Data Analyst Portfolio Project

- Prathyusha K

Professional Background

I have completed my Engineering in Electrical and Electronics engineering in 2021. I am currently working at TCS as a System Engineer. Inspired by the role of data in optimizing system performance, I am transitioning into a Data Analytics career to leverage my analytical skills and technical expertise in a new context. As a System Engineer, I've had the privilege of honing my technical skills and problem-solving abilities in optimizing complex systems. However, over time, I've become increasingly fascinated by the transformative power of data and its ability to uncover valuable insights.

My decision to transition into a data analytics career is driven by my desire for continuous growth and my aspiration to contribute to strategic decision-making through data analysis. I see this transition as a natural progression that aligns perfectly with my long-term goal of making a meaningful impact within a data-driven environment.

In order to gain some experience in data analysis, I have done some real time projects to showcase my skills.

Table Of Contents

Project 1 - Data Analytics Process

Project Description

We use Data Analytics in everyday life without even knowing it.
For eg : Going to a market to buy something. We follow steps like Plan, Prepare, Process, Analyse, Share, Act.

My 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.

My examples of such real-life situations where we can use data analytics are:

- Buying a car
- Investing in a House
- College selection
- Monthly expenses

Buying a Car Analysis

1.Plan: A person wants to buy a car that meets their specific needs and budget.

2.Prepare: The person researches on cars using various online resources, such as car review websites and dealership websites, to identify cars that meet their criteria.

3.Process: Based on the research, the information(data) collected is cleaned and processed to ensure accuracy, such as verifying the reliability ratings ,price and cost of ownership for each car.

4.Analyze: The person performs an analysis of the car data to identify trends, such as which cars have the highest safety ratings and which cars have the lowest cost of ownership.

5.Share: The person shares the results of their analysis with a trusted friend or family member to get feedback on their findings.

6.Act: The person uses the insights gained from their analysis to make an informed decision on which car to buy, such as negotiating a better price or deciding to buy a different car altogether

Investing in a House Analysis

1.Plan: A person wants to invest in a house and wants to ensure that they are making a good investment.

2.Prepare: The person collects information(data) on the housing market, such as recent trends in property values and interest rates, and identifies specific neighborhoods or areas of interest.

3.Process: The data is cleaned and processed to ensure accuracy and completeness, such as removing the houses with worst facilities or unsocialized neighborhood.

4.Analyze: The person performs an analysis of the housing market data to identify trends and patterns, such as which neighborhoods are appreciating in value and which ones are experiencing a decline.

5.Share: The person shares the results of their analysis with a trusted real estate agent or financial advisor to get feedback on their findings.

6.Act: The person uses the insights gained from their analysis to make an informed decision on which house to invest in, such as deciding to invest in a house in an up-and-coming neighborhood with a high potential for appreciation.

College Selection Analysis

1.Plan: A student wants to select the best college for their higher studies and wants to ensure that they are making an informed decision.

2.Prepare: The student collects data on colleges they are interested in, such as rankings, acceptance rates, tuition costs, location, program offerings, student demographics, and student-to-faculty ratios.

3.Process: The data is cleaned and processed to ensure accuracy and completeness, such as removing colleges with not good facilities, or no placements etc.

4.Analyze: The student performs an analysis of the college data to identify trends and patterns, such as which colleges have the best program offerings in their field of interest, which colleges have the highest graduation rates, and which colleges have the lowest student-to-faculty ratios.

5.Share: The student shares the results of their analysis with trusted mentors, such as teachers, guidance counselors, or family members, to get feedback and insights on their findings.

6.Act: The student uses the insights gained from their analysis to make an informed decision on which college to attend, such as selecting a college with strong program offerings in their field of interest, a high graduation rate, and a low student-to-faculty ratio.

Monthly Expenses Analysis

1.Plan: A person wants to understand their monthly expenses to better manage their budget and identify areas where they can save money.

2.Prepare: The person collects data on their expenses for the month, including rent/mortgage, utilities, groceries, transportation, entertainment, and any other expenses they incurred.

3.Process: The data is cleaned and processed to ensure accuracy and completeness, such as categorizing each expense into a specific budget category and removing unnecessary expenses.

4.Analyze: The person performs an analysis of their monthly expenses to identify patterns and trends, such as which budget categories account for the majority of their spending, where they might be overspending or underspending, and where they can make cuts to reduce their expenses.

5.Share: The person may share the results of their analysis with a financial advisor or a friend/family member for feedback and advice on how to better manage their budget.

6.Act: Based on the insights gained from their analysis, the person may take actions such as creating a budget plan, reducing expenses in certain categories, or finding ways to earn additional income to supplement their budget.

Project 2 – Instagram User Analytics

Project Description

User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.
These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.

Instagram is also such social media app. Using Instagram for business can drive brand awareness, boost sales, and build and track audience engagement. It's an excellent way to find customers where they're already spending time. It can also provide valuable audience insights to use with all your marketing plan strategies.

I’m are working with the product team of Instagram and the product manager has asked me to provide insights on the questions asked by the management team.

Tech stack used : SQL online complier db-Fiddle

Insights

1. Rewarding Most Loyal Users: People who have been using the platform for the longest time.

Query #1 **Execution time: 0ms**

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

The most loyal users are Darby_Herzog, Emilio_Bernier52, Elenor88, Nicole71, Jordyn.Jacobson2.

2. Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.

Query #1 **Execution time: 2ms**

user_id	likes
36	257

The winner for the contest id user_id 36, with 257 likes on a single photo.

3. Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo.

Query #1 **Execution time: 1ms**

username	id
Aniya_Hackett	5
Kasandra_Homenick	7
Jaclyn81	14
Rocio33	21
Maxwell.Halvorson	24
Tierra.Trantow	25
Pearl7	34
Ollie_Ledner37	36
Mckenna17	41
David.Osinski47	45
Morgan.Kassulke	49
Linnea59	53
Duane60	54
Julien_Schmidt	57
Mike.Auer39	66
Franco_Keebler64	68
Nia_Haag	71
Hulda.Macejkovic	74
Leslie67	75
Janelle.Nikolaus81	76
Darby_Herzog	80
Esther.Zulauf61	81
Bartholome.Bernhard	83
Jessyca_West	89

There are many users who haven't posted photos on Instagram. The above listed users are inactive and remainders needs to sent to them to post their 1st photo.

4. Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.

Query #1 Execution time: 1ms

tag_name	count
lol	1
party	1
photography	1
beach	1
smile	1

The top most hashtags used by the people on the platform are “lol”,” party”,” photography”,” beach”,” smile”.

5. Launch AD Campaign: The team wants to know, which day would be the best day to launch ADs.

Query #1 Execution time: 0ms

total	date
2	2017-02-06
2	2016-05-06
2	2017-03-30
2	2017-01-01
2	2017-01-23

Most of the users register on the above days. So, these are the best days to launch the AD campaigns.

6. **User Engagement:** Are users still as active and post on Instagram or they are making fewer posts

Query #1 Execution time: 1ms

average_posts_per_user	average_photos_per_user
0.7400	2.5700

By the above result, we can confirm that the user's engagement in the platform is active and they are posting many photos.

7. **Bots & Fake Accounts:** The investors want to know if the platform is crowded with fake and dummy accounts

Query #1 Execution time: 2ms

user_id	total_likes
5	257
14	257
21	257
24	257
36	257
41	257
54	257

The above result shows the user_id of the bots who have liked every photo in the platform.

Conclusion

The below are some insights obtained by analysing the data:

1. There are many loyal users in this platform, so the marketing team can reward them or make them participate in some campaigns.

2. There are few days of a week where most of the users register, so the team can utilise these days and launch the AD campaigns.

3. By analysing the data, we can see that Instagram is performing well and is not becoming redundant like Facebook and the user's engagement is active. There are not may fake accounts/bots, so the investors can be at ease.

Project 3 – Operation Analytics and Investigating Metric Spike

Project Description

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

I'm working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which I must derive certain insights out of it and answer the questions asked by different departments.

Tech stack used : SQL online complier db-Fiddle

Insights

Case Study 1 (Job Data)

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

Query SQL

```sql
SELECT
  ds,
  CONCAT(ds, ' ', LPAD(EXTRACT(HOUR FROM ds), 2, '0')) AS hour,
  COUNT(*) AS jobs_reviewed
FROM
  job_data
WHERE
  ds >= '2020-11-01' AND ds < '2020-12-01'
  AND EXTRACT(MONTH FROM ds) = 11
  AND EXTRACT(YEAR FROM ds) = 2020
GROUP BY
  ds,
  hour
ORDER BY
  ds,
  hour;
```

ds	hour	jobs_reviewed
2020-11-25 00:00:00	2020-11-25 00:00:00 00	1
2020-11-26 00:00:00	2020-11-26 00:00:00 00	1
2020-11-27 00:00:00	2020-11-27 00:00:00 00	1
2020-11-28 00:00:00	2020-11-28 00:00:00 00	2
2020-11-29 00:00:00	2020-11-29 00:00:00 00	1
2020-11-30 00:00:00	2020-11-30 00:00:00 00	2

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

Query SQL

```sql
SELECT
  t1.ds,
  t1.throughput,
  AVG(t2.throughput) AS rolling_average
FROM
  (SELECT
     ds,
     COUNT(*) / 86400.0 AS throughput
   FROM
     job_data
   GROUP BY
     ds) t1
JOIN
  (SELECT
     ds,
     COUNT(*) / 86400.0 AS throughput
   FROM
     job_data
   GROUP BY
     ds) t2
ON
  t2.ds BETWEEN DATE_SUB(t1.ds, INTERVAL 6 DAY) AND t1.ds
GROUP BY
  t1.ds,
  t1.throughput
ORDER BY
  t1.ds;
```

ds	throughput	rolling_average
2020-11-25 00:00:00	0.0222	0.0222
2020-11-26 00:00:00	0.0179	0.0198
2020-11-27 00:00:00	0.0096	0.0146
2020-11-28 00:00:00	0.0606	0.0176
2020-11-29 00:00:00	0.05	0.0202
2020-11-30 00:00:00	0.05	0.0229

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

Query SQL

```sql
SELECT
  language,
  COUNT(*) * 100.0 / (
    SELECT
      COUNT(*)
    FROM
      job_data
    WHERE
      ds >= DATE_SUB(CURRENT_DATE, INTERVAL 30 DAY)
  ) AS percentage_share
FROM
  job_data
WHERE
  ds >= DATE_SUB(CURRENT_DATE, INTERVAL 30 DAY)
GROUP BY
  language;
```

language	percentage
Italian	12.5
Persian	37.5
French	12.5
Hindi	12.5
Arabic	12.5
English	12.5

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

Query SQL

```
1 select * from
2 (
3 select *,
4 row_number()over(partition by job_id) as rowno
5 from job_data
6 )d
7 where rowno>1;
8
```

ds	job_id	actor_id	event	language	time_spent	org	row_num
2020-11-25 00:00:00	20	1003	transfer	Italian	45	C	2
2020-11-26 00:00:00	23	1004	skip	Persian	56	A	2
2020-11-27 00:00:00	11	1007	decision	French	104	D	2
2020-11-28 00:00:00	25	1002	decision	Hindi	11	B	2
2020-11-28 00:00:00	23	1005	transfer	Persian	22	D	2
2020-11-29 00:00:00	23	1003	decision	Persian	20	C	2
2020-11-30 00:00:00	21	1001	skip	English	15	A	2
2020-11-30 00:00:00	22	1006	transfer	Arabic	25	B	2

Case Study 2 (Investigating metric spike)

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

Query SQL

```sql
SELECT
  DATE_TRUNC('week', e.event_date) AS week_start,
  COUNT(DISTINCT e.user_id) AS weekly_engaged_users
FROM
  events e
GROUP BY
  week_start
ORDER BY
  week_start;

```

week_start	Weekly_engaged_users
17	8019
18	17341
19	17224
20	17911
21	17151
23	18280
24	18413
25	18642
26	19061
27	19881
28	20776
29	20067
30	21533
31	18556
32	16612
33	16145
34	16127
35	784

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

Query SQL

```
1 select *,
2 new_user_activated-lag(new_user_activated) over( order by year_,quarter_ ) as user_growth
3 from(select year(created_at) as PresentYear, quarter(created_at) as PresentQuarter, count(user_id) as
  new_user_activated
4 from users
5 where
6 activated_at is not null and state='active'
7 group by 1,2)a ;
8
```

PresentYear	PresentQuarter	new_user_activated	user_growth
2013	1	470	NULL
2013	2	608	138
2013	3	930	322
2013	4	1275	345
2014	1	1692	417
2014	2	2378	686
2014	3	2028	-350

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

Query SQL

```sql
Select
week_period,
first_value(cohort_retained) over (order by week_period) as cohort_size,
cohort_retained,
cohort_retained / first_value(cohort_retained) over (order by week_period) as pct_retained
From
(select
timestampdiff(week,a.activated_at,b.occurred_at) as week_period,
count(distinct a.user_id) as cohort_retained
From
(select user_id, activated_at
 from users where state='active'group by 1) a
inner join
(select user_id,occurred_at from events )b
 on a.user_id=b.user_id
group by 1) c;

```

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

Query SQL

```sql
Select
device_name,
avg(no_of_users) as avg_weekly_users,
avg(device_used) as avg_times_used
From
(select week(occurred_at) as week,
device as device_name ,
count(distinct user_id) as no_of_users,
count(device) as device_used
from events
where event_name='login'
group by 1,2
order by 1) a
group by 1;

```

device_name	avg_weekly_users	avg_times_used.
acer aspire desktop	26	32.9474
acer aspire notebook	43.1579	56.8421
amazon fire phone	10.5556	13.7778
asus chromebook	43.5263	58.8947
dell inspiron desktop	46.6316	62.7368

E. **Email Engagement:** Users engaging with the email service.

Query SQL

```
Select week, num_users, time_weekly_digest_sent,
time_weekly_digest_sent-lag(time_weekly_digest_sent) over(order by week) as time_weekly_digest_sent_growth,
time_email_open,time_email_open-lag(time_email_open) over(order by week) as time_email_open_growth,
time_email_clickthrough,time_email_clickthrough-lag(time_email_clickthrough) over(order by week) as
time_email_clickthrough_growth
From
(select week(occurred_at)as week,
count(distinct user_id) as num_users,
 sum(if(action='sent_weekly_digest',1,0)) as time_weekly_digest_sent,
sum(if(action='email_open',1,0)) as time_email_open,
sum(if(action='email_clickthrough',1,0)) as time_email_clickthrough
from email
group by 1
order by 1) a;
```

week	num_users	time_weekly_digest_sent	time_weekly_digest_sent_growth	time_email_open	time_email_open_growth	time_email_clickthrough	time_email_clickthrough_growth
17	981	908	NULL	310	NULL	166	NULL
18	2714	2602	1694	912	602	430	264
19	2787	2665	63	972	60	477	47
20	2874	2733	68	1004	32	507	30
21	2926	2822	89	1014	10	443	-64
22	3029	2911	89	987	-27	488	45
23	3134	3003	92	1075	88	538	50
24	3254	3105	102	1155	80	554	16
25	3343	3207	102	1096	-59	530	-24
26	3439	3302	95	1165	69	556	26
27	3543	3399	97	1228	63	621	65
28	3641	3499	100	1250	22	599	-22
29	3734	3592	93	1219	-31	590	-9
30	3866	3706	114	1383	164	630	40
31	3950	3793	87	1351	-32	445	-185

Conclusion

Case Study 1 (Job Data):

- The number of jobs reviewed per hour per day for November 2020 is 75%.
- It is observed that the 7-day rolling average smoothes out the daily variations and provides a more generalized view of the overall trend.
- Persian is the language that is used mostly (37.5%), rest of the languages share 12.7%.

Case Study 2 (Investigating metric spike):

- The activeness of the used increased in week 30, then it got decreased.
- The overall count of weekly engagement per device used is the most for MacBook users and Samsung galaxy users.
- There is an increase in the usage of email and we can clearly see that users are comfortably engaging themselves in email usage.

Project 4 - Hiring Process Analytics

Project Description

Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyze before hiring freshers or any other individual.

Being a Data Analyst, my job is to go through these trends and draw insights out of it for hiring department to work upon.

I’m working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked me to answer certain questions making sense out of that data.

Tech stack used : Microsoft Excel

Insights

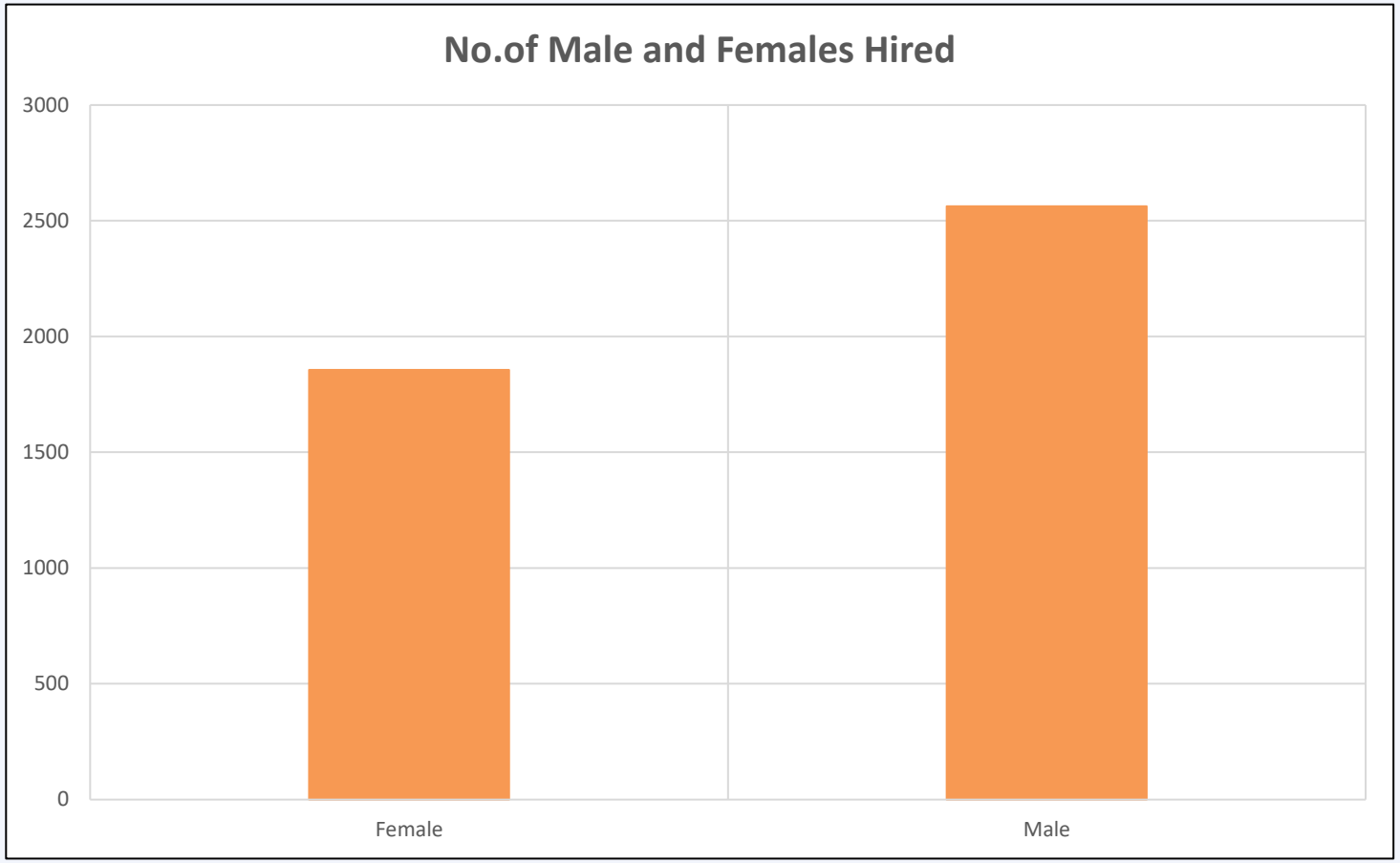
The data set which is given to me is is by the hiring department. I'll follow the pattern from my approach and analyze the dataset. The dataset consists of columns "application_id", "Interview Taken on", "Status", "event_name", "Department", "Post Name", "Offered Salary". By using the data given in the columns I'm going to jot down the insights.

application_id	Interview Taken on	Status	event_name	Department	Post Name	Offered Salary
383422	01-05-2014 11:40	Hired	Male	Service Department	c8	56553
907518	06-05-2014 08:08	Hired	Female	Service Department	c5	22075
176719	06-05-2014 08:08	Rejected	Male	Service Department	c5	70069
429799	02-05-2014 16:28	Rejected	Female	Operations Department	i4	3207
253651	02-05-2014 16:32	Hired	Male	Operations Department	i4	29668
289907	01-05-2014 07:44	Hired	Male	Sales Department	-	85914
959124	06-05-2014 16:27	Rejected	Male	Sales Department	i7	69904
86642	09-05-2014 13:17	Rejected	Male	Sales Department	i7	11758
751029	02-05-2014 13:09	Hired	Female	Service Department	i4	15156
434547	02-05-2014 13:11	Rejected	Female	Service Department	i4	49515
518854	01-05-2014 09:00	Rejected	Male	Service Department	n10	26990
649039	07-05-2014 10:48	Hired	Female	Service Department	b9	200000
199526	07-05-2014 10:50	Hired	Male	Service Department	b9	86787
539803	15-05-2014 09:31	Hired	Male	Finance Department	b9	2308
191009	09-05-2014 12:48	Hired	Female	Service Department	i7	56688
195323	09-05-2014 12:48	Hired	-	Service Department	i7	81757
51318	02-05-2014 08:07	Hired	Male	Service Department	i5	15134
742283	02-05-2014 08:11	Rejected	-	Service Department	i5	100
513166	01-05-2014 22:53	Hired	Female	Operations Department	i1	73579

A. Hiring: Process of intaking of people into an organization for different kinds of positions

I have used Excel function COUNTIFS to fetch the number of Males and Females who were hired.
The output of the function is as follows:

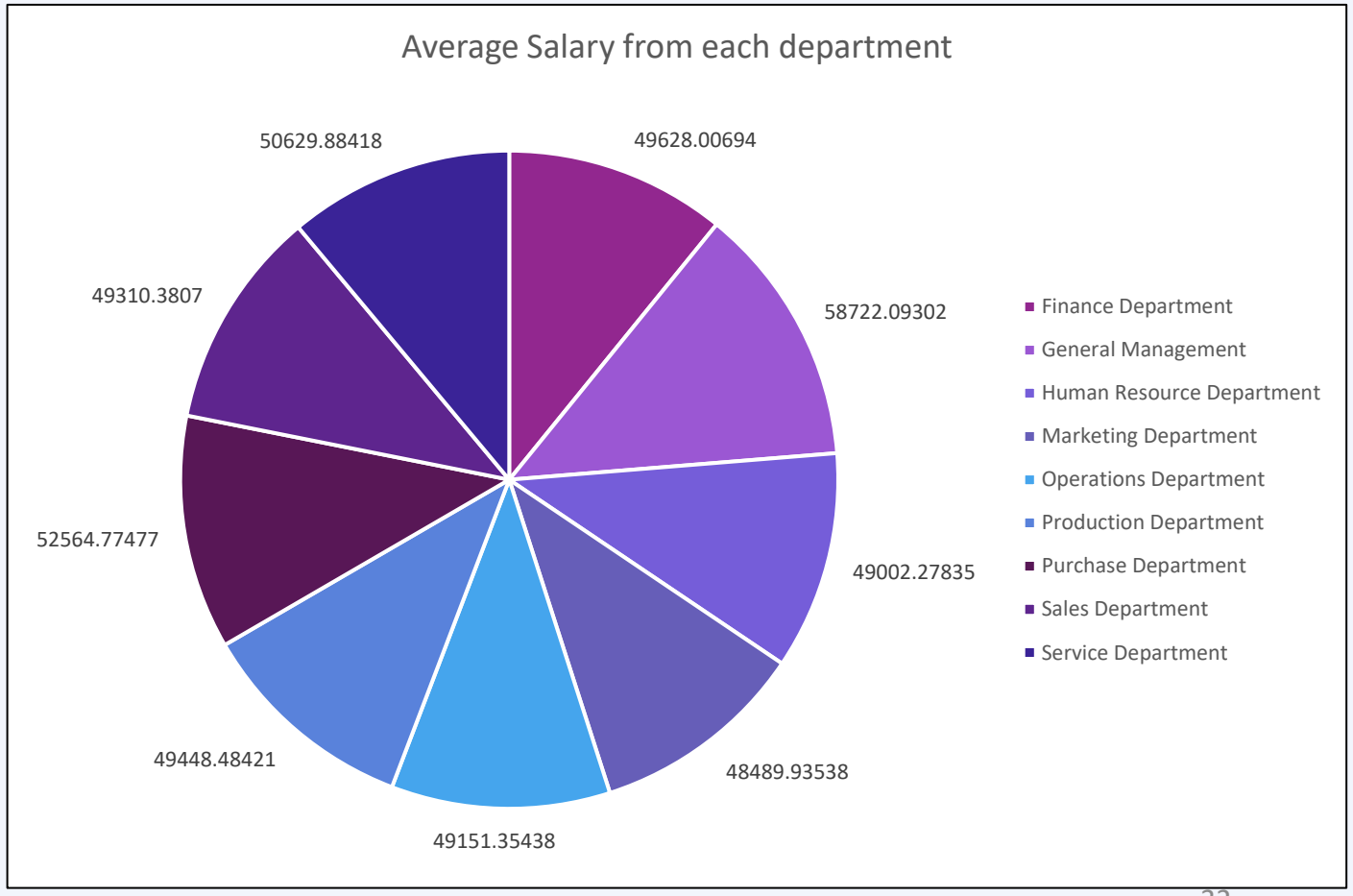
No. of Males and Females Hired:	
Male	2563
Female	1856



B. Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.

I have used Excel function AVERAGEIF to fetch the average salary in each department.
The output of the function is as follows:

Average salary in each department:	
Department	Average Salary
Service Department	50629.88418
Operations Department	49151.35438
Sales Department	49310.3807
Finance Department	49628.00694
General Management	58722.09302
Human Resource Department	49002.27835
Marketing Department	48489.93538
Production Department	49448.48421
Purchase Department	52564.77477



C. Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.

I have used Descriptive statistics to find the mean, median ,standard deviation etc. and based on that I have calculated the class intervals.

The output is as follows:

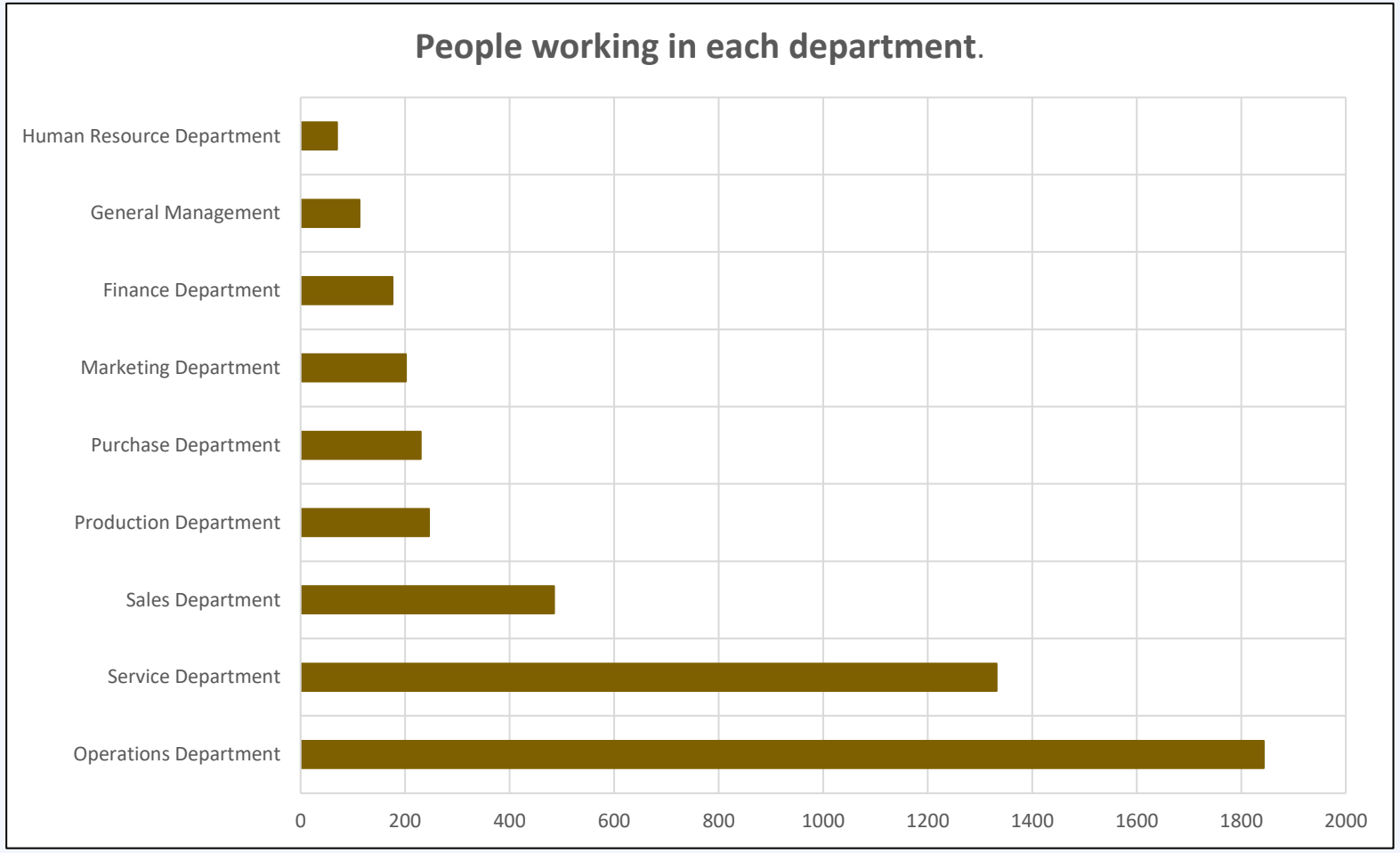
Descriptive statistics of Salary	
Mean	49983.02902
Standard Error	340.8317054
Median	49625
Mode	72843
Standard Deviation	28854.17689
Sample Variance	832563524
Kurtosis	2.610052003
Skewness	0.361578537
Range	399900
Minimum	100
Maximum	400000
Sum	358228369
Count	7167

Bins	15
Range/No.of Bins	26660
Class Intervals	Frequency
100	1
26760	1892
53420	3859
80080	5796
106740	7160
133400	7159
160060	7158
186720	7157
213380	7157
240040	7156
266700	7155
293360	7154
320020	7154
346680	7153
373340	7152
400000	7152
426660	7151

d. Charts and Plots: This is one of the most important part of analysis to visualize the data.

To visualize the data of portion of people working in each department, I have used a Pivot table and plotted a bar graph.

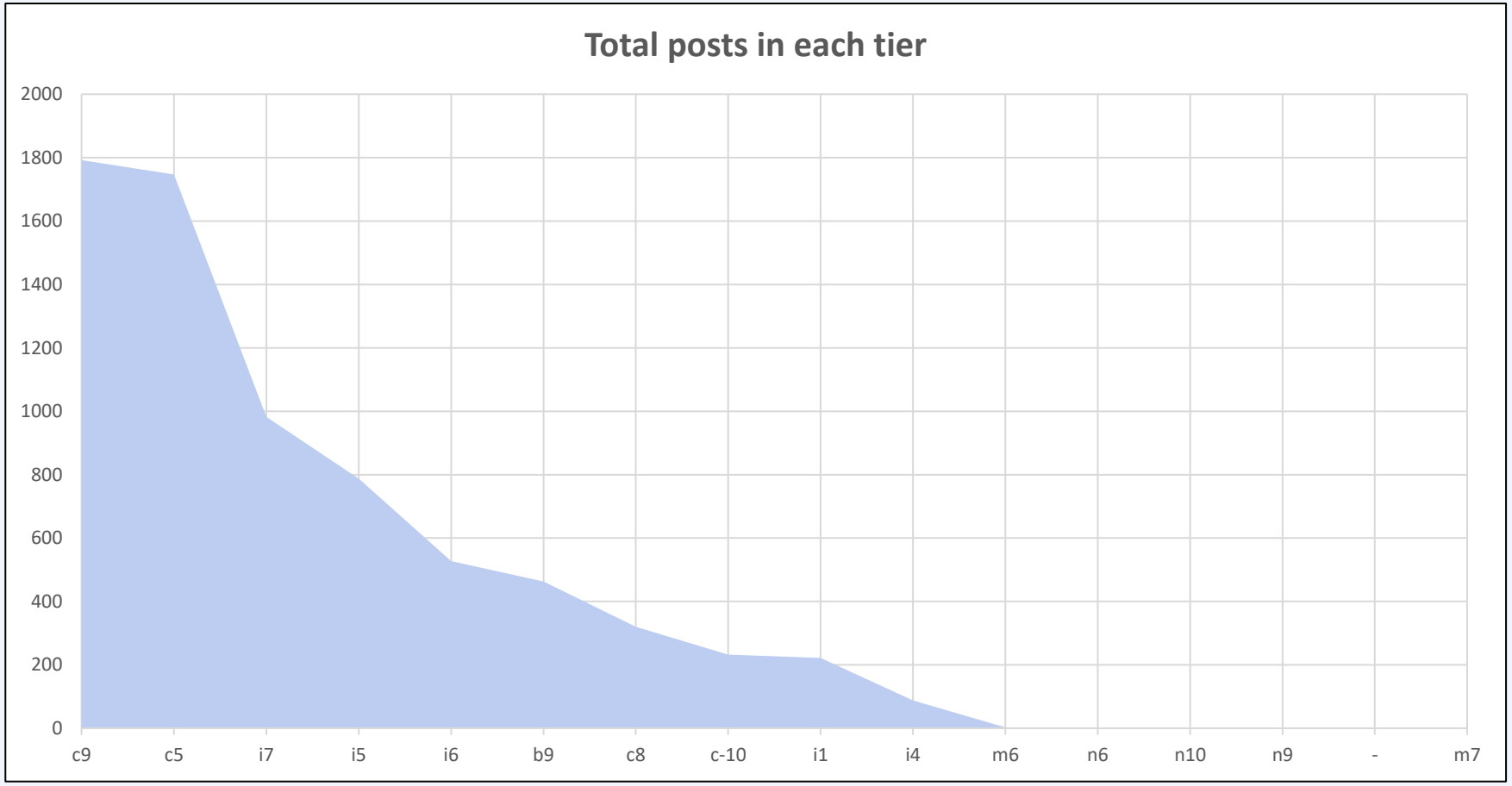
Status	Hired
Row Labels	**Count of application_id**
Operations Department	1843
Service Department	1332
Sales Department	485
Production Department	246
Purchase Department	230
Marketing Department	202
Finance Department	176
General Management	113
Human Resource Department	70
Grand Total	**4697**



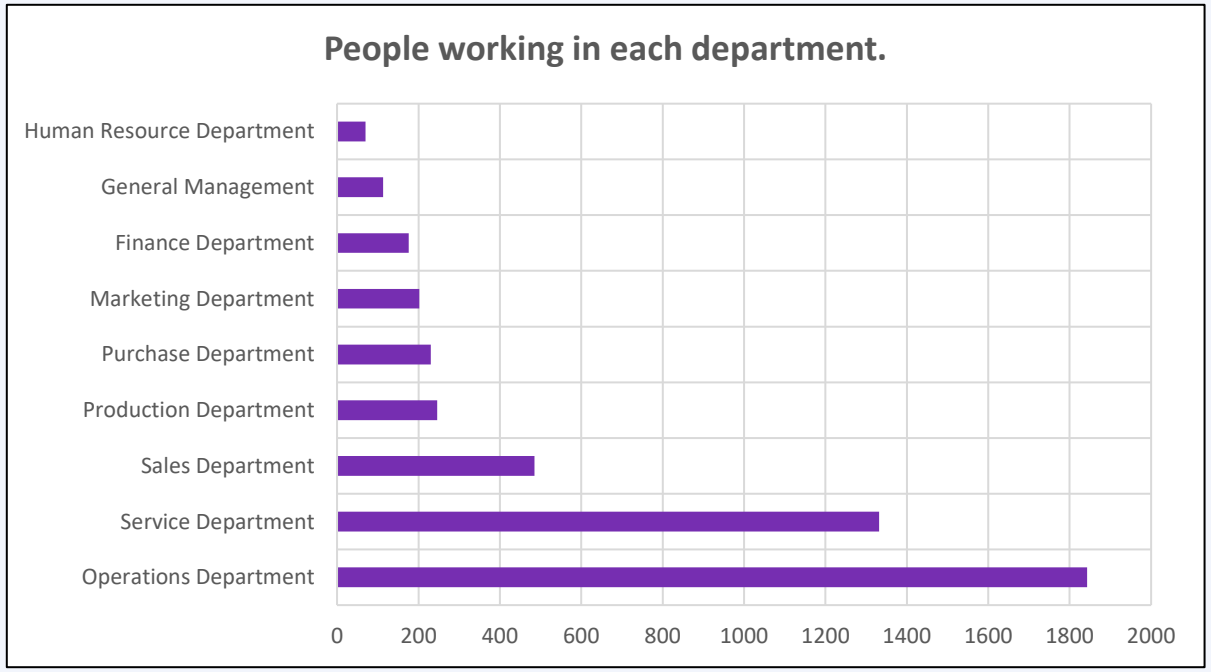
E. Charts: Use different charts and graphs to perform the task representing the data.

