_week_retention
11936	1	0
11939	3	1
11940	1	0
11942	1	0
11944	1	0

4.Weekly Engagement:

```sql
SELECT
extract(year from occurred_at) as year_num,
extract(week from occurred_at) as week_num,
device,
COUNT(distinct user_id) as no_of_users
FROM
events_data
where event_type = 'engagement'
GROUP by 1,2,3
order by 1,2,3;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

year_num	week_num	device	no_of_users
2014	17	acer aspire desktop	2
2014	17	acer aspire notebook	2
2014	17	amazon fire phone	1
2014	17	asus chromebook	3
2014	17	dell inspiron desktop	1

5.Email Engagement:

```
62  SELECT
63  engagements,
64  total_users,
65  engagements/total_users*100 AS engagement_rate
66  from (  SELECT
67              count(distinct(user_id)) AS total_users,
68  COUNT(DISTINCT
69  CASE
70  WHEN action = 'email_open' THEN user_id
71  WHEN action = 'email_clickthrough' THEN user_id END)  AS engagements
72          FROM email_events) as counts;
73
74
75
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

engagements	total_users	engagement_rate
5927	6179	95.9217

Conclusion

Extracted many useful analysis that could help in making data driven decisions making for the business.

HIRING PROCESS ANALYTICS

Project-4

Project Description:

The project aims is to analyze the hiring process of a company and provide a report about the analysis made out of the data provided. The analysis has been done using Excel and statistical techniques.

Approach:

The dataset which contains tables has the information regarding various fields related to company hiring data. I have gone through the dataset closely and then executed various queries to extract the required data from the table and also made graphs, charts and tables out of the datasets.

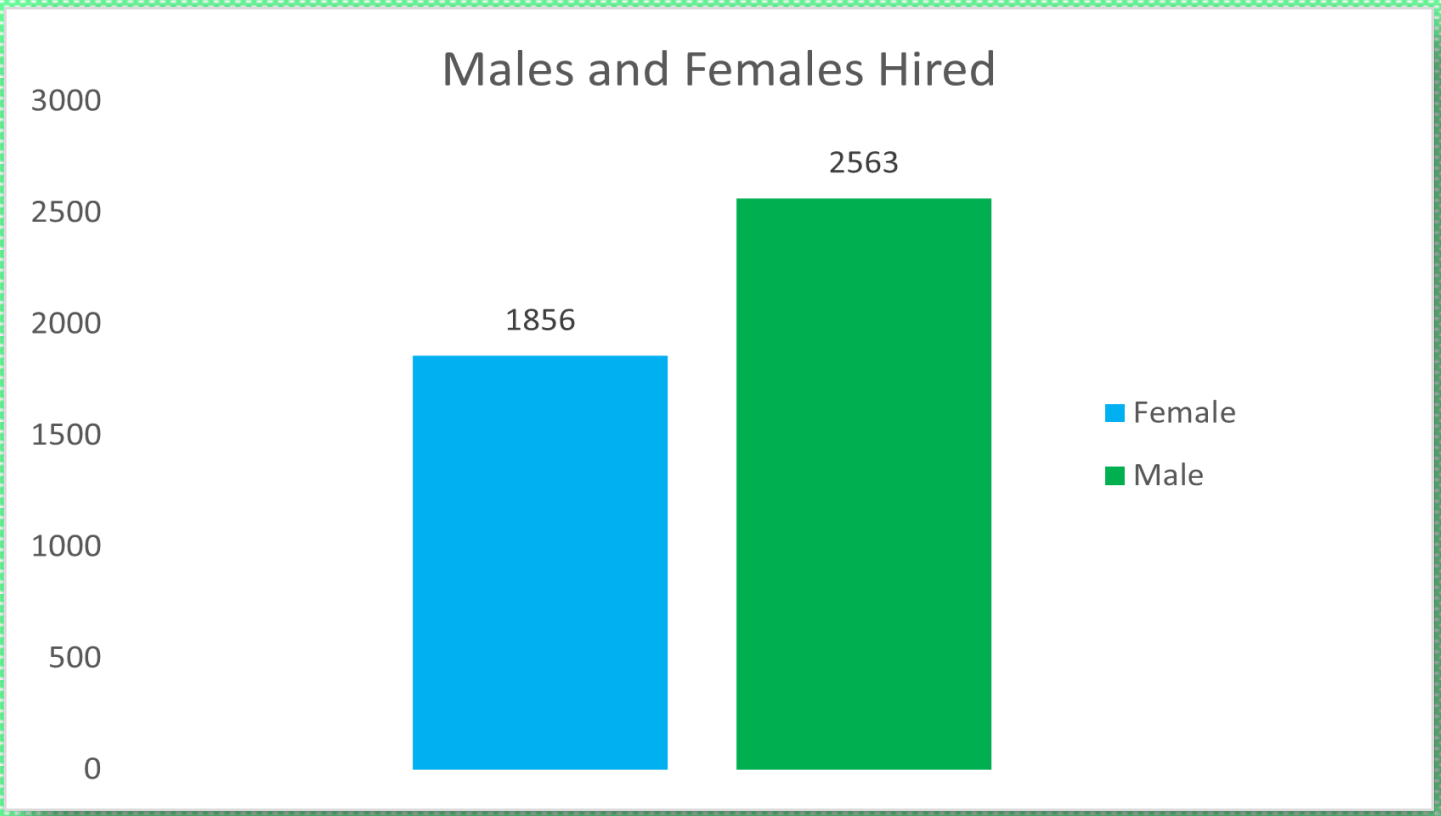
Tech Stack Used:

- PPT – To prepare a detailed report.
- EXCEL – Excel was used to perform entire analysis.

A. Hiring:

How many males and females are Hired?

Status	Hired		
	Column Labels		
	Female	Male	Grand Total
Total no male and female hired	1856	2563	4419



B.Average Salary:

What is the average salary offered in this company?

Offered Salary	(All)
Average of Offered Salary	
49983.03	

C.Class Intervals:

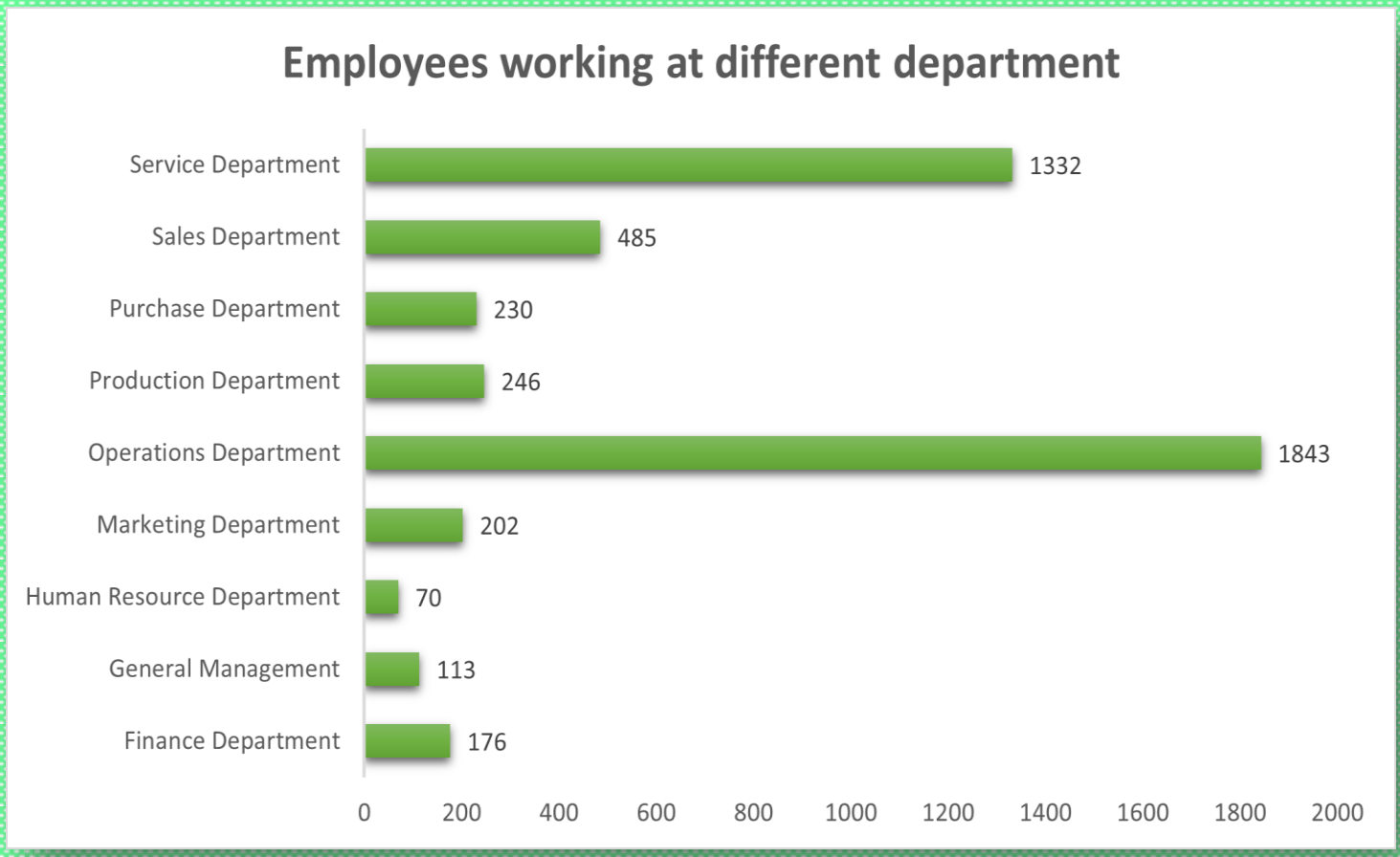
Draw the class intervals for salary in the company?

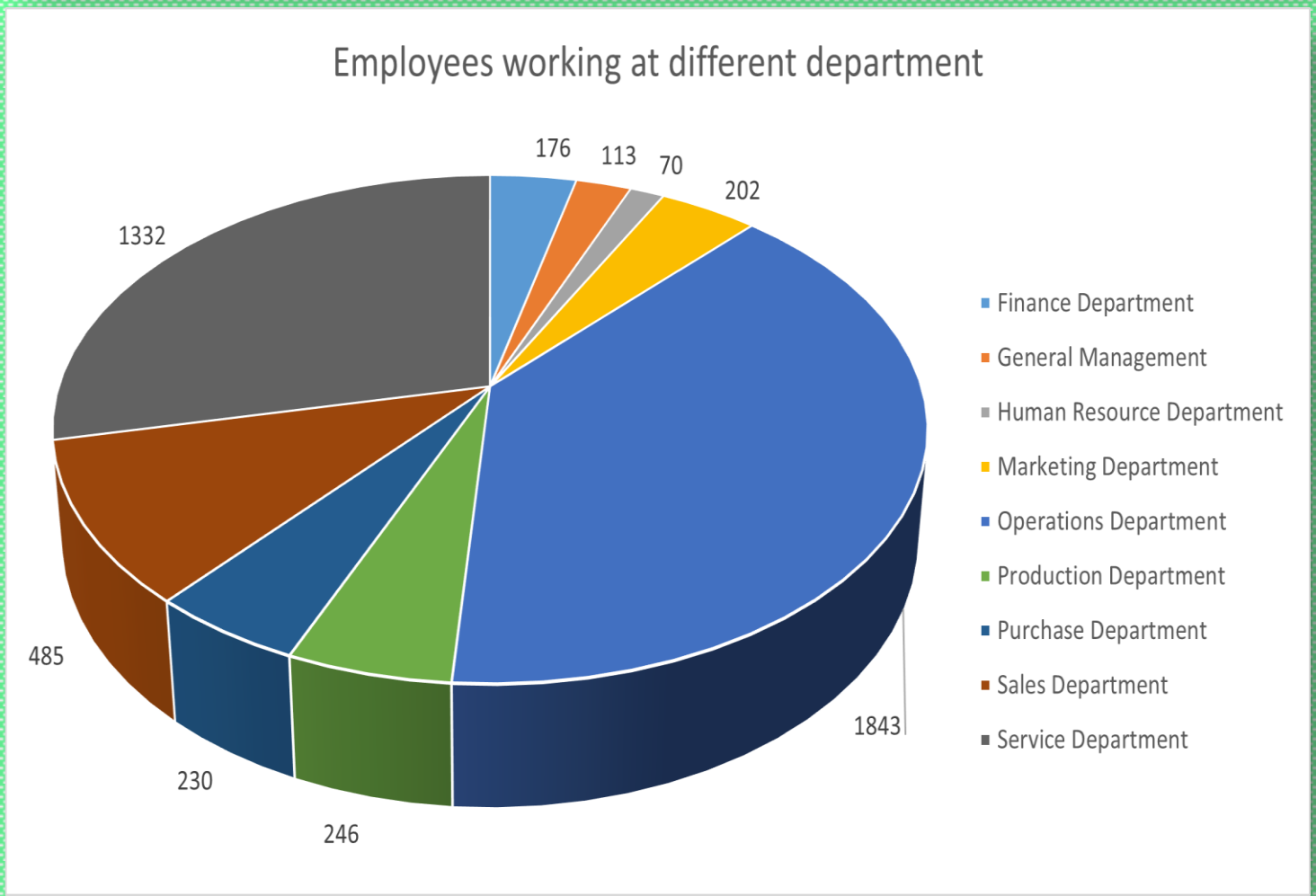
Range	Frequency
1001 - 10000	676
10001 - 20000	732
20001 - 30000	711
30001 - 40000	709
40001 - 50000	780
50001 - 60000	750
60001 - 70000	698
70001 - 80000	734
80001 - 90000	710
90001 - 100000	659

D.Charts and Plots:

Draw Pie Chart / Bar Graph to show proportion of people working different department?

Status	Hired
Row Labels	**Count of Status**
Finance Department	176
General Management	113
Human Resource Department	70
Marketing Department	202
Operations Department	1843
Production Department	246
Purchase Department	230
Sales Department	485
Service Department	1332
Grand Total	**4697**

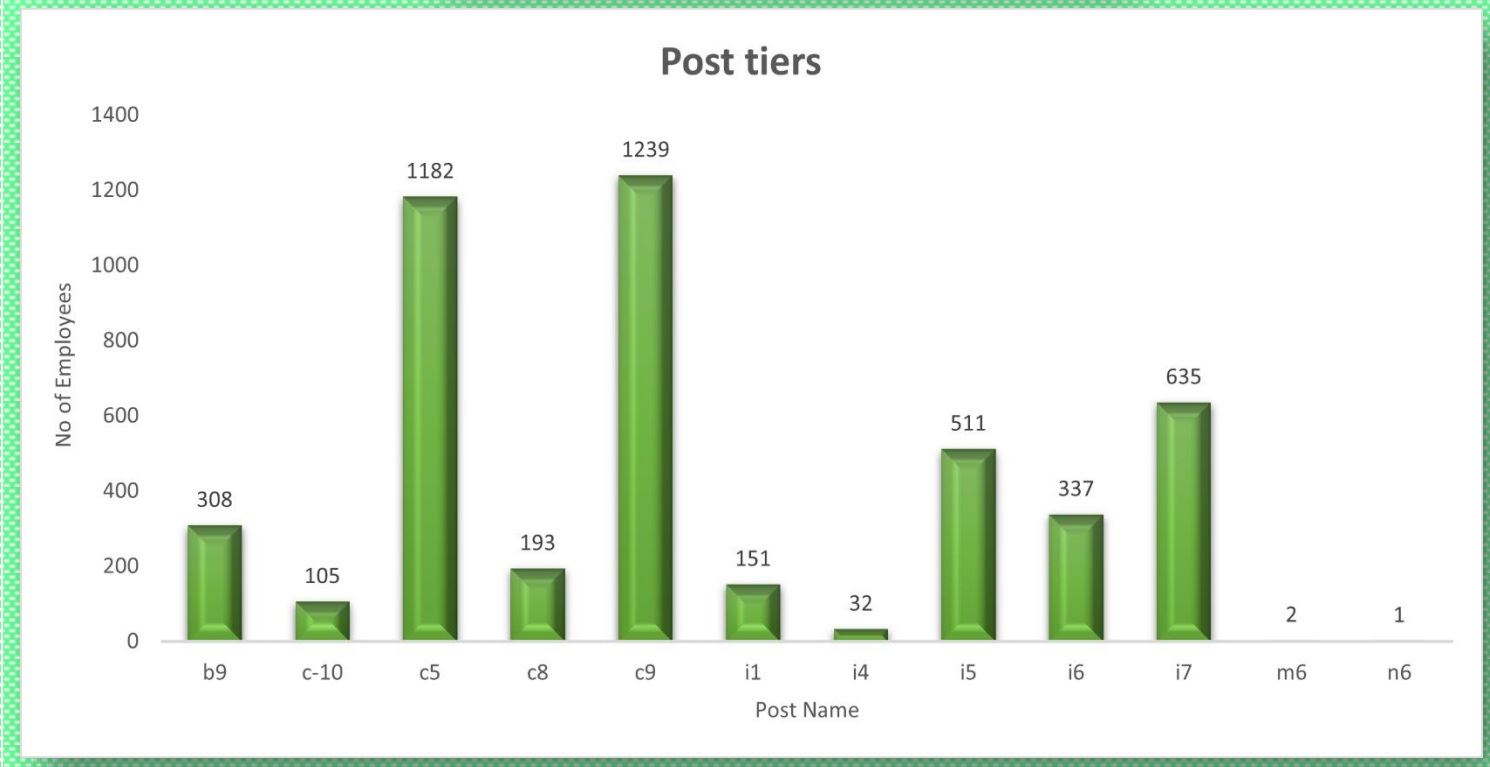

Employees working at different department
Service Department
Sales Department
Purchase Department
Production Department
Operations Department
Marketing Department
Human Resource Department
General Management
Finance Department
1332
485
230
246
1843
202
70
113
176
0
200
400
600
800
1000
1200
1400
1600
1800
2000


Employees working at different department
176
113
70
202
1332
485
230
246
1843
Finance Department
General Management
Human Resource Department
Marketing Department
Operations Department
Production Department
Purchase Department
Sales Department
Service Department

E.Charts:

Represent different post tiers using chart/graph?

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



Conclusion

Overall, this project provided valuable insights into the hiring process and trends within the company, allowing for improvements to be made to ensure a fair and efficient hiring process.

Hiring Process Analytics helps the company to decide the salaries for new freshers joining the company; also it tells requirement of workforce by each department; it also helps the company decide the appraisals and increment for it's current employes.

IMDB Movie Analysis

Project-5

Project Description:

This project is about the IMDB movie analysis based on different criteria. We have IMBD's dataset for different movies, which contains many information and we have to analyze the dataset using the questions that where asked.

Approach:

We have to clean the given datasets which contains tables that contain information regarding various fields related to movies to be rated by IMDB. We have to go through the dataset closely and then executed various queries to extract the required insights from the dataset that was cleaned.

Tech stack used:

PPT – To prepare a detailed report.

EXCEL – Excel was used to perform entire analysis.

Findings:

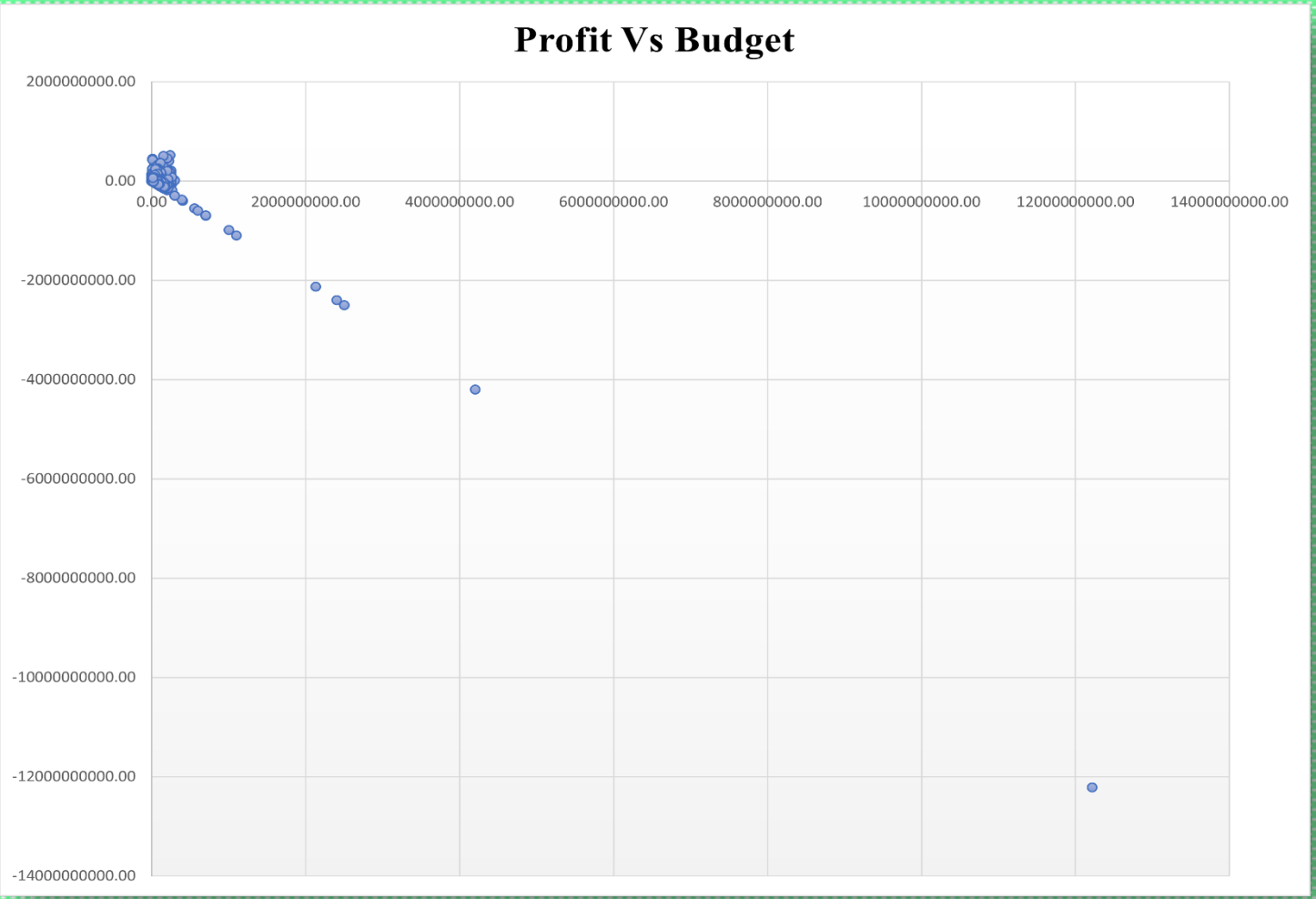
A. Cleaning the data:

1. Dropping unnecessary columns. (Color, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_posts, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes)

2. Remove Blank Cell / Null Value.

3. Removing Duplicate

B. Movies with highest profit:



OUTLIERS:

-12213298588

-4199788333

-2499804112

-2397701809

-2127109510

Top 5 Highest Profit

director_name	actor_1_name	movie_title	title_year	imdb_score	Profit
James Cameron	CCH Pounder	Avatar	2009	7.9	523505847
Colin Trevorrow	Bryce Dallas Howa	Jurassic World	2015	7	502177271
James Cameron	Leonardo DiCapric	Titanic	1997	7.7	458672302
George Lucas	Harrison Ford	Star Wars: Epis	1977	8.7	449935665
Steven Spielberg	Henry Thomas	E.T. the Extra-T	1982	7.9	424449459

C. Top 250:

1.Filtering the data where num_voted_users > 25,000.