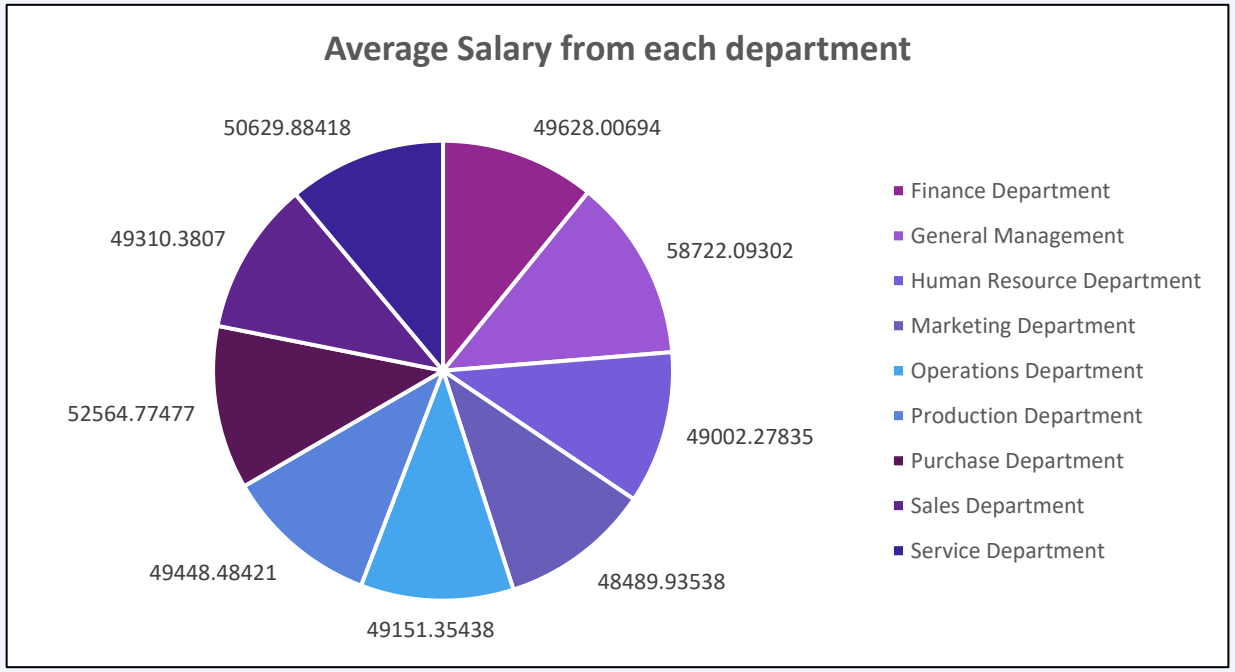
I have used pivot tables to analyze the posts in each tier.

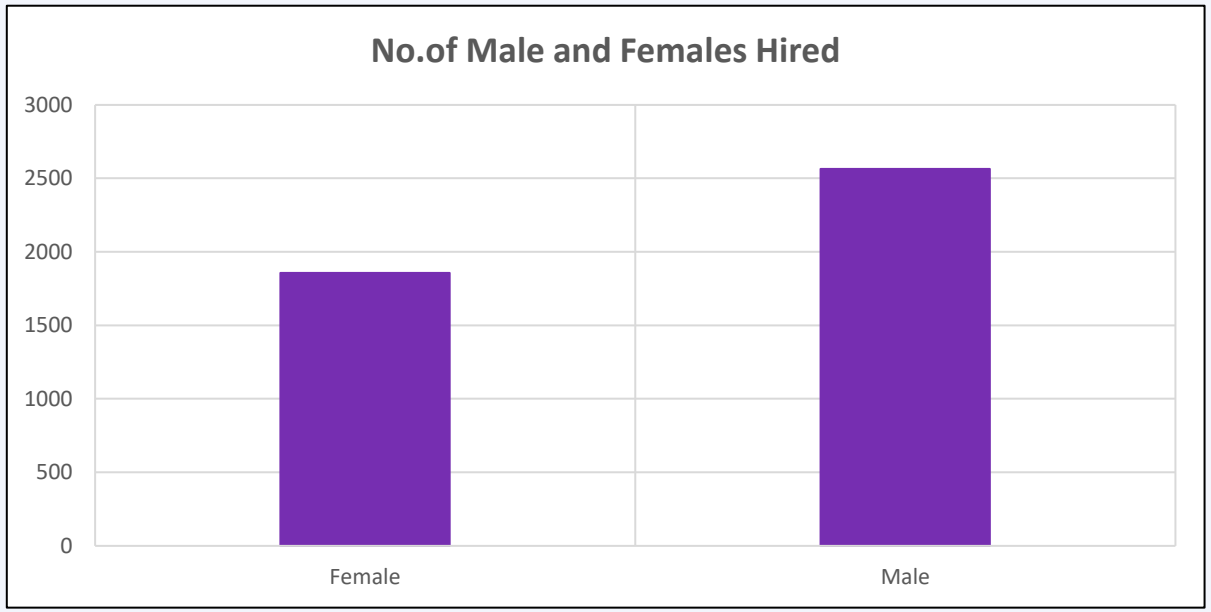
Row Labels	Count of application_id
c9	1792
c5	1747
i7	982
i5	787
i6	527
b9	463
c8	320
c-10	232
i1	222
i4	88
m6	3
n6	1
n10	1
n9	1
m7	1
Grand Total	**7168**

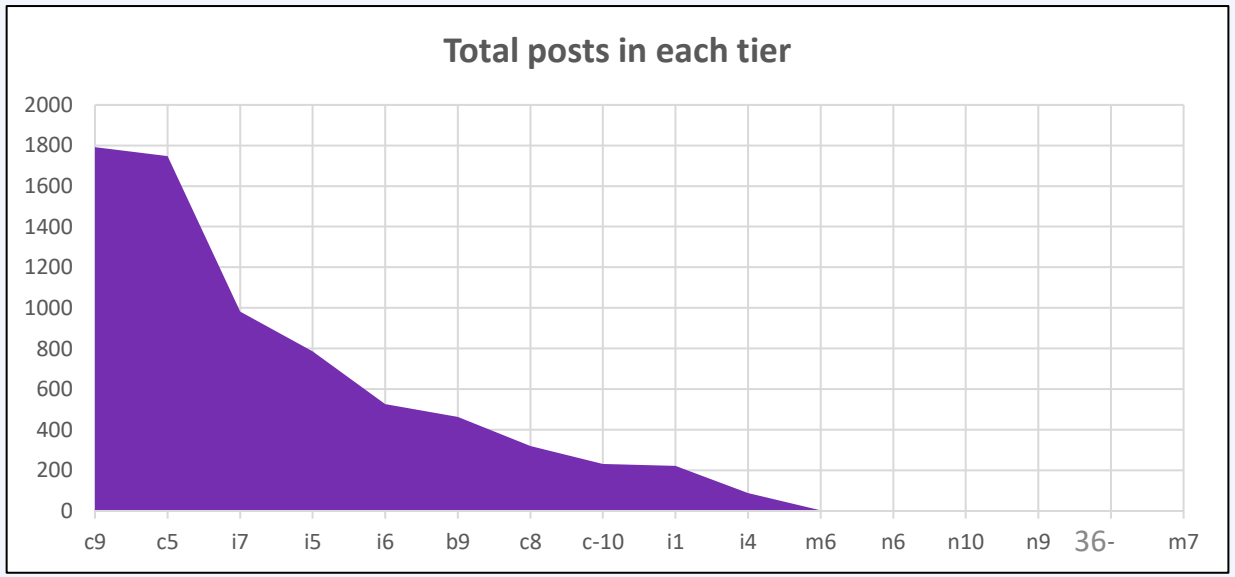


Data visualization(Dashboard)









Conclusion

- Operations department has many people working in it with the average salary of 49151.35438
- From the analysis we can see that Male employees are hired more than Female employees.
- Service department has the highest average salary of 50629.88418
- From the visualization of the analyzed data through graphs, we can see that the highest no. of posts are in “c9”.

Project 5 – IMDB Movie Analysis

Project Description

In this project, I'm provided with dataset having various columns of different IMDB Movies. I'm required to Frame the problem. For this task, I will need to define a problem that I want to shed some light on.

My approach for doing this project is by questioning. We can do this by asking 'What?' This is where you frame the problem i.e.

- What is the problem?
- What do you see happening?
- What is your hypothesis for the cause of the problem?
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Once we have defined a problem, clean the data as necessary, and use the Data Analysis skills to explore the data set and derive insights.

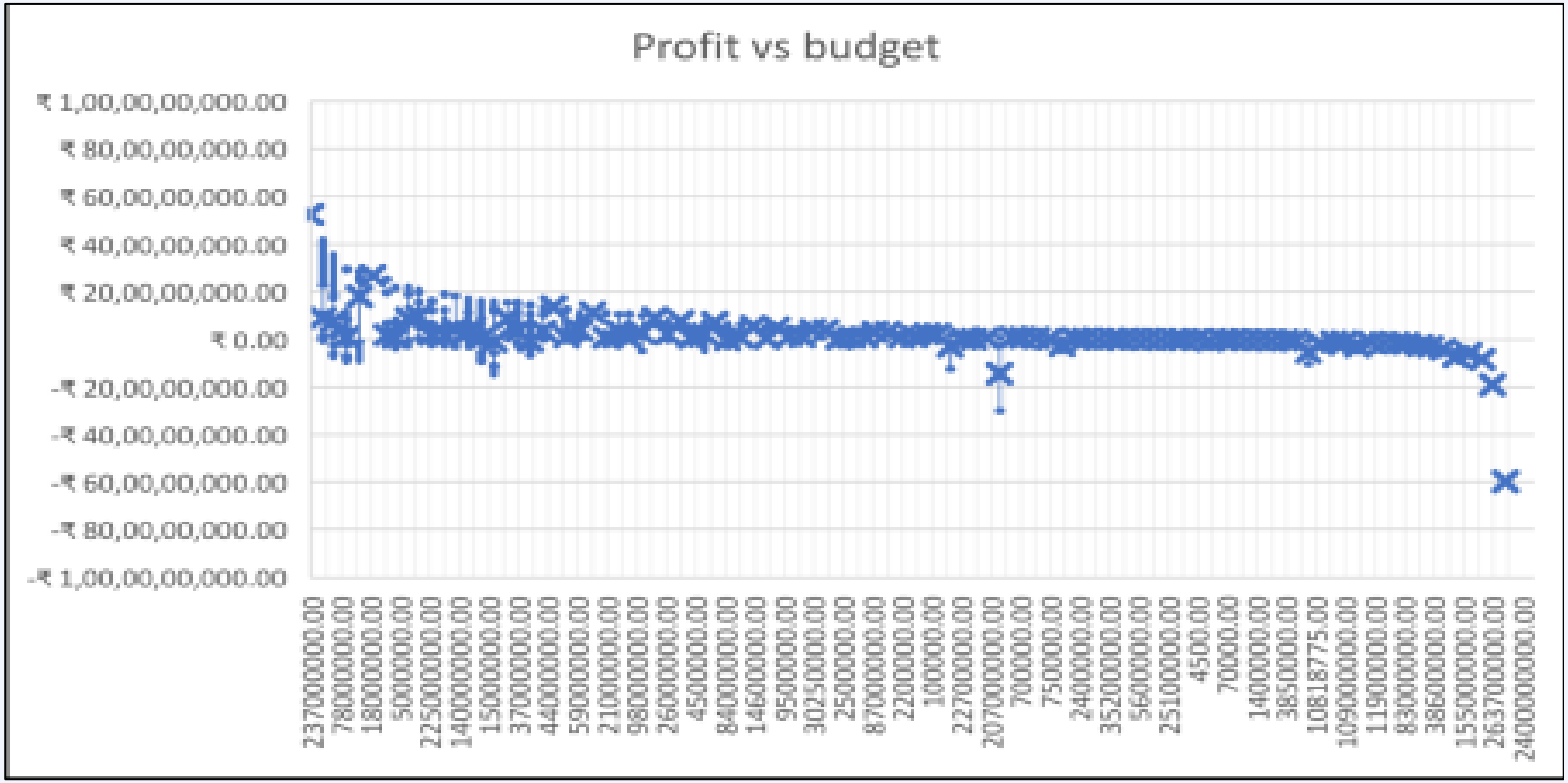
Tech stack used : Microsoft Excel

Insights

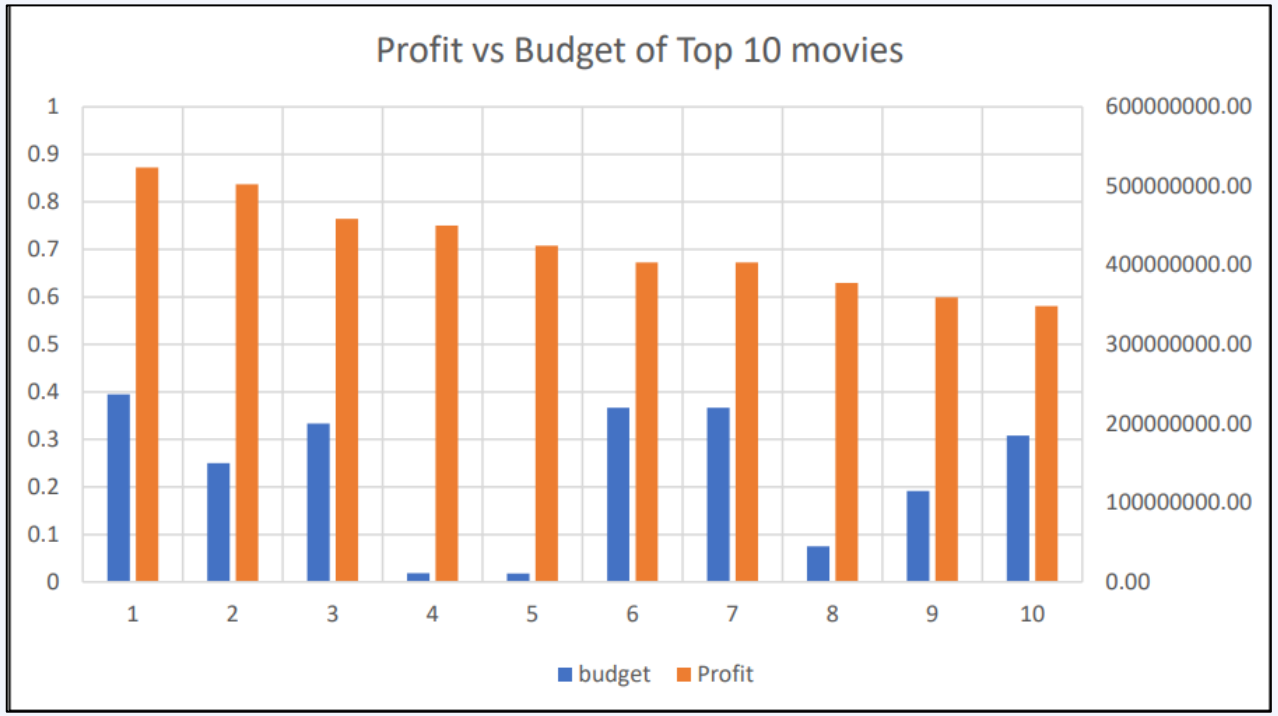
A. Cleaning the data:

The first thing I did to the data was to resize the columns. Then, there were many blank rows and duplicate values, so I have removed them. The movie_title had an extra character “Â “in each cell, so I have removed the extra character. There were many columns which has no use such as “movie_facebook_likes”, “director_facebook_likes”, “actor_1_facebook_likes”, “actor_2_facebook_likes”,” actor_3_facebook_likes”, “cast_total_facebook_likes”, “facenumber_in_poster” etc., so I have deleted the columns. After removing all the columns and rows, the total rows are reduced from 5044 to 3885.

B. Movies with highest profit:

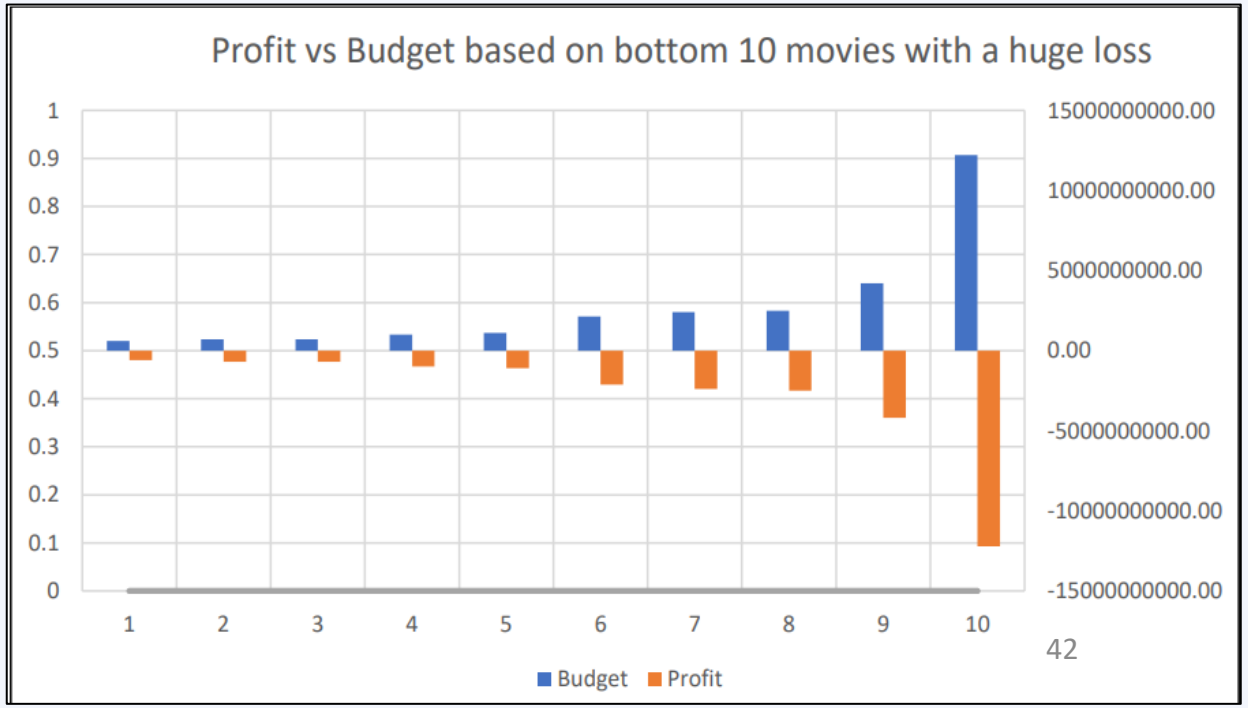


We can also see the difference in the graphs of the Top 10 movies with highest profit and the bottom 10 movies with huge loss.



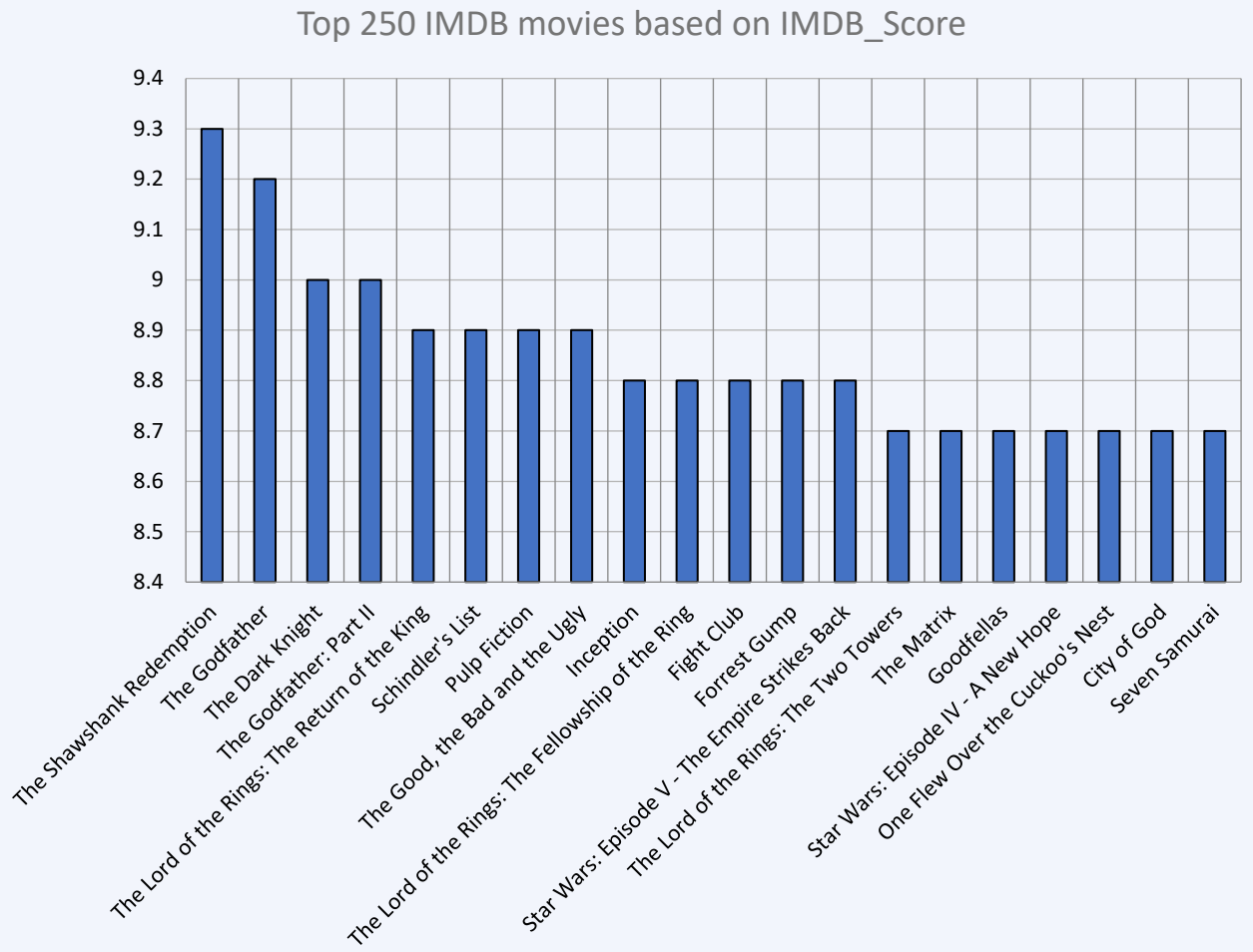
Here in x-axis 1 to 10 represents the Top 10 movies. The movies which had highest profit are Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope, E.T. the Extra-Terrestrial, The Avengers, The Lion King, Star Wars: Episode I - The Phantom Menace, The Dark Knight, The Hunger Games.

Here in x-axis 1 to 10 represents the bottom 10 movies The bottom 10 movies which had huge loss are The Host, Lady Vengeance, Fateless, Princess Mononoke, Steamboy, Akira, Godzilla 2000, Tango, Kabhi Alvida Naa Kehna, Kites didn't make any profit and their profit is less than their budget.



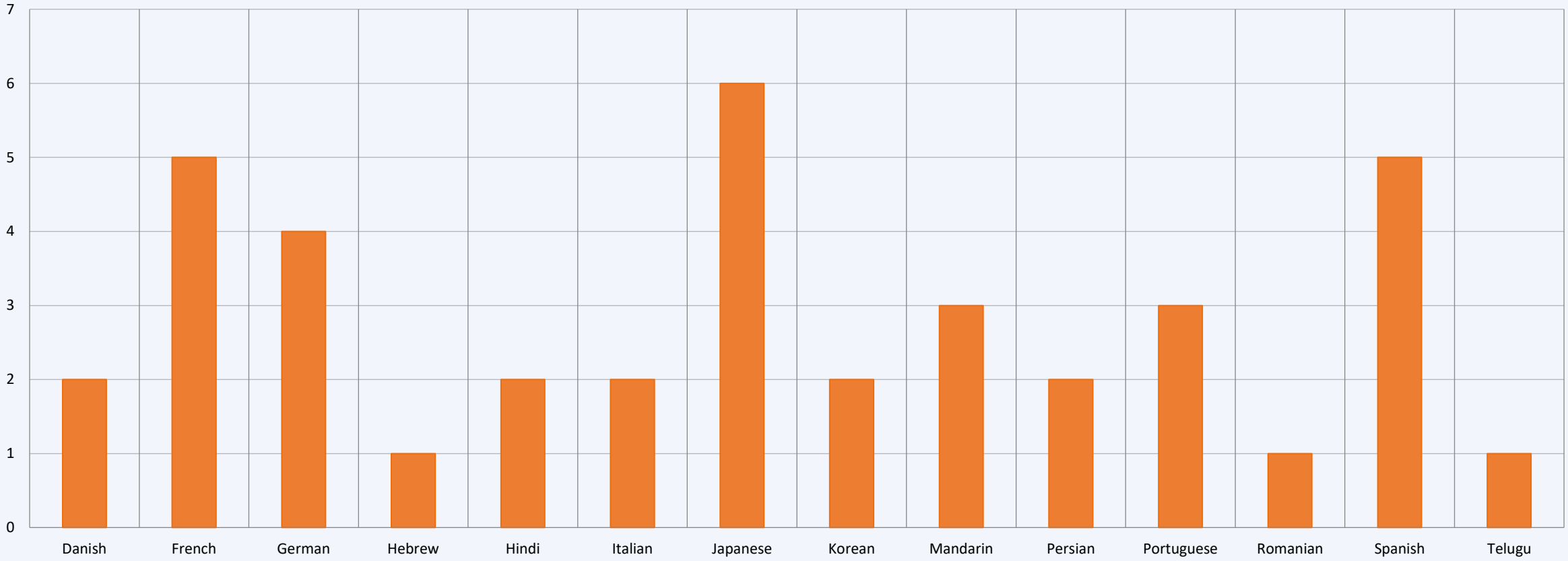
C. Top 250 movies with highest IMDb score

movie_title	imdb_sco	num_voted_use	language	IMDB_Top250
The Shawshank Redemption	9.3	1689764	English	1
The Godfather	9.2	1155770	English	2
The Dark Knight	9	1676169	English	3
The Godfather: Part II	9	790926	English	4
The Lord of the Rings: The Return of the King	8.9	1215718	English	5
Pulp Fiction	8.9	1324680	English	6
Schindler's List	8.9	865020	English	7
The Good, the Bad and the Ugly	8.9	503509	Italian	8
Forrest Gump	8.8	1251222	English	9
Star Wars: Episode V - The Empire Strikes Back	8.8	837759	English	10
The Lord of the Rings: The Fellowship of the Ring	8.8	1238746	English	11
Inception	8.8	1468200	English	12
Fight Club	8.8	1347461	English	13
Star Wars: Episode IV - A New Hope	8.7	911097	English	14
The Lord of the Rings: The Two Towers	8.7	1100446	English	15
The Matrix	8.7	1217752	English	16
One Flew Over the Cuckoo's Nest	8.7	680041	English	17
Goodfellas	8.7	728685	English	18
City of God	8.7	533200	Portuguese	19
Seven Samurai	8.7	229012	Japanese	20
Saving Private Ryan	8.6	881236	English	21
The Silence of the Lambs	8.6	887467	English	22
Se7en	8.6	1023511	English	23
Interstellar	8.6	928227	English	24
The Usual Suspects	8.6	740918	English	25



The highest IMDB score is 9.3 and it is for the movie “The Shawshank Redemption”. The average IMDB score of top 250 movies is nearly 8.6. And we can say that these movies are mostly viewed.

Top Foreign language film

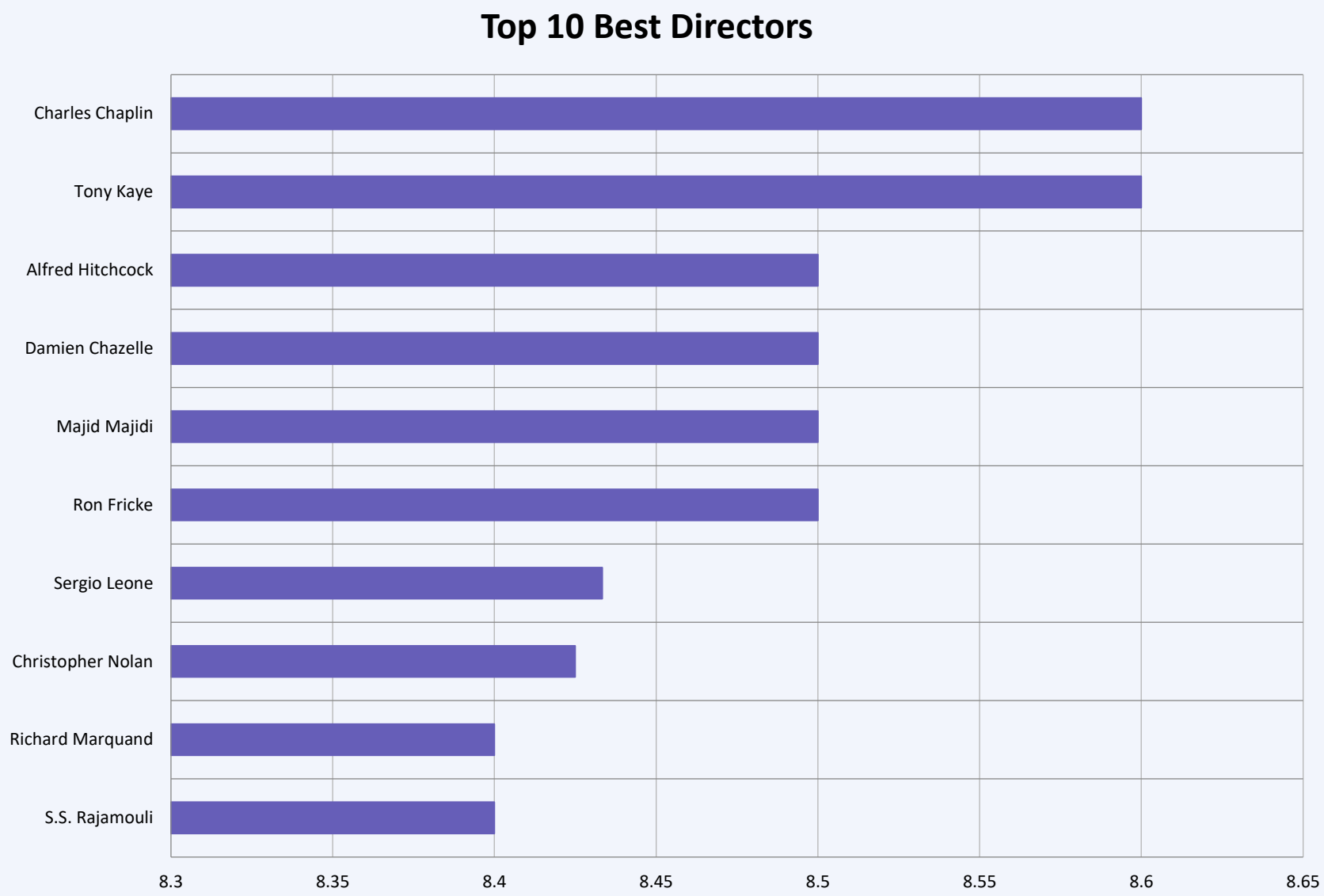


The majority of the movies which are viewed are mostly Japanese, Spanish and French besides English

D. Best Directors with highest IMDB score :

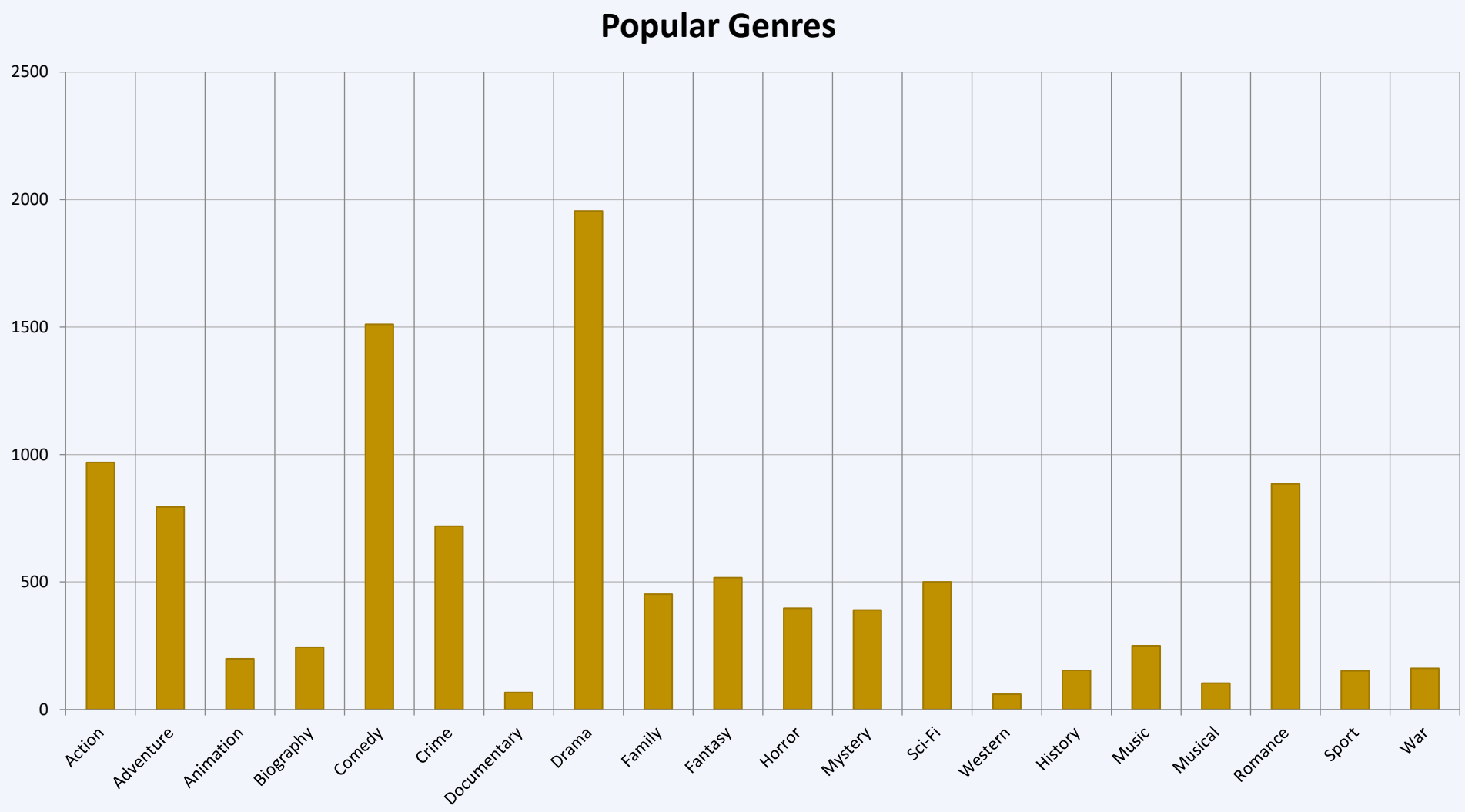
Top 10 Best Directors	
Row Labels	**Average of imdb_score**
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
S.S. Rajamouli	8.4

The highest IMDB score is for the director Charles Chaplin and Tony kaye (8.6).