2.Sort the data using imbd_score column in largest to smallest score.

3.Create column ‘Rank’ for top 250 movies.

4. Use Sequence Formula to Rank.

5.Formula =SEQUENCE(COUNTA(G2:G251),1,1,1)

6. Filter out language by unselecting English to extract the top foreign language movie.

Top 250 Movies

director_name	num_criti	duration	gross	genres	actor_1_nai	movie_title	num_voted_users	num_use	language	country	budget	title_year	imdb_sc	Profit	rank
Frank Darabont	199	142	28341469	Crime|Dra	Morgan Freem	The Shawshank RedemptionÂ	1689764	4144	English	USA	25000000	1994	9.3	3341469	1
Francis Ford Coppo	208	175	134821952	Crime|Dra	Al Pacino	The GodfatherÂ	1155770	2238	English	USA	6000000	1972	9.2	128821952	2
Christopher Nolan	645	152	533316061	Action|Cri	Christian Bale	The Dark KnightÂ	1676169	4667	English	USA	185000000	2008	9	348316061	3
Francis Ford Coppo	149	220	57300000	Crime|Dra	Robert De Niro	The Godfather: Part IIÂ	790926	650	English	USA	13000000	1974	9	44300000	4
Quentin Tarantino	215	178	107930000	Crime|Dra	Bruce Willis	Pulp FictionÂ	1324680	2195	English	USA	8000000	1994	8.9	99930000	5
Peter Jackson	328	192	377019252	Action|Ad	Orlando Bloom	The Lord of the Rings: The Retu	1215718	3189	English	USA	94000000	2003	8.9	283019252	6
Steven Spielberg	174	185	96067179	Biography	Liam Neeson	Schindler's ListÂ	865020	1273	English	USA	22000000	1993	8.9	74067179	7
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugl	503509	780	Italian	Italy	1200000	1966	8.9	4900000	8
Christopher Nolan	642	148	292568851	Action|Ad	Leonardo DiCa	InceptionÂ	1468200	2803	English	USA	160000000	2010	8.8	132568851	9
David Fincher	315	151	37023395	Drama	Brad Pitt	Fight ClubÂ	1347461	2968	English	USA	63000000	1999	8.8	-25976605	10
Robert Zemeckis	149	142	329691196	Comedy|D	Tom Hanks	Forrest GumpÂ	1251222	1398	English	USA	55000000	1994	8.8	274691196	11
Peter Jackson	297	171	313837577	Action|Ad	Christopher Le	The Lord of the Rings: The Fell	1238746	5060	English	New Zealar	93000000	2001	8.8	220837577	12
Irvin Kershner	223	127	290158751	Action|Ad	Harrison Ford	Star Wars: Episode V - The Emp	837759	900	English	USA	18000000	1980	8.8	272158751	13
Lana Wachowski	313	136	171383253	Action|Sci	Keanu Reeves	The MatrixÂ	1217752	3646	English	USA	63000000	1999	8.7	108383253	14
Peter Jackson	294	172	340478898	Action|Ad	Christopher Le	The Lord of the Rings: The Two	1100446	2417	English	USA	94000000	2002	8.7	246478898	15
George Lucas	282	125	460935665	Action|Ad	Harrison Ford	Star Wars: Episode IV - A New	911097	1470	English	USA	11000000	1977	8.7	449935665	16
Martin Scorsese	192	146	46836394	Biography	Robert De Niro	GoodfellasÂ	728685	989	English	USA	25000000	1990	8.7	21836394	17
Milos Forman	149	133	112000000	Drama	Scatman Croth	One Flew Over the Cuckoo's Ne	680041	760	English	USA	4400000	1975	8.7	107600000	18
Fernando Meirelle	214	135	7563397	Crime|Dra	Alice Braga	City of GodÂ	533200	749	Portuguese	Brazil	3300000	2002	8.7	4263397	19
Akira Kurosawa	153	202	269061	Action|Ad	Takashi Shimur	Seven SamuraiÂ	229012	596	Japanese	Japan	2000000	1954	8.7	-1730939	20
David Fincher	216	127	100125340	Crime|Dra	Morgan Freem	Se7enÂ	1023511	1080	English	USA	33000000	1995	8.6	67125340	21
Christopher Nolan	712	169	187991439	Adventure	Matthew McC	InterstellarÂ	928227	2725	English	USA	165000000	2014	8.6	22991439	22
Jonathan Demme	185	138	130727000	Crime|Dra	Anthony Hopki	The Silence of the LambsÂ	887467	916	English	USA	19000000	1991	8.6	111727000	23
Steven Spielberg	219	169	216119491	Action|Dra	Tom Hanks	Saving Private RyanÂ	881236	2277	English	USA	70000000	1998	8.6	146119491	24
Tony Kaye	162	101	6712241	Crime|Dra	Ethan Suplee	American History XÂ	782437	1420	English	USA	7500000	1998	8.6	-787759	25
Bryan Singer	162	106	23272306	Crime|Dra	Kevin Spacey	The Usual SuspectsÂ	740918	1182	English	USA	6000000	1995	8.6	17272306	26

Top Foreign Language Movies

director_name	num_critic_fo	duration	gross	genres	actor_1_name	movie_title	num_voted	num_use	language	country	budget	title_year	imdb_score	Profit	rank
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the UglyÂ	503509	780	Italian	Italy	1200000	1966	8.9	4900000	8
Fernando Meirelle	214	135	7563397	Crime|Drama	Alice Braga	City of GodÂ	533200	749	Portuguese	Brazil	3300000	2002	8.7	4263397	19
Akira Kurosawa	153	202	269061	Action|Adventure|	Takashi Shimura	Seven SamuraiÂ	229012	596	Japanese	Japan	2000000	1954	8.7	-1730939	20
Hayao Miyazaki	246	125	10049886	Adventure|Animati	Bunta Sugawara	Spirited AwayÂ	417971	902	Japanese	Japan	19000000	2001	8.6	-8950114	27
Florian Henckel vo	215	137	11284657	Drama|Thriller	Sebastian Koch	The Lives of OthersÂ	259379	407	German	Germany	2000000	2006	8.5	9284657	46
Majid Majidi	46	89	925402	Drama|Family	Bahare Seddiqi	Children of HeavenÂ	27882	130	Persian	Iran	180000	1997	8.5	745402	47
Jean-Pierre Jeunet	242	122	33201661	Comedy|Romance	Mathieu Kassovitz	AmÃ©lieÂ	534262	1314	French	France	77000000	2001	8.4	-43798339	54
Chan-wook Park	305	120	2181290	Drama|Mystery|Th	Min-sik Choi	OldboyÂ	356181	809	Korean	South Kore	3000000	2003	8.4	-818710	56
Hayao Miyazaki	174	134	2298191	Adventure|Animati	Minnie Driver	Princess MononokeÂ	221552	570	Japanese	Japan	2.4E+09	1997	8.4	-2397701809	57
Wolfgang Peterser	96	293	11433134	Adventure|Drama|	JÃ¼rgen Prochnow	Das BootÂ	168203	426	German	West Germ	14000000	1981	8.4	-2566866	60
Asghar Farhadi	354	123	7098492	Drama|Mystery	Shahab Hosseini	A SeparationÂ	151812	264	Persian	Iran	500000	2011	8.4	6598492	61
S.S. Rajamouli	44	159	6498000	Action|Adventure|	Tamannaah Bhatia	Baahubali: The BeginningÂ	62756	410	Telugu	India	18026148	2015	8.4	-11528148	62
Oliver Hirschbiegel	192	178	5501940	Biography|Drama|	Thomas Kretschmann	DownfallÂ	248354	564	German	Germany	13500000	2004	8.3	-7998060	80
Thomas Vinterberg	349	115	610968	Drama	Thomas Bo Larsen	The HuntÂ	170155	249	Danish	Denmark	3800000	2012	8.3	-3189032	84
Fritz Lang	260	145	26435	Drama|Sci-Fi	Brigitte Helm	MetropolisÂ	111841	413	German	Germany	6000000	1927	8.3	-5973565	86
Guillermo del Torc	406	112	37623143	Drama|Fantasy|W	Ivana Baquero	Pan's LabyrinthÂ	467234	1083	Spanish	Spain	13500000	2006	8.2	24123143	96
Hayao Miyazaki	212	119	4710455	Adventure|Animati	Christian Bale	Howl's Moving CastleÂ	214091	330	Japanese	Japan	24000000	2004	8.2	-19289545	105
Juan JosÃ© Camp	262	129	20167424	Drama|Mystery|Th	Ricardo DarÃn	The Secret in Their EyesÂ	131831	231	Spanish	Argentina	2000000	2009	8.2	18167424	107
Denis Villeneuve	226	139	6857096	Drama|Mystery|W	Lubna Azabal	IncendiesÂ	80429	156	French	Canada	6800000	2010	8.2	57096	109
Alejandro G. IÃ±Ã	157	115	5383834	Drama|Thriller	Adriana Barraza	Amores PerrosÂ	173551	361	Spanish	Mexico	2000000	2000	8.1	3383834	148
Katsuhiro Ã”tomo	150	124	439162	Action|Animation|	Mitsuo Iwata	AkiraÂ	106160	430	Japanese	Japan	1.1E+09	1988	8.1	-1099560838	150
JosÃ© Padilha	142	115	8060	Action|Crime|Dran	Wagner Moura	Elite SquadÂ	81644	107	Portuguese	Brazil	4000000	2007	8.1	-3991940	151
Thomas Vinterberg	98	105	1647780	Drama	Ulrich Thomsen	The CelebrationÂ	65951	258	Danish	Denmark	1300000	1998	8.1	347780	152

D. Best Directors:

Top 10 Directors	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.4
Christopher Nolan	8.4
Asghar Farhadi	8.4
Marius A. Markevi	8.4

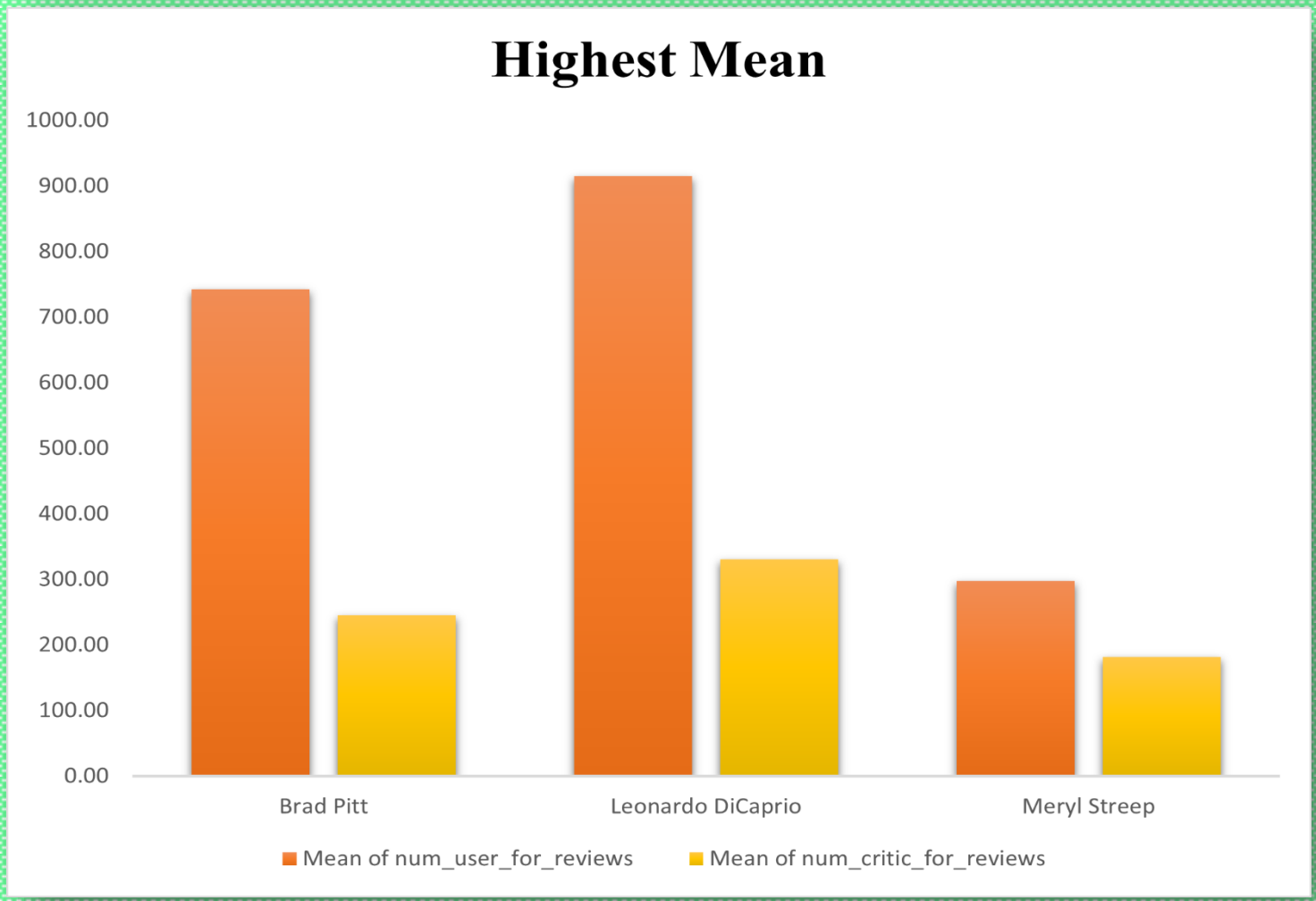
E. Popular Genres:

Popular Genres	Average of imdb_score
Crime|Drama|Fantasy|Mystery	8.5
Adventure|Animation|Drama|Family|Musical	8.5
Adventure|Drama|Thriller|War	8.4
Adventure|Animation|Fantasy	8.4
Action|Adventure|Drama|Fantasy|War	8.4
Documentary|War	8.3
Documentary|Drama|Sport	8.3
Biography|Drama|History|Music	8.3
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
Adventure|Drama|War	8.3

F. Charts:

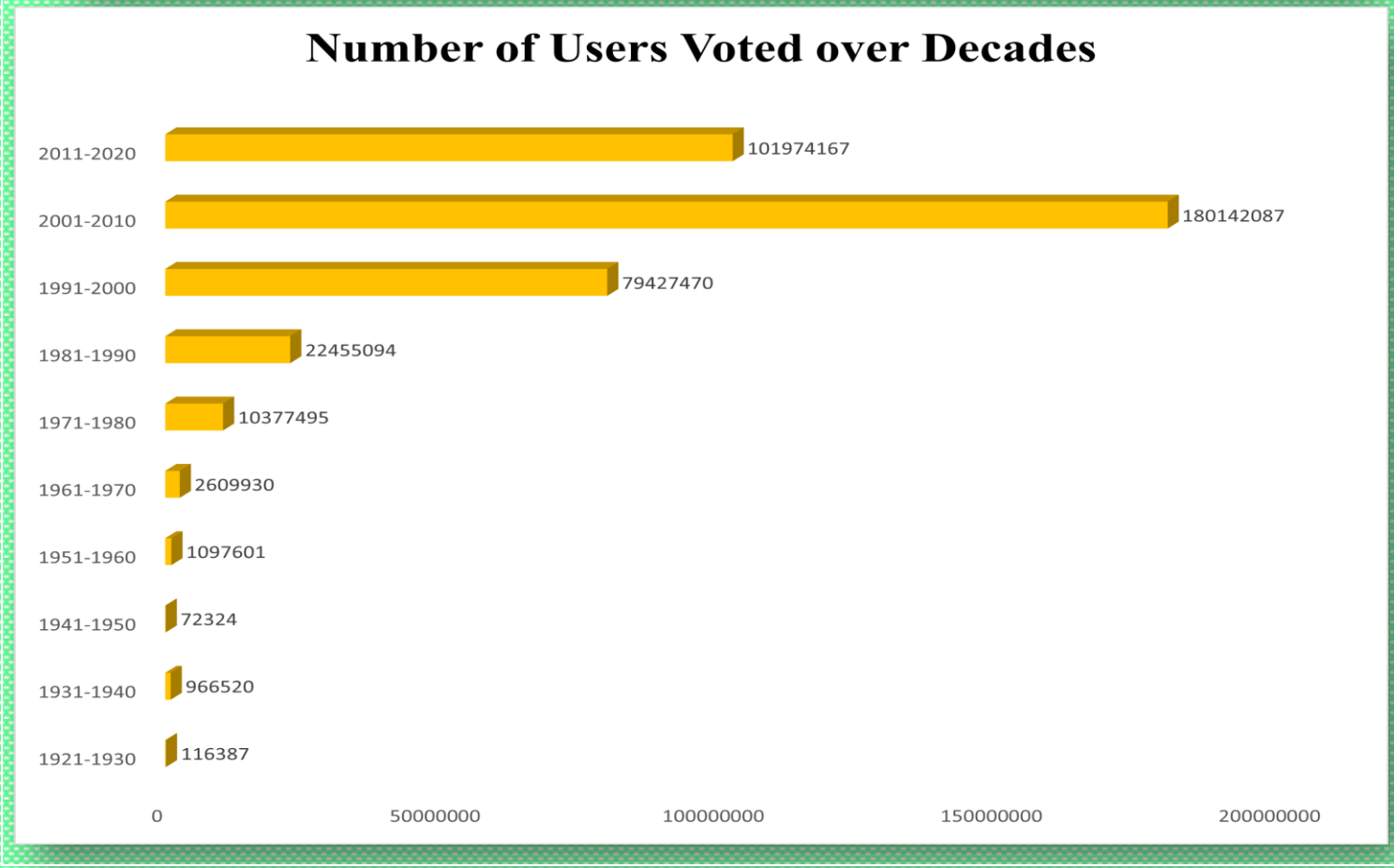
1.Highest Mean

actor_1_name	Mean of num_user_for_reviews	Mean of num_critic_for_reviews
Brad Pitt	742.35	245
Leonardo DiCaprio	914.48	330.19
Meryl Streep	297.18	181.45



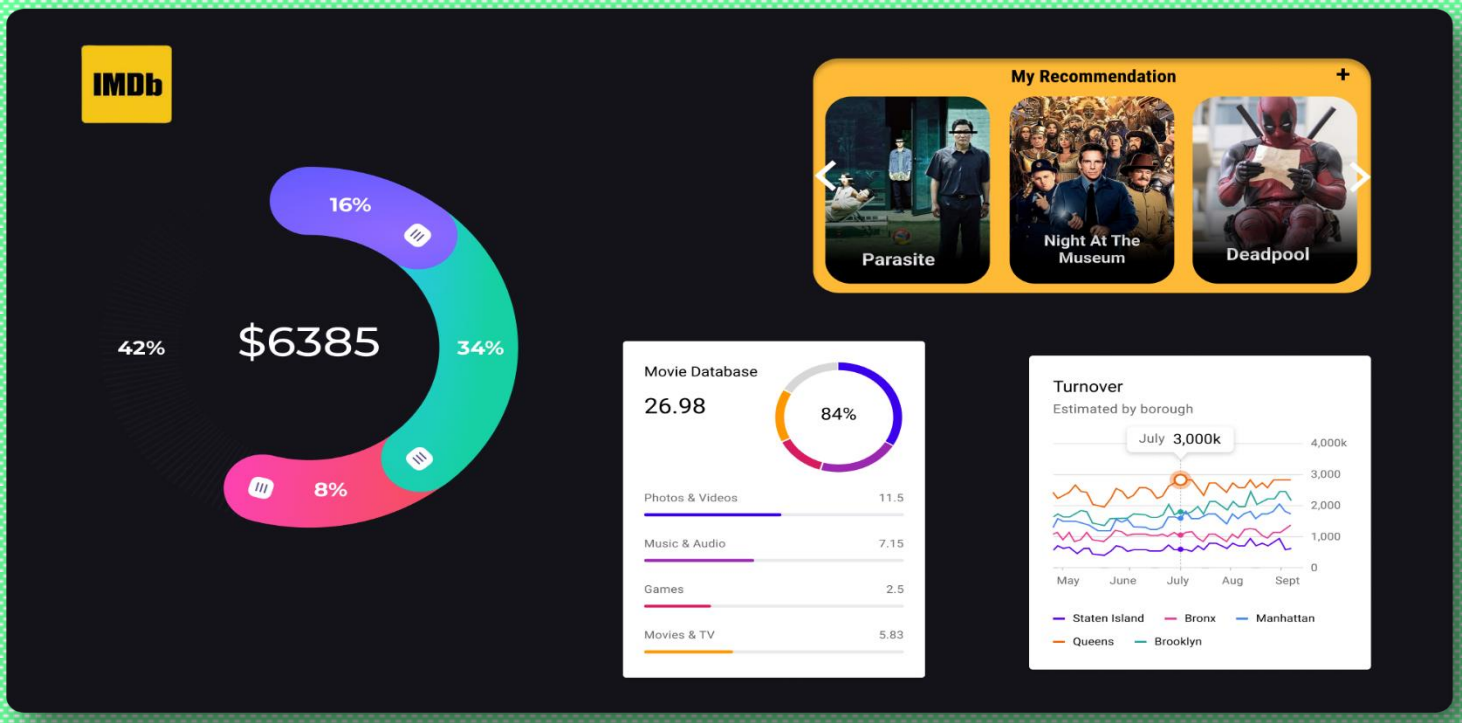
2. Voted users over decades

Decade	Number of Users Voted
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2609930
1971-1980	10377495
1981-1990	22455094
1991-2000	79427470
2001-2010	180142087
2011-2020	101974167



Conclusion

Overall, this project provided valuable insights into IMDB Movie Analysis such analysis is done not only by Movie makers before movie production, but it is also done by various investors, stakeholders, theatre outlet owners, allowing for improvements to be made to ensure a fair and efficient Profits.