E. Popular Genres:

Genre	Count
Action	969
Adventure	794
Animation	199
Biography	244
Comedy	1511
Crime	719
Documentary	67
Drama	1955
Family	452
Fantasy	517
Horror	397
Mystery	390
Sci-Fi	500
Western	60
History	154
Music	250
Musical	103
Romance	885
Sport	152
War	161



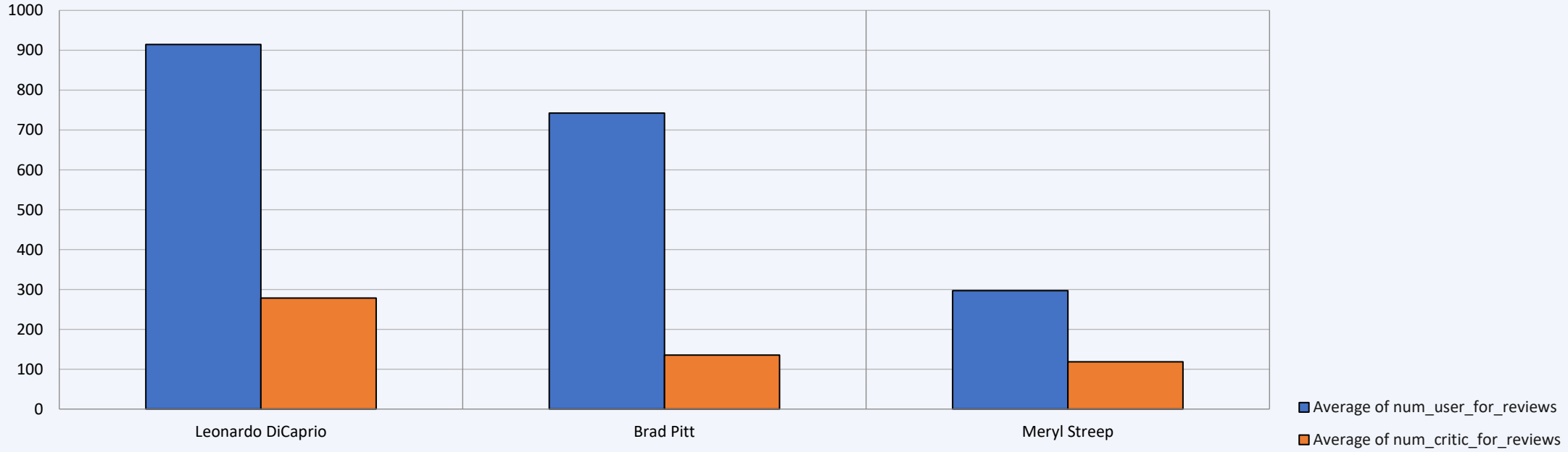
The most popular genres are Drama and Comedy.

F. Charts: Movies by Actors.

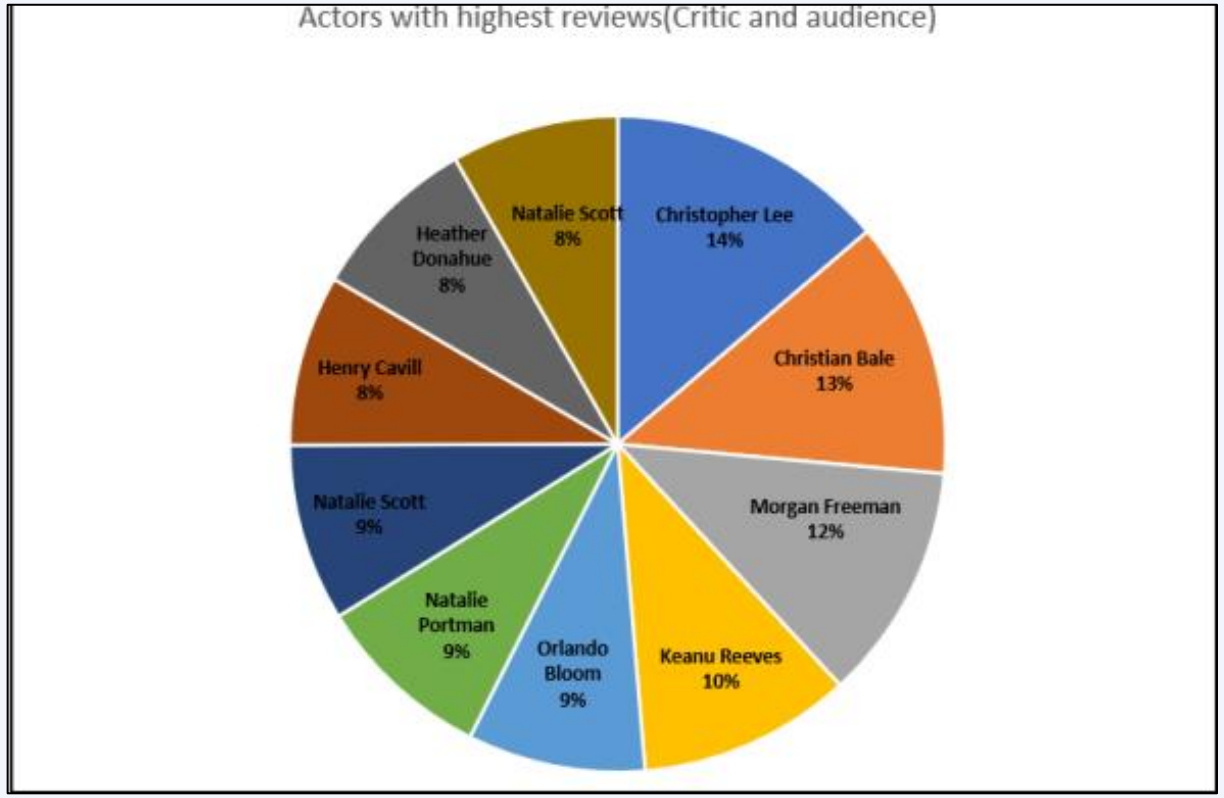
Meryl Streep	Leonardo DiCaprio	Brad Pitt
A Prairie Home Companion	Blood Diamond	Babel
Hope Springs	Body of Lies	By the Sea
It's Complicated	Catch Me If You Can	Fight Club
Julie & Julia	Django Unchained	Fury
Lions for Lambs	Gangs of New York	Interview with the Vampire: The Vampire Chronicles
One True Thing	Inception	Killing Them Softly
Out of Africa	J. Edgar	Mr. & Mrs. Smith
The Devil Wears Prada	Marvin's Room	Ocean's Eleven
The Hours	Revolutionary Road	Ocean's Twelve
The Iron Lady	Romeo + Juliet	Seven Years in Tibet
The River Wild	Shutter Island	Sinbad: Legend of the Seven Seas
	The Aviator	Spy Game
	The Beach	The Assassination of Jesse James by the Coward Robert Ford
	The Departed	The Curious Case of Benjamin Button
	The Great Gatsby	The Tree of Life
	The Man in the Iron Mask	Troy
	The Quick and the Dead	True Romance
	The Revenant	
	The Wolf of Wall Street	
	Titanic	

Row Labels	Average of num_user_for_reviews	Average of num_critic_for_reviews
Leonardo DiCaprio	914.476	278.524
Brad Pitt	742.353	135.118
Meryl Streep	297.182	118.727
Grand Total	**716.1836735**	**192.8979592**

Leonardo DiCaprio is the audience and critic favorite actor

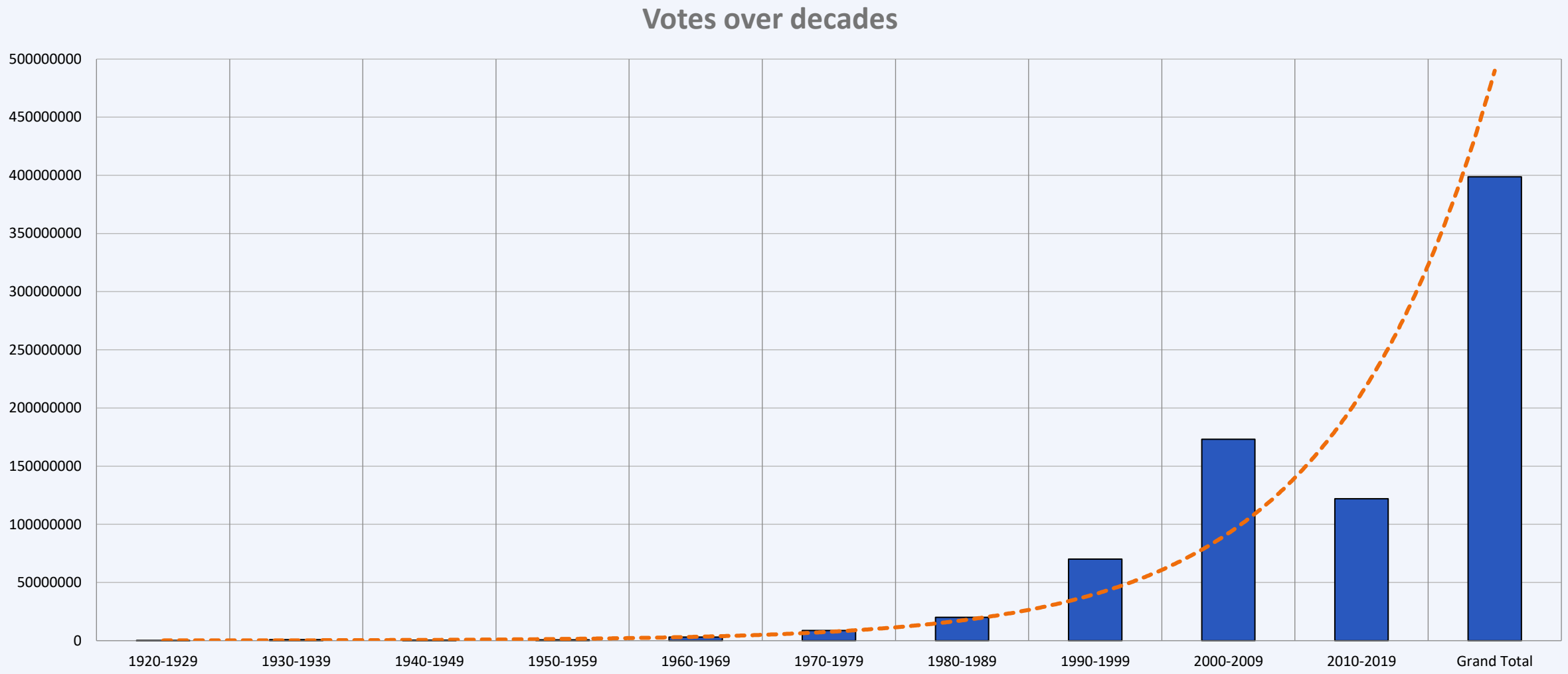


actor_1_name	num_critic_for_reviews	num_user_for_reviews	Mean
Christopher Lee	673	5060	2866.5
Christian Bale	602	4667	2634.5
Morgan Freeman	723	4144	2433.5
Keanu Reeves	733	3646	2189.5
Orlando Bloom	462	3189	1825.5
Natalie Portman	133	3516	1824.5
Natalie Scott	43	3597	1820
Henry Cavill	493	3018	1755.5
Heather Donahue	82	3400	1741
Natalie Portman	143	3286	1714.5



These are the top 10 highest reviewed actors base on the mean on num_critc_for_reviews and num_user_for_reviews.

Votes over decades



Conclusion

From this analysis, we can say that,

- The movies which had highest profit are Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope, E.T. the Extra-Terrestrial.
- The top 10 best directors are Charles Chaplin, Tony Kaye, Alfred Hitchcock, Damien Chazelle, Majid Majidi, Ron Fricke, Sergio Leone, Christopher Nolan, Richard Marquand, S.S. Rajamouli.
- The most popular genres are Drama and Comedy.
- Leonardo DiCaprio turns out to be the audience and critic favorite actor.
- The votes over decades are exponentially increasing indicating that the audience are keeping their keen interests in movies.

Project 6 - Bank Loan Case Study

Project Description

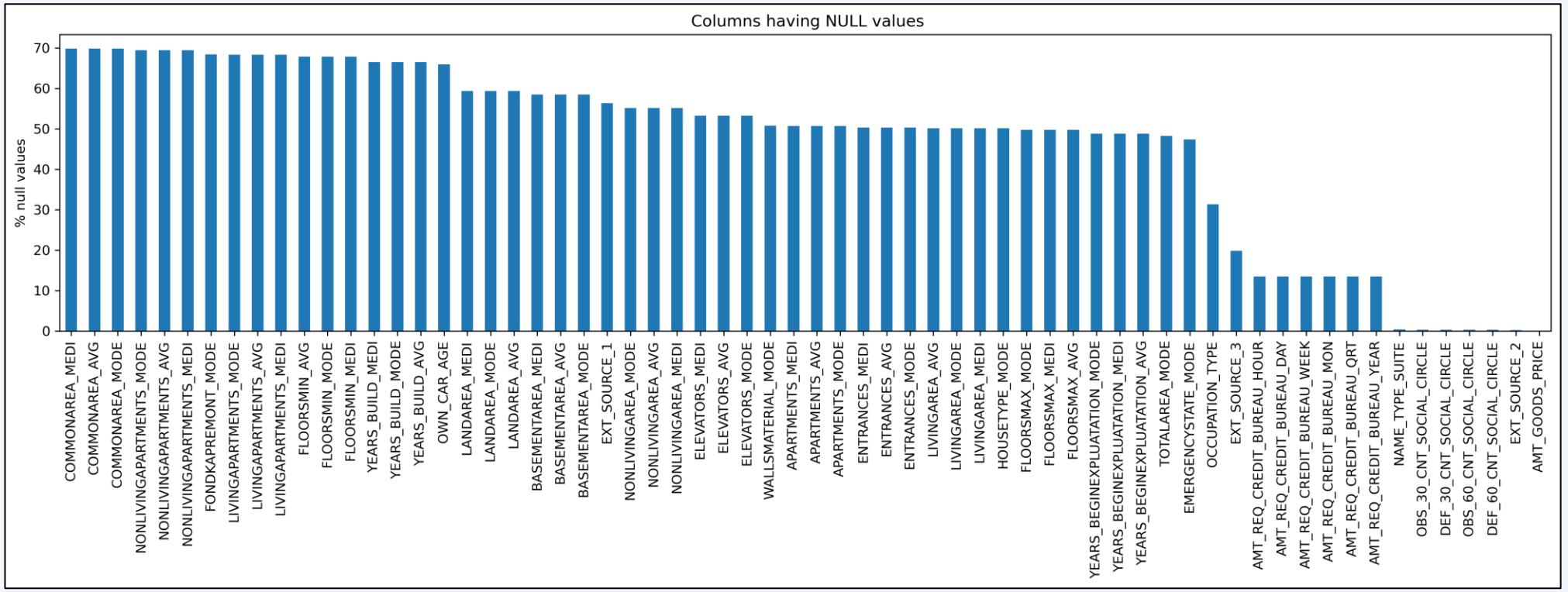
- This project aims to give an idea of applying EDA in a real business scenario. This project helps us to develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- The project aims to understand the driving factors (or driver variables) behind the loan default. i.e the variables which are strong in loan default.
- I have been given 2 datasets for which analysis needs to be done.
    - `**application_data.csv**` contains all the information of the client at the time of application.
      The data is about whether a client has payment difficulties.
    - `**previous_application.csv**` contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Tech stack used : Jupyter Notebook(python)

Insights

To make analysis, I have plotted the percentage of null values in each column, so that I can get a clear idea on how to proceed with the data cleaning.

```
In [14]: # graphical representation of columns having % null values

plt.figure(figsize= (20,4),dpi=300)
col_null_percentage(df_App)[col_null_percentage(df_App)>0].plot(kind = 'bar')
plt.title (' Columns having NULL values')
plt.ylabel('% null values')
plt.show()
```



```
In [15]: # Extracting the columns with null values more than 40%

Null_column_40 = col_null_percentage(df_App)[col_null_percentage(df_App)>40]
print("Number of columns having null value more than 40% :", len(Null_column_40.index))
print(Null_column_40)

Number of columns having null value more than 40% : 49
COMMONAREA_MEDI                69.87
COMMONAREA_AVG                 69.87
COMMONAREA_MODE                69.87
NONLIVINGAPARTMENTS_MODE       69.43
NONLIVINGAPARTMENTS_AVG        69.43
NONLIVINGAPARTMENTS_MEDI       69.43
FONDKAPREMONT_MODE             68.39
LIVINGAPARTMENTS_MODE          68.35
LIVINGAPARTMENTS_AVG           68.35
LIVINGAPARTMENTS_MEDI          68.35
FLOORSMIN_AVG                  67.85
FLOORSMIN_MODE                 67.85
FLOORSMIN_MEDI                 67.85
YEARS_BUILD_MEDI               66.50
```

After careful observation , I have extracted the columns in which the percentage of null values is greater than 40%. So we can see from the output that there are 49 columns which has null value percentage greater than 40%

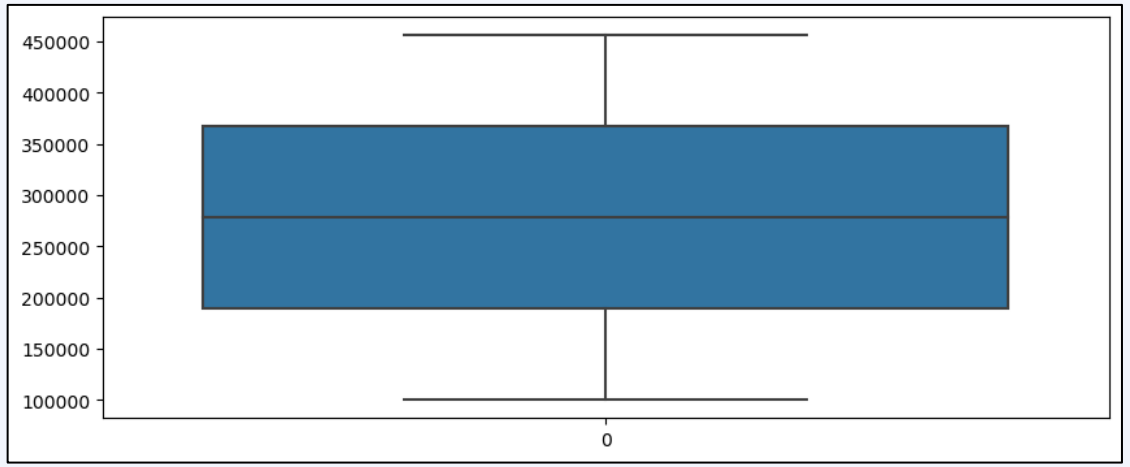
```
In [17]: # Dropping the columns with null values more than 40%

df_App.drop(columns = Null_column_40.index, inplace = True)

In [18]: df_App.shape

Out[18]: (307511, 73)
```

I have used drop() function to drop the columns which has null value percentage more than 40%.

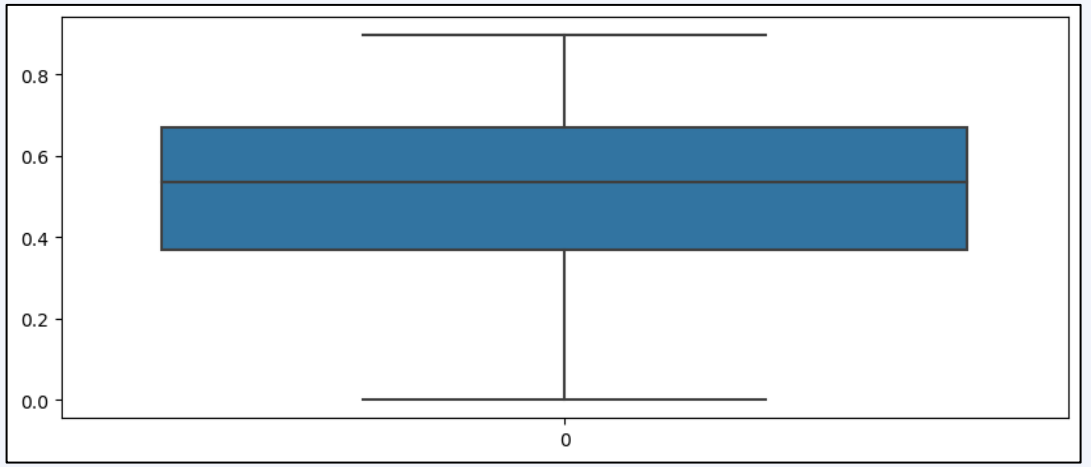
```
In [19]: # Extracting the columns having <20% null values

Null_column_20 = col_null_percentage(df_App)[col_null_percentage(df_App)<20]
print("Number of columns having null value less than 20% :", len(Null_column_20.index))
print(Null_column_20)

Number of columns having null value less than 20% : 72
EXT_SOURCE_3                  19.83
AMT_REQ_CREDIT_BUREAU_YEAR    13.50
AMT_REQ_CREDIT_BUREAU_QRT     13.50
AMT_REQ_CREDIT_BUREAU_MON     13.50
AMT_REQ_CREDIT_BUREAU_WEEK    13.50
AMT_REQ_CREDIT_BUREAU_DAY     13.50
AMT_REQ_CREDIT_BUREAU_HOUR    13.50
NAME_TYPE_SUITE                0.42
OBS_30_CNT_SOCIAL_CIRCLE       0.33
DEF_30_CNT_SOCIAL_CIRCLE       0.33
OBS_60_CNT_SOCIAL_CIRCLE       0.33
DEF_60_CNT_SOCIAL_CIRCLE       0.33
EXT_SOURCE_2                   0.21
```

I have extracted the columns in which the percentage of null values is less than 20%. As we can see, there are 72 columns in which the null value percentage is less than 20%.

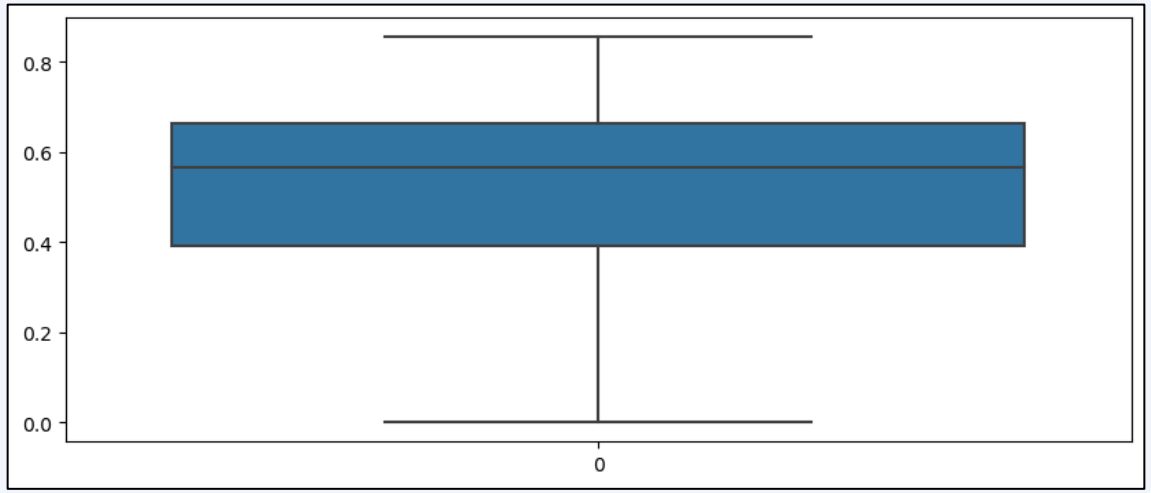
In [26]:
```
# Box plot for SK_ID_CURR column to check the outliers
plt.figure(figsize=(10,4))
sns.boxplot(df_App['SK_ID_CURR'])
plt.show()
```



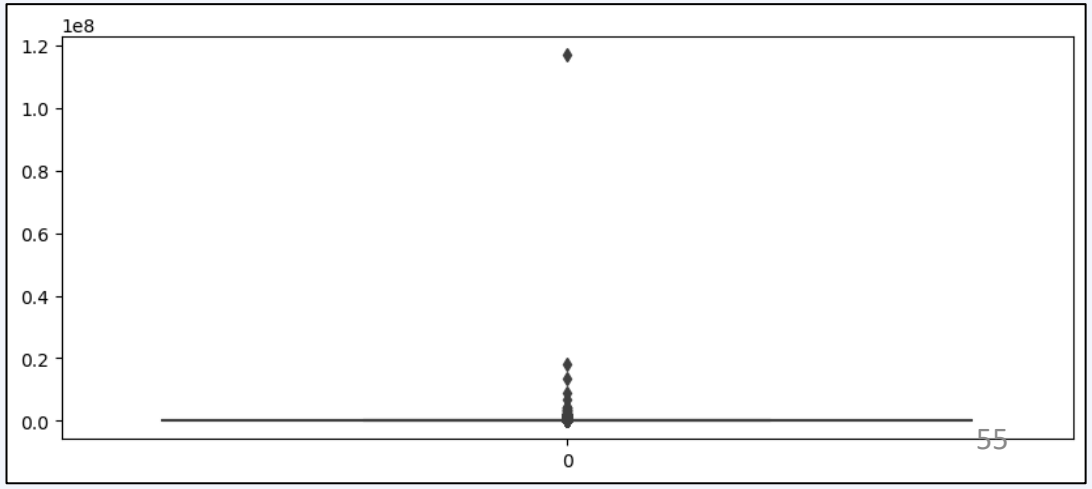
In [27]:
```
# Box plot for EXT_SOURCE_3 column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_App['EXT_SOURCE_3'])
plt.show()
```



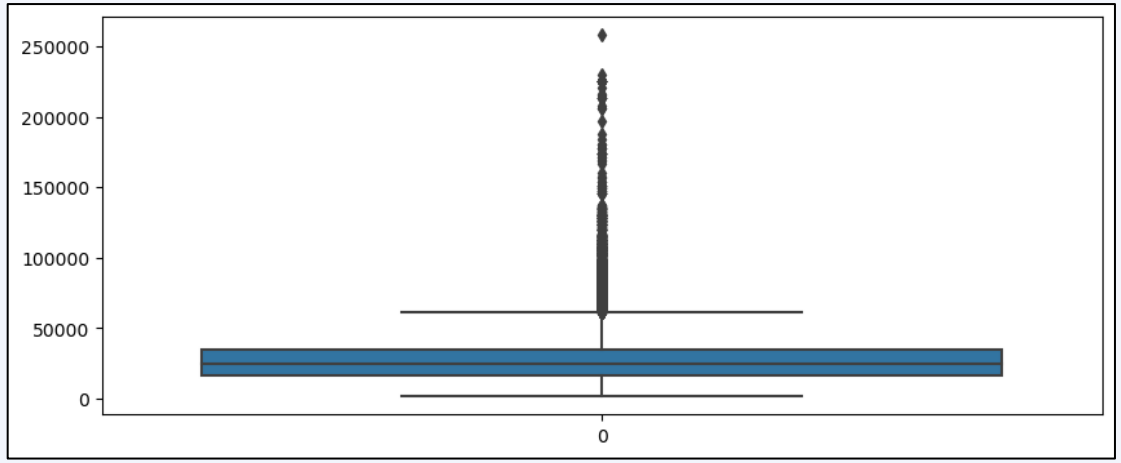
In [28]:
```
# Box plot for EXT_SOURCE_2 column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_App['EXT_SOURCE_2'])
plt.show()
```



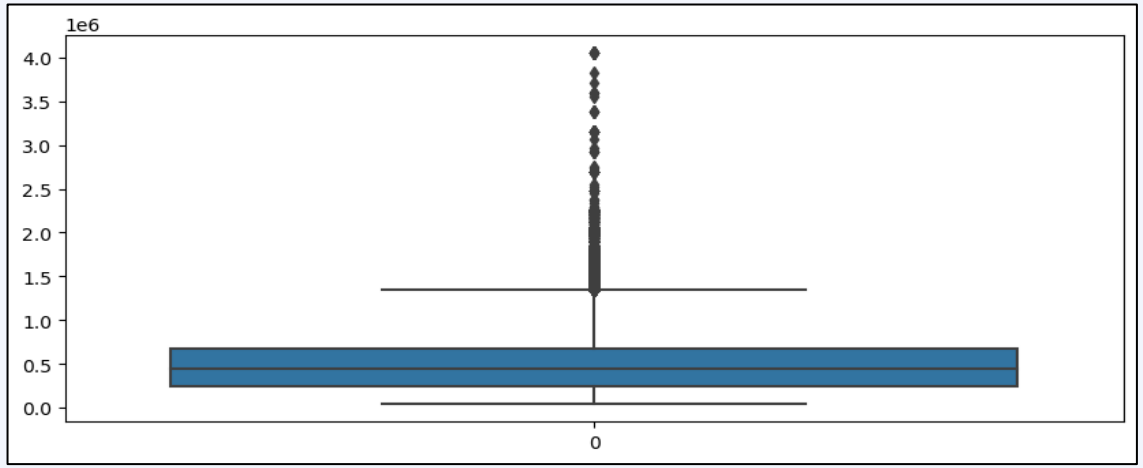
In [29]:
```
# Box plot for AMT_INCOME_TOTAL column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_App['AMT_INCOME_TOTAL'])
plt.show()
```



```
In [30]: # Box plot for AMT_ANNUITY column to check the outliers

         plt.figure(figsize=(10,4))
         sns.boxplot(df_App['AMT_ANNUITY'])
         plt.show()
```




```
In [31]: # Box plot for AMT_GOODS_PRICE column to check the outliers

         plt.figure(figsize=(10,4))
         sns.boxplot(df_App['AMT_GOODS_PRICE'])
         plt.show()
```



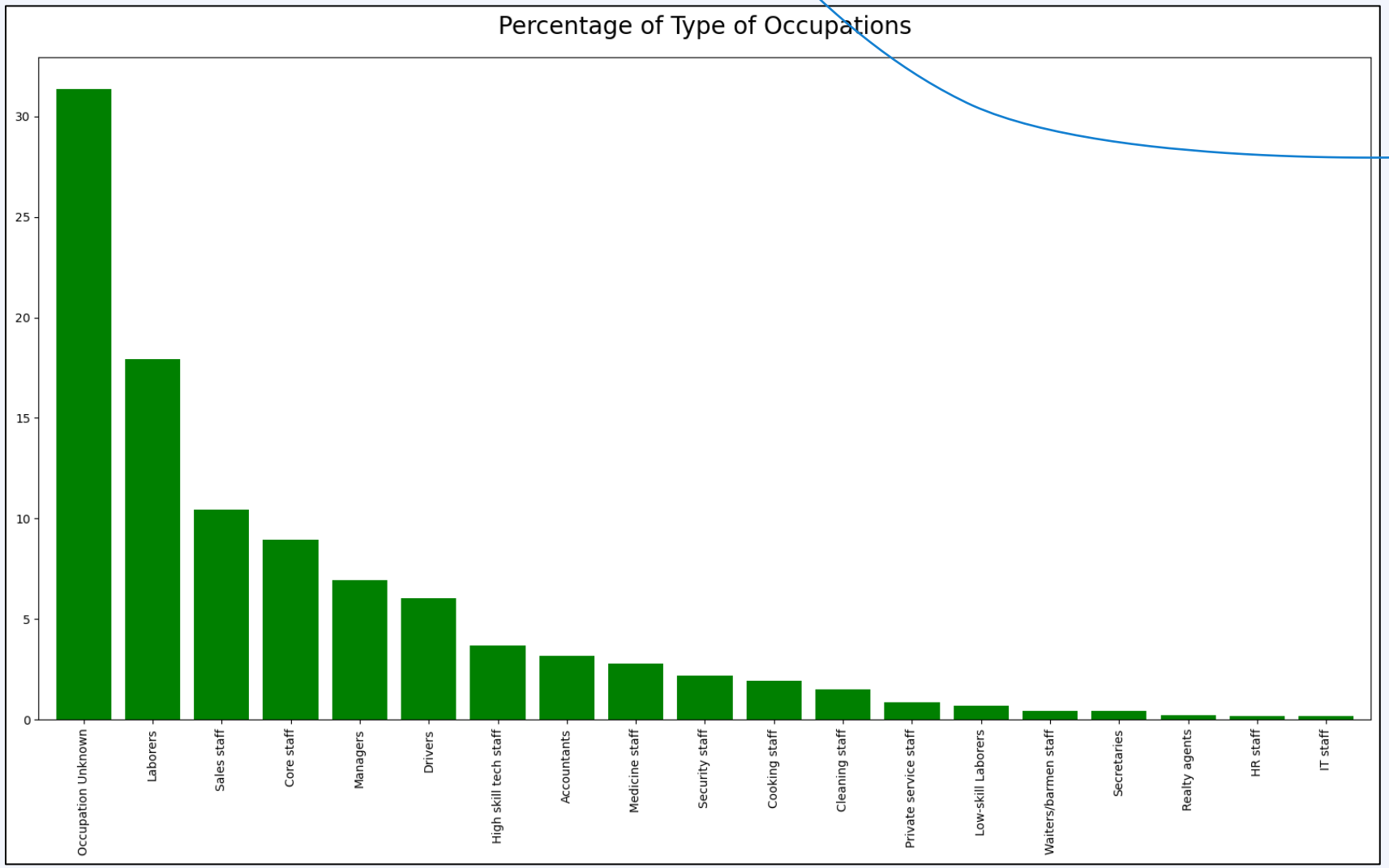
From the box plots, we can say that,
For the columns “EXT_SOURCE_3’, ‘EXT_SOURCE_2’, ‘SK_ID_CURR’, there are no outliers present. It can be assumed that the missing values can be replaced with median values.

As for the columns “AMT_GOODS_PRICE’, ‘AMT_INCOME_TOTAL’, ‘AMT_ANNUITY’, there are significant number of outliers present in the data. So, the missing values should be replaced with median values.

```
In [40]: df_App['OCCUPATION_TYPE'].value_counts(normalize = True)*100
```

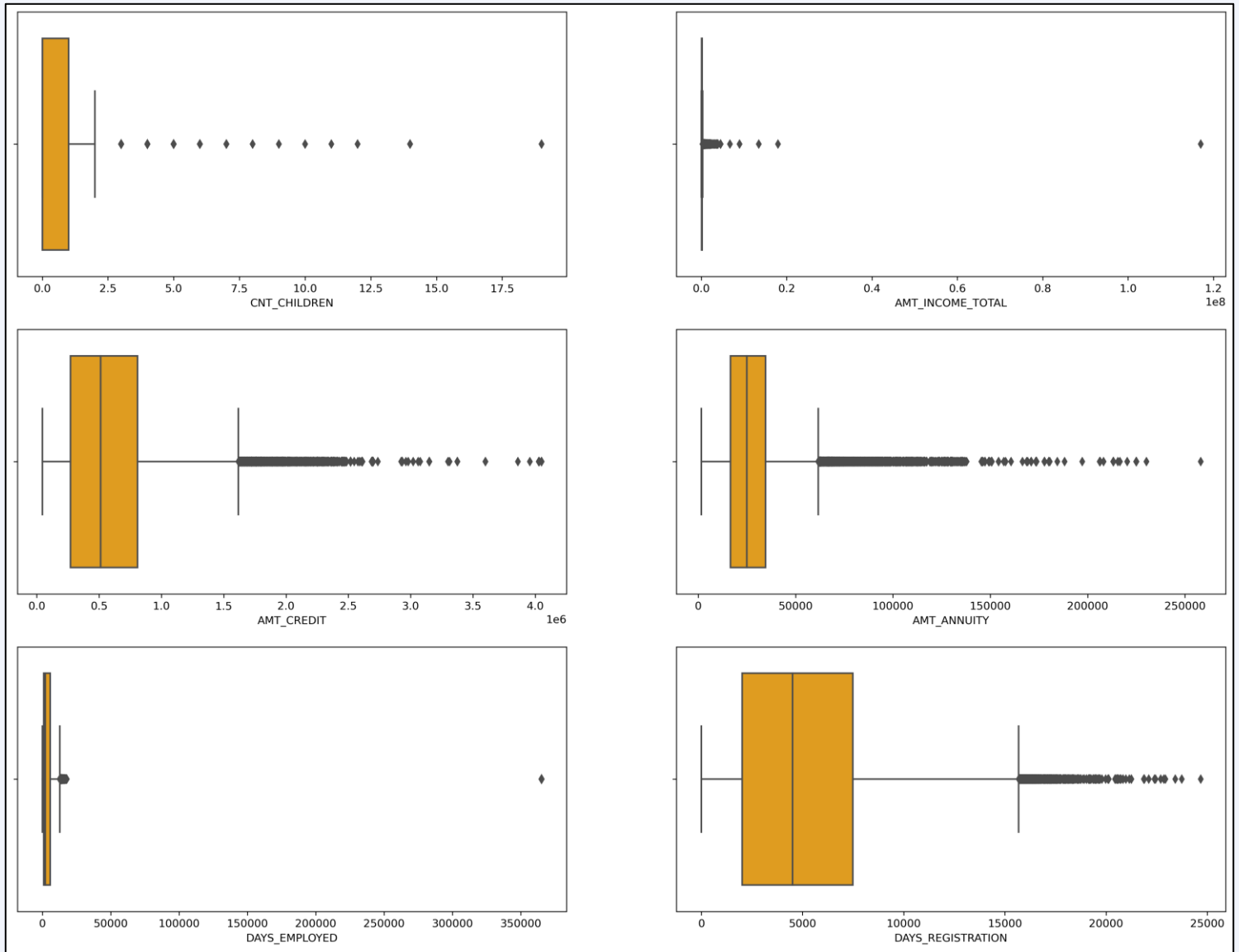
```
Out[40]: Occupation Unknown      31.345545
         Laborers                17.946025
         Sales staff             10.439301
         Core staff               8.965533
         Managers                 6.949670
         Drivers                  6.049540
         High skill tech staff    3.700681
         Accountants              3.191105
         Medicine staff           2.776161
         Security staff           2.185613
```

```
In [41]: #Analyzing the occupation column by ploting the graph

plt.figure(figsize = [20,10])
(df_App["OCCUPATION_TYPE"].value_counts(normalize=True)*100).plot.bar(color= "green",width = .8)
plt.title("Percentage of Type of Occupations", fontdict={"fontsize":20}, pad =20)
plt.show()
```



For the OCCUPATION_TYPE column, I have replaced null values with Occupation unknown and plotted a bar graph to get the clear idea on the number of occupations.

```
In [46]: # Box plot for selected columns

features = ['CNT_CHILDREN', 'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','DAYS_EMPLOYED', 'DAYS_REGISTRATION']

plt.figure(figsize = (20, 15), dpi=300)
for i in enumerate(features):
    plt.subplot(3, 2, i[0]+1)
    sns.boxplot(x = i[1], data = df_App, color = "Orange")
plt.show()
```



I have plotted box plots for the selected columns to find out the outliers.

```
In [52]: # Dividing the dataset into two dataset of  target=1(client with payment difficulties) and target=0(all other)

target0=df_App.loc[df_App["TARGET"]==0]
target1=df_App.loc[df_App["TARGET"]==1]
```

```
In [53]: # insights from number of target values

percentage_defaulters= round(100*len(target1)/(len(target0)+len(target1)),3)

percentage_nondefaulters=round(100*len(target0)/(len(target0)+len(target1)),3)

print('Count of target0:', len(target0))
print('Count of target1:', len(target1))


print('People who paid their loan are: ', percentage_nondefaulters, '%' )
print('People who did not pay their loan are: ', percentage_defaulters, '%' )
```

```
Count of target0: 282686
Count of target1: 24825
People who paid their loan[in percentage] are:  91.927 %
People who did not paid their loan[in percentage] are:  8.073 %
```

Dividing the dataset into two based on the target column for further analysis and found out the percentage of people who paid the loan and the people who didn't pay the loan.

```
In [54]: # Calculating Imbalance percentage, since the majority is target0 compared to target1

imbalance_percentage = round(len(target0)/len(target1),3)

print('Imbalance Percentage:', imbalance_percentage)
```

```
Imbalance Percentage: 11.387
```

Found out the imbalance percentage using the described formula. The Imbalance percentage is 11.387.

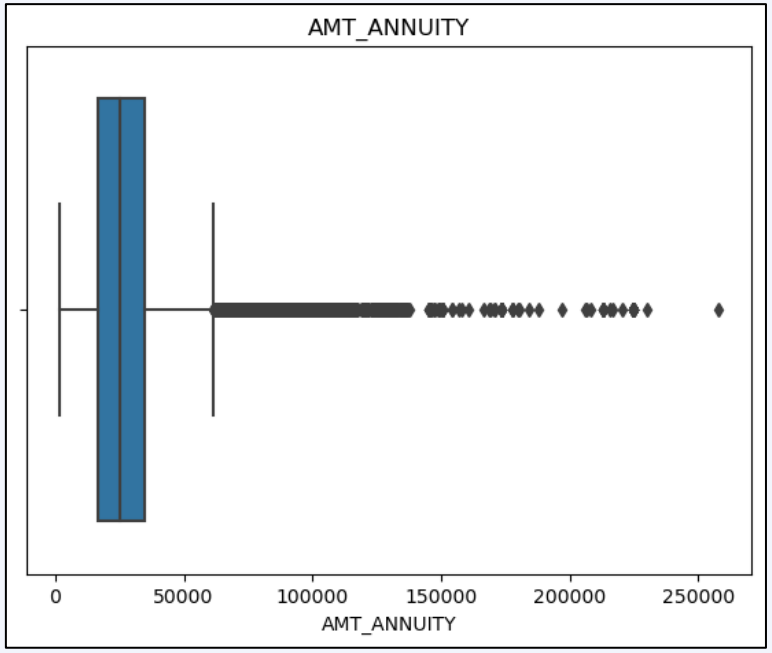
Univariate analysis (application_data.csv)

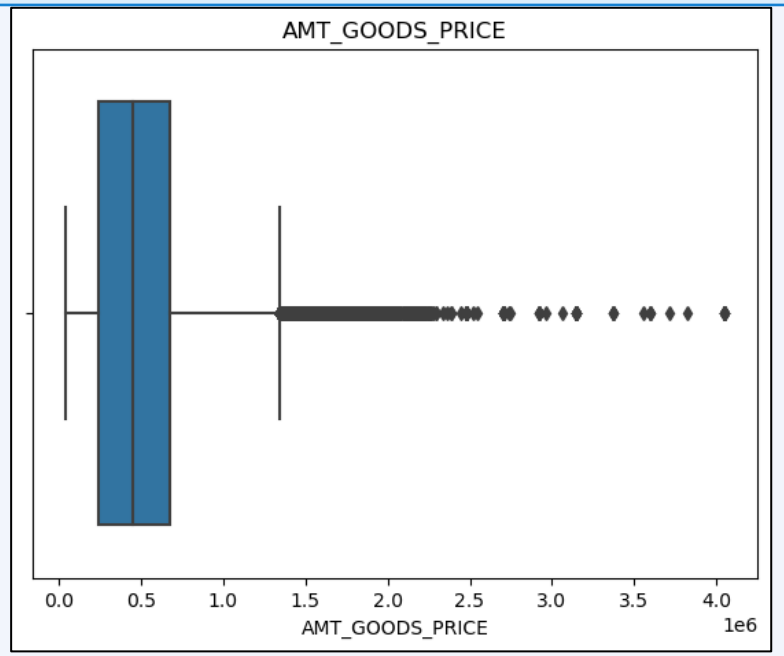
Univariate analysis is done for the data that has only one kind of variable. The main purpose of univariate analysis is to describe the data and find the patterns exists within it. I have categorized the data as Numerical and Categorical and added the respective columns in each category to perform the univariate analysis.

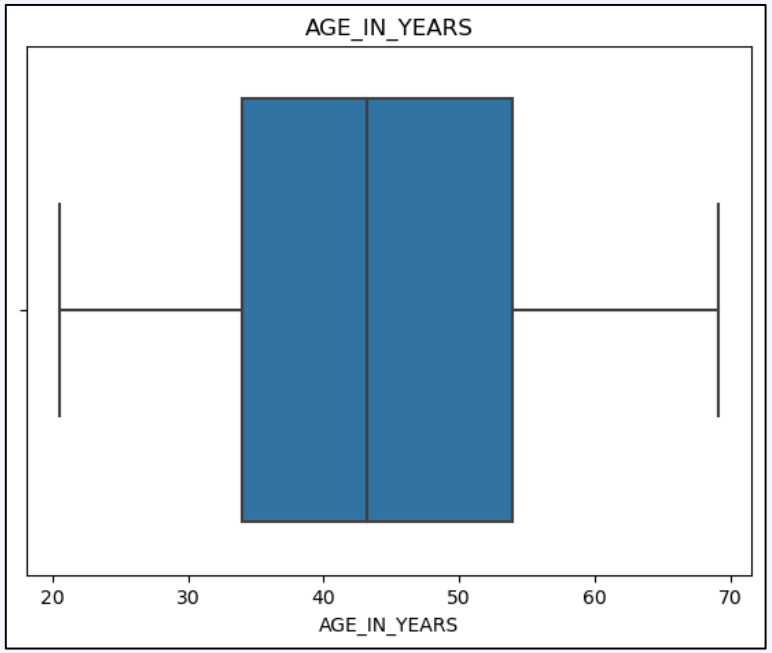
In [55]:
```
Numerical_data_columns = ['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'AGE_IN_YEARS', 'EMPLOYMENT_YEARS','AMT_CREDIT', 'CNT_FAM_MEMBERS']
Categorical_data_columns = ['NAME_TYPE_SUITE','NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE'
```

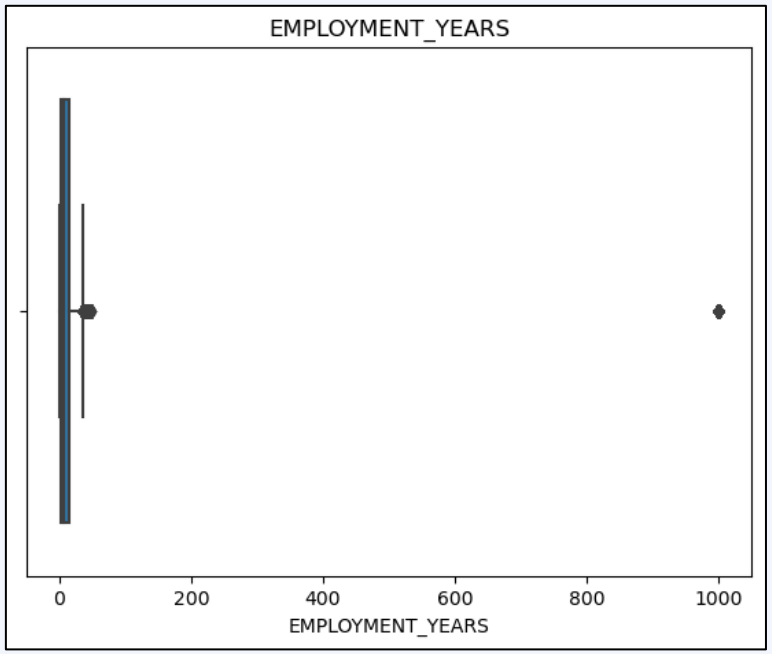
In [56]:
```
def Univariate_Analysis_Num(df,col):
    plt.figure(figsize = (15,5))
    plt.subplot(1,2,1)
    sns.boxplot(data= df,x=col,orient='h')
    plt.title(col)
    plt.show()

def Univariate_Analysis_Cat(df,col):
    plt.figure(figsize =[25,10])
    df[col].value_counts().plot.barh(width =0.6)
    plt.title(col)
    plt.show()
```

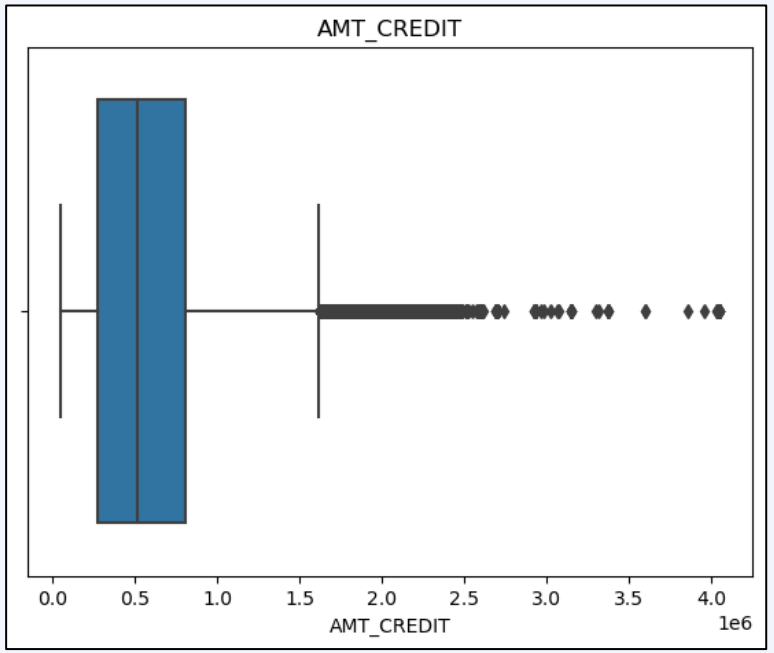
Univariate analysis for numerical data

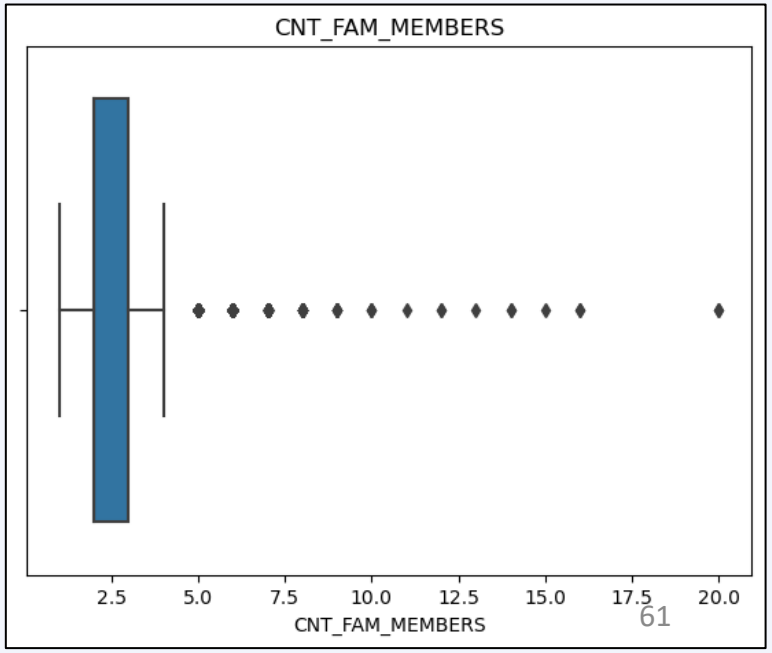




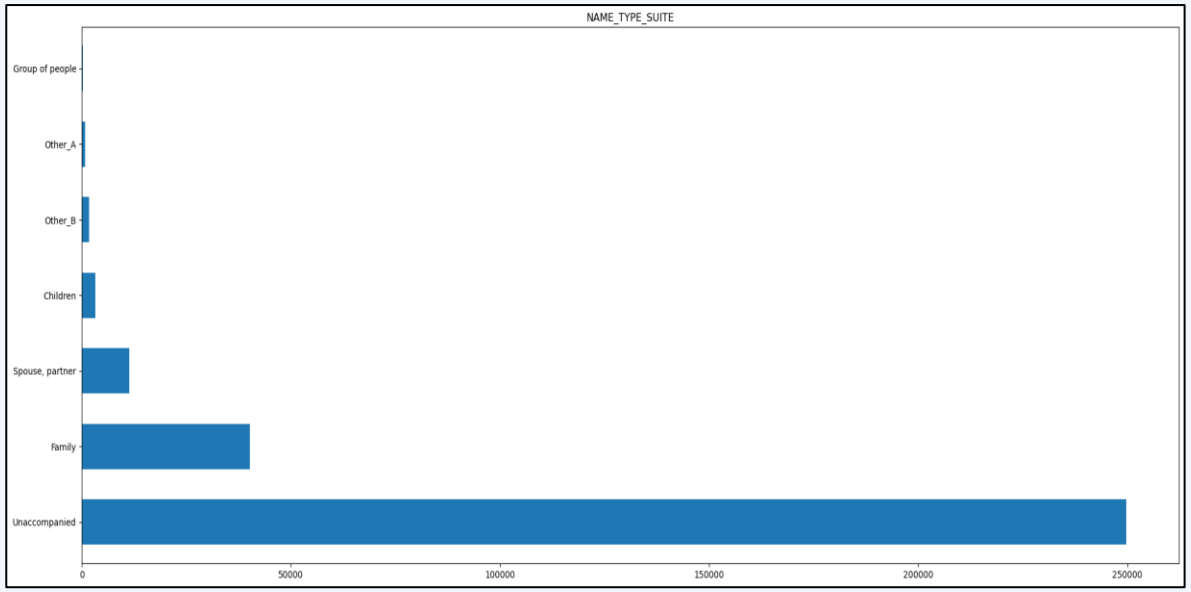


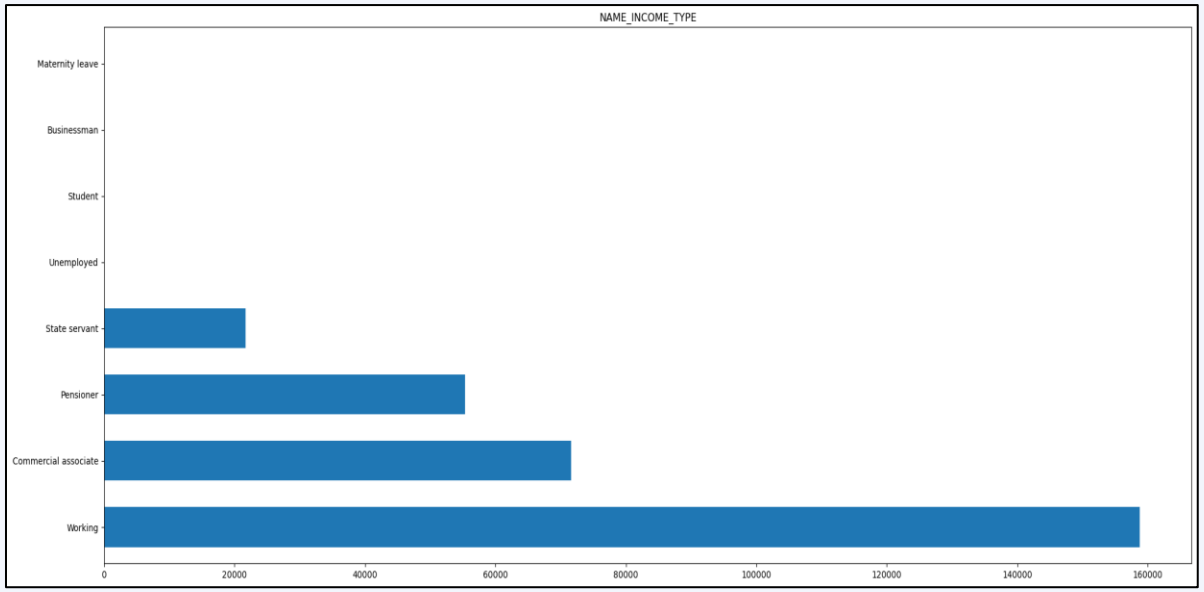


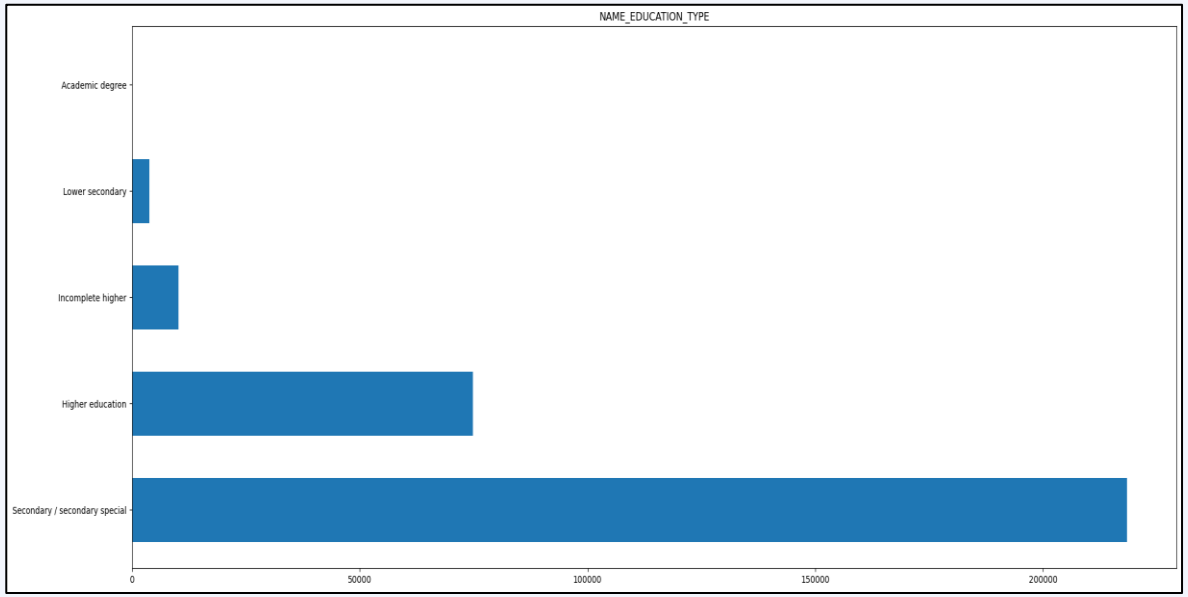


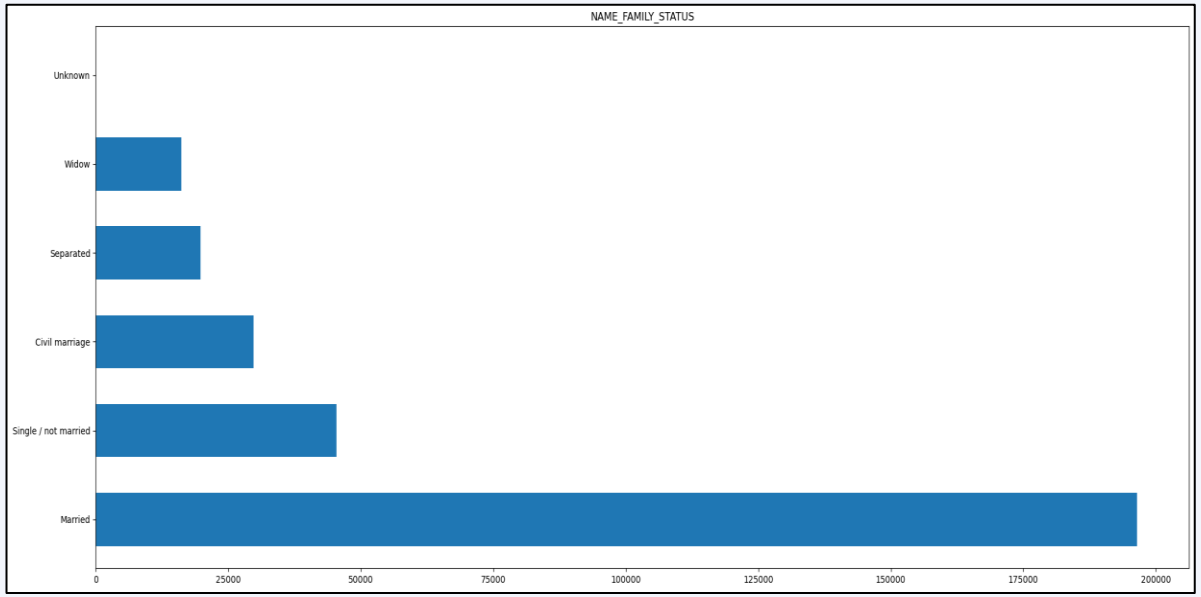


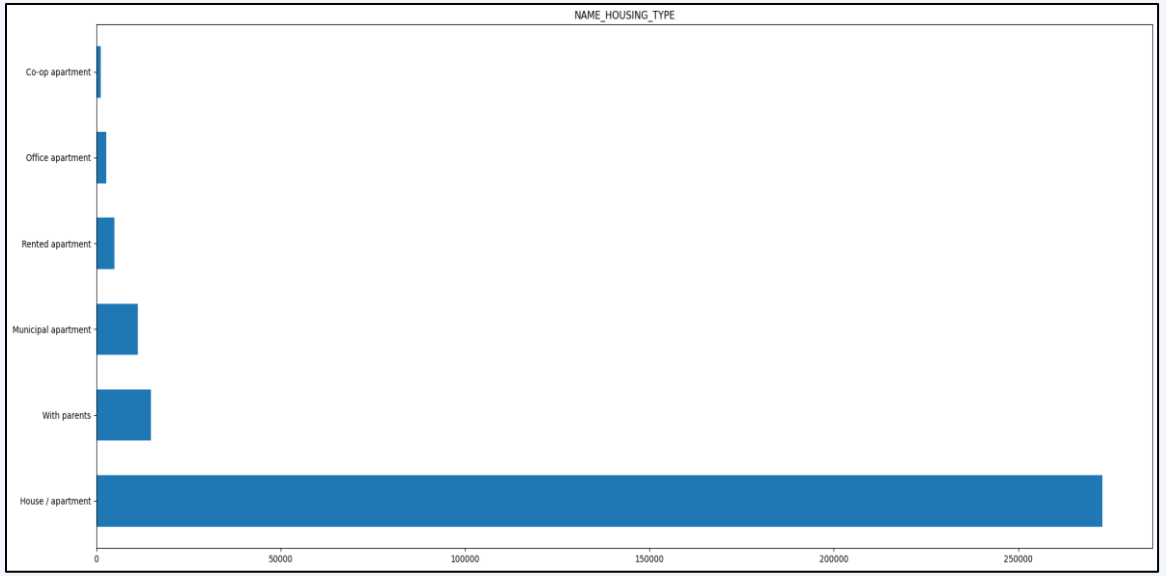
Univariate analysis for categorical data