Bank Loan Case Study

Project-6

Project Description:

This project is about a case study on Bank Loan. We have to follow EDA (Exploratory Data Analysis) based on our analysis, we will get the insights for required questions that were raised.

Business Objectives:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make

decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

Approach:

I have imported the dataset to perform analysis after that we have to handle the data by clean, replace inappropriate values and preprocess the data, then after modification of dataset we had to perform necessary methods to get the desired output/result.

Tech/Tools used:

PPT – To prepare a detailed report.

EXCEL – Excel was used to perform entire analysis.

Insights:

A. Identify Missing Data and Deal with it Appropriately:

1. The dataset have been cleaned and analyzed in the excel,

2. To clean the data the percentage of null values needs to be analyzed and those columns that have more than 50% of the null data have to be dropped,

3. And those columns with less than 50% of the null data have to be replaced with mean or median or the highest occurring variables,

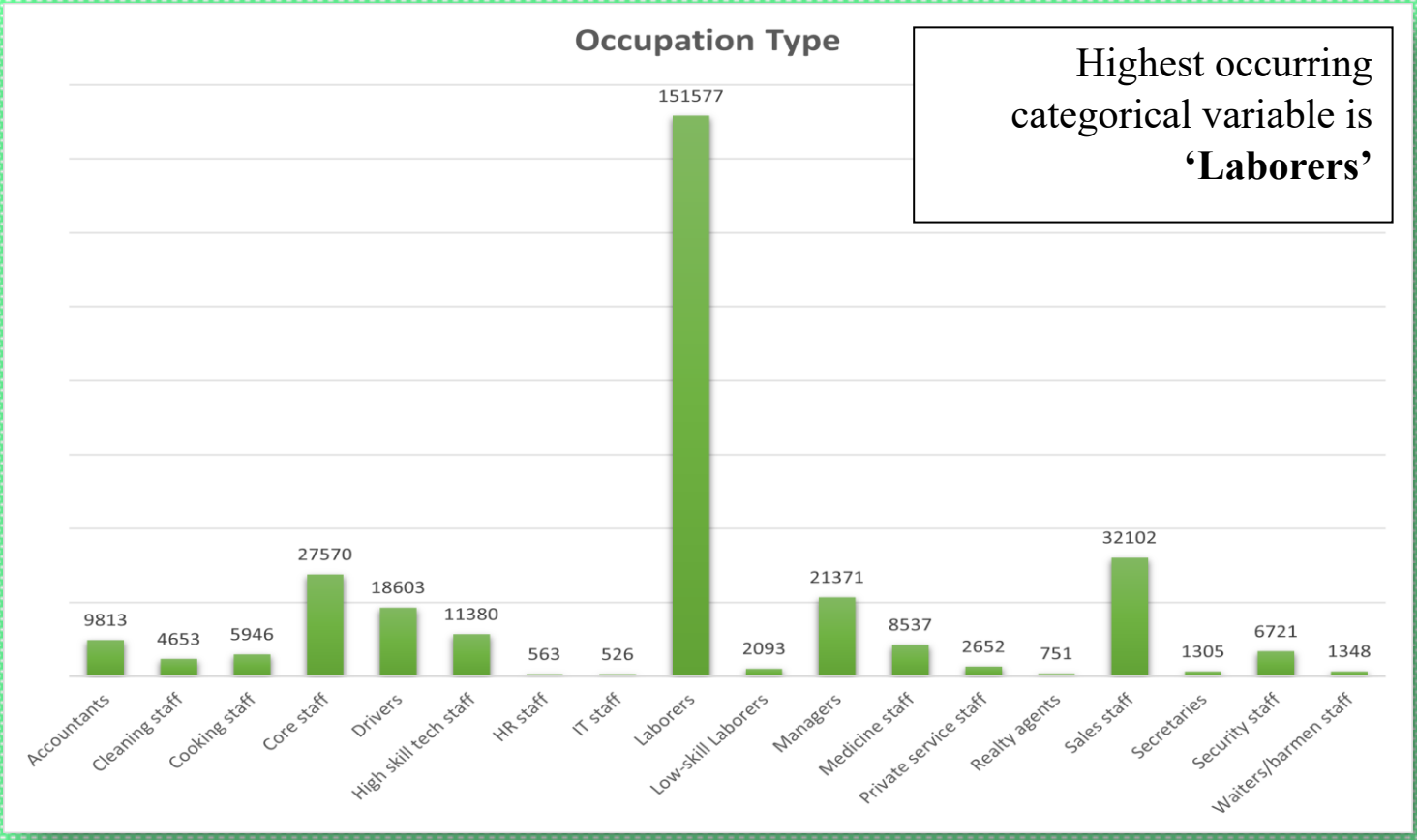
4. 50% is the threshold that I have chosen.

ALL THE COLUMN NAME WHICH ARE IN THE FOLLOWING TABLE ARE NEED TO BE DROPPED DOWN AS THEY HAVE NULL VALUES GREATER THAN OR EQUAL TO 50% AND THE COLUMN THAT ARE IRRELEVANT COLUMNS FOR DOING OUR ANALYSIS.

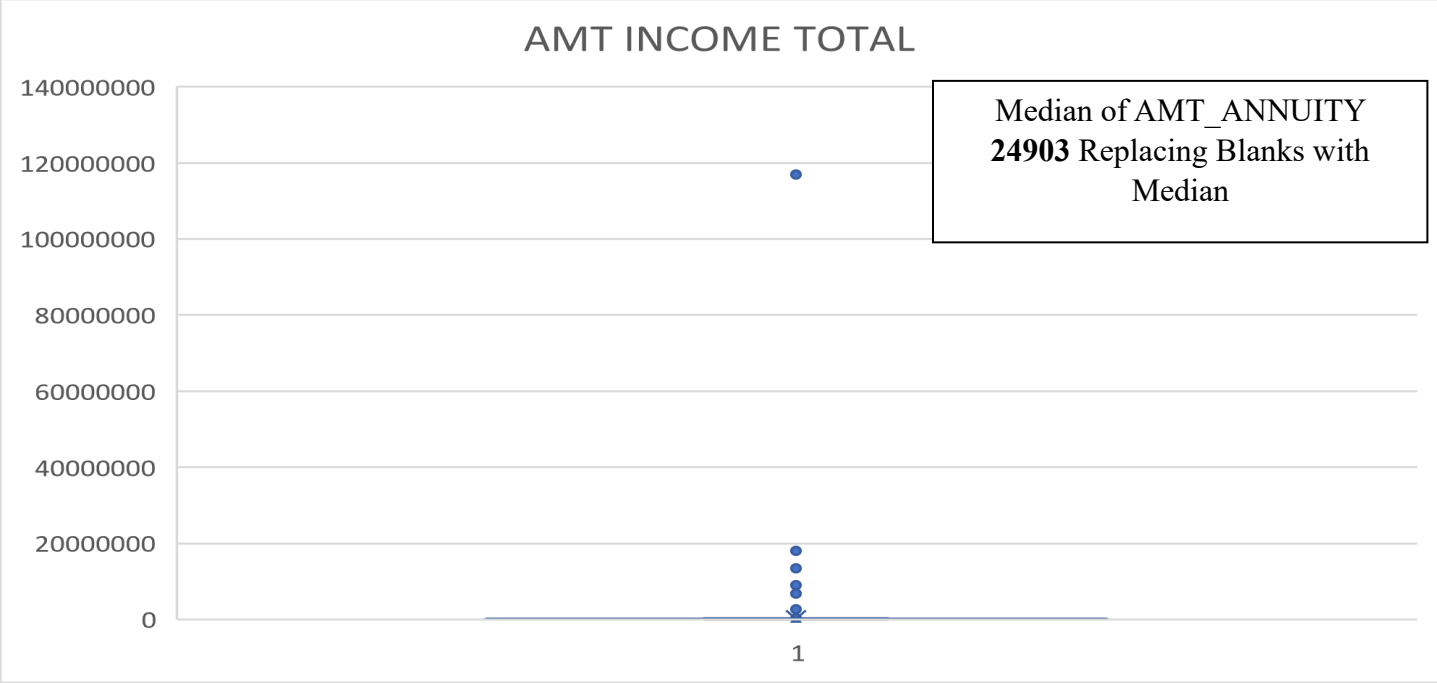
Column name	Total number of null values	Percentage of null value in that column	ROUND PER
OWN_CAR_AGE	202930	65.99113528	66
EXT_SOURCE_1	173379	56.38139774	56
APARTMENTS_AVG	156061	50.74972928	51
BASEMENTAREA_AVG	179943	58.51595553	59
YEARS_BUILD_AVG	204488	66.49778382	66
COMMON_AREA_AVG	214865	69.87229725	70
ELEVATORS_AVG	163891	53.29597966	53
ENTRANCES_AVG	154828	49.70488861	50
FLOORSMAX_AVG	153021	49.76114676	50
FLOORSMIN_AVG	208642	67.84862981	68
LANDAREA_AVG	182590	59.37673774	59
LIVINGAPARTMENTS_AVG	210199	68.35495316	68
LIVINGAREA_AVG	154350	50.19332642	50
NONLIVINGAPARTMENTS_AVG	213514	69.43296337	69
NONLIVINGAREA_AVG	169682	55.17916432	55
APARTMENTS_MODE	156061	50.74972928	51
BASEMENTAREA_MODE	179943	58.51595553	59
YEARS_BUILD_MODE	204488	66.49778382	66
COMMON_AREA_MODE	214865	69.87229725	70
ELEVATORS_MODE	163891	53.29597966	53
ENTRANCES_MODE	154828	50.34876801	50
FLOORSMAX_MODE	153020	49.76082156	50
FLOORSMIN_MODE	208642	67.84862981	68
LANDAREA_MODE	182590	59.37673774	59
LIVINGAPARTMENTS_MODE	210199	68.35495316	68
LIVINGAREA_MODE	154350	50.19332642	50
NONLIVINGAPARTMENTS_MODE	213514	69.43296337	69
NONLIVINGAREA_MODE	169682	55.17916432	55
APARTMENTS_MEDIAN	156061	50.74972928	51
BASEMENTAREA_MEDIAN	179943	58.51595553	59
YEARS_BUILD_MEDIAN	204488	66.49778382	66
COMMON_AREA_MEDIAN	214865	69.87229725	70
ELEVATORS_MEDIAN	163891	53.29597966	53
ENTRANCES_MEDIAN	154828	50.34876801	50
FLOORSMAX_MEDIAN	153020	49.76082156	50
FLOORSMIN_MEDIAN	208642	67.84862981	68
LANDAREA_MEDIAN	182590	59.37673774	59
LIVINGAPARTMENTS_MEDIAN	210199	68.35495316	68
LIVINGAREA_MEDIAN	154350	50.19332642	50
NONLIVINGAPARTMENTS_MEDIAN	213514	69.43296337	69
NONLIVINGAREA_MEDIAN	169682	55.17916432	55
FONDKAPREMONT_MODE	210295	68.38617155	68
HOUSETYPE_MODE	154297	50.17609126	50
WALLSMATERIAL_MODE	156341	50.84078293	5

Column name	Total number of null values	Percentage of null value	ROUND PER
FLAG_MOBIL	1	0.000325192	0
FLAG_EMPLOY_PHONE	55387	18.01138821	18
FLAG_WORK_PHONE	0	0	0
FLAG_CONT_MOBILE	0	0	0
FLAG_PHONE	0	0	0
FLAG_EMAIL	0	0	0
CNT_FAMILY_MEMBERS	2	0.000650383	0
REGION_RATING_CLENT	0	0	0
REGION_RATING_CLENT_W_CITY	0	0	0
EXT_SOURCE_3	60965	19.82530706	20
YEAR_BEGINEXPLUATATION_AVG	150008	48.78134441	49
YEAR_BEGINEXPLUATATION_MODE	150007	48.78101922	49
YEAR_BEGINEXPLUATATION_MEDIAN	150007	48.78101922	49
TOTAL_AREA_MODE	148431	48.26851722	48
EMERGENCYSTATE_MODE	145755	47.39830445	47
DAYS_LAST_PHONE_CHANGE	1	0.000325192	0
FLAG_DOC_2	0	0	0
FLAG_DOC_3	0	0	0
FLAG_DOC_4	0	0	0
FLAG_DOC_5	0	0	0
FLAG_DOC_6	0	0	0
FLAG_DOC_7	0	0	0
FLAG_DOC_8	0	0	0
FLAG_DOC_9	0	0	0
FLAG_DOC_10	0	0	0
FLAG_DOC_11	0	0	0
FLAG_DOC_12	0	0	0
FLAG_DOC_13	0	0	0
FLAG_DOC_14	0	0	0
FLAG_DOC_15	0	0	0
FLAG_DOC_16	0	0	0
FLAG_DOC_17	0	0	0
FLAG_DOC_18	0	0	0
FLAG_DOC_19	0	0	0
FLAG_DOC_20	0	0	0
FLAG_DOC_21	0	0	0

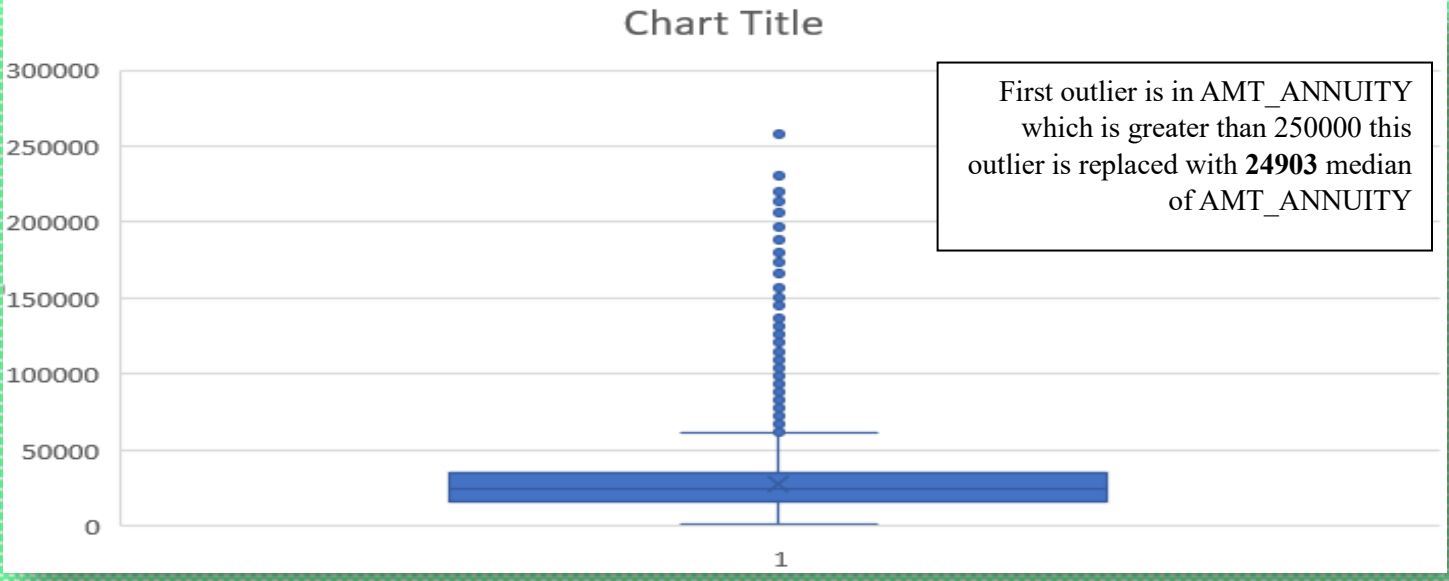
Replacing Blank cell in Occupation_Type column of the Application Dataset with the highest occurring variable.



Replacing Blanks in AMT_ANNUTIY column of the Application Dataset with the median of the AMT_ANNUITY as there exists outliers in the AMT_ANNUITY column.