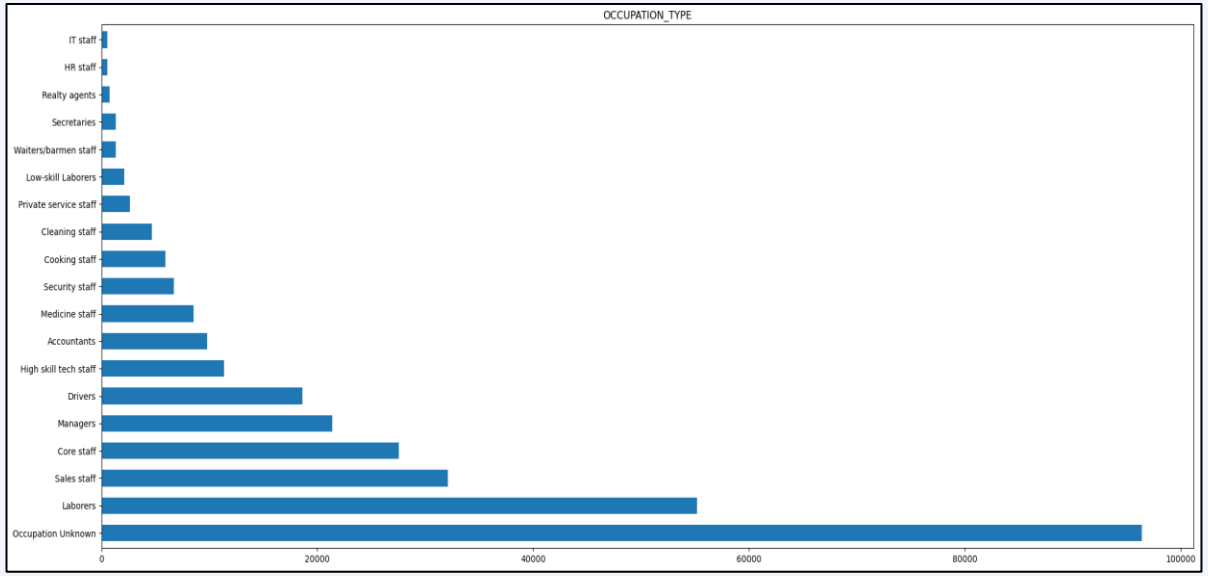


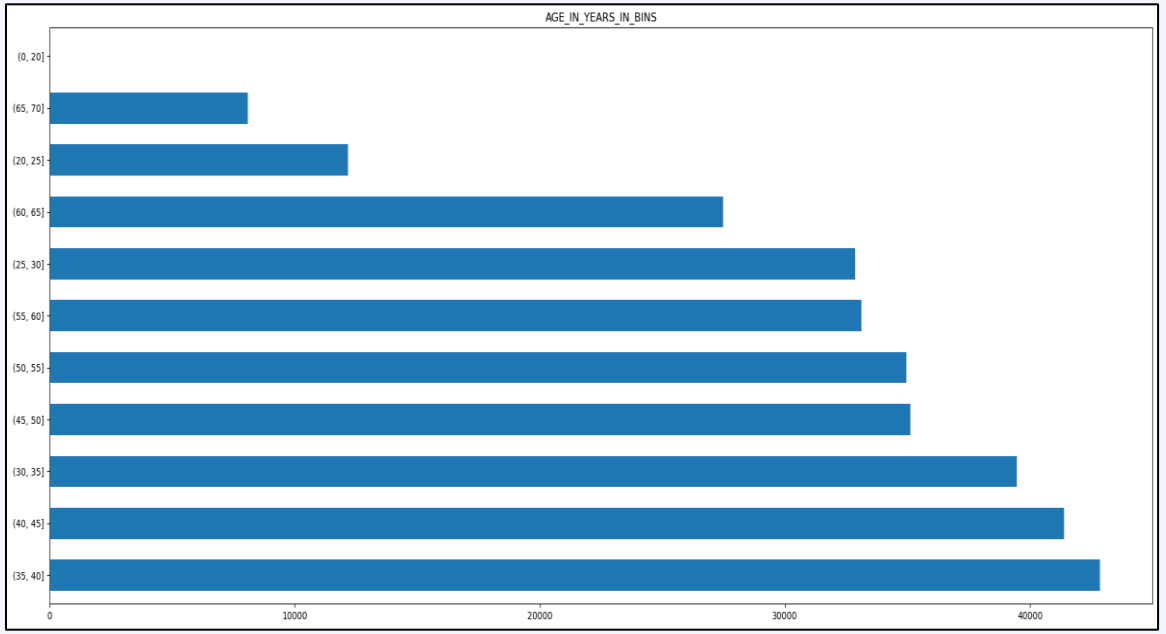


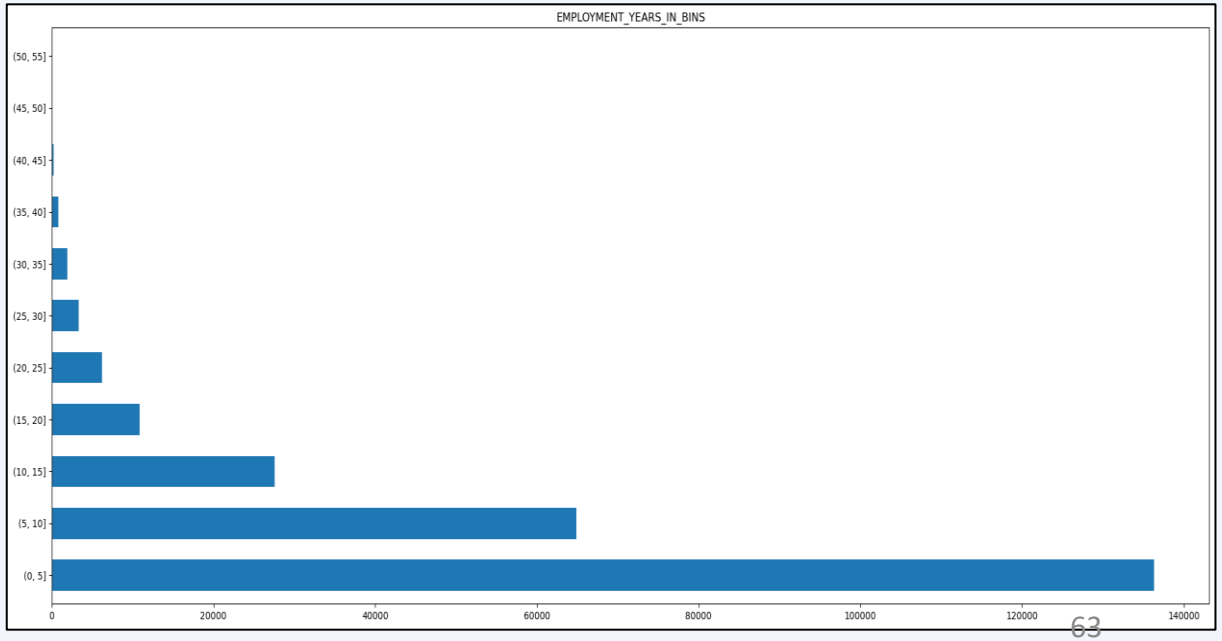


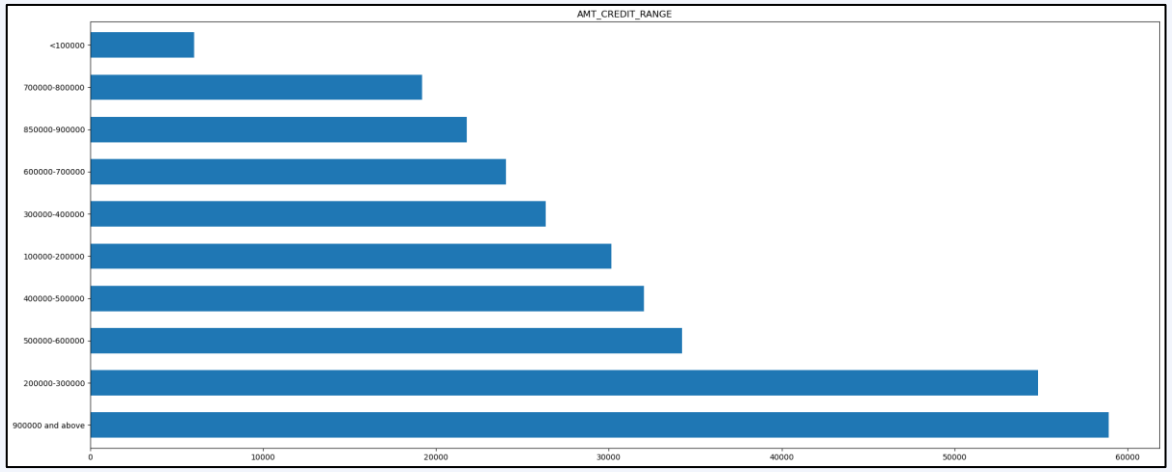


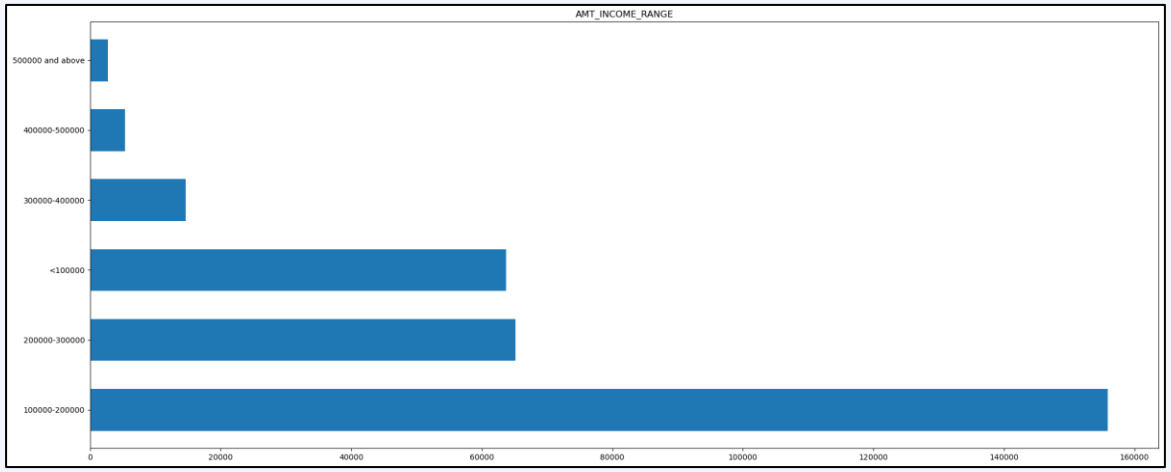

NAME_HOUSING_TYPE
Co-op apartment
Office apartment
Rented apartment
Municipal apartment
With parents
House / apartment
0
50000
100000
150000
200000
250000


OCCUPATION_TYPE
IT staff
HR staff
Realty agents
Secretaries
Waiters/barmen staff
Low-skill Laborers
Private service staff
Cleaning staff
Cooking staff
Security staff
Medicine staff
Accountants
High skill tech staff
Drivers
Managers
Core staff
Sales staff
Laborers
Occupation Unknown
0
20000
40000
60000
80000
100000


AGE_IN_YEARS_IN_BINS
(0, 20]
(65, 70]
(20, 25]
(60, 65]
(25, 30]
(55, 60]
(50, 55]
(45, 50]
(30, 35]
(40, 45]
(35, 40]
0
10000
20000
30000
40000


EMPLOYMENT_YEARS_IN_BINS
(50, 55]
(45, 50]
(40, 45]
(35, 40]
(30, 35]
(25, 30]
(20, 25]
(15, 20]
(10, 15]
(5, 10]
(0, 5]
0
20000
40000
60000
80000
100000
120000
140000





Insights from univariate analysis:

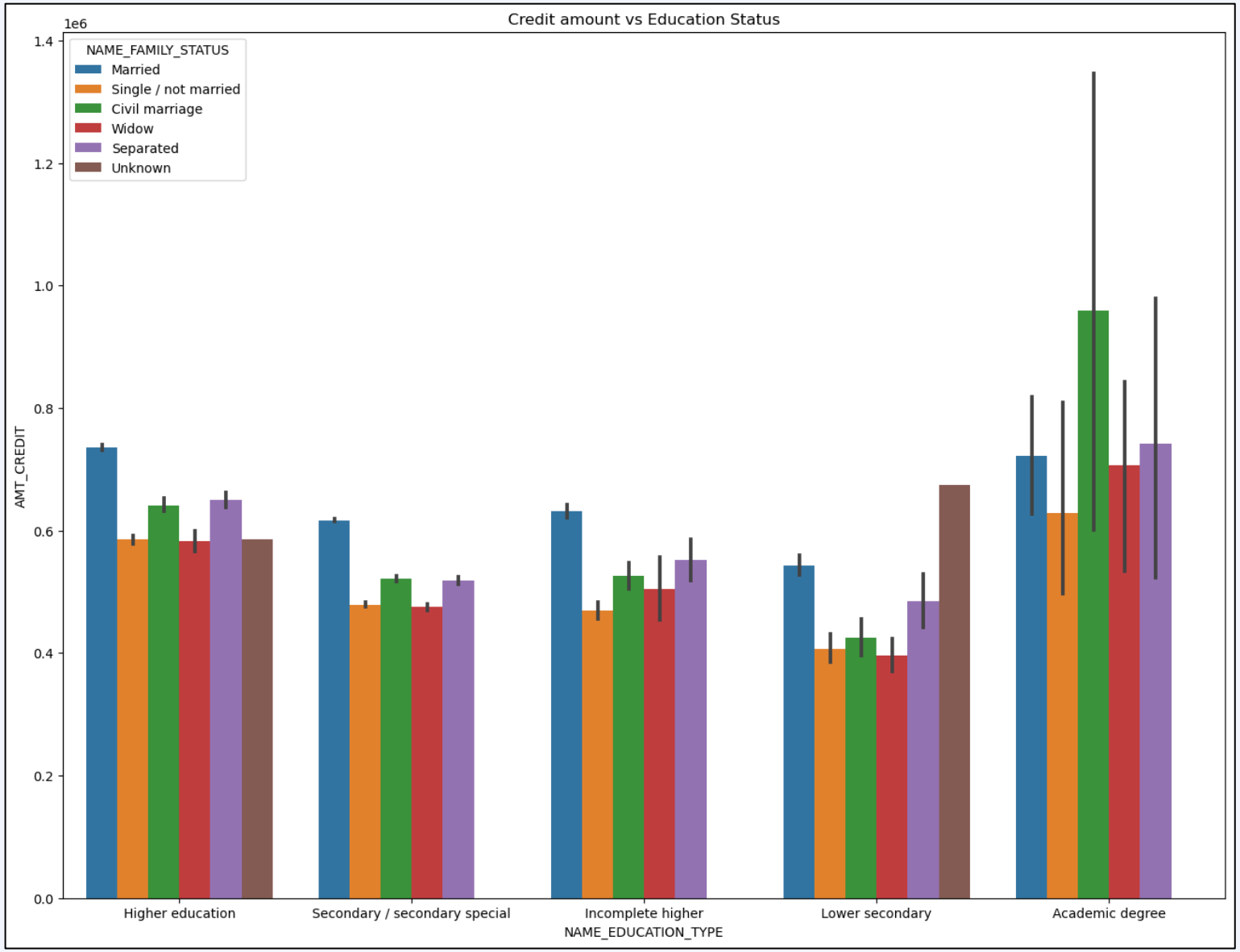
- Some outliers are observed in In 'AMT_ANNUITY','AMT_GOODS_PRICE','AMT_CREDIT', 'CNT_FAM_MEMBERS' in the dataset
- There is single high value data point as outlier present in EMPLOYEMENT_YEARS. Removal this point will drastically impact the box plot for further analysis.

- The people having income 100000-200000 are having higher number of loans.
- Working, State servant and Commercial associates have higher default percentage ..
- People in the age group of 35-40 are having more loans compared to other age groups.

Bivariate analysis (application_data.csv)

For Target 0

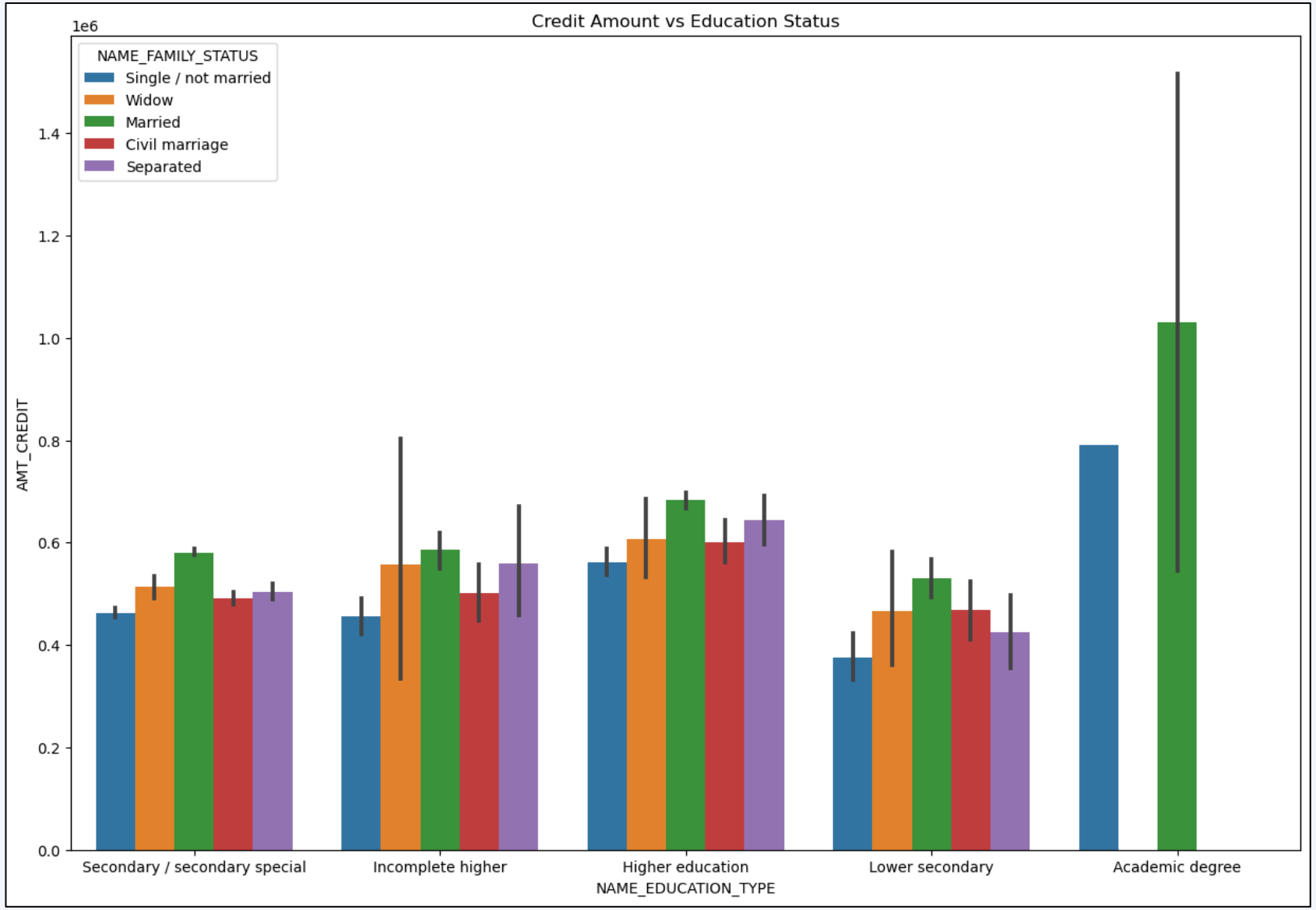
```
In [59]: # Box plotting for Credit amount

         plt.figure(figsize=(16,12))
         plt.xticks(rotation=0)
         sns.barplot(data =target0, x='NAME_EDUCATION_TYPE',y='AMT_CREDIT', hue ='NAME_FAMILY_STATUS',orient='v')
         plt.title('Credit amount vs Education Status')
         plt.show()
```



The family status of 'civil marriage', 'married' and 'separated' in Academic degree education are having higher amount of credits than others.
Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
.

For Target 1

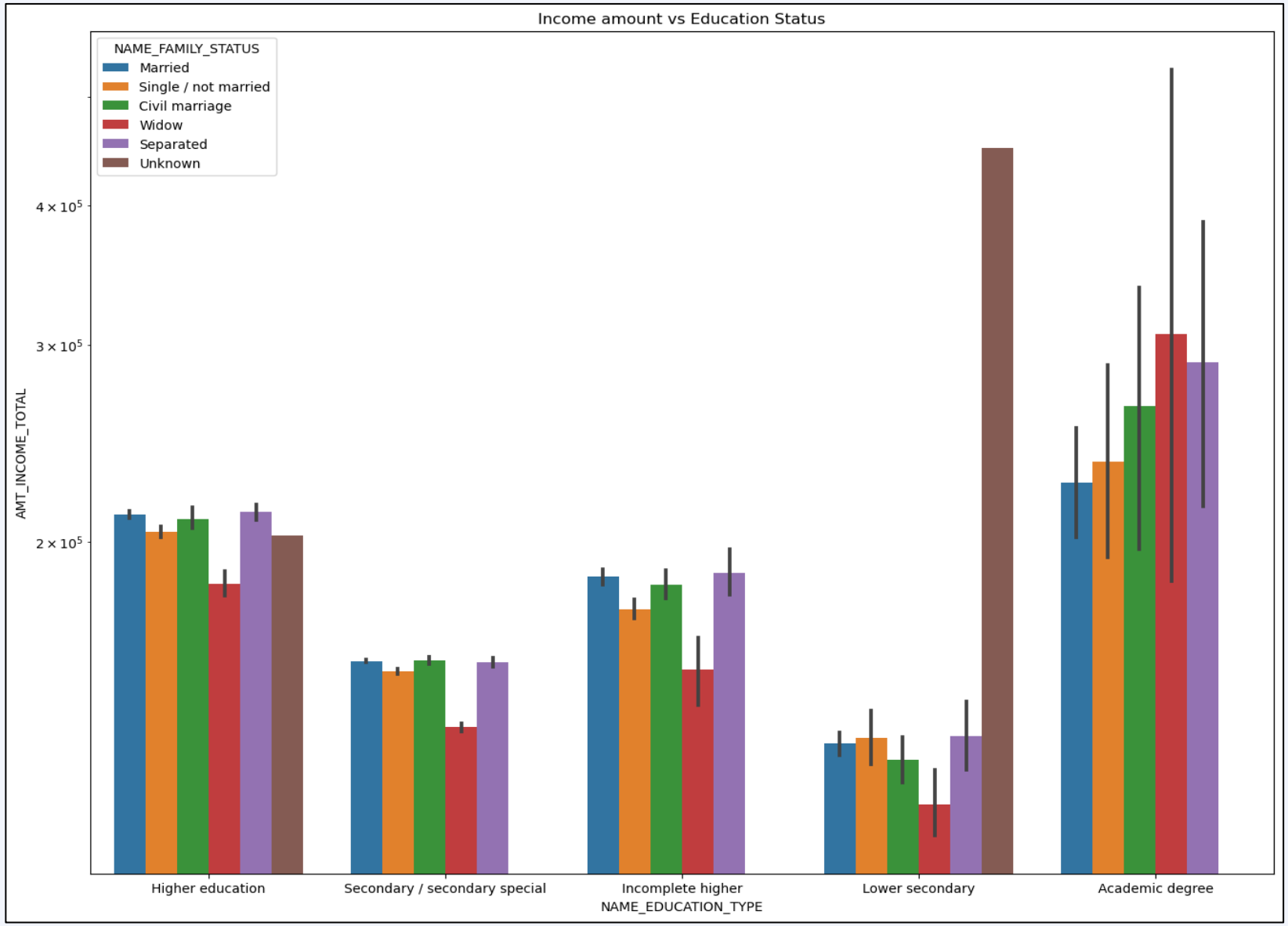
```
In [61]: # Box plotting for credit amount

         plt.figure(figsize=(15,10))
         plt.xticks(rotation=0)
         sns.barplot(data =target1, x='NAME_EDUCATION_TYPE',y='AMT_CREDIT', hue ='NAME_FAMILY_STATUS',orient='v')
         plt.title('Credit Amount vs Education Status')
         plt.show()
```



Observations are Quite similar with Target 0
Family status of 'married' and ‘single/not married' of Academic degree education are having higher amount of credits than others.
Most of the outliers are from Education type 'Higher education' and 'Secondary'.
.

For target 0

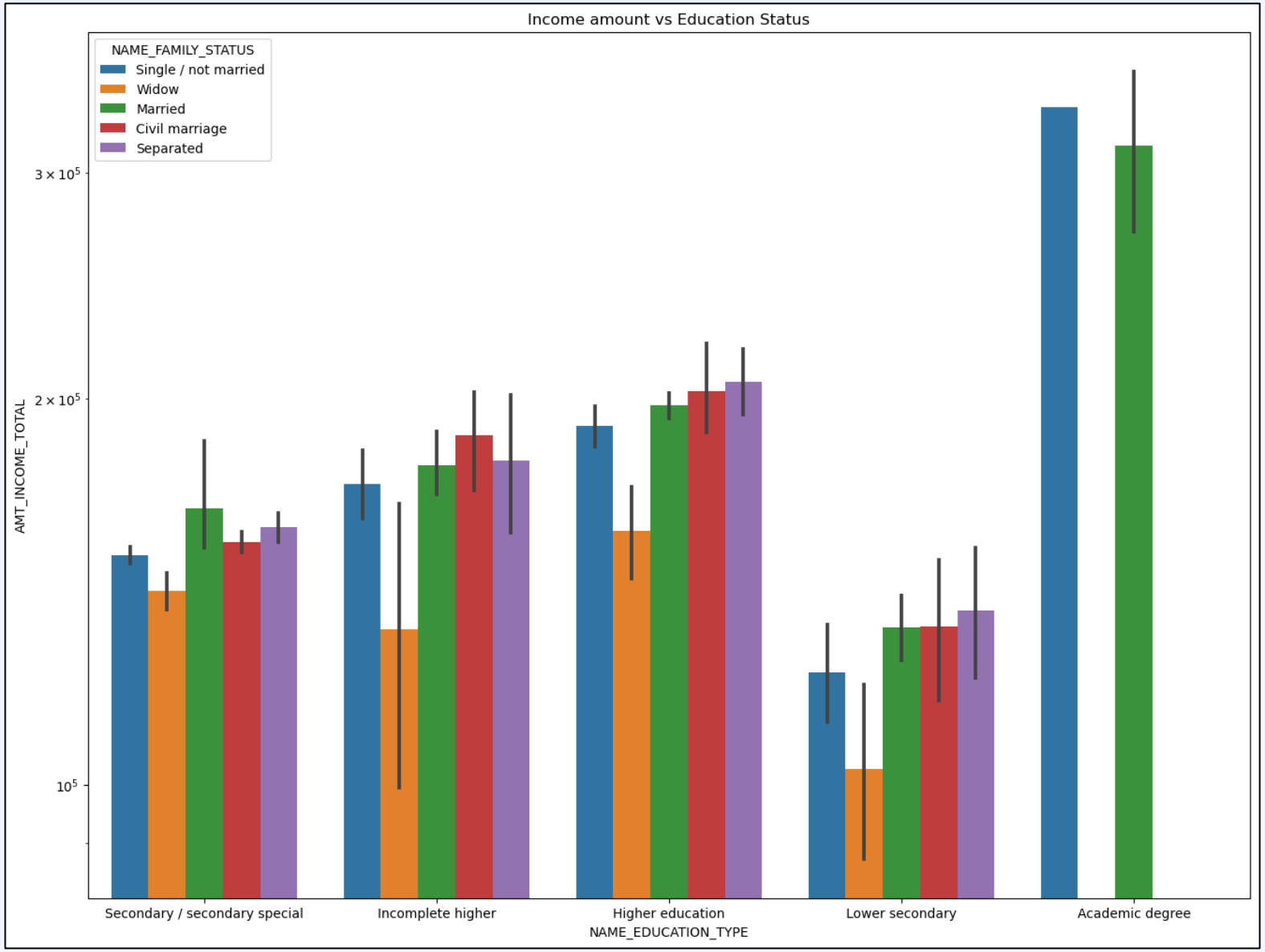
```
In [63]: # Box plotting for Income amount in logarithmic scale

         plt.figure(figsize=(16,12))
         plt.xticks(rotation=0)
         plt.yscale('log')
         sns.barplot(data =target0, x='NAME_EDUCATION_TYPE',y='AMT_INCOME_TOTAL', hue ='NAME_FAMILY_STATUS',orient='v')
         plt.title('Income amount vs Education Status')
         plt.show()
```



In Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers.
There are less outliers for Academic degree but there income amount is little higher that of Higher education.
Lower secondary of civil marriage family status are have less income amount than others.
.

for Target 1

```
In [64]: # Box plotting for Income amount in logarithmic scale

plt.figure(figsize=(16,12))
plt.xticks(rotation=0)
plt.yscale('log')
sns.barplot(data =target1, x='NAME_EDUCATION_TYPE',y='AMT_INCOME_TOTAL', hue ='NAME_FAMILY_STATUS',orient='v')
plt.title('Income amount vs Education Status')
plt.show()
```



Observations are quite similar to Target0,
In Education type 'Higher education' the income amount is mostly equal with family status There are less outliers for Academic degree but there income amount is little higher that of Higher education.
Lower secondary are have less income amount than others.
.

Correlation

Getting top 10 correlation between variables

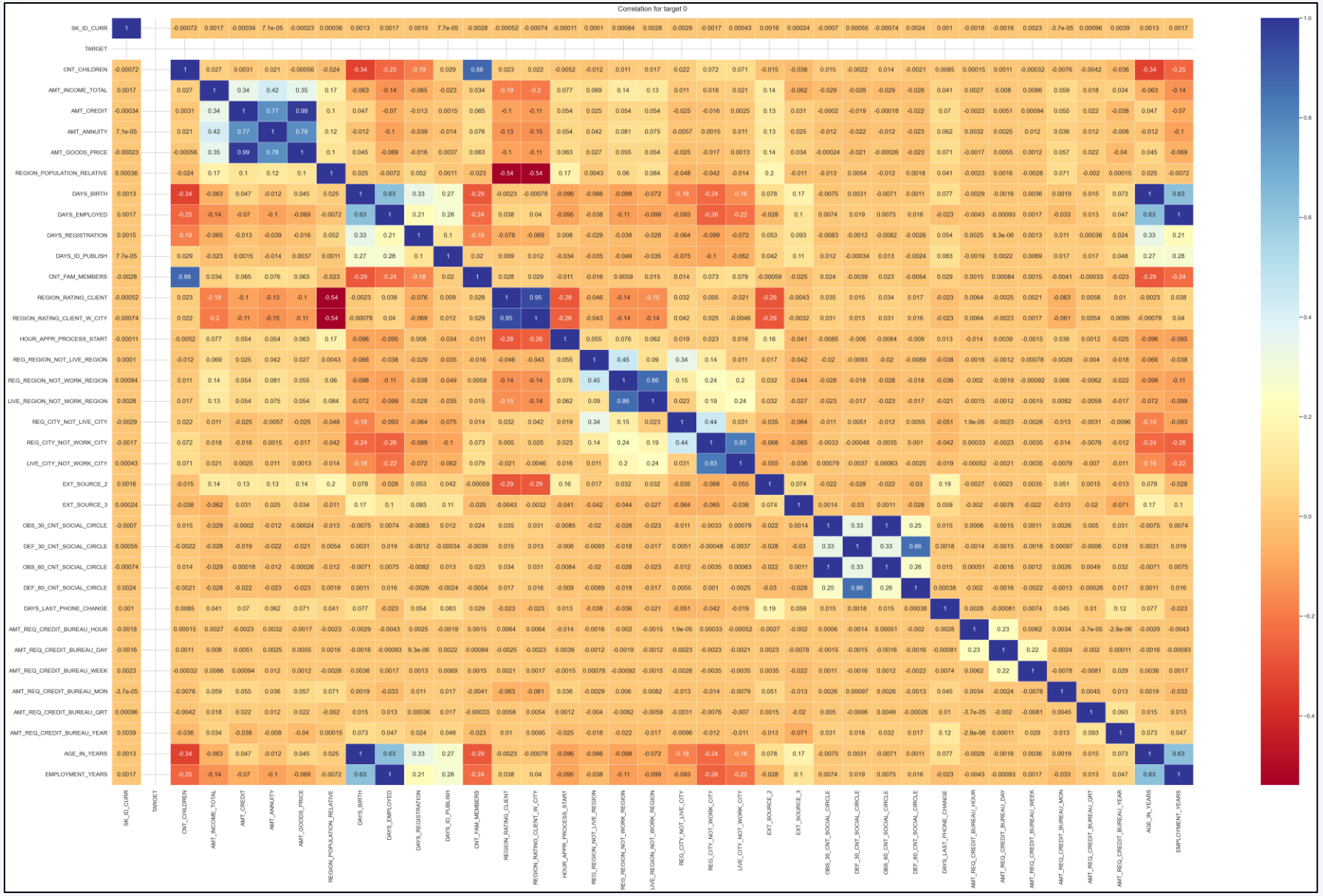
In [122]:
```
# Top 10 correlated variables: target 0 dataframe

corr = target0.corr()
corr_target0 = corr.where(np.triu(np.ones(corr.shape), k=1).astype(np.bool))
corr_target0 = corr_target0.unstack().reset_index()
corr_target0.columns = ['Var1', 'Var2', 'Correlation']
corr_target0.dropna(subset = ['Correlation'], inplace = True)
corr_target0['Correlation'] = round(corr_target0['Correlation'], 2)
corr_target0['Correlation'] = abs(corr_target0['Correlation'])
corr_target0.sort_values(by = 'Correlation', ascending = False).head(10)
```

Out[122]:

	Var1	Var2	Correlation
986	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
1341	EMPLOYMENT_YEARS	DAYS_EMPLOYED	1.00
1303	AGE_IN_YEARS	DAYS_BIRTH	1.00
226	AMT_GOODS_PRICE	AMT_CREDIT	0.99
531	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
446	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
1024	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
683	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
797	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
227	AMT_GOODS_PRICE	AMT_ANNUITY	0.78

To find out the correlation between the variables, I have considered target 0 data first and found out the top 10 correlation and analyzed using a heatmap.

In [123]:
```
#plotting heatmap to see the correlation for target 0

fig = plt.figure(figsize=(70,40))
sns.heatmap(target0.corr(), cmap="RdYlBu",annot=True,linewidth =.7)
plt.title("Correlation for target 0")
plt.show()
```



From the correlation analysis it can be said that the highest corelation (1) is between (OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE) and (EMPLOYMENT_YEARS with DAYS_EMPLOYED) which is same for both the data set.

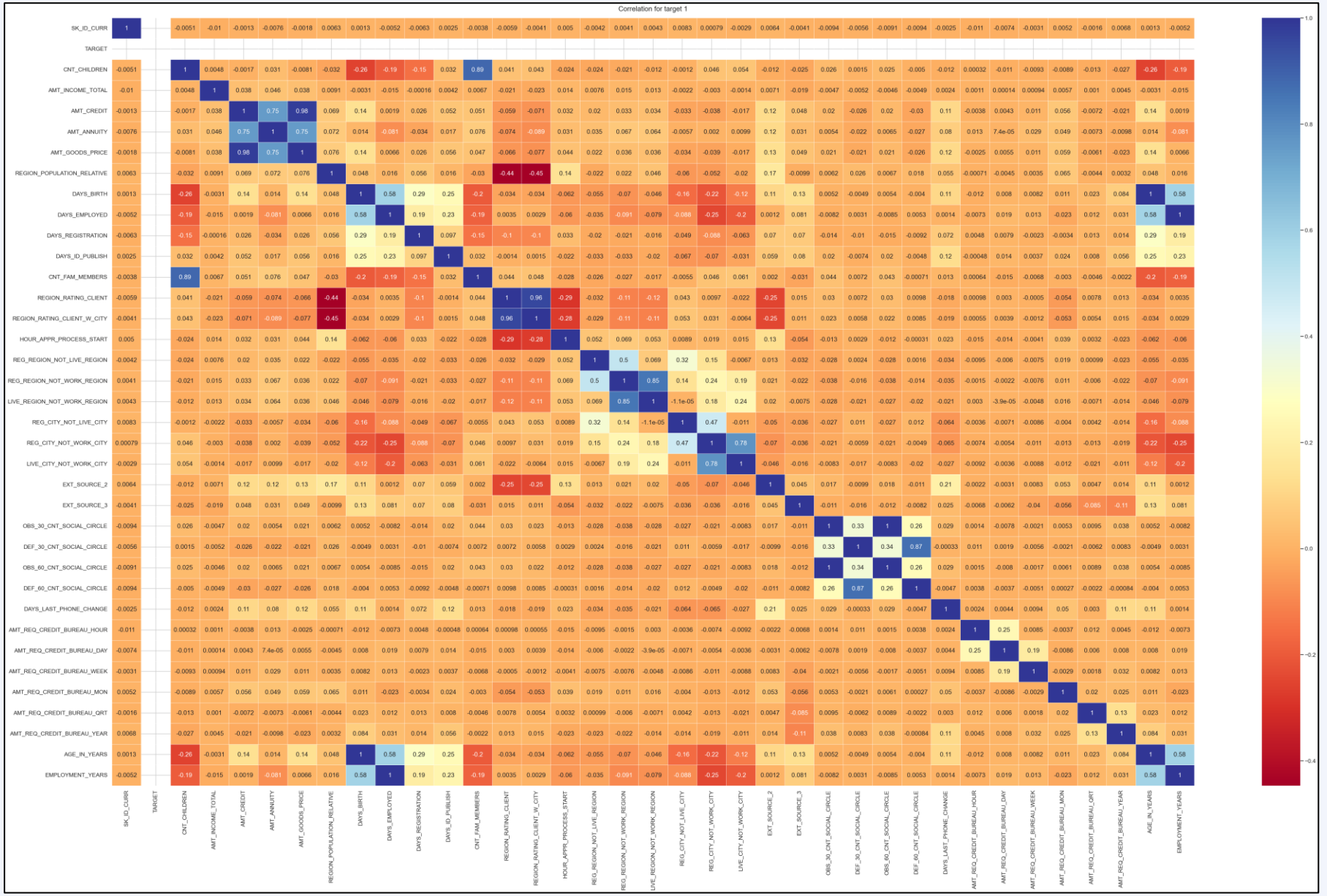
In [125]:
```
# Top 10 correlated variables: target 1 dataframe

corr = target1.corr()
corr_target1 = corr.where(np.triu(np.ones(corr.shape), k=1).astype(np.bool))
corr_target1 = corr_target1.unstack().reset_index()
corr_target1.columns = ['Var1', 'Var2', 'Correlation']
corr_target1.dropna(subset = ['Correlation'], inplace = True)
corr_target1['Correlation'] = round(corr_target1['Correlation'], 2)
corr_target1['Correlation'] = abs(corr_target1['Correlation'])
corr_target1.sort_values(by = 'Correlation', ascending = False).head(10)
```

To find out the correlation between the variables, I have considered target 1 data first and found out the top 10 correlation and analyzed using a heatmap.