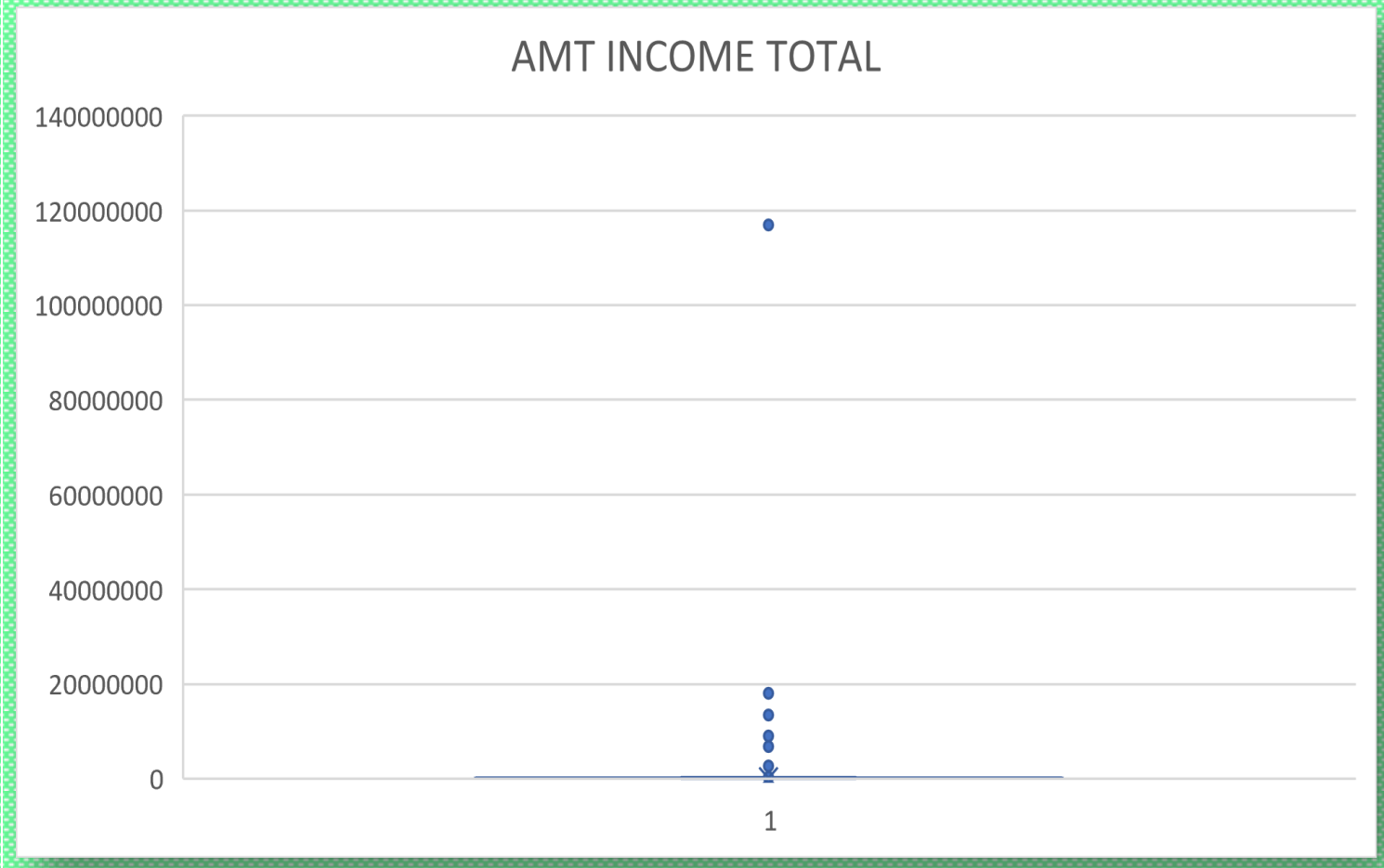


B. Identify Outliers in the Dataset:



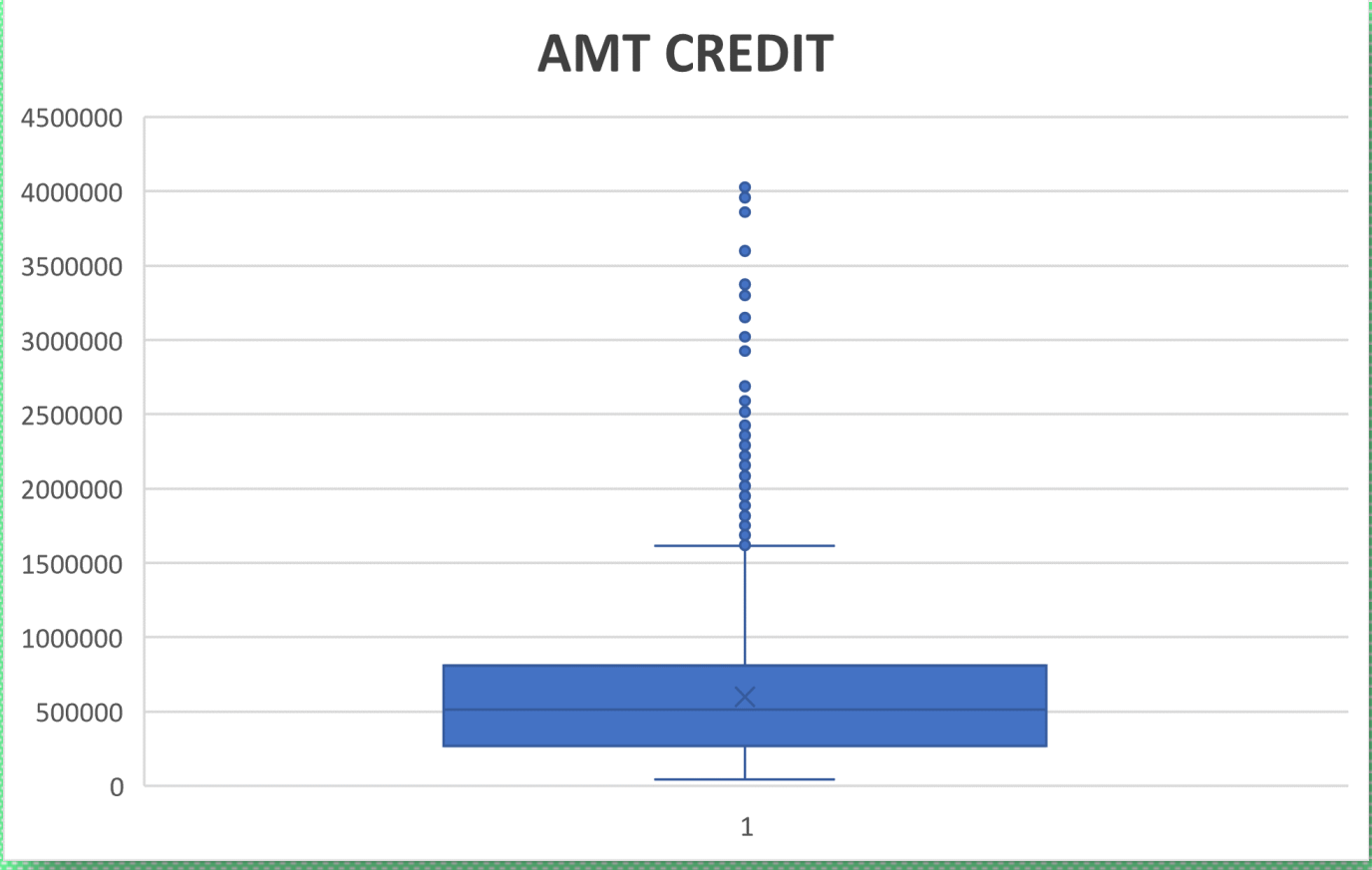
Here we can observe that there is huge difference between the 25%, 50% and 75% quartile and this is due to presence of outliers But since the amount of total income varies from person to person we will not remove the outlier.

	Quartiles at AMT_INCOME_TOTAL
MIN	25650
25%	112500
50%	147150
75%	202500
MAX	117000000



	AMT_CREDIT
	Quartiles at AMT_CREDIT
MIN	45000
25%	270000
50%	513531
75%	808650
MAX	4050000

From the chart it is clear that outliers lie in the 98% and near the max side of box plot and also there is a significant difference between the 75% quartile and the max value and this is due the presence of the outliers, But since the amount of credit varies from person to person hence we will not remove the outliers.



	DAYS_EMPLOYED
	Quartiles at DAYS_EMPLOYED
MAX	17912.00
75%	2760.00
50%	1213.00
25%	289.00
MIN	365243.00

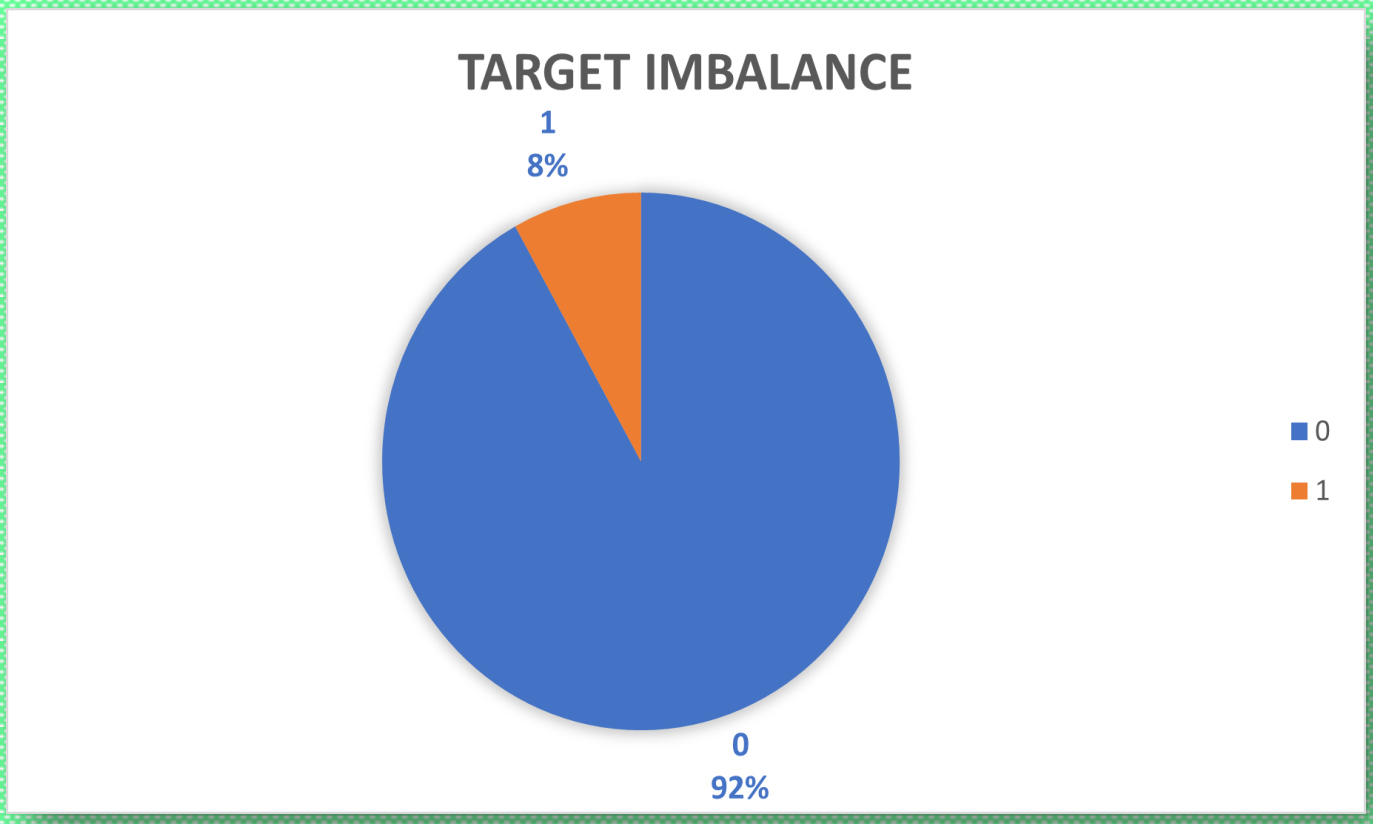
There exists only 1 outlier i.e. + or - 365243

Replace with median **1213.00**

C. Analyze Data Imbalance:

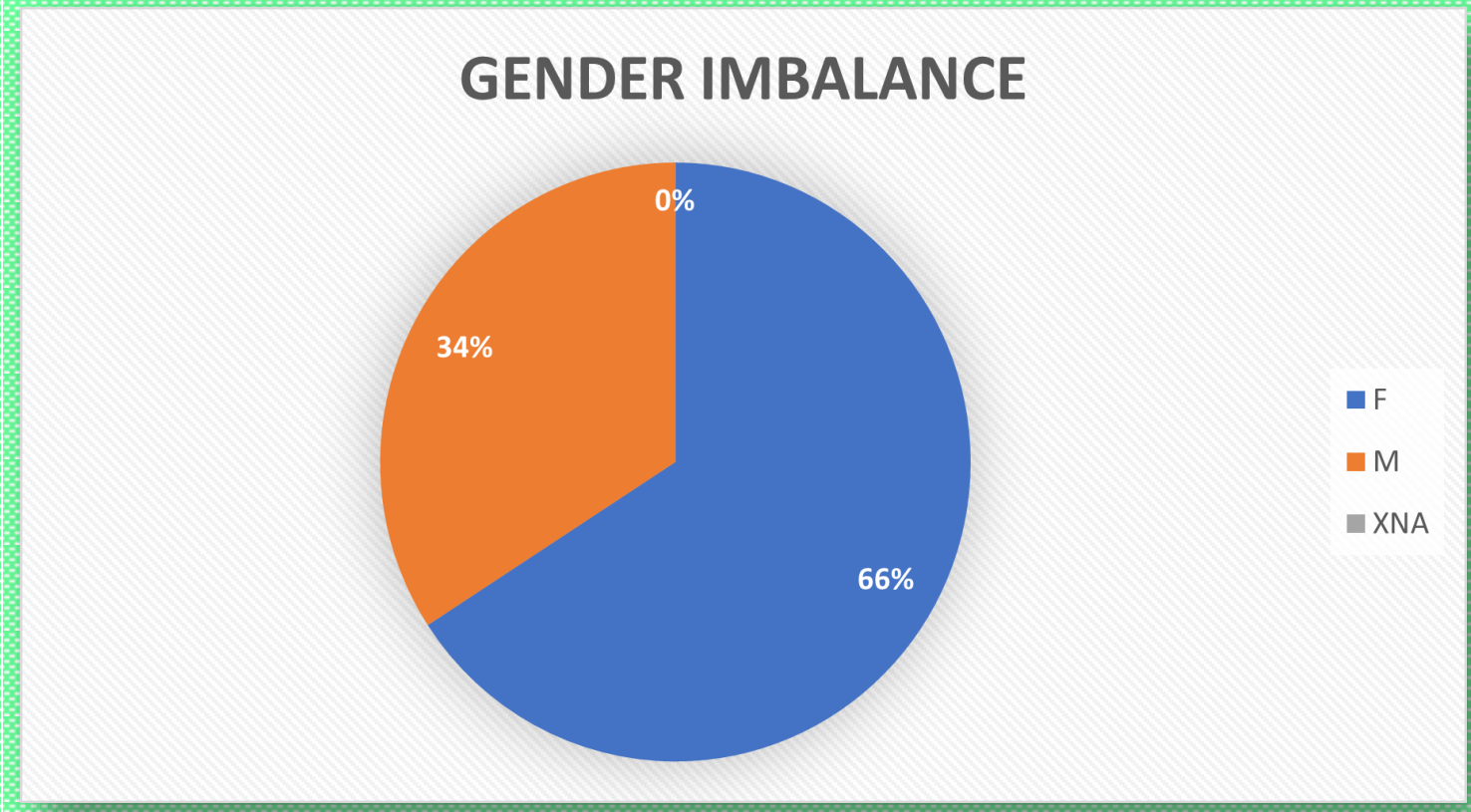
Target Imbalace	Count of TARGET
0	282686
1	24825
Grand Total	307511

The Target Imbalance Pie chart shows that almost 92% of the total clients had no problem during payment while 8% of the clients had some or the other problem.

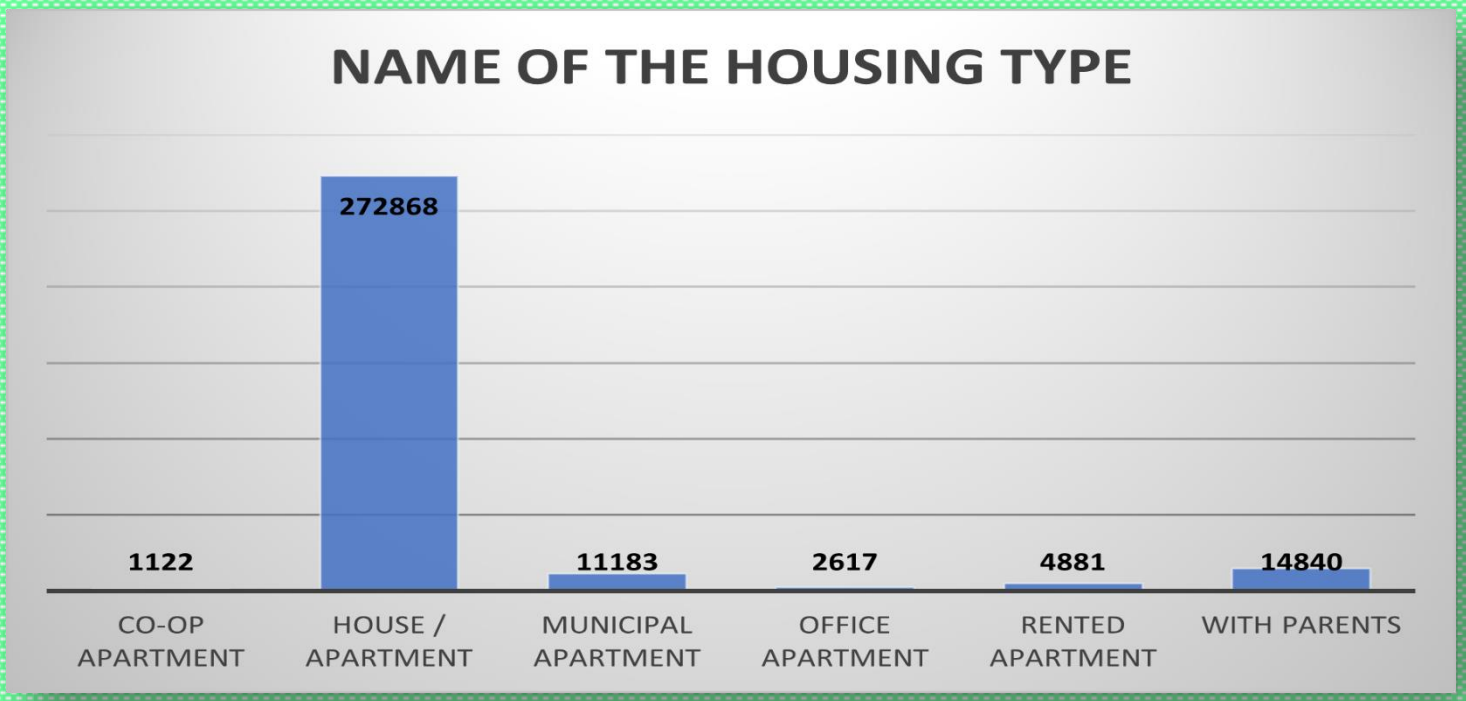


GENDER IMBALANCE	Count of CODE_GENDER
F	202448
M	105059
XNA	4
Grand Total	307511

From the GENDER IMBALANCE pie chart we can infer that almost 66% of the clients are female and 34% of the clients are Male and the 4 of the applicants have gender as XNA which can be ignored.



NAME OF THE HOUSING TYPE	Count of NAME OF THE HOUSING TYPE
Co-op apartment	1122
House / apartment	272868
Municipal apartment	11183
Office apartment	2617
Rented apartment	4881
With parents	14840
Grand Total	307511

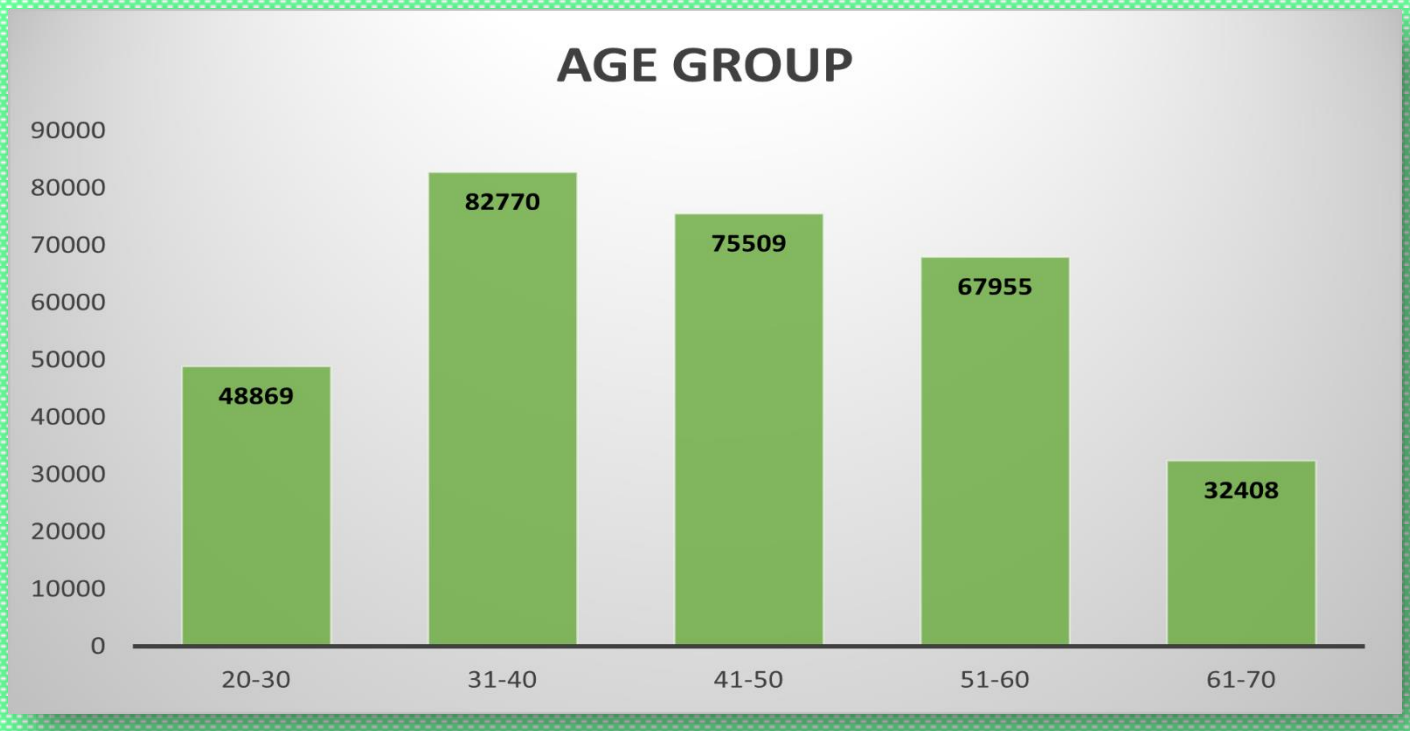


From the bar graph. The bank can target those groups who do not have their own apartment i.e. the bank may consider the people living in Co-op apartment, Municipal Apartment, Rented Apartment and people living with their parent.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

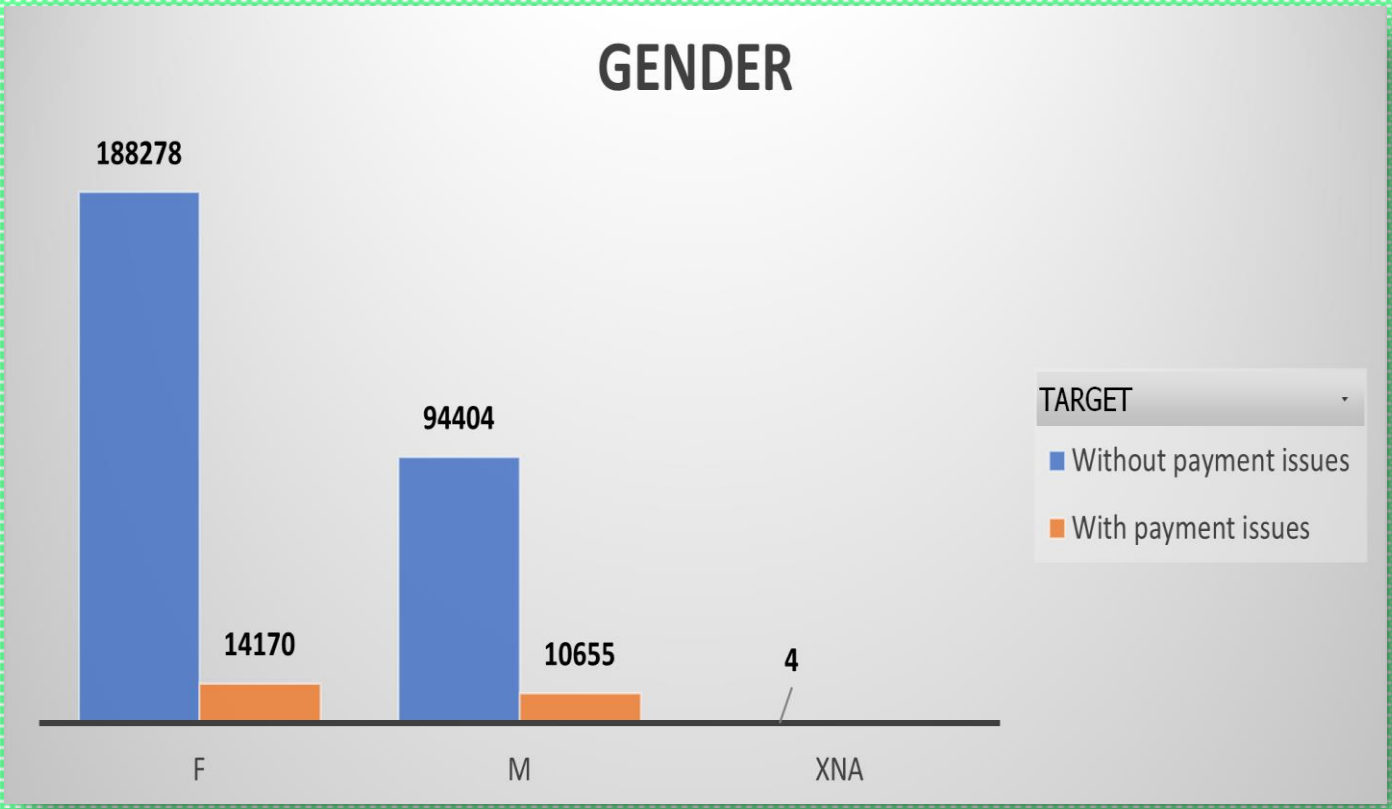
UNIVARIATE

AGE GROUPS	Count of YEARS BIRTH RANGE
20-30	48869
31-40	82770
41-50	75509
51-60	67955
61-70	32408
Grand Total	307511



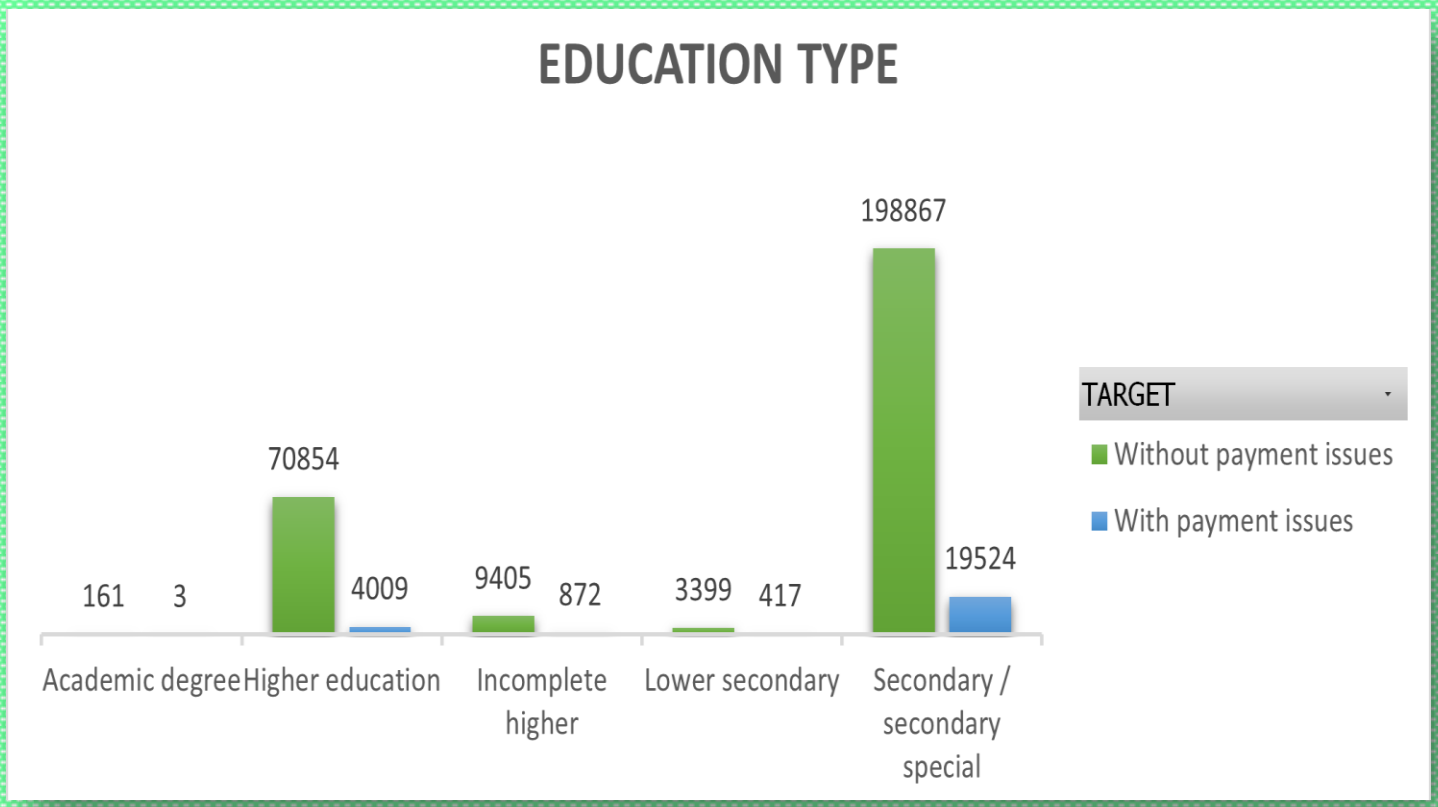
From the bar graph we can infer that most of the applicants belong to the **Age Group 31-40.**

GENDER	Without payment issues	With payment issues	Grand Total
F	188278	14170	202448
M	94404	10655	105059
XNA	4		4
Grand Total	282686	24825	307511



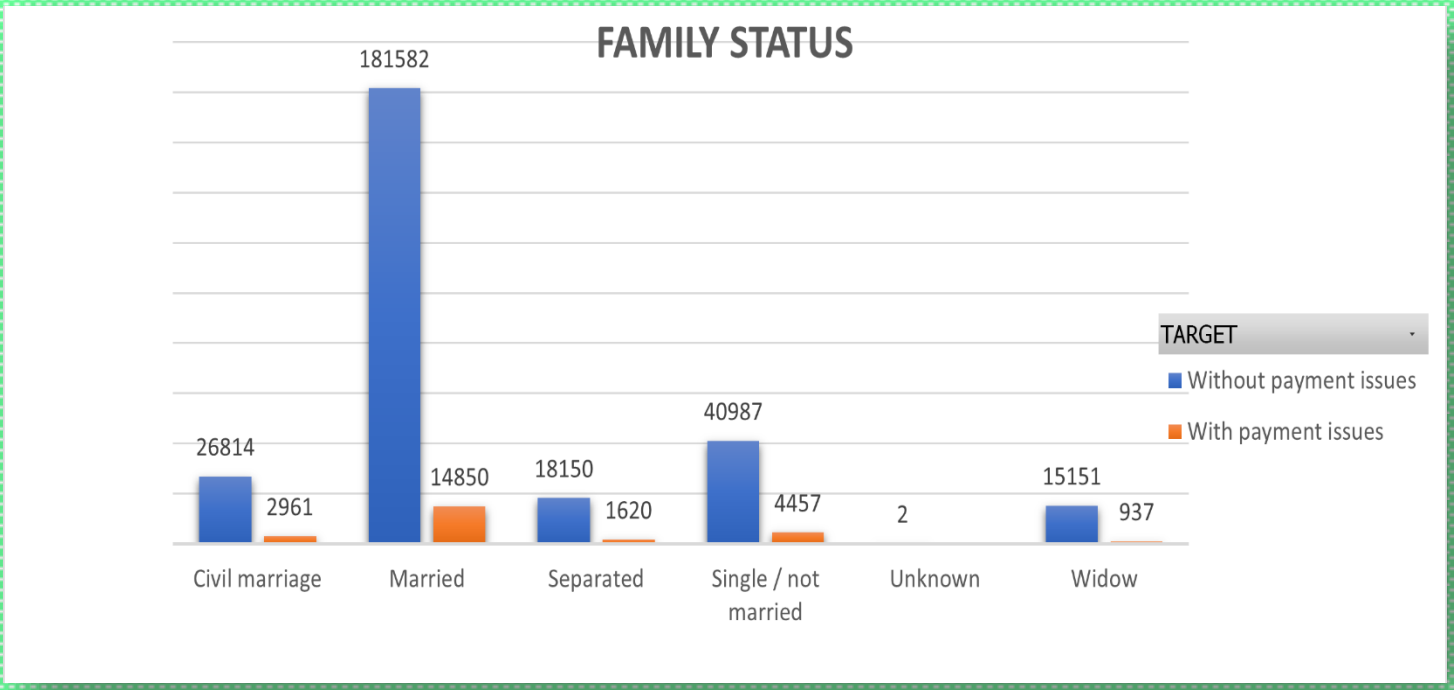
From the Bar graph we can infer that Clients with GENDER = 'F' have the highest number of non-defaulters i.e. 188278-14170 = **174108**.

EDUCATION TYPE	Without payment issues	With payment issues	Grand Total
Academic degree	161	3	164
Higher education	70854	4009	74863
Incomplete higher	9405	872	10277
Lower secondary	3399	417	3816
Secondary / secondary special	198867	19524	218391
Grand Total	**282686**	**24825**	**307511**



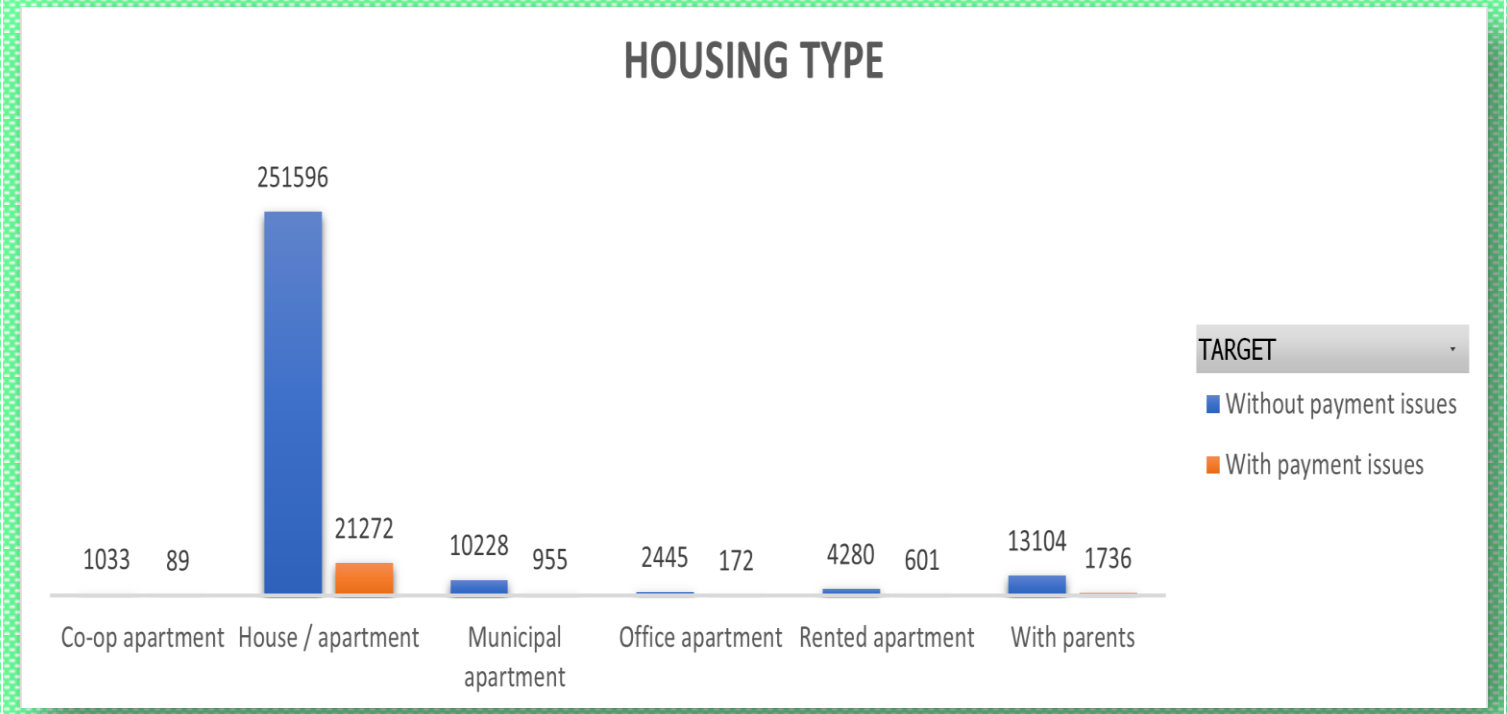
From the above Bar graph we can infer that clients having EDUCATION TYPE = 'SECONDARY/SECONDARY SPECIAL' have the highest count for Non defaulters.

FAMILY STATUS	Without payment issues	With payment issues	Grand Total
Civil marriage	26814	2961	29775
Married	181582	14850	196432
Separated	18150	1620	19770
Single / not married	40987	4457	45444
Unknown	2		2
Widow	15151	937	16088
Grand Total	282686	24825	307511



From the above Bar chart we can infer that clients having FAMILY STATUS = 'MARRIED' have the highest count of Non defaulters.

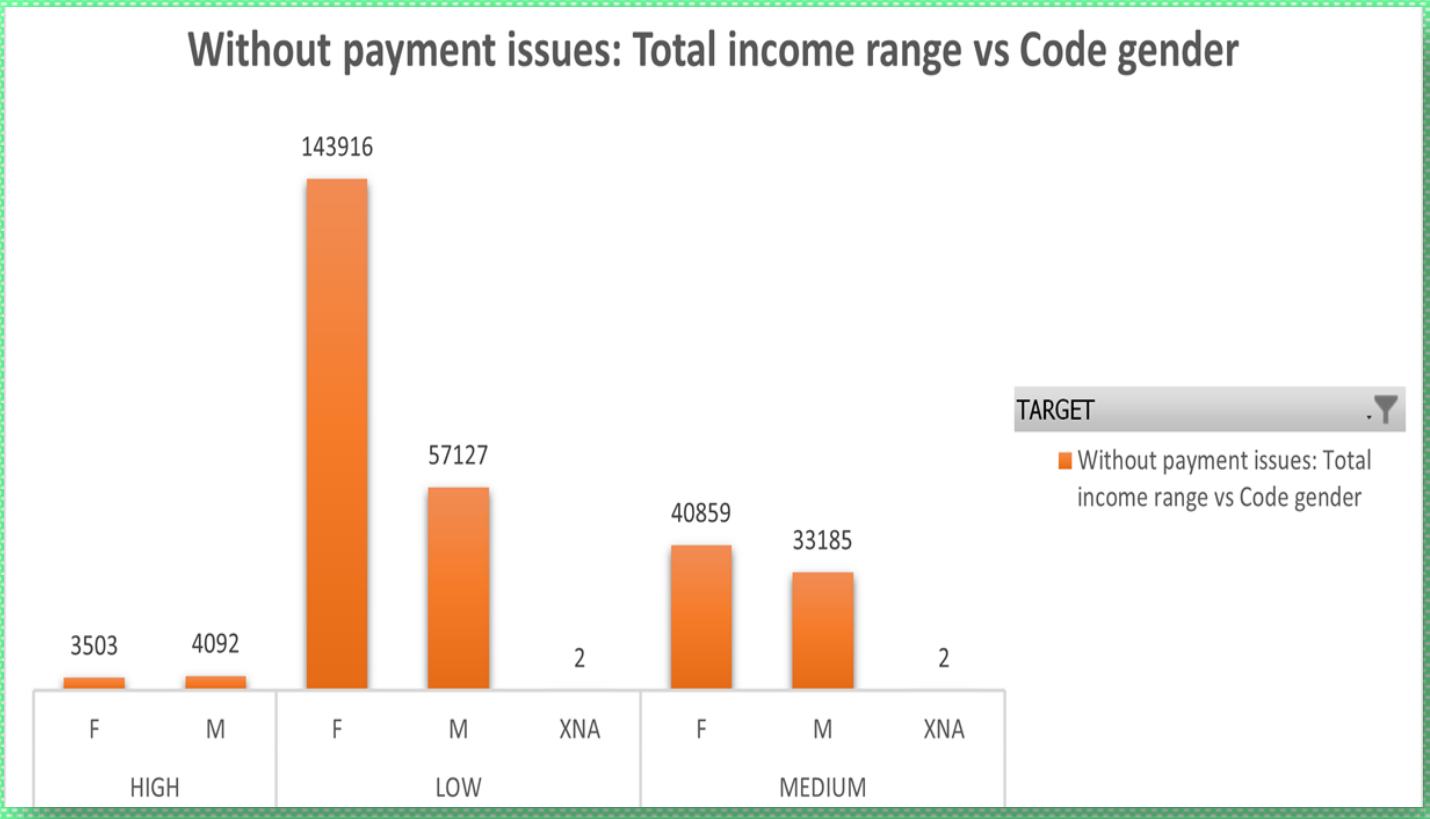
HOUSING TYPE	Without payment issues	With payment issues	Grand Total
Co-op apartment	1033	89	1122
House / apartment	251596	21272	272868
Municipal apartment	10228	955	11183
Office apartment	2445	172	2617
Rented apartment	4280	601	4881
With parents	13104	1736	14840
Grand Total	282686	24825	307511



From the above Bar Chart we can infer that clients having HOUSING TYPE = **'House/Apartment'** have the highest count of Non-defaulters.

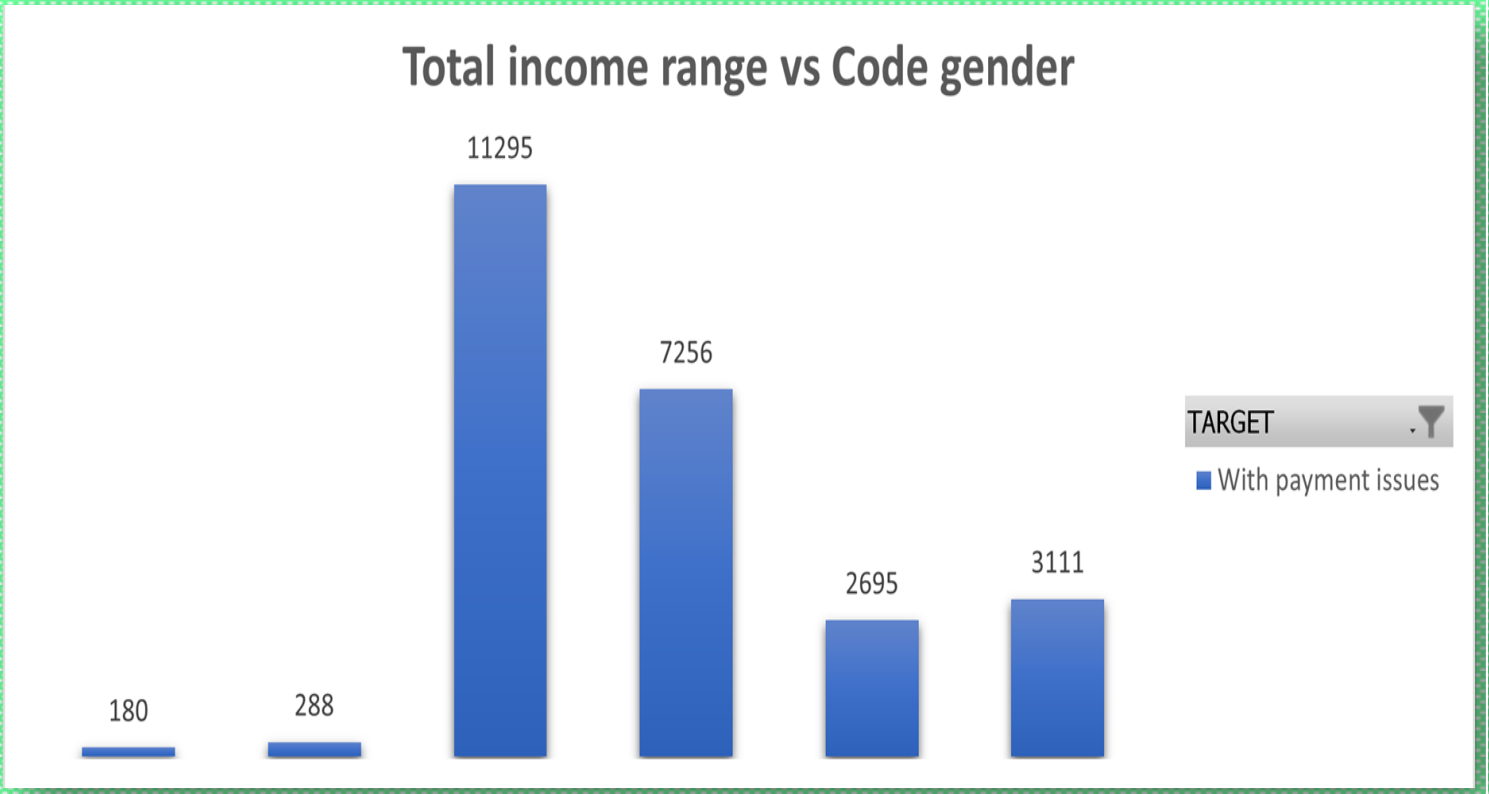
Bivariate

Total income range and Code gender	Without payment issues: Total income range vs Code gender	Grand Total
HIGH	7595.00	7595.00
F	3503.00	3503.00
M	4092.00	4092.00
LOW	201045.00	201045.00
F	143916.00	143916.00
M	57127.00	57127.00
XNA	2.00	2.00
MEDIUM	74046.00	74046.00
F	40859.00	40859.00
M	33185.00	33185.00
XNA	2.00	2.00
Grand Total	282686.00	282686.00



From the bar graph we can infer that Females belonging to Low income group are the highest number of clients with no payment issues.

Total income range and Code gender	With payment issues	Grand Total
HIGH	468	468
F	180	180
M	288	288
LOW	18551	18551
F	11295	11295
M	7256	7256
MEDIUM	5806	5806
F	2695	2695
M	3111	3111
Grand Total	24825	24825



From the bar graph we can infer that Females belonging to Low income group are the highest number of clients with payment issues.