Out[125]:

	Var1	Var2	Correlation
986	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
1341	EMPLOYMENT_YEARS	DAYS_EMPLOYED	1.00
1303	AGE_IN_YEARS	DAYS_BIRTH	1.00
226	AMT_GOODS_PRICE	AMT_CREDIT	0.98
531	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
446	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
1024	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
683	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
797	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78
189	AMT_ANNUITY	AMT_CREDIT	0.75

In [126]:
```
#plotting heatmap to see the correlation for target 1

fig = plt.figure(figsize=(70,40))
sns.heatmap(target1.corr(), cmap="RdYlBu",annot=True,linewidth =.7)
plt.title("Correlation for target 1")
plt.show()
```



From the correlation analysis it can be said that the highest corelation (1) is between (OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE) and (EMPLOYMENT_YEARS with DAYS_EMPLOYED) which is same for both the data set.

```
In [174]: # Reading Previous Application dataset

In [175]: df_PreApp = pd.read_csv(r'previous_application.csv')

In [176]: df_PreApp.head()
```

Out[176]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN

```
In [177]: df_PreApp.shape
Out[177]: (1670214, 37)
```

I have read the previous_application_data.csv file and then checked the no. of rows and columns in the data.

Then found out the null values in each column using isnull() function.

```
In [179]: df_PreApp.describe()
```

Out[179]:

	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AM
count	1.670214e+06	1.670214e+06	1.297979e+06	1.670214e+06	1.670213e+06	7.743700e+05	
mean	1.923089e+06	2.783572e+05	1.595512e+04	1.752339e+05	1.961140e+05	6.697402e+03	
std	5.325980e+05	1.028148e+05	1.478214e+04	2.927798e+05	3.185746e+05	2.092150e+04	
min	1.000001e+06	1.000010e+05	0.000000e+00	0.000000e+00	0.000000e+00	-9.000000e-01	
25%	1.461857e+06	1.893290e+05	6.321780e+03	1.872000e+04	2.416050e+04	0.000000e+00	
50%	1.923110e+06	2.787145e+05	1.125000e+04	7.104600e+04	8.054100e+04	1.638000e+03	
75%	2.384280e+06	3.675140e+05	2.065842e+04	1.803600e+05	2.164185e+05	7.740000e+03	
max	2.845382e+06	4.562550e+05	4.180581e+05	6.905160e+06	6.905160e+06	3.060045e+06	

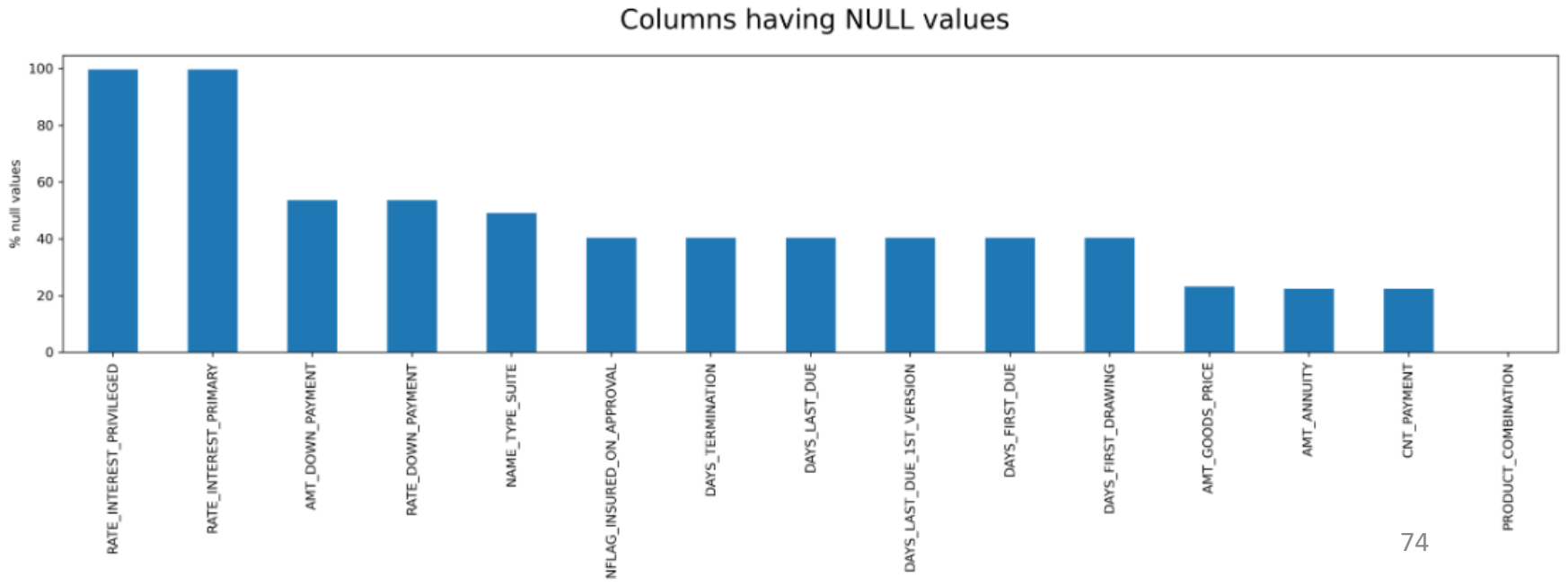
```
In [180]: #Finding out the null values in the columns
          df_PreApp.isnull().sum()
Out[180]: SK_ID_PREV              0
          SK_ID_CURR              0
          NAME_CONTRACT_TYPE      0
          AMT_ANNUITY        372235
          AMT_APPLICATION         0
          AMT_CREDIT              1
          AMT_DOWN_PAYMENT   895844
          AMT_GOODS_PRICE    385515
```

```
In [181]: # Calling the function and getting the percentage of missing values from the cloumns

          col_null_percentage(df_PreApp)

Out[181]: RATE_INTEREST_PRIVILEGED      99.64
          RATE_INTEREST_PRIMARY         99.64
          AMT_DOWN_PAYMENT              53.64
          RATE_DOWN_PAYMENT             53.64
          NAME_TYPE_SUITE               49.12
          NFLAG_INSURED_ON_APPROVAL     40.30
          DAYS_TERMINATION              40.30
          DAYS_LAST_DUE                 40.30
          DAYS_LAST_DUE_1ST_VERSION     40.30
          DAYS_FIRST_DUE                40.30
          DAYS_FIRST_DRAWING            40.30
          AMT_GOODS_PRICE               23.08
```

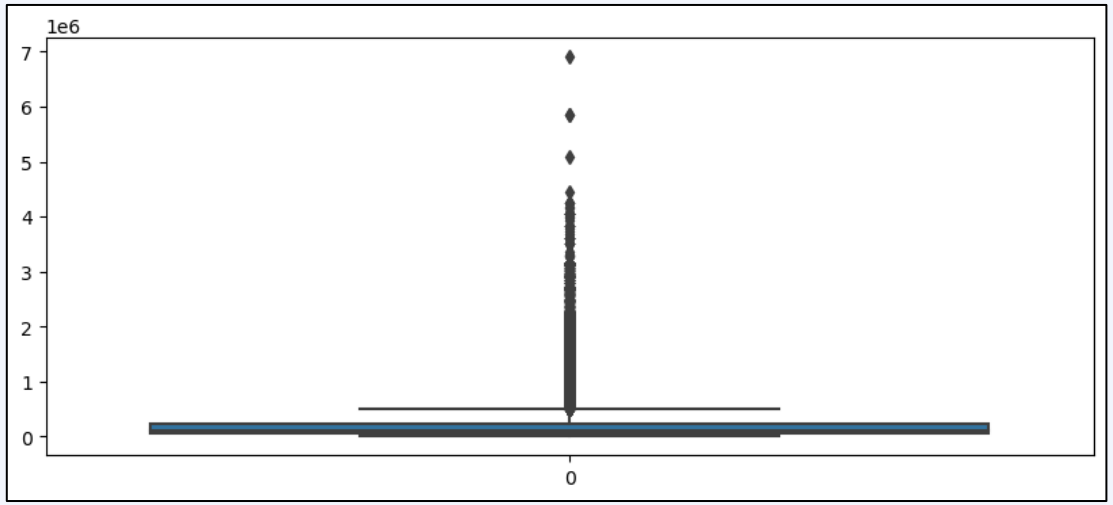
Then using col_null_percentage() function which I have defined above, I have extracted the percentage of null values in the columns.

Then plotted a bar graph representing the columns having null values.

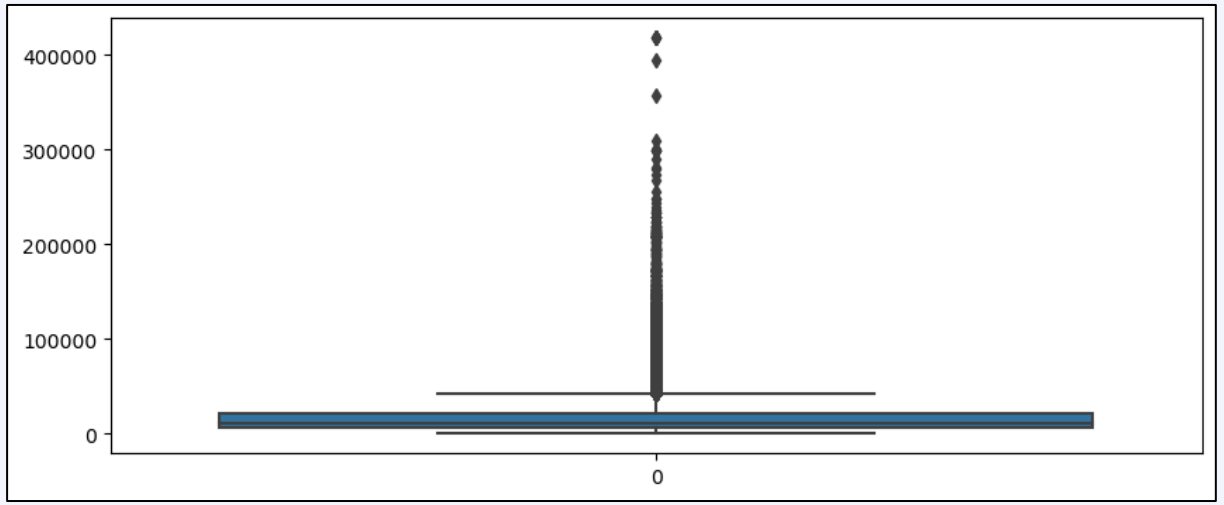
```
In [182]: # graphical representation of columns having % null values

          plt.figure(figsize= (20,4),dpi=300)
          col_null_percentage(df_PreApp)[col_null_percentage(df_PreApp)>0].plot(kind = 'bar')
          plt.title (' Columns having NULL values')
          plt.ylabel('% null values')
          plt.show()
```



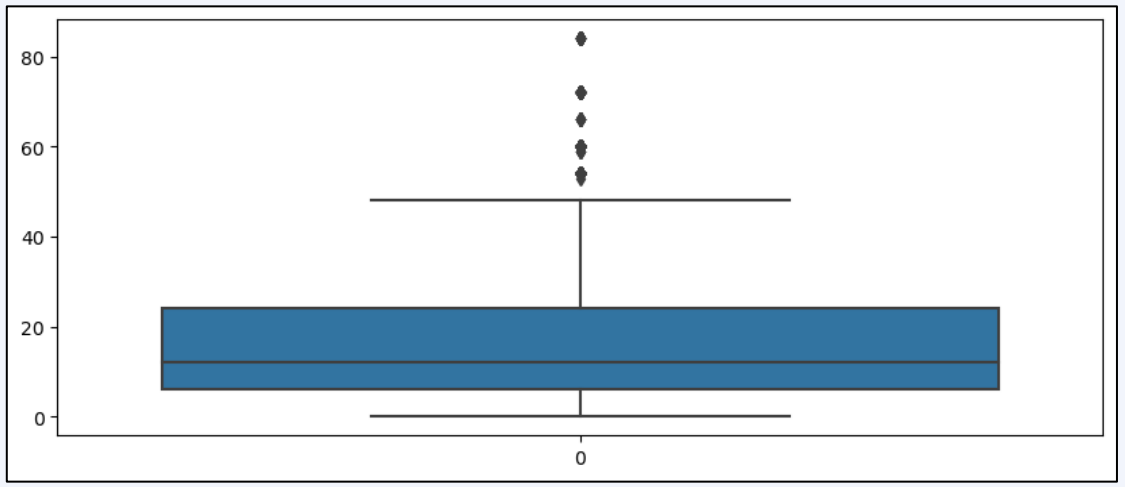
In [187]:
```
# Box plot for AMT_GOODS_PRICE  column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_PreApp['AMT_GOODS_PRICE'])
plt.show()
```



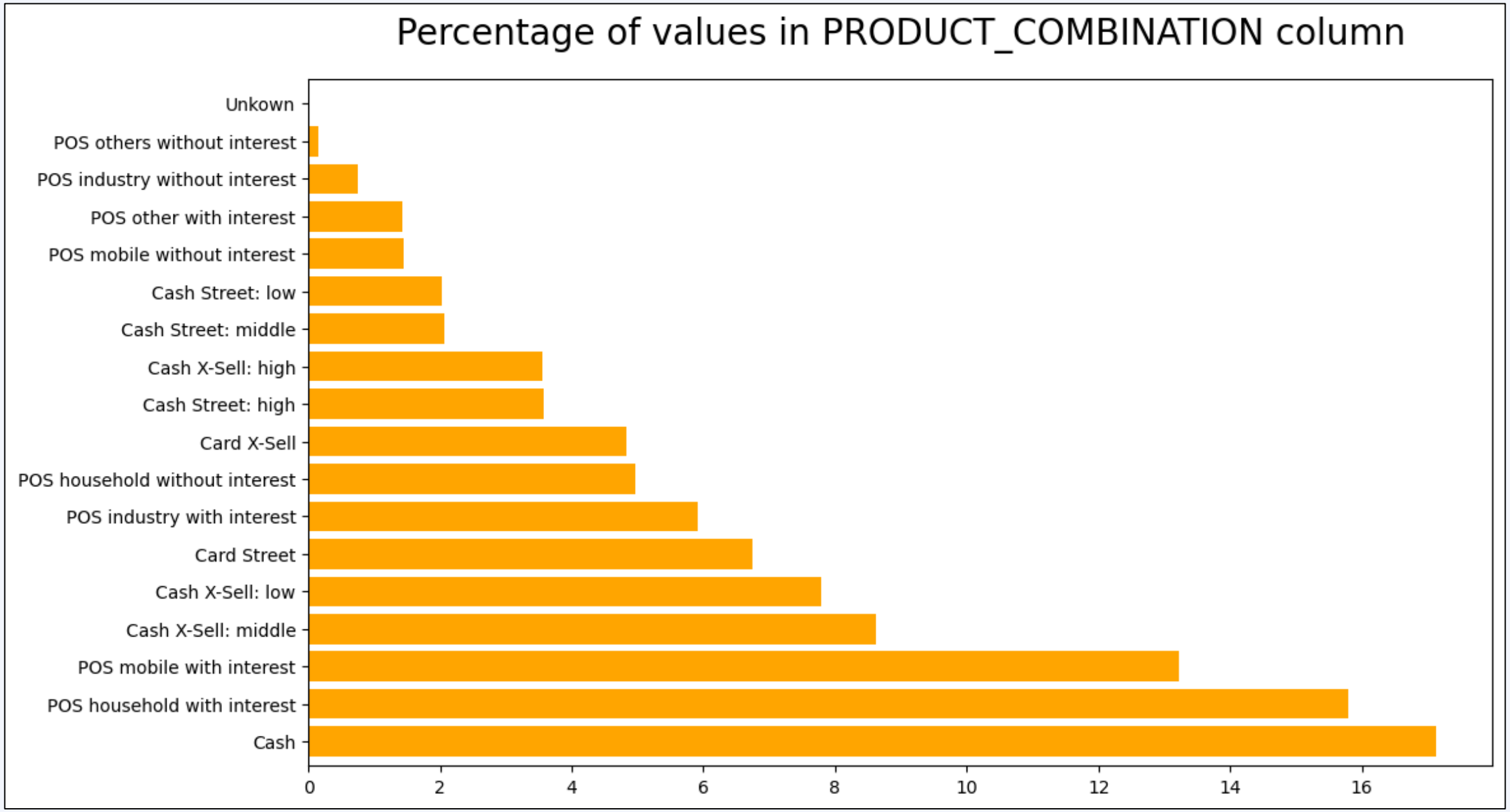
In [188]:
```
# Box plot for AMT_ANNUITY column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_PreApp['AMT_ANNUITY'])
plt.show()
```



In [189]:
```
# Box plot for CNT_PAYMENT column to check the outliers

plt.figure(figsize=(10,4))
sns.boxplot(df_PreApp['CNT_PAYMENT'])
plt.show()
```

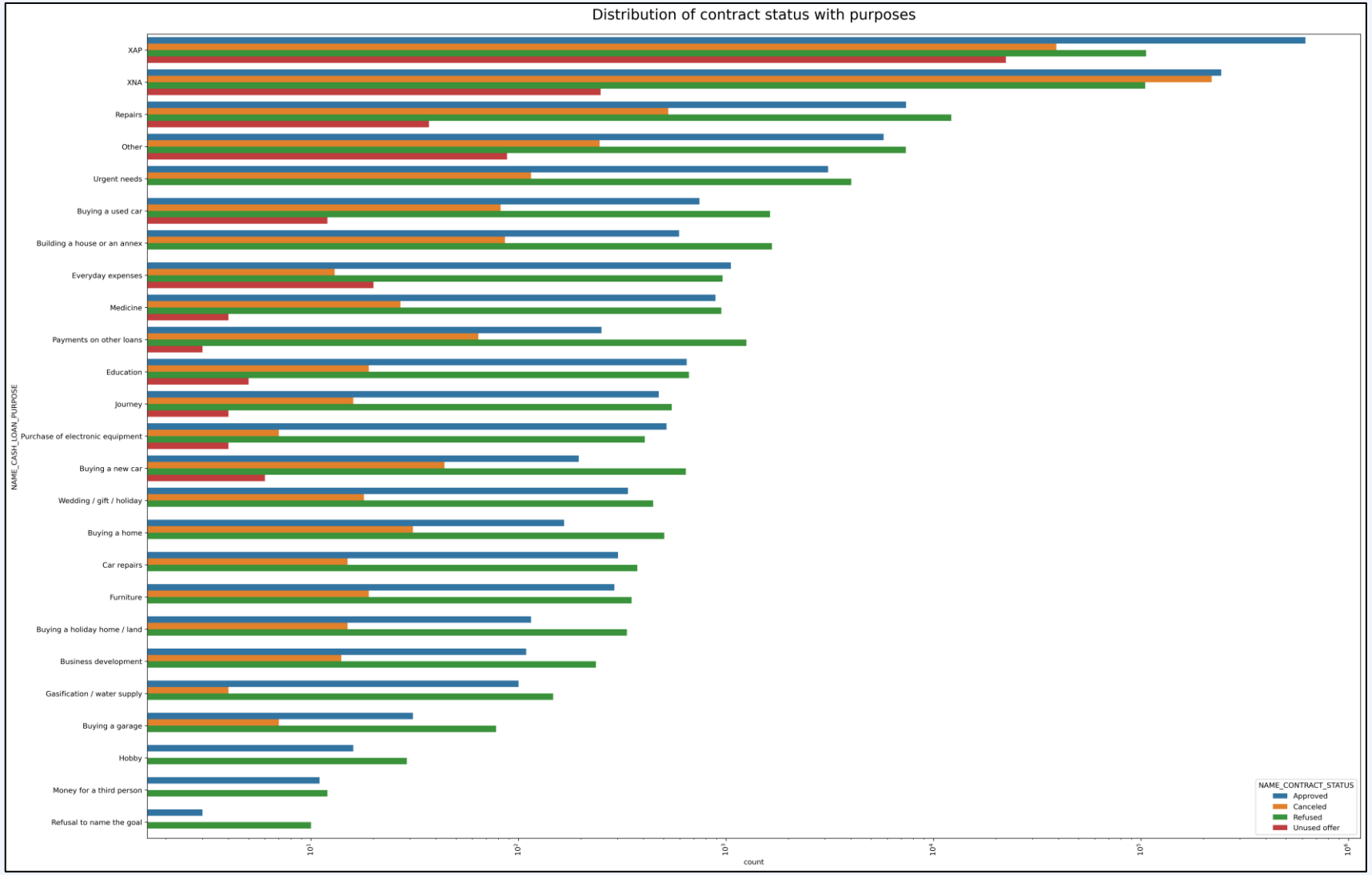


Plotted box plots for certain columns to find out the outliers in the data.


Percentage of values in PRODUCT_COMBINATION column
Unkown
POS others without interest
POS industry without interest
POS other with interest
POS mobile without interest
Cash Street: low
Cash Street: middle
Cash X-Sell: high
Cash Street: high
Card X-Sell
POS household without interest
POS industry with interest
Card Street
Cash X-Sell: low
Cash X-Sell: middle
POS mobile with interest
POS household with interest
Cash
0
2
4
6
8
10
12
14
16

Univariate analysis

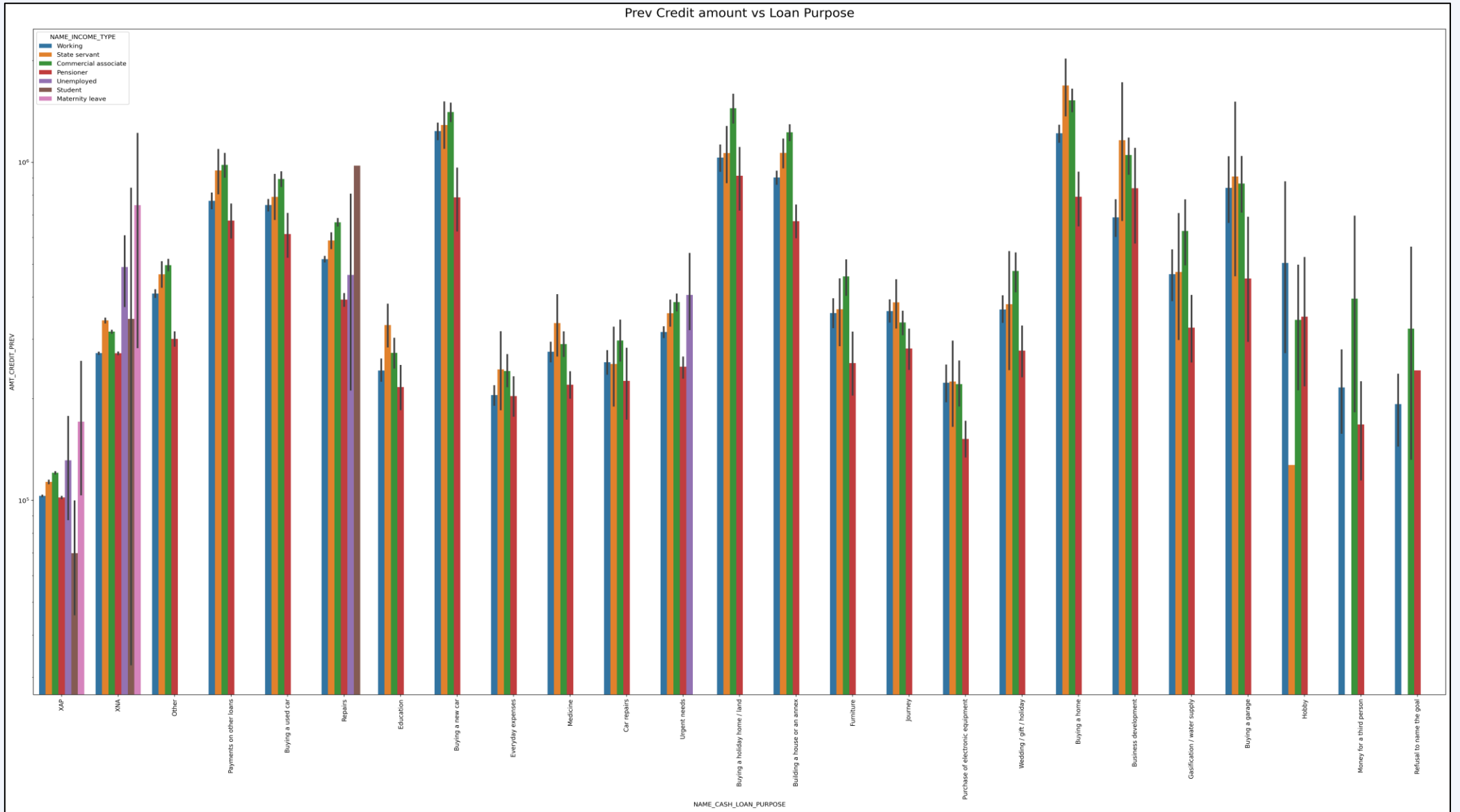
```
In [203]: # Distribution of contract status in logarithmic scale

          plt.figure(figsize=(30,20),dpi = 300)
          plt.rcParams["axes.labelsize"] = 10
          plt.rcParams['axes.titlesize'] = 20
          plt.rcParams['axes.titlepad'] = 20
          plt.xticks(rotation=90)
          plt.xscale('log')
          plt.title('Distribution of contract status with purposes')
          ax = sns.countplot(data = df_merge, y= 'NAME_CASH_LOAN_PURPOSE',
                             order=df_merge['NAME_CASH_LOAN_PURPOSE'].value_counts().index,hue = 'NAME_CONTRACT_STATUS')
```



There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education'. Loan purposes with 'Repairs' are facing more difficulties in payment on time.

Bivariate analysis

```
In [212]: # Bar plotting for Credit amount in logarithmic scale

          plt.figure(figsize=(40,20),dpi = 350)
          plt.xticks(rotation=90)
          plt.yscale('log')
          sns.barplot(data =df_merge, x='NAME_CASH_LOAN_PURPOSE',hue='NAME_INCOME_TYPE',y='AMT_CREDIT_PREV',orient = 'v')
          plt.title('Prev Credit amount vs Loan Purpose')
          plt.show()
```



```
In [206]: # Bisecting the "df_merge" dataframe based on Target value 0 and 1 for correlation and other analysis

          Repayers = df_merge[df_merge['TARGET']==0]
          Defaulters = df_merge[df_merge['TARGET']==1]
```

Dividing the merged dataframe based on target value as we have done above.

Finding correlation of target0(i.e, repayers)

```
In [207]: # Getting  top 10 correlation for the Repayers dataframe

corr_repayer = Repayers.corr()
correlation_repayer = corr_repayer.where(np.triu(np.ones(corr_repayer.shape),k=1).astype(np.bool_)).unstack().reset_index()
correlation_repayer.columns =['VAR1','VAR2','Correlation']
correlation_repayer.dropna(subset = ["Correlation"], inplace = True)
correlation_repayer["Correlation"]=correlation_repayer["Correlation"].abs()
correlation_repayer.sort_values(by='Correlation', ascending=False, inplace=True)
correlation_repayer.head(10)
```

Out[207]:

	VAR1	VAR2	Correlation
1139	EMPLOYMENT_YEARS	DAYS_EMPLOYED	1.000000
1099	AGE_IN_YEARS	DAYS_BIRTH	1.000000
758	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998579
1358	AMT_GOODS_PRICEx	AMT_APPLICATION	0.987533
198	AMT_GOODS_PRICE_	AMT_CREDIT	0.986402
1319	AMT_CREDIT_PREV	AMT_APPLICATION	0.975725
1359	AMT_GOODS_PRICEx	AMT_CREDIT_PREV	0.971650
519	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.944356
430	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878475
798	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.863126

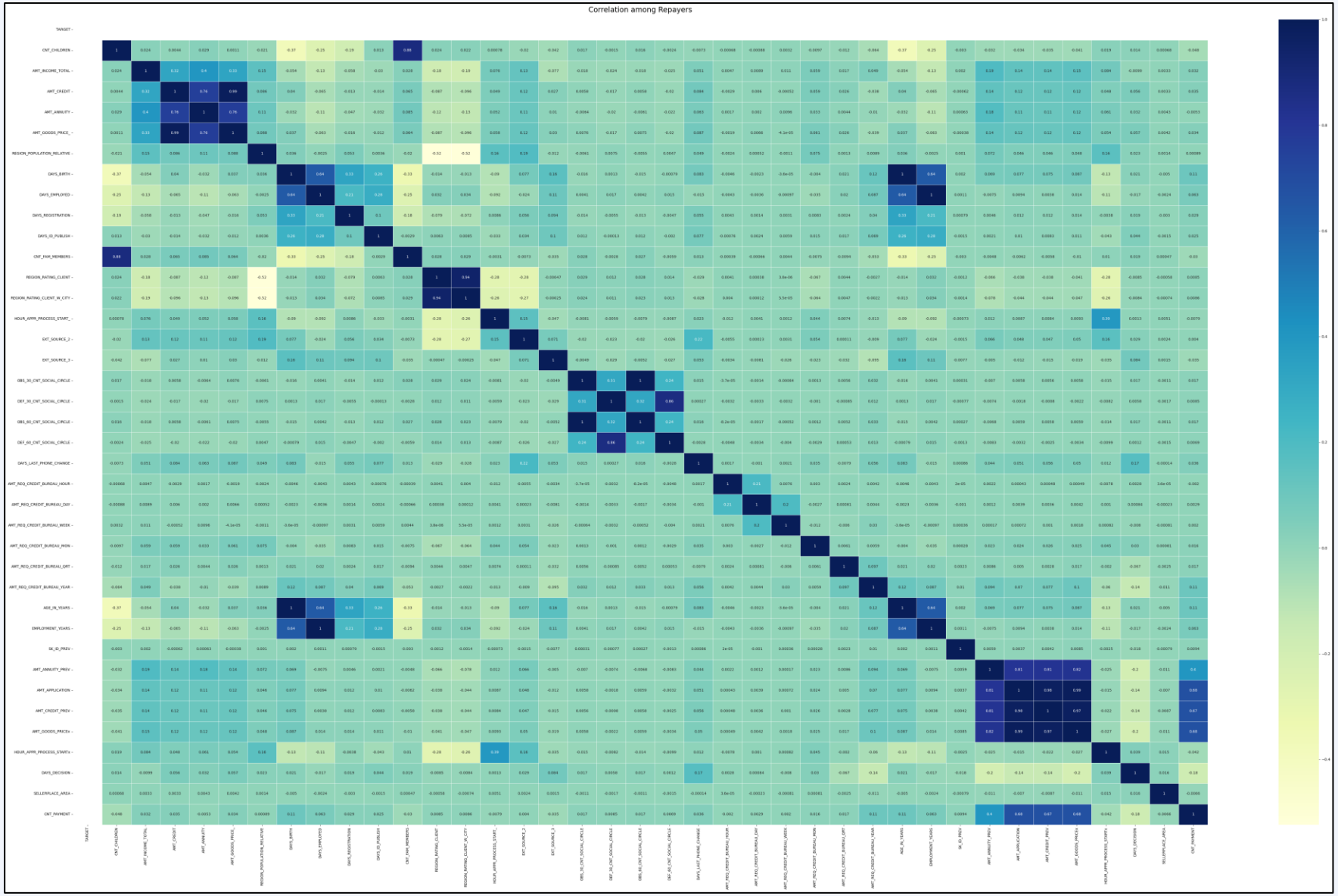
Finding correlation of target1(i.e, defaulters)

```
In [208]: # Getting  top 10 correlation for the Defaulters dataframe

corr_defaulter = Defaulters.corr()
correlation_defulter = corr_defaulter.where(np.triu(np.ones(corr_defaulter.shape),k=1).astype(np.bool_)).unstack().reset_index()
correlation_defulter.columns =['VAR1','VAR2','Correlation']
correlation_defulter.dropna(subset = ["Correlation"], inplace = True)
correlation_defulter["Correlation"]= correlation_defulter["Correlation"].abs()
correlation_defulter.sort_values(by='Correlation', ascending=False, inplace=True)
correlation_defulter.head(10)
```

Out[208]:

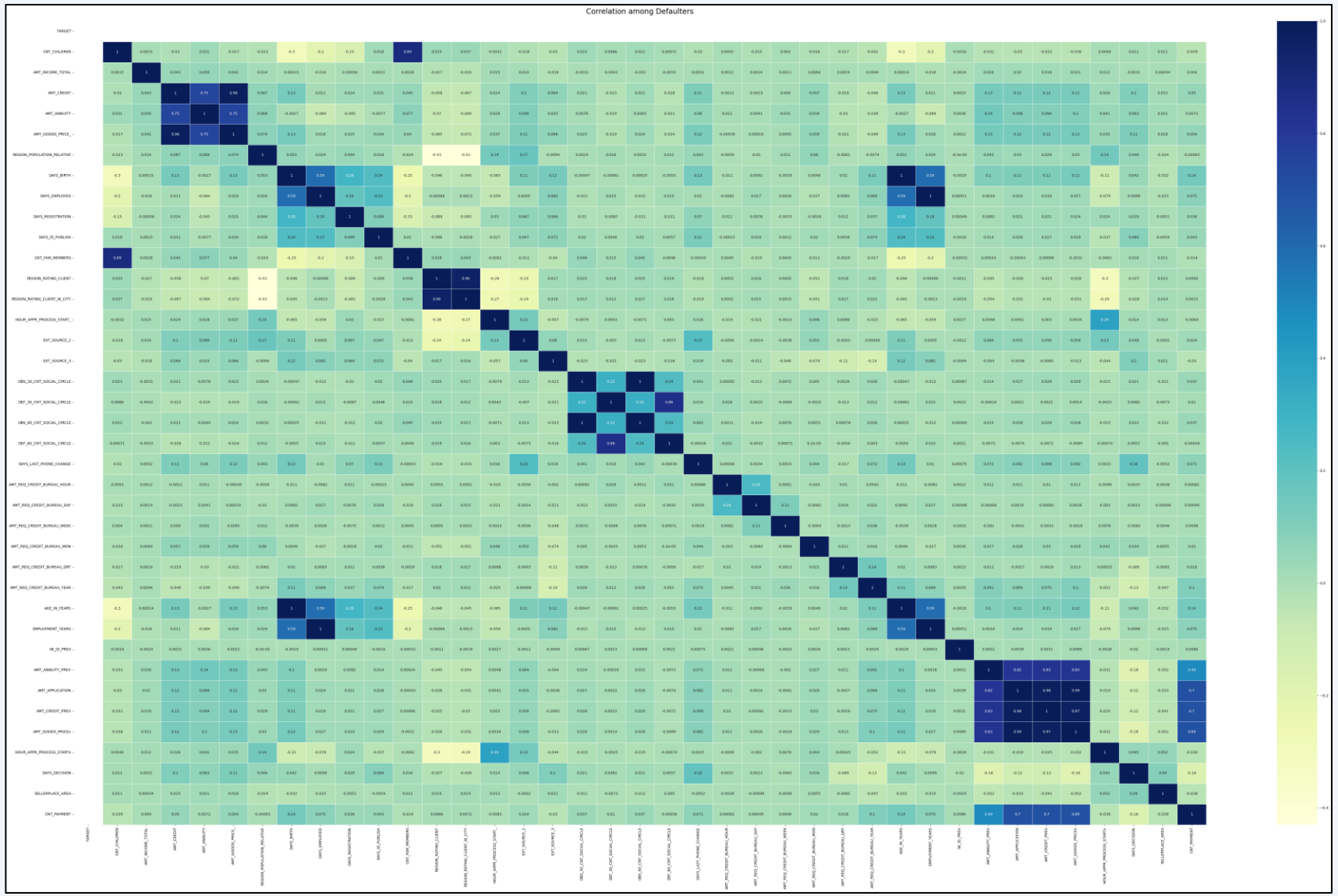
	VAR1	VAR2	Correlation
1139	EMPLOYMENT_YEARS	DAYS_EMPLOYED	1.000000
1099	AGE_IN_YEARS	DAYS_BIRTH	1.000000
758	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998379
1358	AMT_GOODS_PRICEx	AMT_APPLICATION	0.985655
198	AMT_GOODS_PRICE_	AMT_CREDIT	0.982525
1319	AMT_CREDIT_PREV	AMT_APPLICATION	0.975377
1359	AMT_GOODS_PRICEx	AMT_CREDIT_PREV	0.969027
519	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956483
430	CNT_FAM_MEMBERS	CNT_CHILDREN	0.886300
798	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.858301

```
In [209]: #plotting heatmap to see the correlation among Repayers

          fig = plt.figure(figsize=(70,40))
          sns.heatmap(Repayers.corr(), cmap="YlGnBu",annot=True,linewidth =.7)
          plt.title("Correlation among Repayers")
          plt.show()
```



From the correlation analysis it can be said that the highest corelation (1) is between (EMPLOYMENT_YEARS with DAYS_EMPLOYED) and (AGE_IN_YEARS with DAYS_BIRTH) which is same for both the data set.

```
In [210]: #plotting heatmap to see the correlation among defaulters

          fig = plt.figure(figsize=(70,40))
          sns.heatmap(Defaulters.corr(), cmap="YlGnBu",annot=True,linewidth =.7)
          plt.title("Correlation among Defaulters")
          plt.show()
```



From the correlation analysis it can be said that the highest corelation (1) is between (EMPLOYMENT_YEARS with DAYS_EMPLOYED) and (AGE_IN_YEARS with DAYS_BIRTH) which is same for both the data set.