Previous Application Dataset:

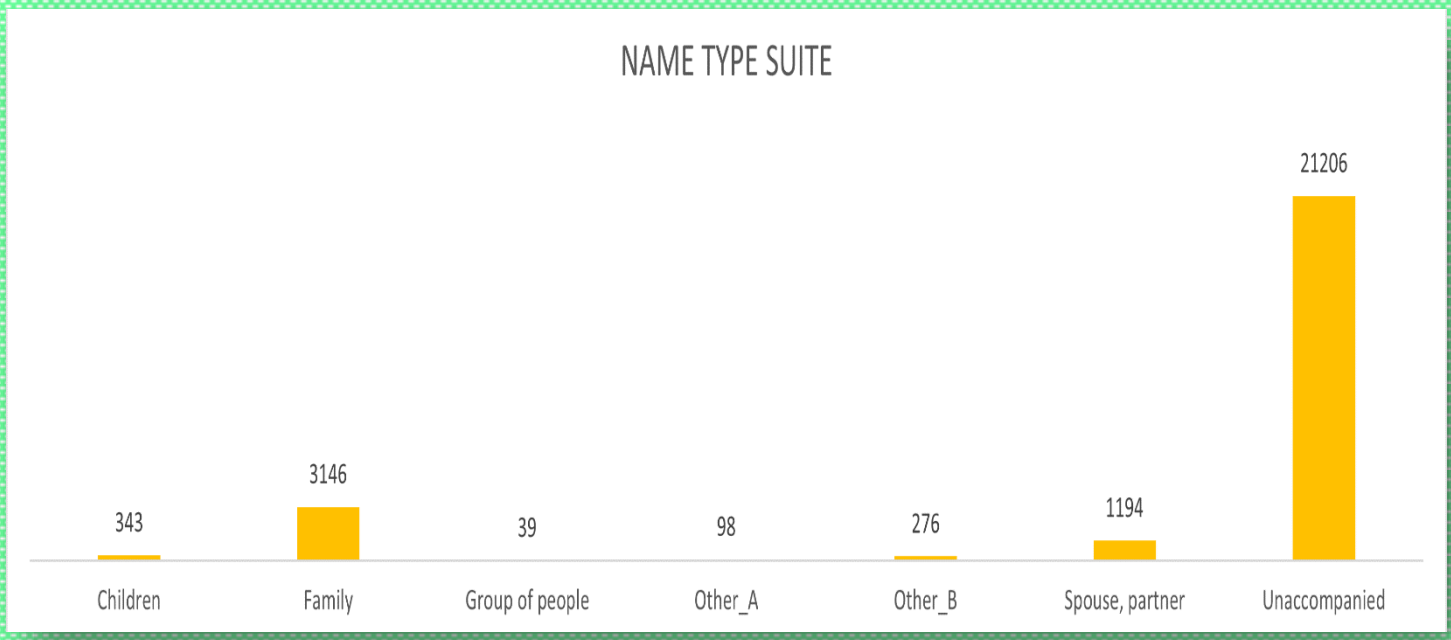
The following columns of the previous application datasets need to be dropped as they are irrelevant for the analysis:

- HOUR_APPR_PROCESS_START
- WEEKDAY_APPR_PROCESS_START_PREV
- FLAG_LAST_APPL_PER_CONTRACT
- NFLAG_LAST_APPL_IN_DAY
- SK_ID_CURR
- WEEKDAY_APPR_PROCESS_START

Removing the rows with the values 'XNA' &'XAP' from the column:
NAME_TYPE_SUITE

In the column **AMT_ANNUITY** replace the blank rows with **Median of AMT_ANNUITY '21340'.**

NAME TYPE SUITE	Count of NAME_TYPE_SUITE
Children	343
Family	3146
Group of people	39
Other_A	98
Other_B	276
Spouse, partner	1194
Unaccompanied	21206
Grand Total	26302



From the bar graph we can infer that Unaccompanied has the highest count therefore replace blank rows with Unaccompanied.

Name Contract Status	Approved	Canceled	Refused	Unused offer	Grand Total	
Building a house or an annex	434	60	1188		1682	
Business development	78	12	164		254	
Buying a garage	28	5	51		84	
Buying a holiday home / land	91	13	230		334	
Buying a home	130	23	393		546	
Buying a new car	139	29	465	4	637	
Buying a used car	552	57	1166	9	1784	
Car repairs	223	14	256		493	
Education	481	14	476	4	975	
Everyday expenses	732	8	740	7	1487	
Furniture	210	15	250		475	
Gasification / water supply	75	3	125		203	
Hobby	11		20		31	
Journey	329	10	404	2	745	
Medicine	676	25	696	5	1402	
Money for a third person	10		6		16	
Other	4106	186	5310	62	9664	
Payments on other loans	189	45	973	3	1210	
Purchase of electronic equipment	357	4	280	3	644	
Refusal to name the goal	1		7		8	
Repairs	5385	381	8973	28	14767	
Urgent needs	2228	83	2998		5309	
Wedding / gift / holiday	248	10	336		594	
Grand Total	**16713**	**997**	**25507**	**127**	**43344**	

From the graph and table we can infer that Name of Contract status i.e. **Repairs** work has the highest count of Approved and Refused Loans.

Conclusion

From the above analysis, we can infer that what kind of peoples can repay loan, what kinds of loan people prefer to take, people taking loans come from which background, what is their source of income, for what type of people the loan applications are refused and based on which conditions.

Impact of Car Features on Price and Profitability

Project-7

Project Description:

The aim of this project is to analyze the impact of car features on price and profitability in the automotive industry. By analyzing the dataset containing information on various car models and their specifications and their prices.

Approach:

Firstly I have cleaned the dataset to ensure accurate and reliable results for the analysis. Then I have performed analytical methods and Visualization Methods. These techniques will help to get insights about the Impact of Car Features on Price and Profitability.

Tech-Stack Used:

PPT – To prepare a detailed report.

EXCEL – Excel was used to perform entire analysis.

Insight:

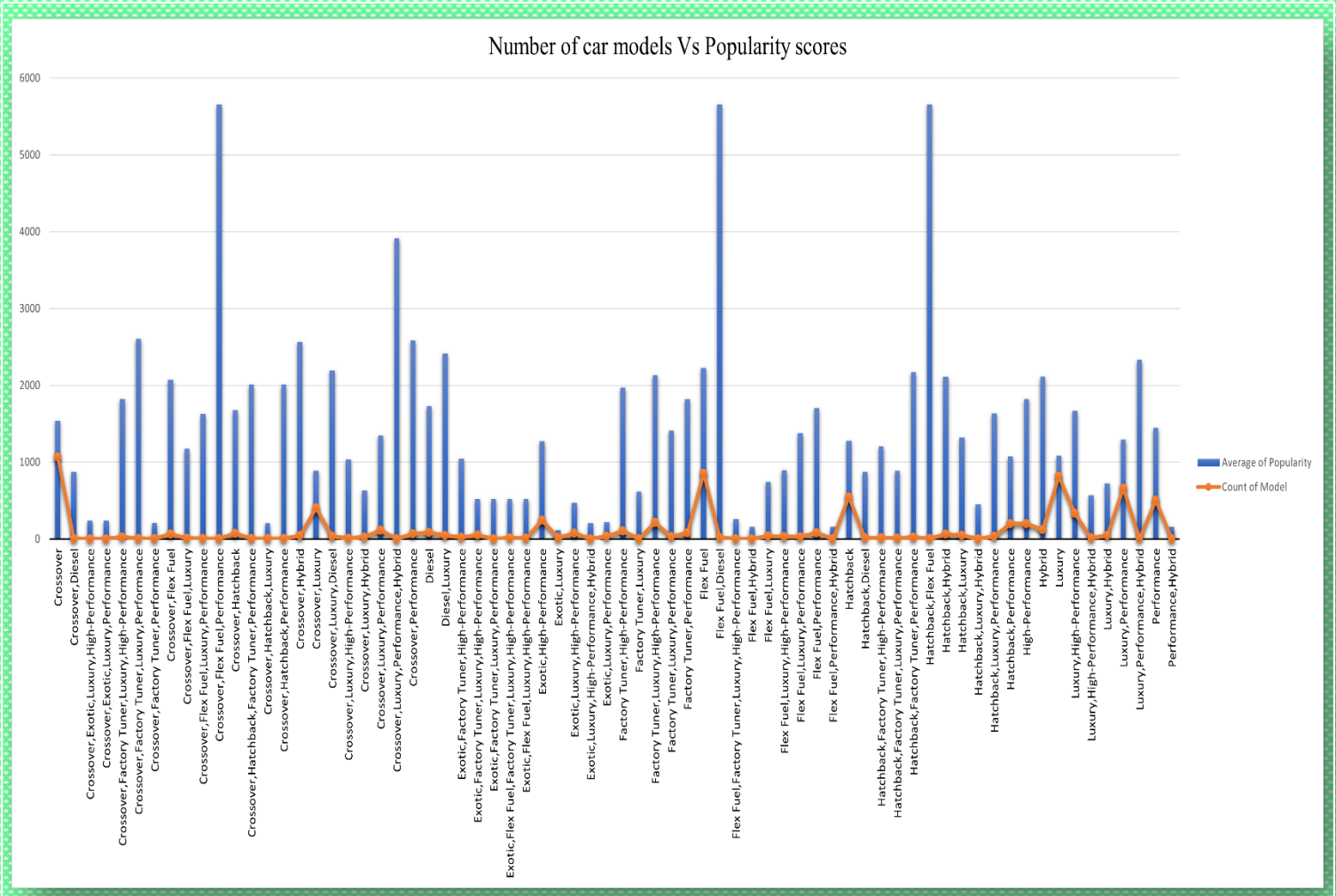
1.How does the popularity of a car model vary across different market categories?

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

Pivot Table have been created to show 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.

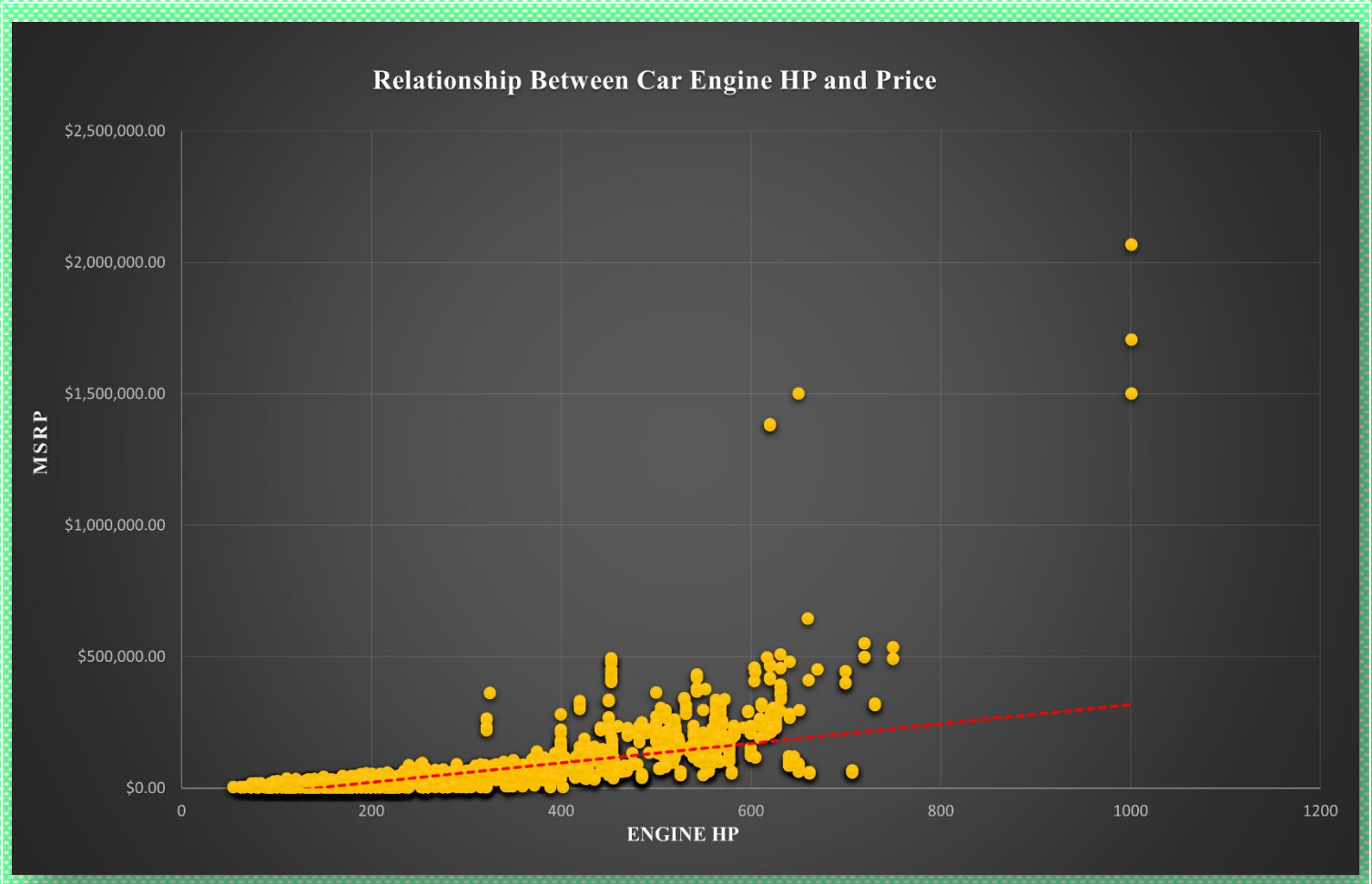
Chart have been created to show the relationship between market category and popularity. From the chart below we can infer that crossover, flex fuel, diesel, hatchback, and performance are the most popular market categories for car models.



2.What is the relationship between a car's engine power and its price?

Task 2: 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.

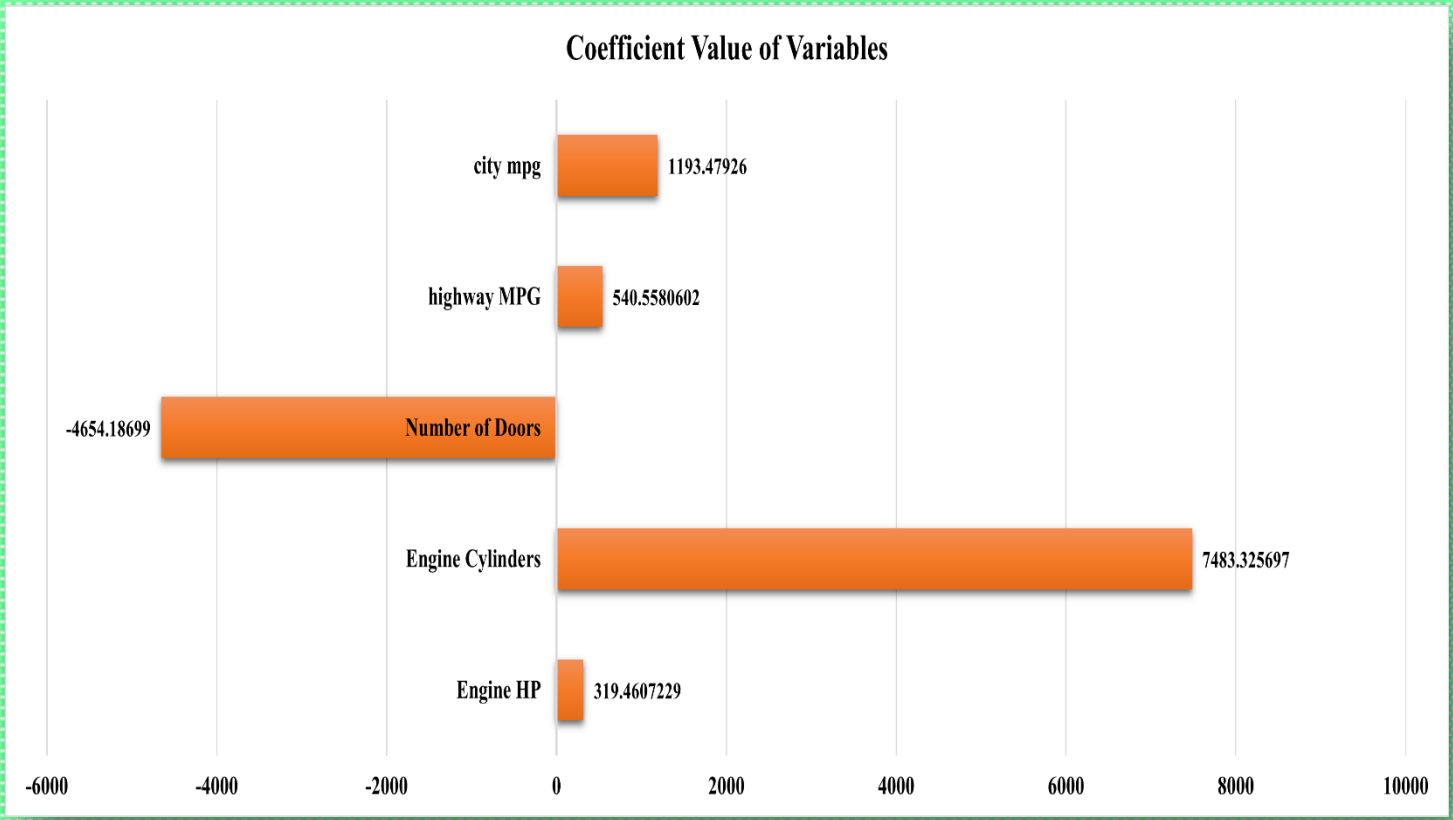
Scatter chart have been created in following slide from the chart we can infer that the horse power of the engine increases then the price also increases.



3.Which car features are most important in determining a car's price?

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

Varibles	Coefficients
Engine HP	319.4607229
Engine Cylinders	7483.325697
Number of Doors	-4654.18699
highway MPG	540.5580602
city mpg	1193.47926



From the chart we infer that **Engine cylinders** are one of the most important features that determine the price of a car.

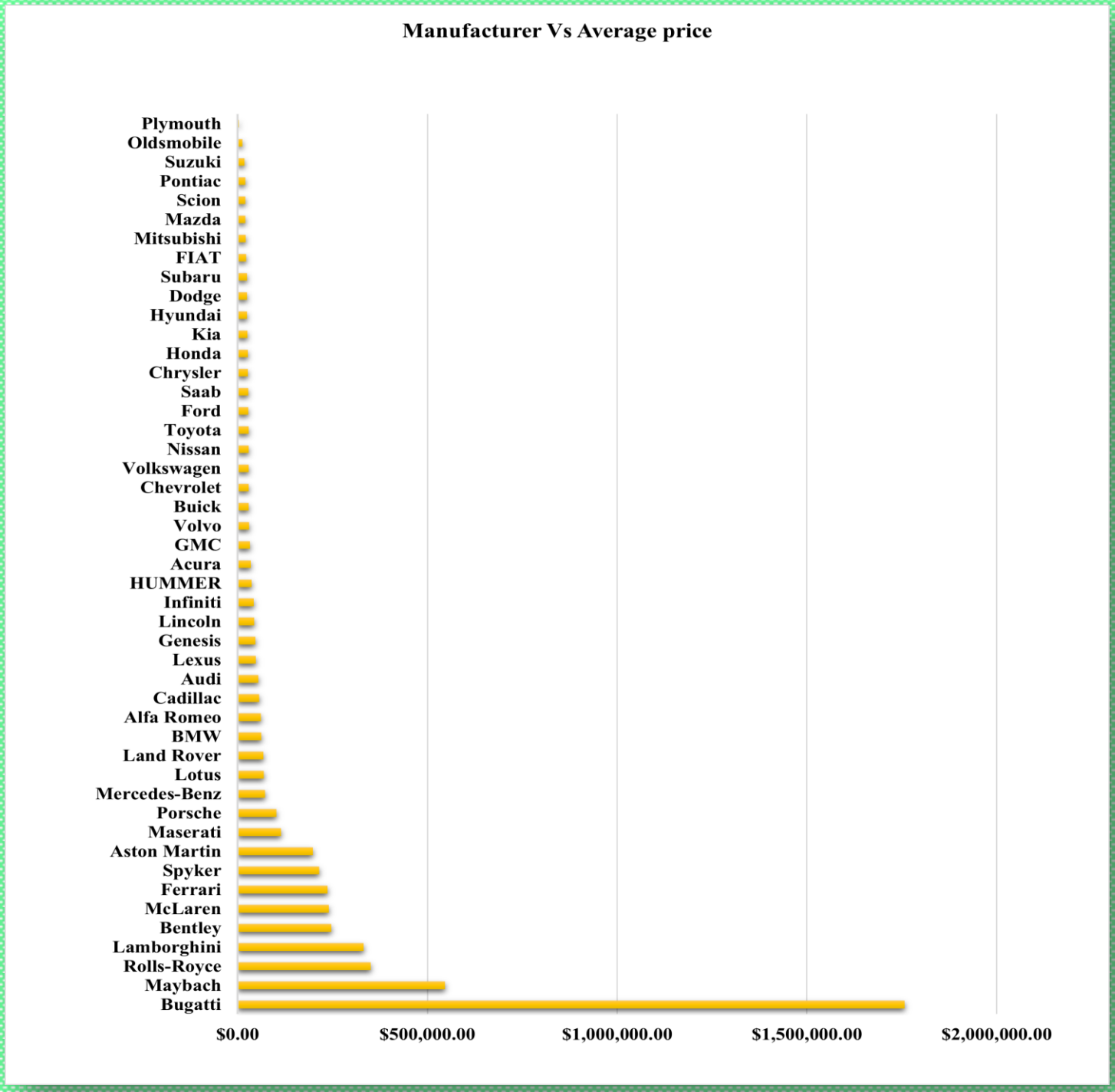
4.How does the average price of a car vary across different manufacturers?

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.