Conclusion

1. Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments. Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.

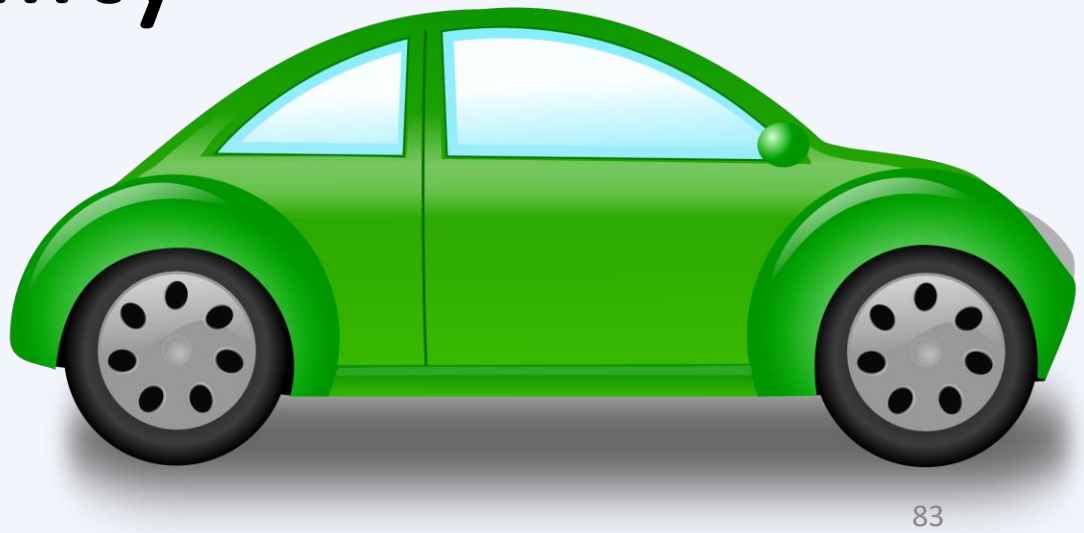
2. Although having higher number of rejection in loan purposes with 'Repairs' there are observed difficulties in payment on time.

3. Bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type ‘with parents’ , ‘House\apartment’ and ‘municipal apartment’ for successful payments.

4. People who get loan for 3-6 Lakhs tend to be default more than others and so increasing the interest rate is better for them.

5. The people having income 100000-200000 are having higher number of loans.

Project 7 - Analysing the Impact of Car Features on Price and Profitability



Project Description

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.

For the given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?

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

My approach for this project is to going through the data set and understanding the dataset thoroughly and then start analysing it. Firstly, when we check the dataset, it can be seen that the dataset contains over 11,000 car models information. Understanding the meaning of each column in the dataset is important to jot the insights for the data analysis.

Tech stack used : Microsoft Excel

Insights

Data cleaning:

- Firstly, I have gone through the whole data and understood that there were duplicates in it, so I have removed them. Second thing is, there where null values in some columns such as “Engine Fuel Type”, “Engine HP”, “Engine Cylinders” so I have replaced the null values with “unknown” and median values, so that it will be helpful while visualizing the data. After this process, I have set some goals to do the data analysis. Those are, Exploring trends in car features and pricing over time, Comparing the fuel efficiency of different types of cars, Investigating the relationship between a car's features and its popularity, Predicting the price of a car based on its features and market category. By obtaining the answers for these, we can have a clear vision on how to analyze and visualize the data.

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

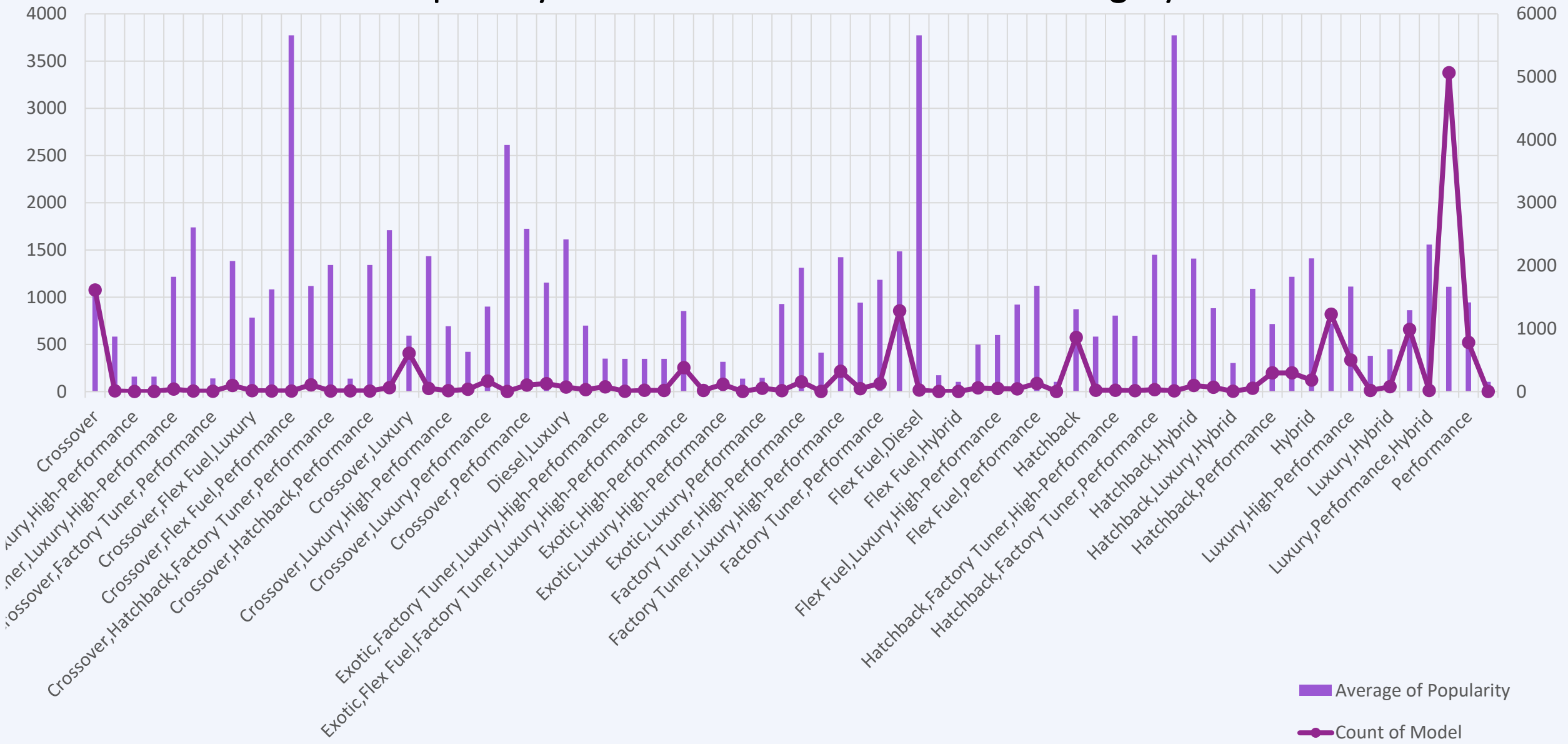
Market Category	Count of Model	Average of Popularity
Crossover	1075	1556.168372
Crossover,Diesel	7	873
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.461538
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Factory Tuner,Performance	4	210
Crossover,Flex Fuel	64	2073.75
Crossover,Flex Fuel,Luxury	10	1173.2
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover,Flex Fuel,Performance	6	5657
Crossover,Hatchback	72	1675.694444
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Luxury	7	204
Crossover,Hatchback,Performance	6	2009
Crossover,Hybrid	42	2563.380952
Crossover,Luxury	406	889.2142857
Crossover,Luxury,Diesel	34	2149.411765
Crossover,Luxury,High-Performance	9	1037.222222

Used Pivot table to analyse this insight. The no.of car models in each market category and their corresponding popularity is extracted and a graph is plotted.

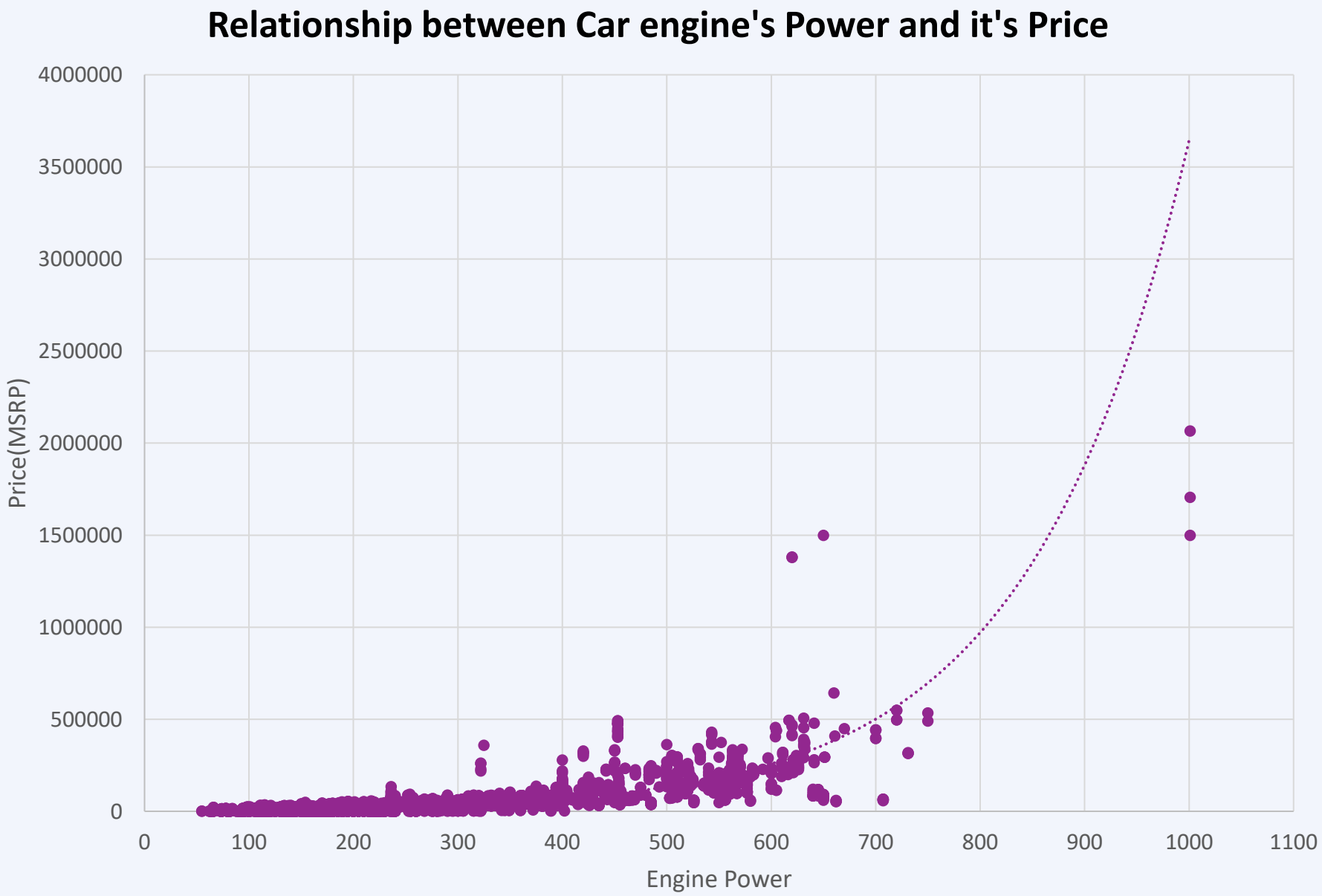
From the graph we can say that,

Hatchback, Flex Fuel, crossover are the highest populated and have the highest number of cars

Popularity of car model based on Market category



Insight Required: What is the relationship between a car's engine power and its price?



From the scatter chart we can say that,

The cars that have high engine power have higher prices. Thus, as the car’s power increases car’s price will also increase.

Insight Required: Which car features are most important in determining a car's price?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.679393674
R Square	0.461575764
Adjusted R Square	0.461287116
Standard Error	45164.91774
Observations	11199

ANOVA

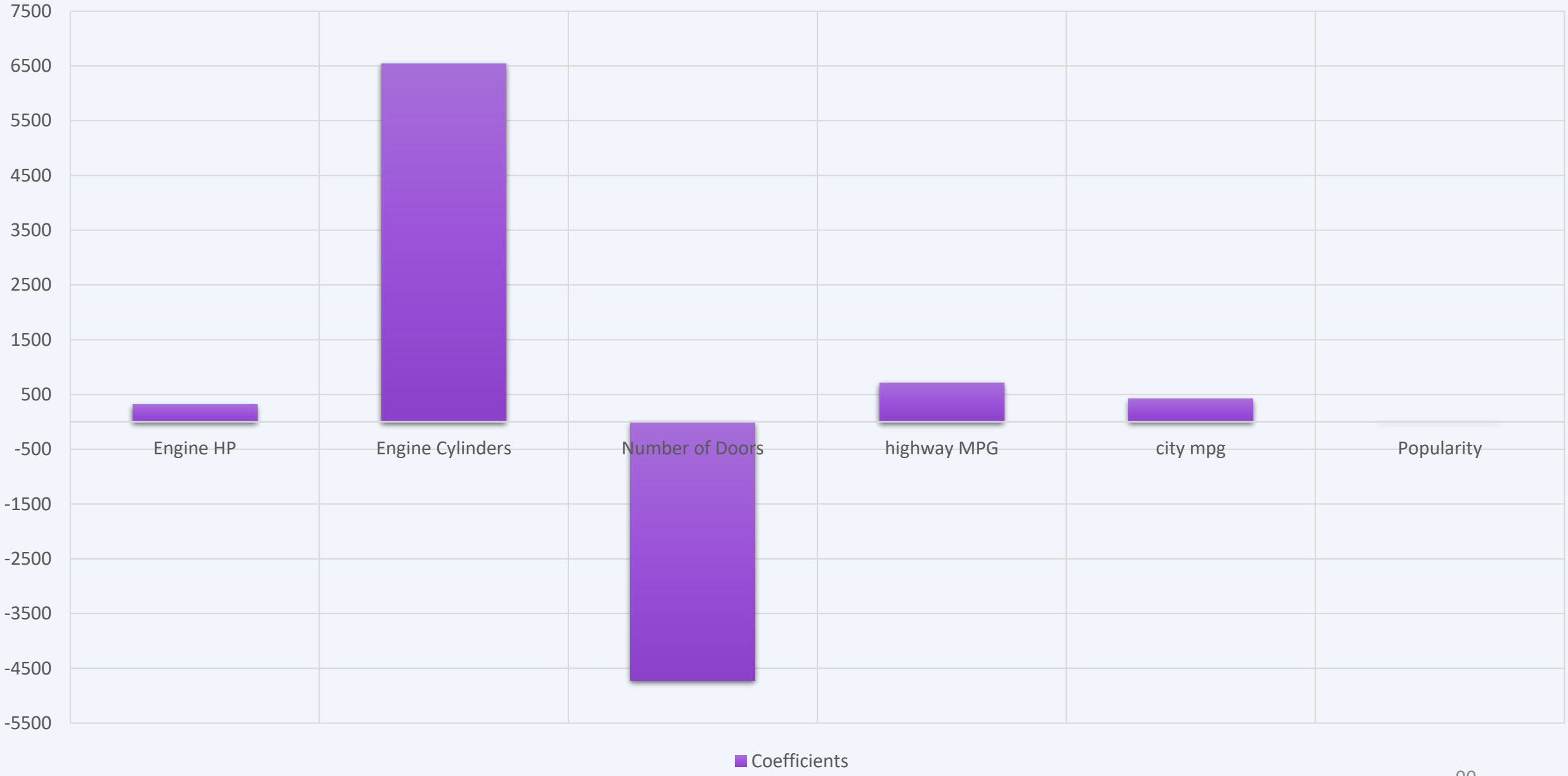
	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.95717E+13	3.26195E+12	1599.097143	0
Residual	11192	2.28302E+13	2039869795		
Total	11198	4.24019E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-80802.1188	3559.332189	-22.70148289	1.401E-111	-87779.03622	-73825.20138	-87779.03622	-73825.20138
Engine HP	316.859532	6.258747685	50.62666653	0	304.5912852	329.1277788	304.5912852	329.1277788
Engine Cylinders	6541.145155	445.1162651	14.69536314	1.94913E-48	5668.63895	7413.651361	5668.63895	7413.651361
Number of Doors	-4720.586227	495.3515027	-9.529770679	1.90094E-21	-5691.562338	-3749.610115	-5691.562338	-3749.610115
highway MPG	710.6823433	106.9848231	6.64283328	3.21979E-11	500.9732641	920.3914225	500.9732641	920.3914225
city mpg	415.2977565	100.9680929	4.113158371	3.93074E-05	217.3825272	613.2129857	217.3825272	613.2129857
Popularity	-3.464137718	0.29610483	-11.69902468	1.9644E-31	-4.04455529	-2.883720145	-4.04455529	-2.883720145

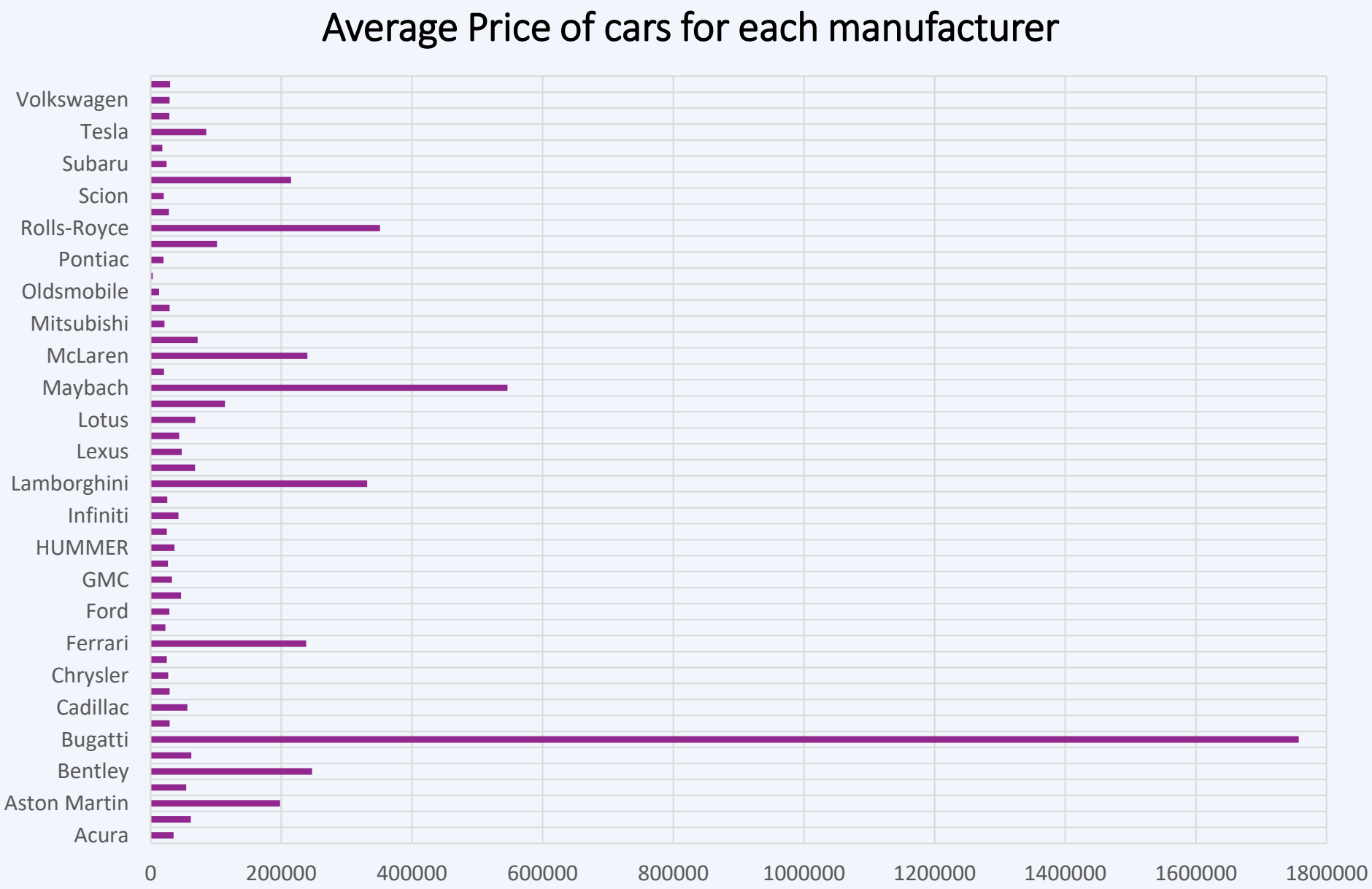
For the regression analysis, the variables Engine HP, Engine Cylinders, Number of doors, highway MPG, city mpg, Popularity are considered.

From the graph we can say that,
Engine cylinders have the positive relationship with the car’s price stating that it is the most important feature determining the car’s price. Number of doors have negative relationship with the car’s price.

Coefficients determining the Car's price



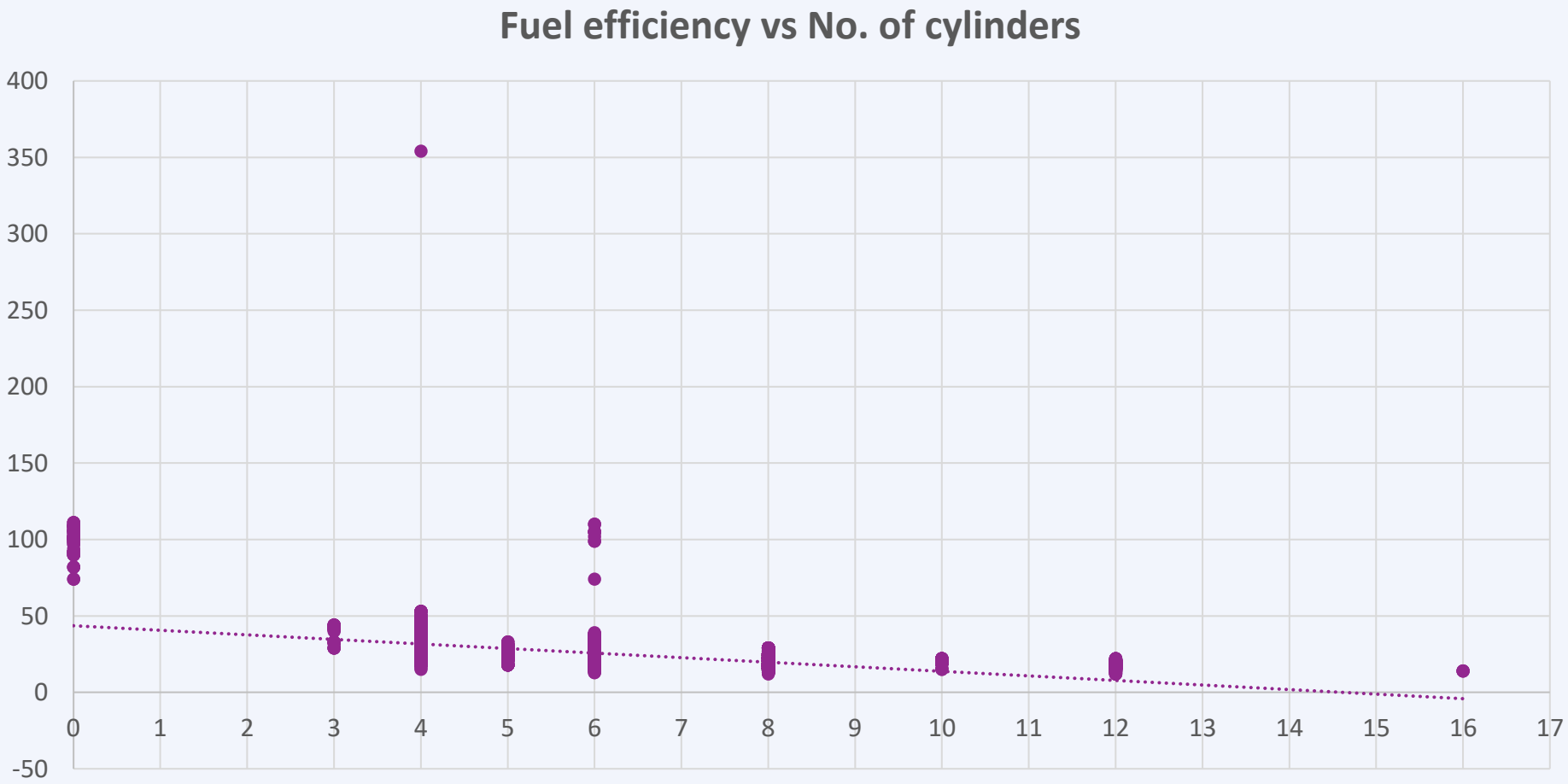
Insight Required: How does the average price of a car vary across different manufacturers?



From the bar chart we can say that,

Bugatti has the highest average price followed by Maybatch.

Insight Required: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?



The graph indicates negative trendline, which means decrease in the highway MPG with increase in the no. of cylinders.

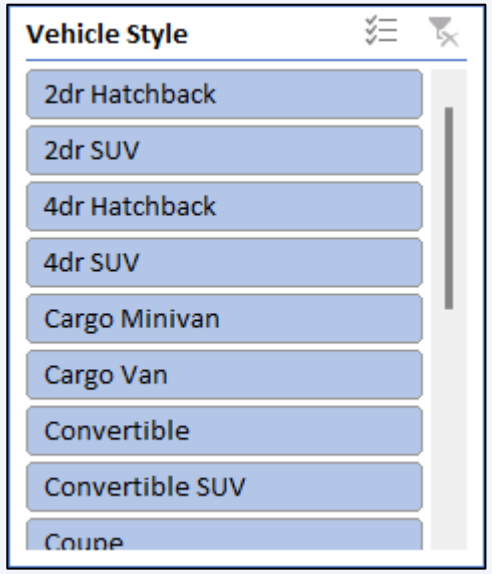
	Engine Cylinders	*highway MPG*
Engine Cylinders	1	
highway MPG	-0.596246019	1

After performing the correlation analysis, it can be seen that there is a negative correlation between Engine cylinders and highway MPG. This means that as the number of cylinders increases, fuel efficiency increases.

Building the Dashboard:

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

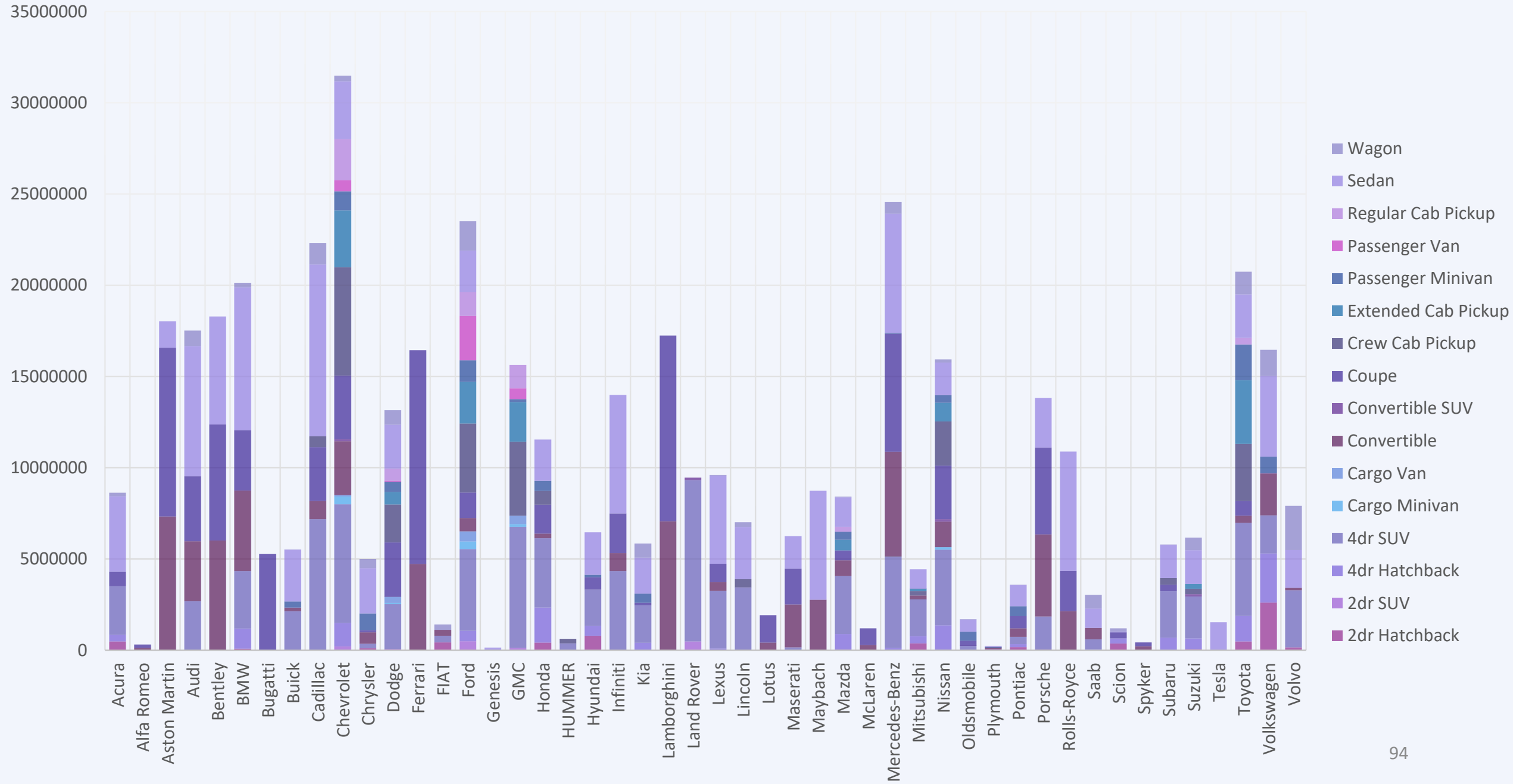
Sum of MSRP	Vehicle style																
Brand	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
Acura	480917		357440	2663505					793748						4134552	201360	8631522
Alfa Romeo							129800		178200								308000
Aston Martin							7321655		9258845						1448735		18029235
Audi	4000			2674900			3291405		3556290						7144348	847350	17518293
Bentley							6012870		6356760						5920900		18290530
BMW	80097		1103100	3160950			4403171		3304051						7829700	259600	20140669
Bugatti									5271671								5271671
Buick				2141770			179325		18534			330065			2838590	8212	5516496
Cadillac				7182555			985607		2953574	599150					9416847	1184100	22321833
Chevrolet	8000	193310	1287260	6509468	420150	74688	2953245	106300	3504525	5927617	3117951	1047240	599670	2260032	3177797	300675	31487928
Chrysler	98805			250545			630105		114510			922295			2479859	501075	4997194
Dodge	38000	12000	16000	2462875	60520	338497	6000		2973842	2072780	684682	557425	70708	653408	2409585	793055	13149377
Ferrari							4723811		11713289								16437100
FIAT	420715			369305			327965									287570	1405555
Ford	24000	467873	567615	4482771	415630	556351	730007		1398144	3782518	2285584	1179285	2429898	1299240	2279348	1623565	23521829
Genesis															139850		139850
GMC		128319		6633919	142750	460085				4062482	2175866	150630	599670	1284328			15638049
Honda	413200		1919260	3800589			252135		1588705	750215		553185			2264390		11541679



From the stacked column we can say that,

Chevrolet and Mercedes-Benz have the highest contribution to the car's price.