Manufacturers	Average Price
Bugatti	$1,757,223.67
Maybach	$546,221.88
Rolls-Royce	$351,130.65
Lamborghini	$331,567.31
Bentley	$247,169.32
McLaren	$239,805.00
Ferrari	$237,383.82
Spyker	$214,990.00
Aston Martin	$198,123.46
Maserati	$113,684.49
Porsche	$101,622.40
Mercedes-Benz	$72,135.03
Lotus	$68,377.14
Land Rover	$68,067.09
BMW	$62,162.56
Alfa Romeo	$61,600.00
Cadillac	$56,368.27
Audi	$54,574.12
Lexus	$47,549.07
Genesis	$46,616.67
Lincoln	$43,560.01
Infiniti	$42,640.27

Manufacturers	Average Price
HUMMER	$36,464.41
Acura	$35,087.49
GMC	$32,444.09
Volvo	$29,724.68
Buick	$29,034.19
Chevrolet	$29,000.22
Volkswagen	$28,947.37
Nissan	$28,856.42
Toyota	$28,758.77
Ford	$28,522.86
Saab	$27,879.81
Chrysler	$26,722.96
Honda	$26,608.88
Kia	$25,318.75
Hyundai	$24,926.26
Dodge	$24,857.05
Subaru	$24,240.67
FIAT	$22,206.02
Mitsubishi	$21,316.35
Mazda	$20,106.56
Scion	$19,932.50
Pontiac	$19,800.04
Suzuki	$18,021.05
Oldsmobile	$12,843.80
Plymouth	$3,296.87



From the chart we can infer that **Bugatti** has the highest average price.

5.What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

Task 5. A: 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.

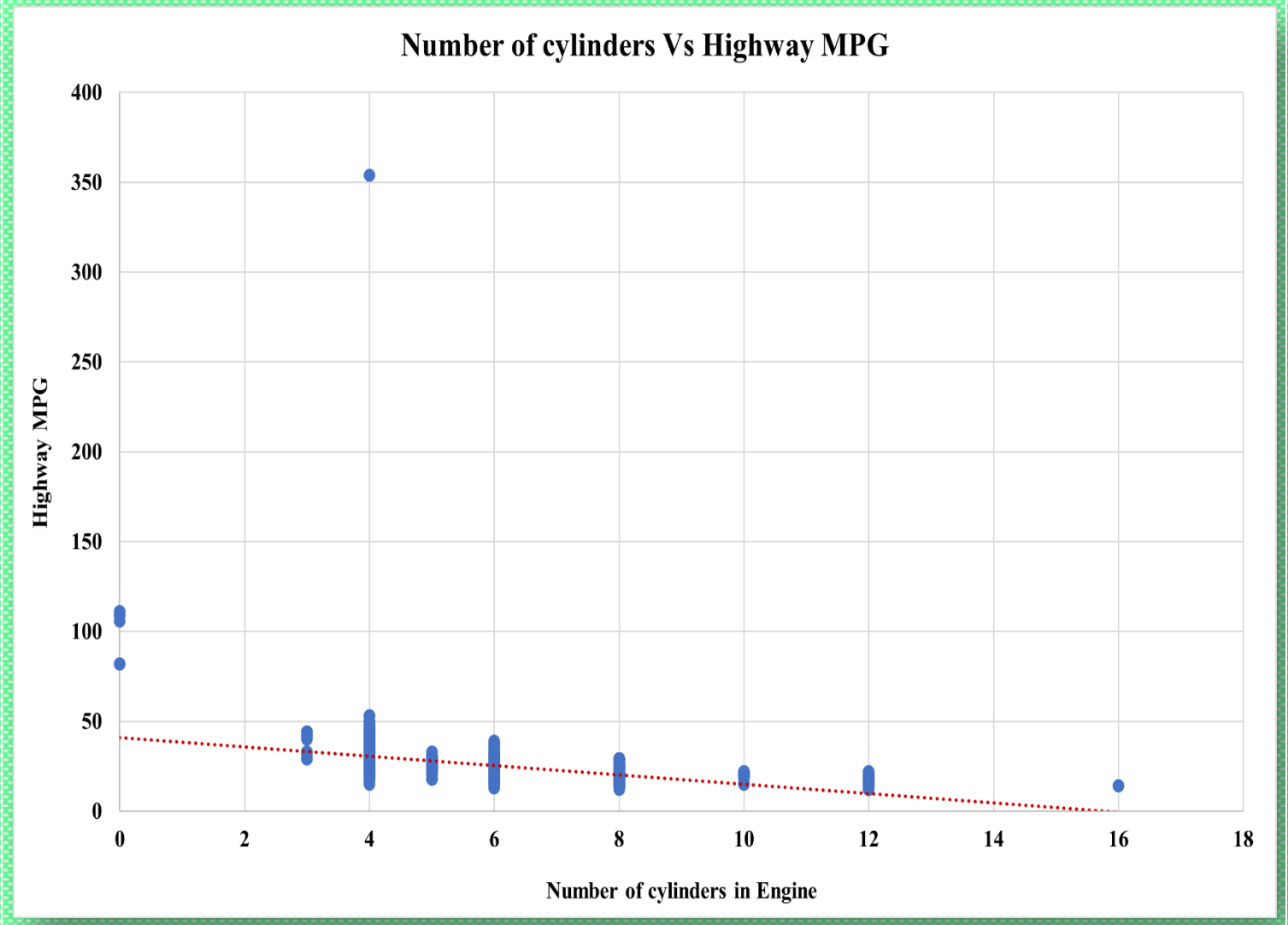
Task 5. B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

A. Created a **scatter chart** with the number of cylinders on the x-axis and highway MPG on the y-axis and a trendline have been plotted.

B. Correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship have been calculated using the formula:

=CORREL(A2:A11098,B2:B11098)

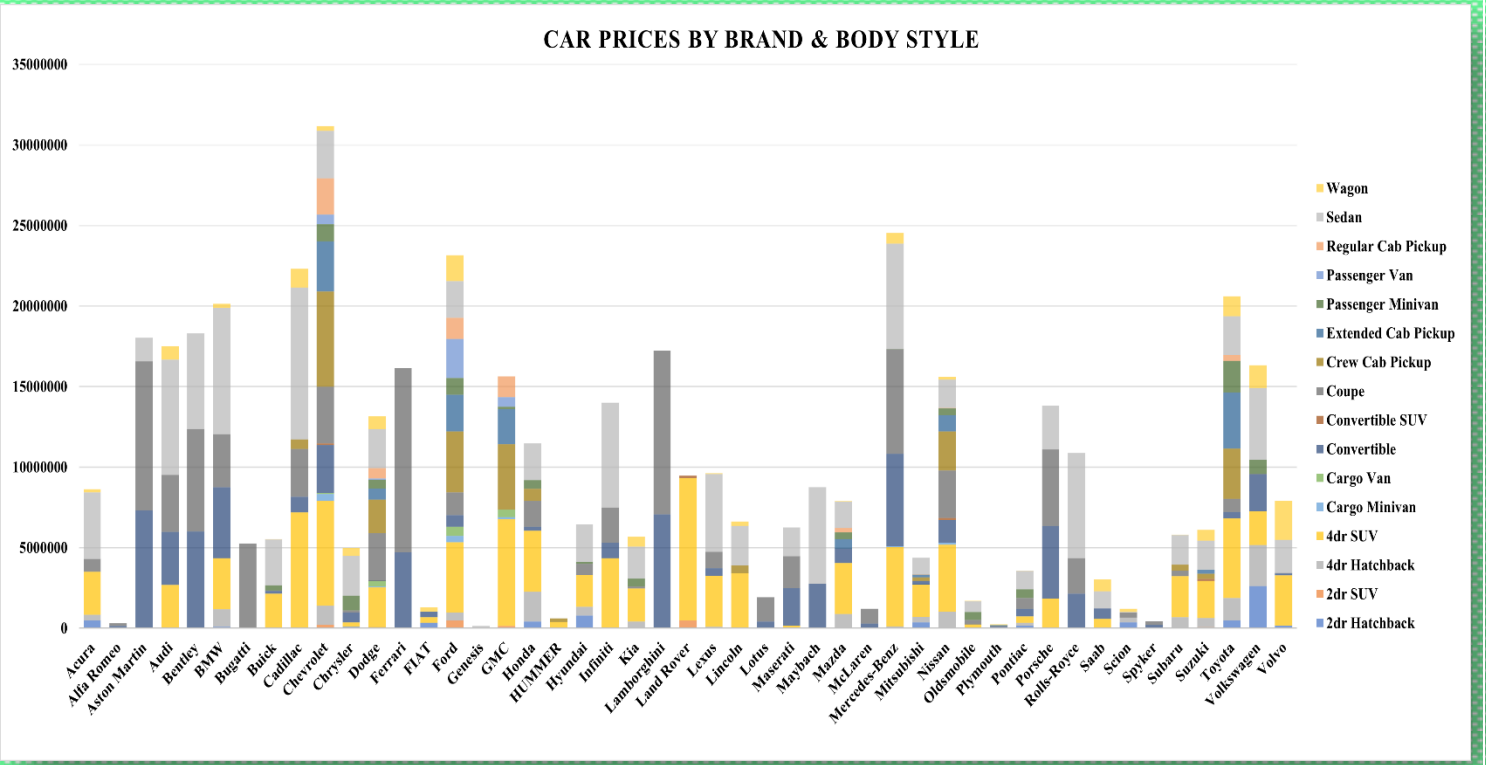
Correlation Coefficient	-0.614703



From the chart we can infer that if the **number of cylinders increases then the highway mpg will decrease.**

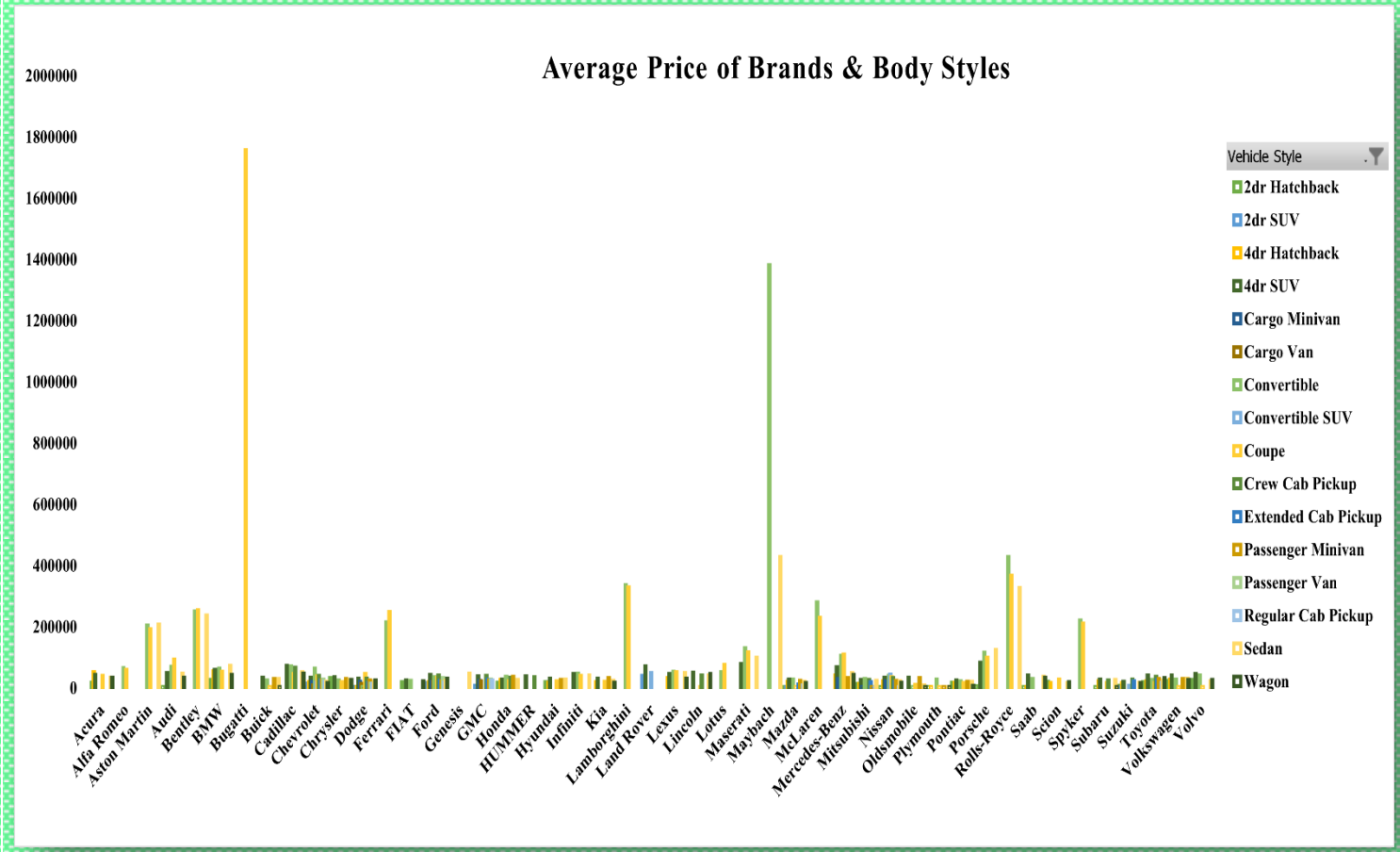
Building the Dashboard:

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



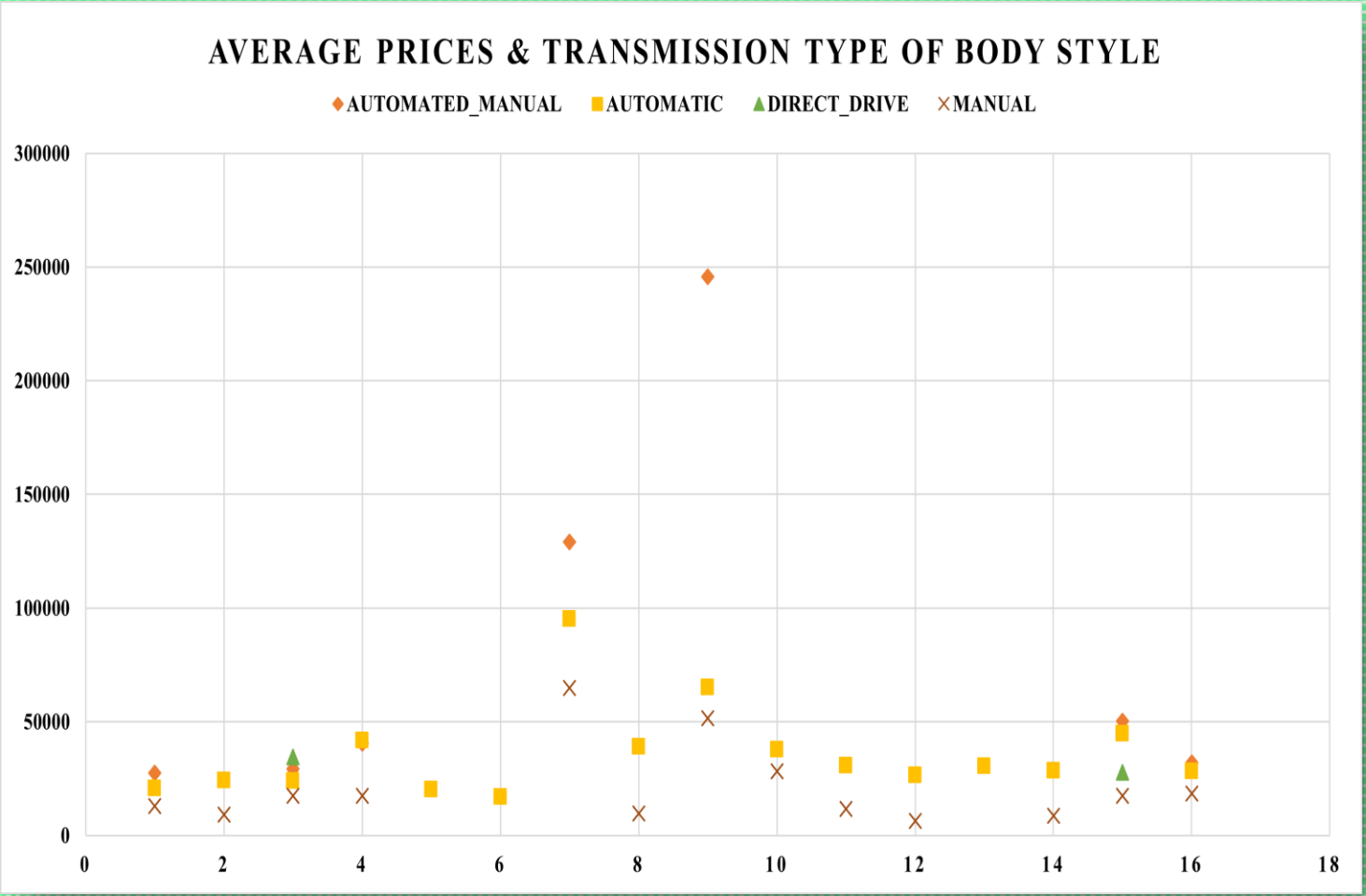
From the chart we can infer that **Chevrolet** has the highest price distribution by body style.

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



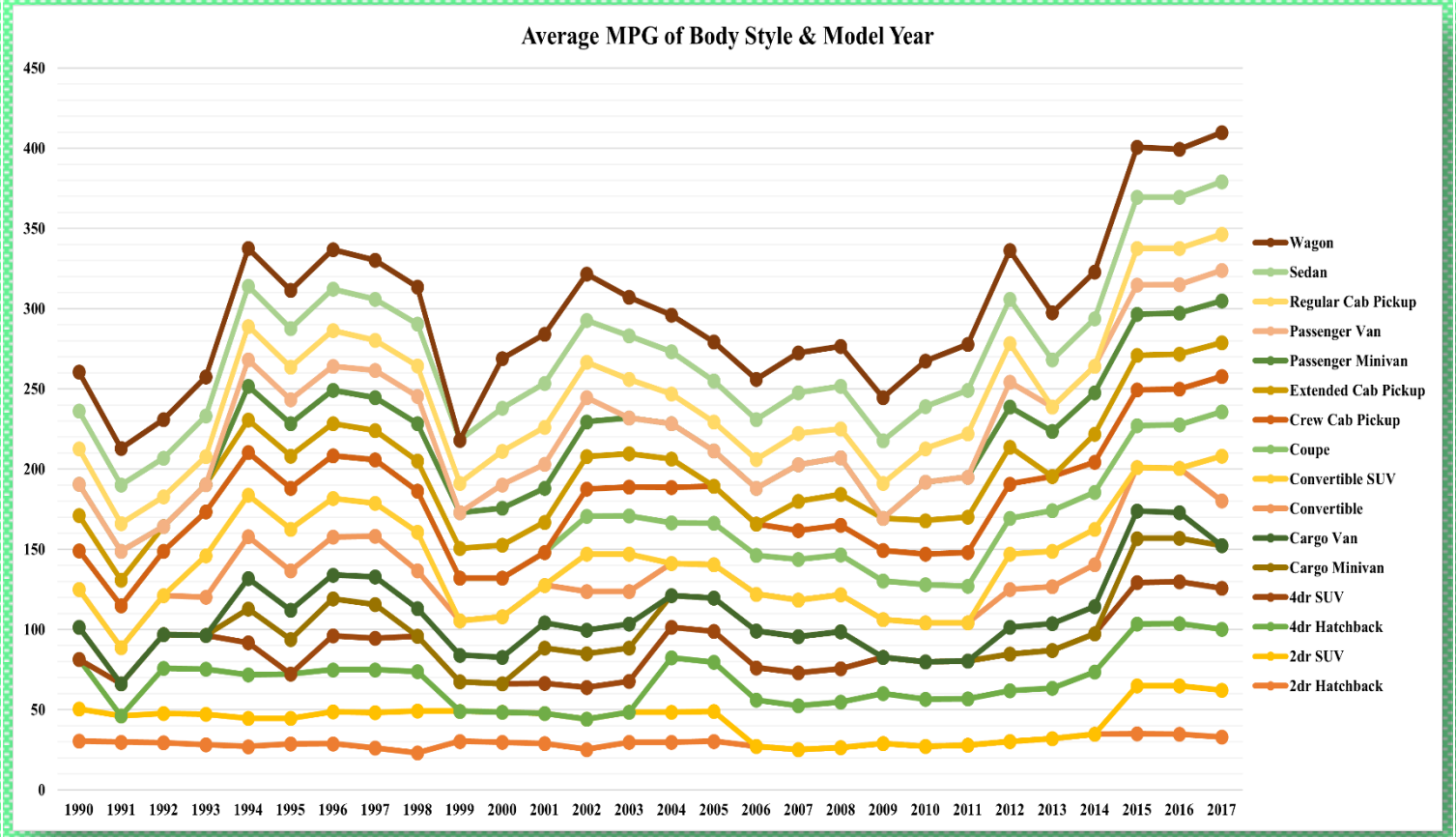
From the chart we can infer that **Bugatti has the highest average price** and **Plymouth has the lowest average price.**

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



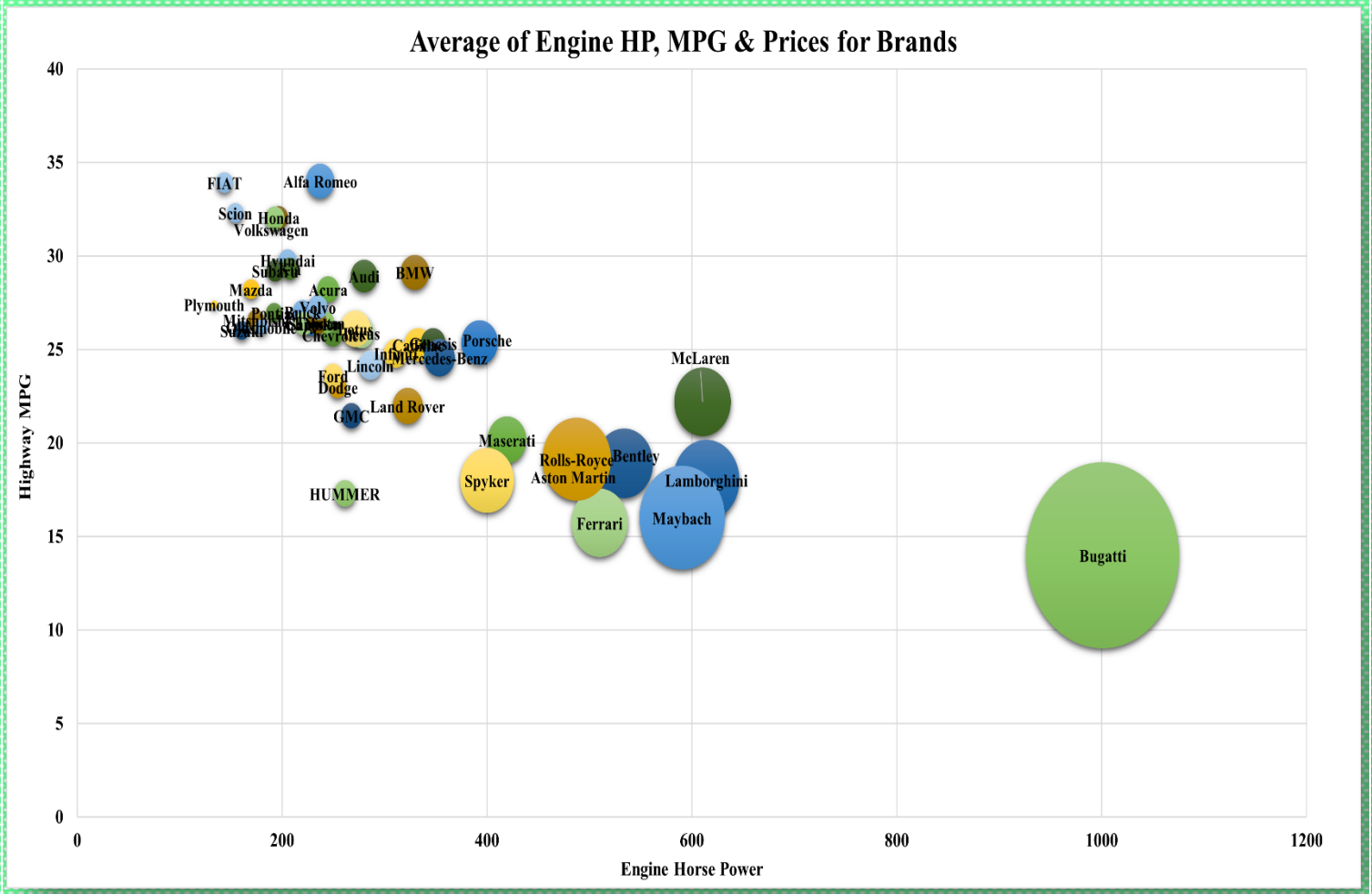
From the chart we can infer that **automated manual** is the most expensive transmission.

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



From the chart we can infer that fuel efficiency of cars increased across different body styles.

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



From the chart we can infer that if the engine horse power increases then highway mpg will decrease and the price also will increase hence making Bugatti more expensive.

Conclusion

From the above analysis, we can infer that the provided dataset has several meaningful insights which can help the car manufacturers to optimize pricing and product development decisions to maximize profitability while meeting consumer demand.

ABC CALL VOLUME TREND ANALYSIS

PROJECT-8

Project Description:

In this project I have provided with a dataset of a Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. You'll be provided with a dataset that spans 23 days and includes various details and the goal is to attract, engage, and delight customers, turning them into loyal advocates for the business. Additionally, the project involves addressing the issue of unanswered calls during the night and proposing a manpower plan for that time period as well.

Approach:

Firstly I have cleaned the dataset to ensure accurate and reliable results for the analysis. Then I have performed analytical methods and Visualization Methods. These techniques will help to get insights about the ABC Call Volume Trend Analysis.

Tech-Stack Used:

PPT – To prepare a detailed report.

EXCEL – Excel was used to perform entire analysis.

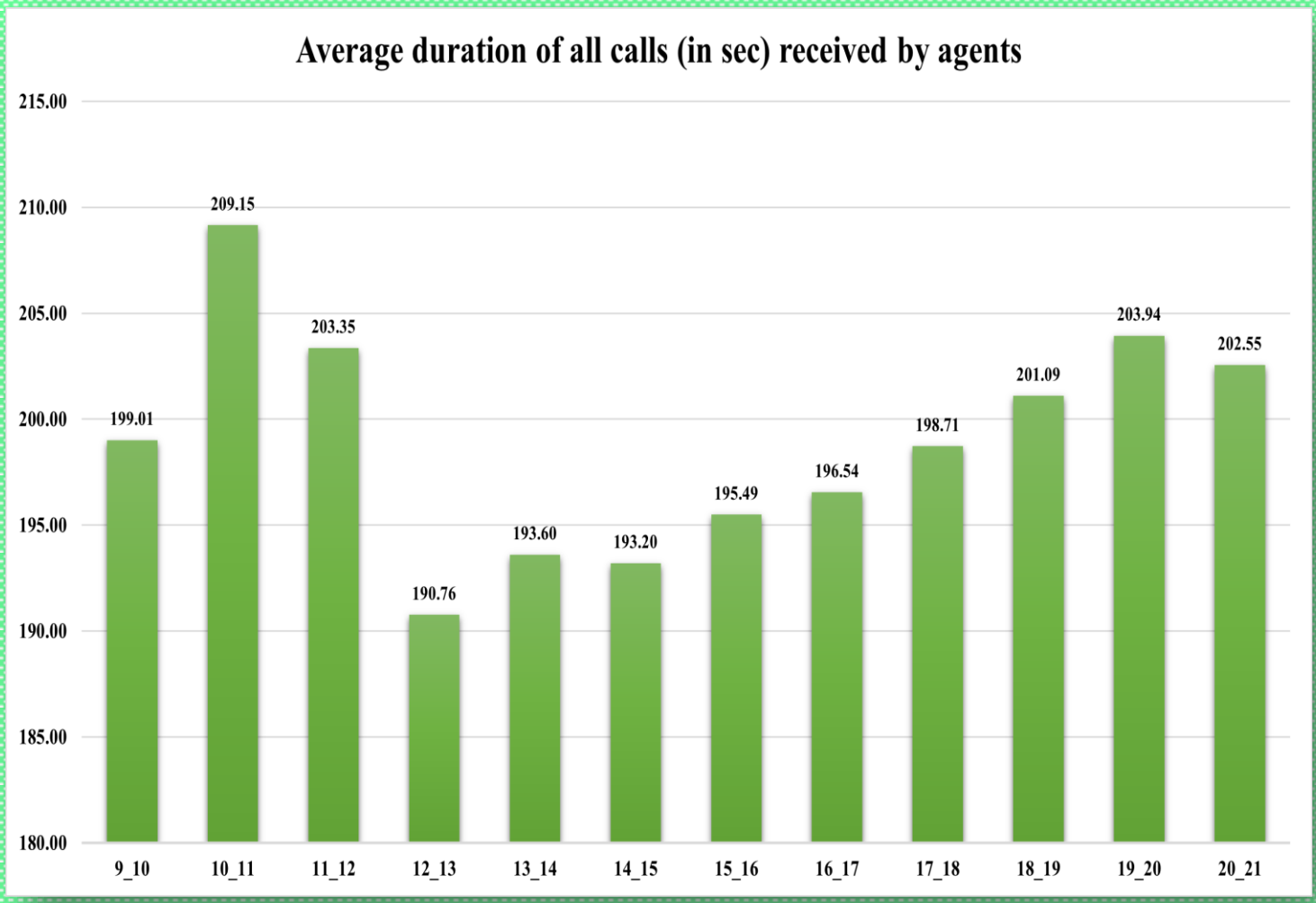
Insights:

1.Average Call Duration:

What is the average duration of calls for each time bucket?

From the above bar chart we can infer that time bucket 10_11 i.e. 10AM to 11AM has the highest of average of calls in seconds i.e. **209.15**

Time bucket	Average of Call Seconds (s)
9_10	199.01
10_11	209.15
11_12	203.35
12_13	190.76
13_14	193.60
14_15	193.20
15_16	195.49
16_17	196.54
17_18	198.71
18_19	201.09
19_20	203.94
20_21	202.55

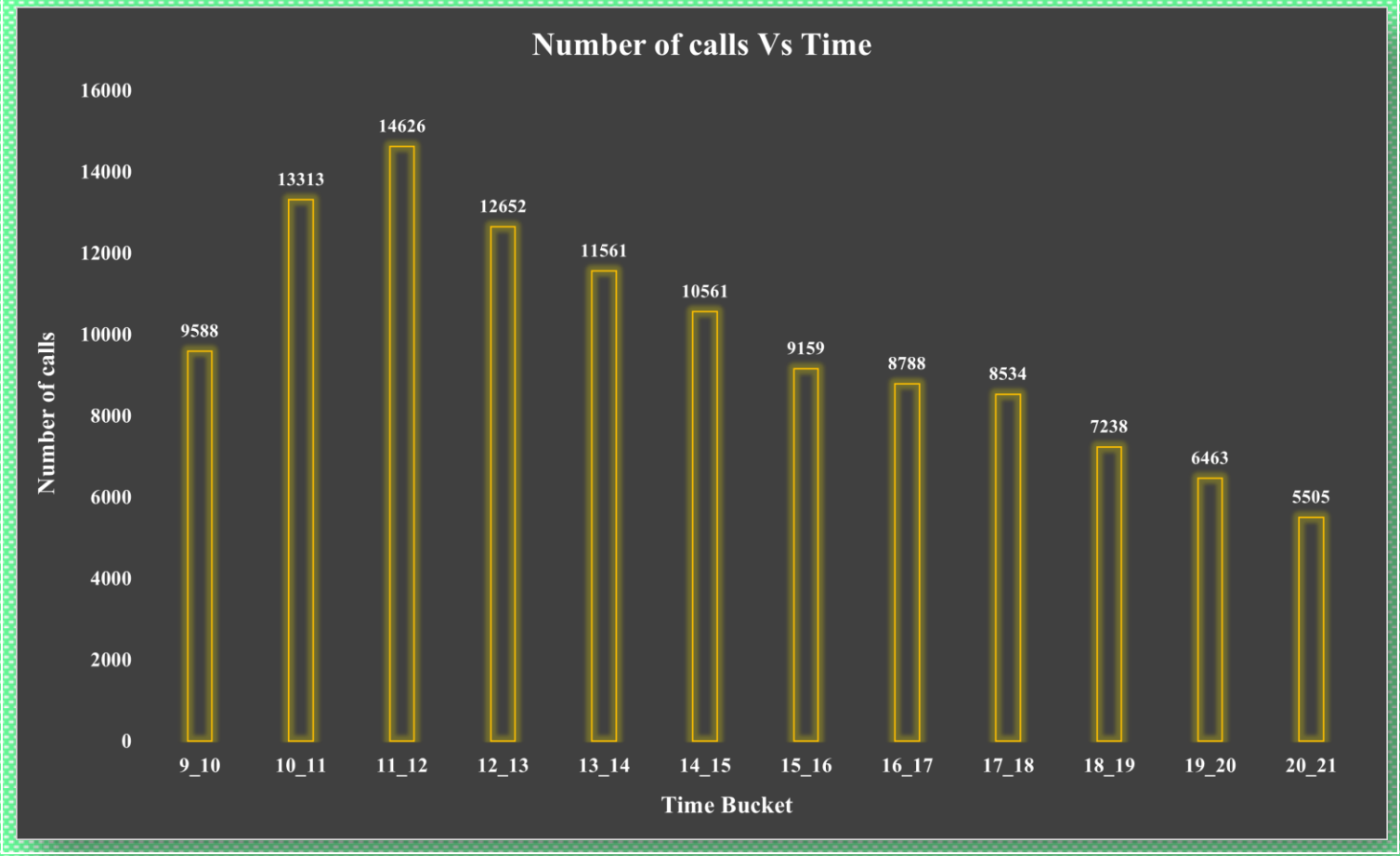


2.Call Volume Analysis:

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

From the bar chart we can infer that time bucket 11_12 i.e. 11AM to 12PM has the highest count for total number incoming calls i.e. **14626**

Time bucket	Number of calls
9_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505



3.Manpower Planning:

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

Assumptions:		
Agents working Hrs	9	Hrs
Break	1.5	Hrs
Agents working Hrs	7.5	Hrs
Agents Time spent on calls	4.5	Hrs

Insights:		
Avg Call Volume per day(9:00 AM to 9:00 PM)	5130	Nos
Average call duration (9:00 AM to 9:00 PM)	199	Seconds
Total call duration for 90% Calls	255	Hrs
Agents required per day	57	Persons

Time Bucket	Count of Time	Percentage of calls inbound	Agents Required
9_10	9588	8%	5
10_11	13313	11%	6
11_12	14626	12%	7
12_13	12652	11%	6
13_14	11561	10%	6
14_15	10561	9%	5
15_16	9159	8%	4
16_17	8788	7%	4
17_18	8534	7%	4
18_19	7238	6%	3
19_20	6463	5%	3
20_21	5505	5%	3
Grand Total	117988	100%	57

The minimum number of agents required to reduce the abandon rate to 10% is **57 agents.**

4.Night Shift Manpower Planning:

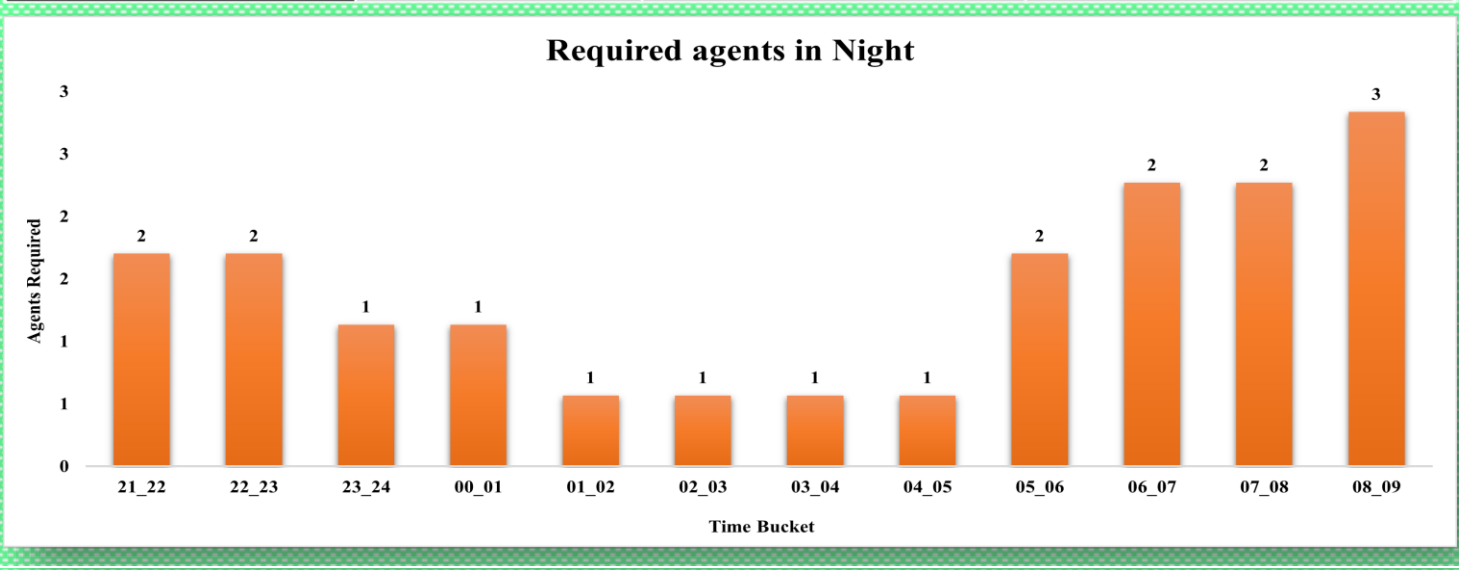
Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

Assumptions:		
Agents working Hrs	9	Hrs
Break	1.5	Hrs
Agents working Hrs	7.5	Hrs
Agents Time spent on calls	4.5	Hrs

Insights:		
Avg Call Volume per day(9:00 AM to 9:00 PM)	5130	Nos
Average call duration (9:00 AM to 9:00 PM)	199	Seconds
Total call duration for 90% Calls	255	Hrs
Agents required per day	57	Persons

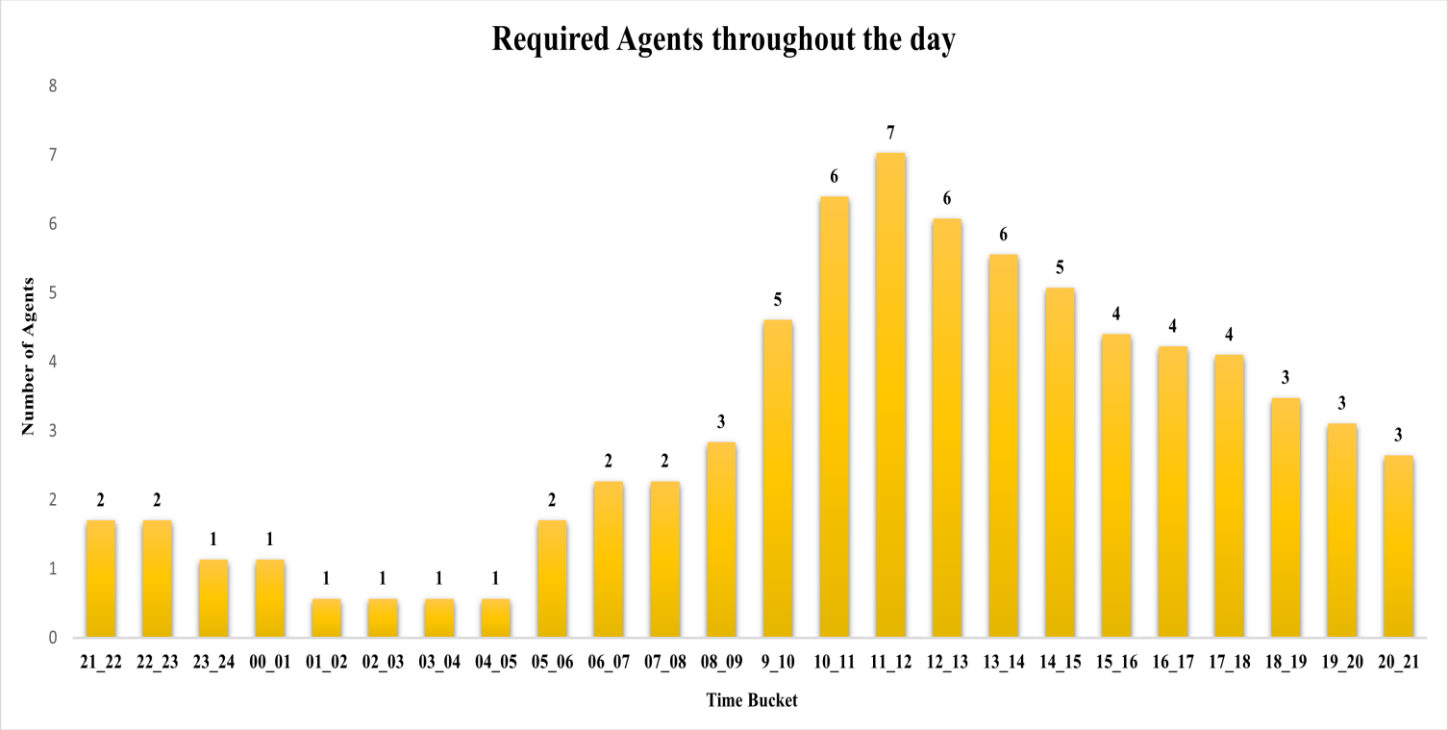
Agents required (9:00 PM to 9:00 AM):		
Avg Call Volume (9:00 PM to 9:00 AM)	1539	Nos
Total Call duration (9:00 PM to 9:00AM)	77	Hours
Agents required (9:00 PM to 9:00AM)	17	Persons

Time_Bucket	Calls in nights	Percentage of calls	Required agents in night
21_22	3	10%	2
22_23	3	10%	2
23_24	2	7%	1
00_01	2	7%	1
01_02	1	3%	1
02_03	1	3%	1
03_04	1	3%	1
04_05	1	3%	1
05_06	3	10%	2
06_07	4	13%	2
07_08	4	13%	2
08_09	5	17%	3
Total	30	100%	17



Therefore the ABC company needs **17 Agents** to answer the consumer calls at night time by keeping the answered rate to 90% / Abandon rate to 10%.

Time_Bucket	Required Agents throughout the day
21_22	2
22_23	2
23_24	1
00_01	1
01_02	1
02_03	1
03_04	1
04_05	1
05_06	2
06_07	2
07_08	2
08_09	3
9_10	5
10_11	6
11_12	7
12_13	6
13_14	6
14_15	5
15_16	4
16_17	4
17_18	4
18_19	3
19_20	3
20_21	3



Therefore the ABC company needs **74 Agents per day** to answer the consumer calls for day as well as the night time keeping the answered rate to 90% / Abandon rate to 10%.

CONCLUSION

From the above analysis, we can infer that the provided dataset has several meaningful insights which can help the ABC company to optimize customer service department and it can help improve the customer satisfaction.

Link for all the projects:

https://drive.google.com/drive/folders/1HNYuAyzY5zMzVKOtDuT8rvizwe4oov8d?usp=sharing

Link for Portfolio:

https://drive.google.com/drive/folders/1MyoBEz7u9Pel8-hSEEluqP3ujrO9dNKp?usp=sharing