CAR PRICES BY BRAND AND BODY STYLE



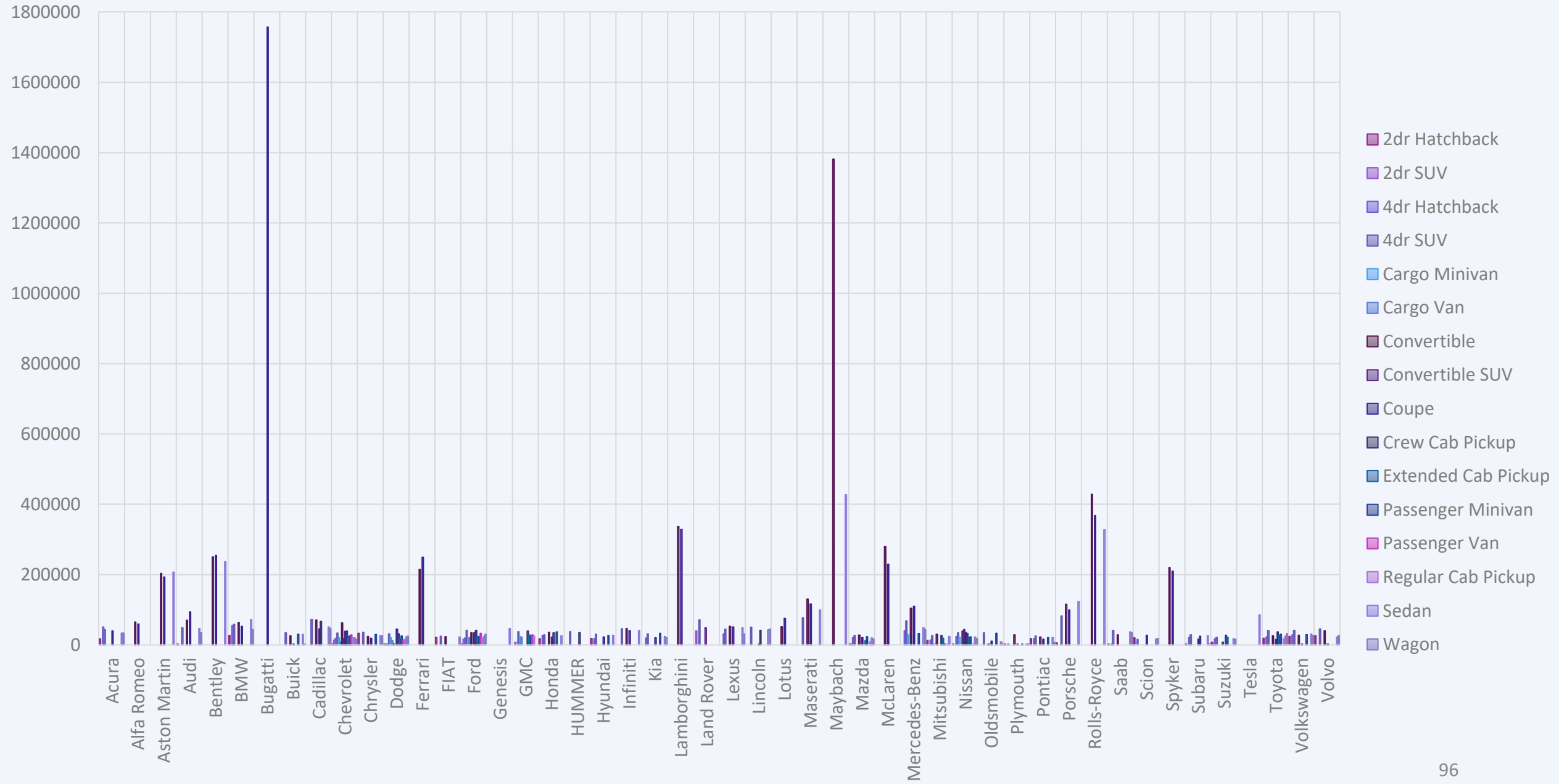
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

Average of MSRP	Body Style															
Brand	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	Grand Total
Acura	17175.60714		51062.85714	42959.75806					39687.4						33614.2439	33560	35087.4878
Alfa Romeo							64900		59400								61600
Aston Martin							203379.3056		192892.6042						206962.1429		198123.4615
Audi	2000			48634.54545			70029.89362		93586.57895						46391.87013	33894	54574.1215
Bentley							250536.25		254270.4						236836		247169.3243
BMW	26699		55155	58536.11111			63814.07246		52445.25397						71832.11009	43266.66667	62162.55864
Bugatti									1757223.667								1757223.667
Buick				33996.34921			25617.85714		2059.333333			30005.90909			29568.64583	2053	29034.18947
Cadillac				72551.06061			70400.5		45439.6	66572.22222					51178.5163	47364	56368.26515
Chevrolet	2000	13807.85714	18930.29412	33553.95876	20007.14286	8298.666667	62835	17716.66667	38939.16667	39255.74172	24170.16279	24934.28571	28555.71429	19824.84211	20635.04545	15825	29074.72576
Chrysler	32935			35792.14286			24234.80769		19085			29751.45161			26103.77895	26372.36842	26722.96257
Dodge	2000	2000	2000	31175.63291	20173.33333	12536.92593	2000		45058.21212	31405.75758	16301.95238	25337.5	14141.6	14850.18182	22519.48598	24782.96875	24857.04537
Ferrari							214718.6818		249218.9149								238218.8406
FIAT	21035.75			24620.33333			23426.07143									22120.76923	22670.24194
Ford	2000	16133.55172	19572.93103	41507.13889	19791.90476	20605.59259	34762.2381		34101.07317	41566.13187	23808.16667	23123.23529	32836.45946	17797.80822	23258.65306	30066.01852	28511.30788
Genesis															46616.66667		46616.66667
GMC		7128.833333		37479.76836	23791.66667	21908.80952				39062.32692	27895.71795	25105	28555.71429	25182.90196			32444.08506
Honda	17216.66667		26656.38889	28575.85714			36019.28571		21763.08219	34100.68182		36879			26027.47126		26655.14781
HUMMER				37749						34629.28571							36464.41176
Hyundai	18363.95349		17629.33333	30218.0303					22126.45161			26615			27666.5119		24926.26255

From the clustered chart we can say that,

The couple style Bugatti and the Convertible style of Maybach have the highest average car prices.

Price(MSRP) of Car Brands with respect to Body Style



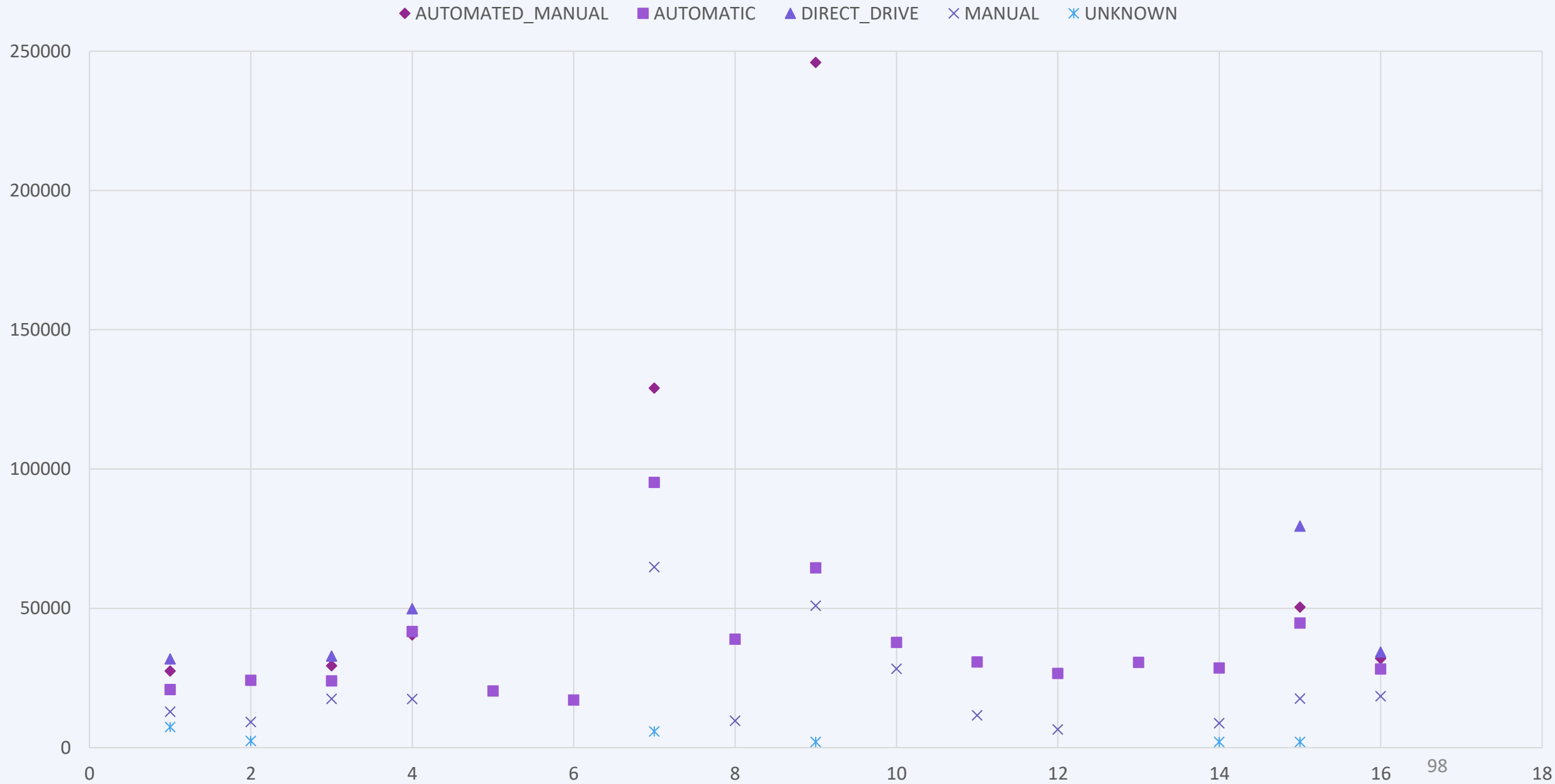
Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

Average of MSRP	Transmission Type				
Body Style	AUTOMATED_MANUAL	AUTOMATIC	DIRECT_DRIVE	MANUAL	UNKNOWN
2dr Hatchback	27470.41667	20784.09901	31800	12840.65556	7361.5
2dr SUV		24153.60606		9173.018519	2371
4dr Hatchback	29347.04545	23888.73529	32799.72973	17500.36364	
4dr SUV	40451.15385	41638.26534	49800	17422.08791	
Cargo Minivan		20315.59322			
Cargo Van		17019.29762			
Convertible	129082.2339	95153.3131		64794.34437	5783.5
Convertible SUV		38925.5		9594.8	
Coupe	245977.4252	64523.41955		50901.4973	2000
Crew Cab Pickup		37718.95307		28233.10811	
Extended Cab Pickup		30711.45251		11553.29707	
Passenger Minivan		26589.50919		6510	
Passenger Van		30578.06612			
Regular Cab Pickup		28536.8239		8759.454054	2000
Sedan	50385.39326	44671.35638	79512.25	17557.26441	2000
Wagon	31985.27778	28219.45742	34250	18398.57813	
Grand Total	**108718.9873**	**41816.12431**	**47351.25**	**28267.91989**	**3647.833333**

From the scatter chart we can say that,

The automated convertible and automatic are highly contributing in MSRP.

RELATIONSHIP BETWEEN MSRP(PRICE) AND TRANSMISSION TYPE



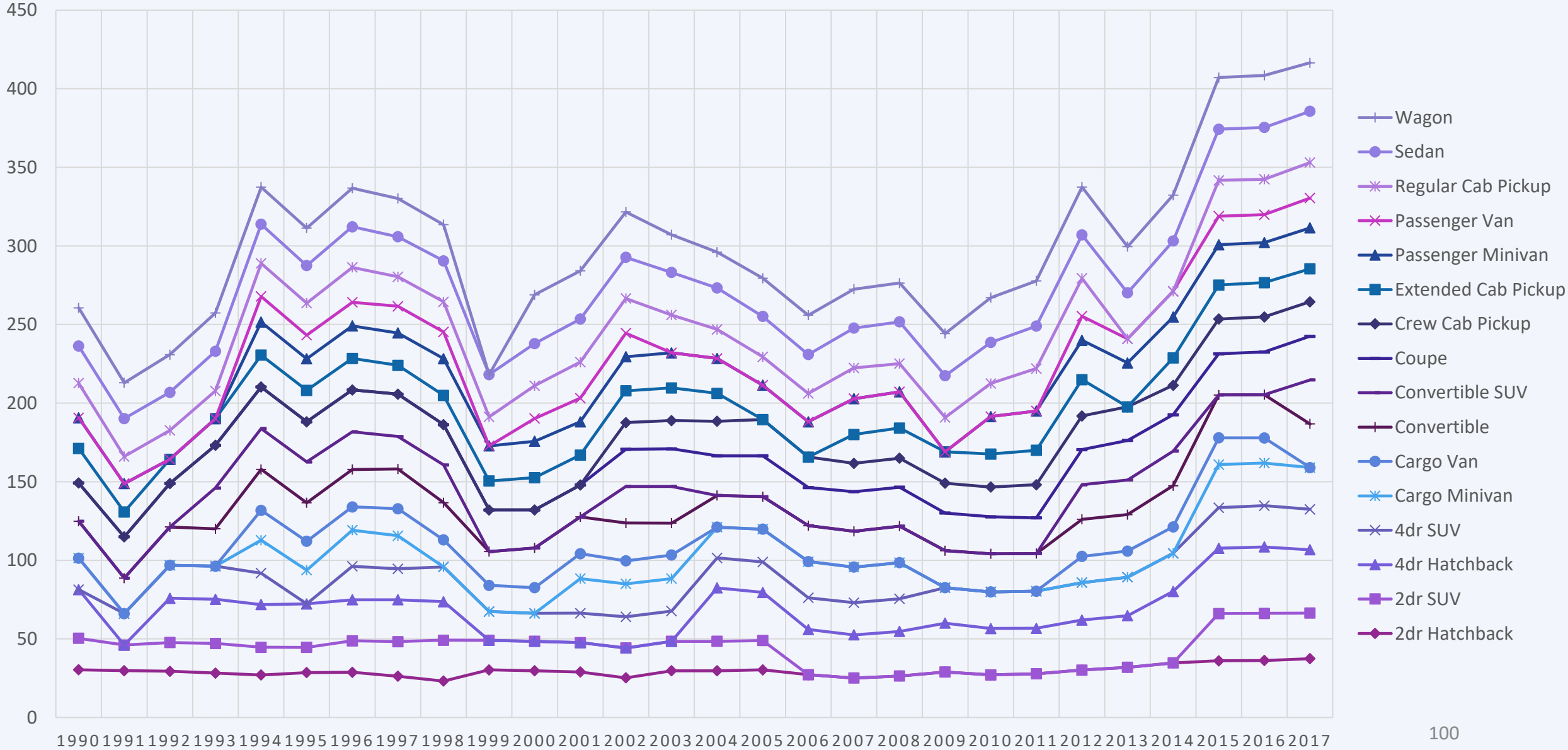
Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

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
1990	30.4	20	31		20		23.5		24.27272727		22	19.5		22	23.65384615	24.3
1991	29.83333333	16.25		20			22.625		26.25		15.83333333	18		17.28571429	24.11111111	22.72727273
1992	29.39285714	18.28571429	28.16666667	21			24.375		27.63636364		15.4			18.42857143	24.175	24
1993	28.25925926	18.85714286	28.125	21			23.81818182	26	27.13636364		16.90909091			17.625	25.2	24.41666667
1994	27.05263158	17.625	27.14285714	20	21	19	26	26	26.42857143		20.28571429	21	16.33333333	21	24.90625	23.63636364
1995	28.6	16	27.66666667		21.5	18.33333333	24.5	26	25.51724138		20	20.1	15	20.375	23.88461538	23.88888889
1996	28.8	20	26.125	21.25	23	14.8	23.8	24	26.6		20	20.77777778	15	22.2	25.83333333	24.57142857
1997	26.25	22	26.66666667	19.7	21	17.2	25.28571429	20.66666667	26.92857143		18.35714286	20.55555556	17	18.78571429	25.42105263	24.4
1998	23.2	26	24.5	22.11111111		17.2	23.66666667	24	25.64285714		18.625	23.33333333	17	19.15151515	26.1	23
1999	30.33333333	18.75		18.3		16.66666667	21.5		26.5		18.42307692	22.33333333		18.42857143	26.875	
2000	29.72727273	18.75		17.73333333		16.4	25.28571429		24.16666667		20.5	23.16666667	14.5	20.83333333	26.86363636	31
2001	29	18.66666667		18.72727273	22	15.8	23.4375		20.29411765		19	21.2	15	23	27.37735849	30.625
2002	25.25	19		19.79411765	21	14.6	24.07142857	23.28571429	23.6	17	20.22222222	21.6875	15	22.06666667	26.14	28.88888889
2003	29.75	18.75		19.22857143	20.66666667	15	20.23076923	23.4	23.87878788	18	20.77777778	22.2972973		24.08333333	27.05769231	24
2004	29.71428571	18.75	34	19.04081633	19.6		20.1		25.26666667	22	17.75	22.2		18.46153846	26.36231884	22.8
2005	30.33333333	18.66666667	30.6	19.33333333	20.85714286		20.72727273		26	23		21.88888889		18	25.75409836	24.27777778
2006	27.25		28.75	20.19444444	23		22.85714286		24.25925926	19.38461538		22.45		18	24.75	25
2007	25.09090909		27.45454545	20.46296296	22.66666667		22.76		25.2	18.03333333	18.38983051	22.75		19.57692308	25.30769231	24.8
2008	26.42857143		28.33333333	20.765625	23		23.19047619		24.78947368	18.43181818	19.22222222	23		18	26.52173913	24.71428571
2009	29		31	22.59139785			23.55		23.89473684	19.02941176	19.95454545			21.85714286	26.54545455	26.84848485
2010	27.125		29.5	23.25454545			24.26315789		23.52173913	18.94594595	21	23.85714286		21	26.12307692	28.47826087
2011	27.83333333		28.93103448	23.58333333			23.94444444		22.67857143	21.1	21.9	25		27	27.01298701	28.73333333
2012	30.21428571		31.76190476	23.84444444		16.66666667	23.57692308	22	22.37142857	21.43333333	23.0625	25	15.33333333	24.125	27.60493827	30.44117647
2013	31.90909091		32.8627451	24.47368421		16.66666667	23.18181818	22	25.18604651	21.31818182		28	15.33333333		29.234375	29.39473684
2014	34.75		45.46808511	24.2231405		16.85714286	26.15789474	22	23.16883117	18.71428571	17.4	26	16.375		31.99065421	29.25
2015	36.10294118	30	41.57638889	25.76951673	27.5	17	27.22377622		26.16931217	22.12121212	21.65934066	25.65384615	18.14285714	22.74285714	32.64007092	32.83333333
2016	36.26530612	30	42.28	26.1965812	27.11111111	16	27.51666667		27.10989011	22.37593985	21.78409091	25.5	17.71428571	22.52941176	33	33.08333333
2017	37.4375	29	40.29411765	25.73739496	26.5		27.80263158	28	27.71724138	21.96153846	20.98684211	26.05555556	19	22.52941176	32.61589404	30.86486486
Grand Total	**31.33737864**	**19.52747253**	**37.81146305**	**24.51725555**	**23.69491525**	**16.54761905**	**25.48043185**	**23.71428571**	**25.63744681**	**21.11908397**	**20.22613065**	**23.58868895**	**17.25619835**	**20.85507246**	**30.17727752**	**28.40213523**

From the line chart we can say that,

Fuel efficiency (highway MPG) is increased over years of all cars across various body styles..

FUEL EFFICIENCY OF CARS AROSS DIFFERENT BODY STYLES AND MODEL YEARS



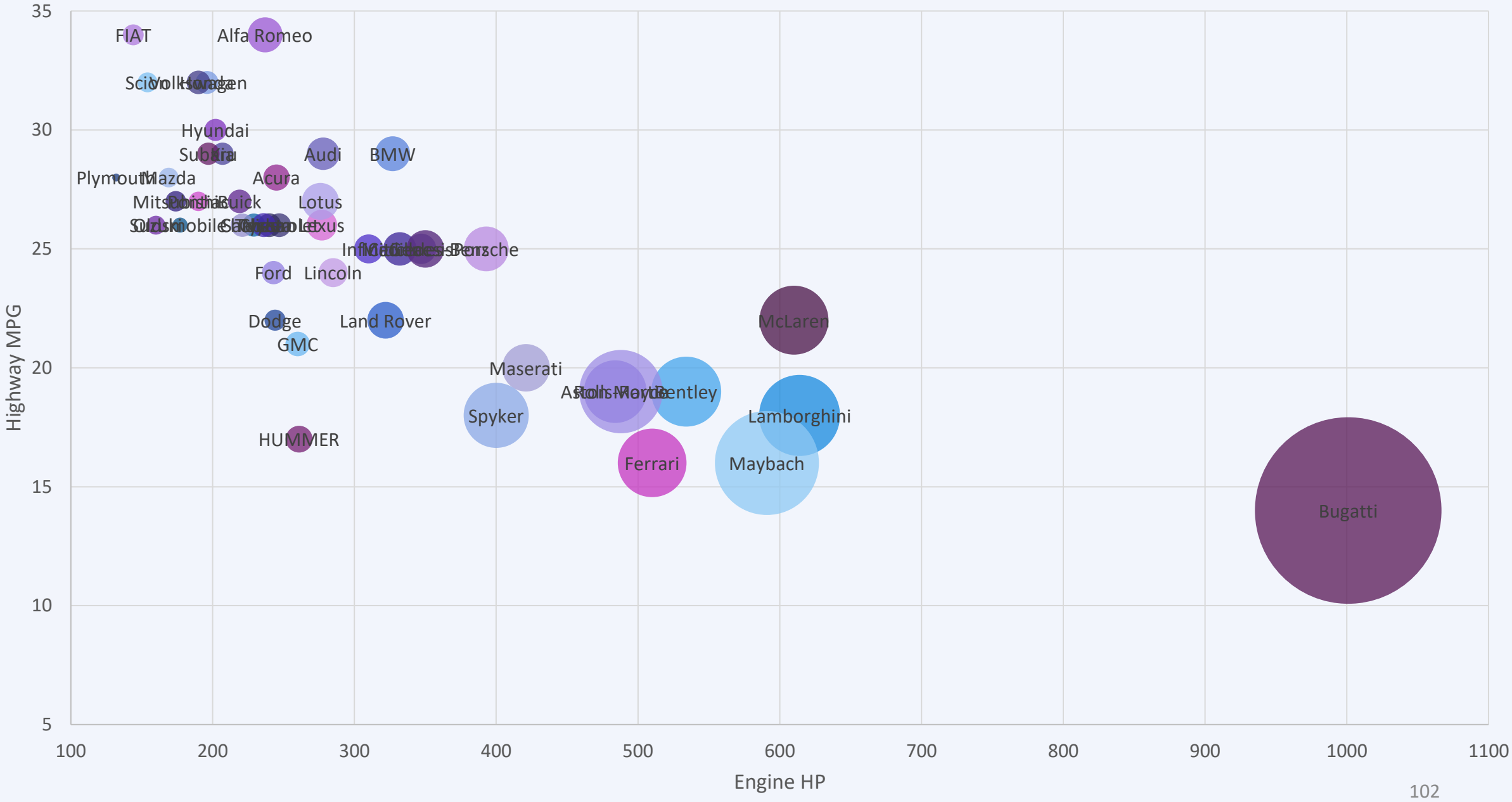
Task 5: How does the car's horsepower, MPG, and price vary across different Brands?

Car Brand	Average of Engine HP	Average of Highway MPG	Average of MSRP
Acura	245	28	34,888
Alfa Romeo	237	34	61,600
Aston Martin	484	19	1,97,910
Audi	278	29	53,452
Bentley	534	19	2,47,169
BMW	327	29	61,547
Bugatti	1001	14	17,57,224
Buick	219	27	28,207
Cadillac	332	25	56,231
Chevrolet	247	26	28,273
Chrysler	229	26	26,723
Dodge	244	22	22,390
Ferrari	510	16	2,37,384
FIAT	144	34	22,206
Ford	243	24	27,393
Genesis	347	25	46,617
GMC	260	21	30,493
Honda	196	32	26,630
HUMMER	261	17	36,464
Hyundai	202	30	24,597
Infiniti	310	25	42,394
Kia	207	29	25,112

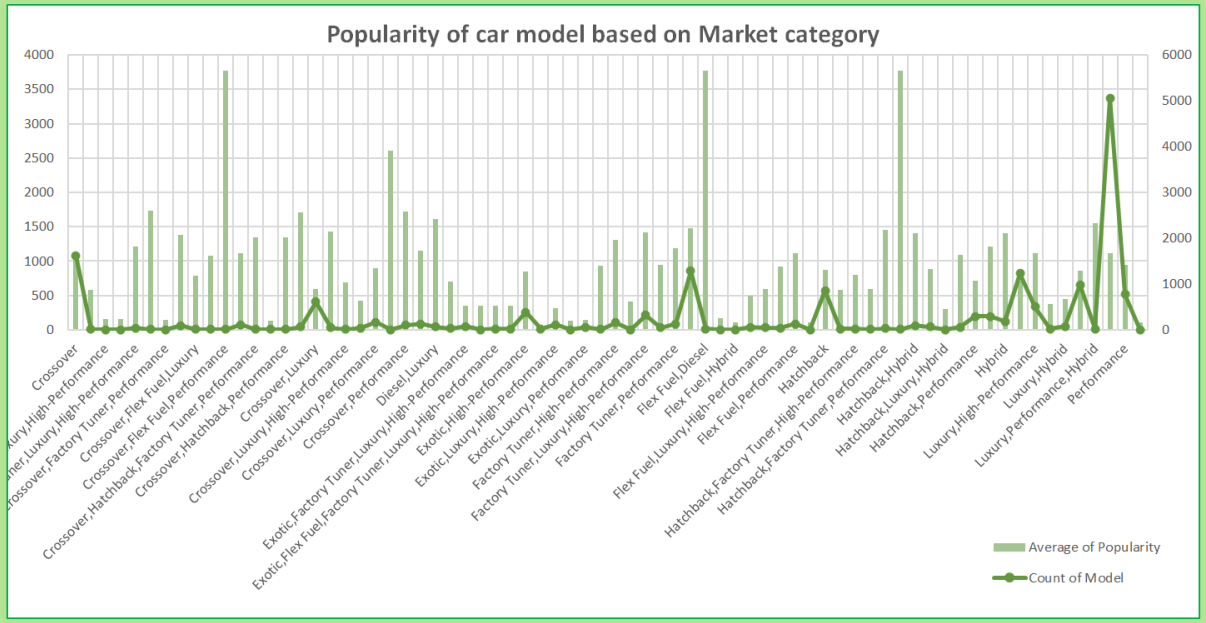
From the Bubble chart we can say that,

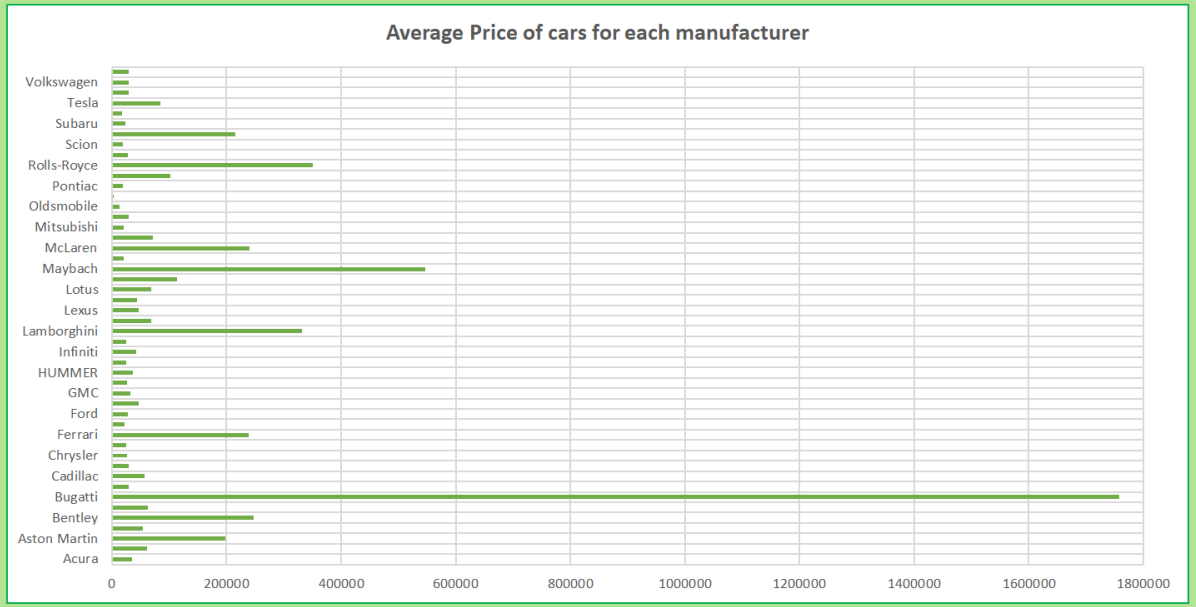
If Engine HP increases, Highway MPG will decrease and the price(MSRP) will also increase.

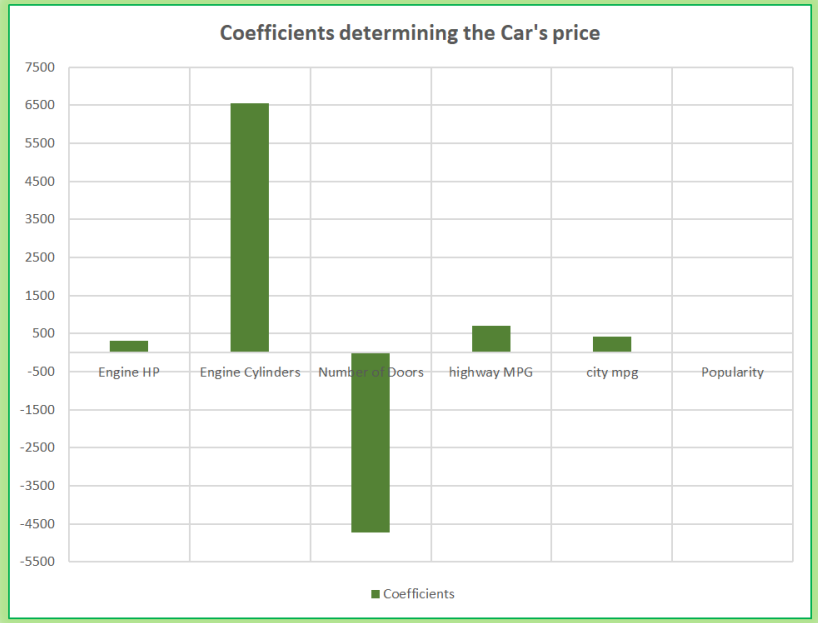
Relation between Engine HP, MPG and Car price across different brands

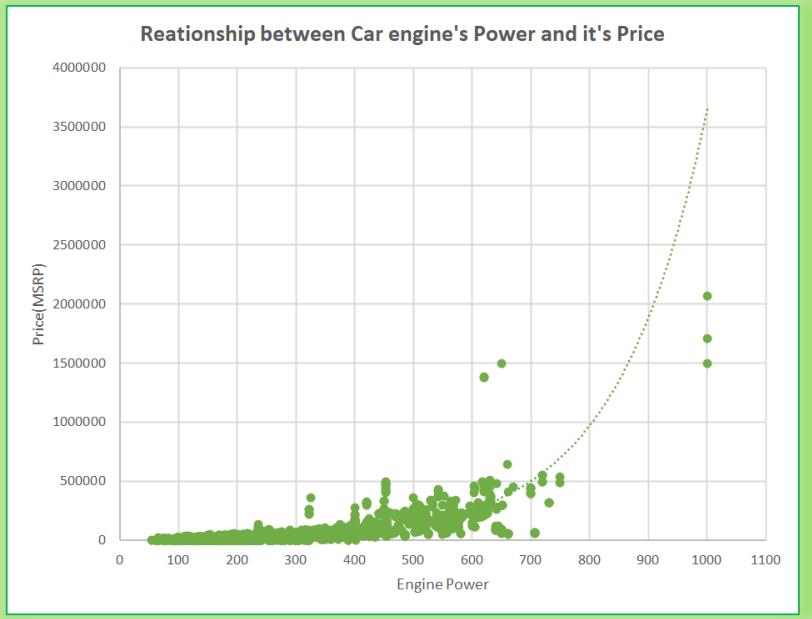


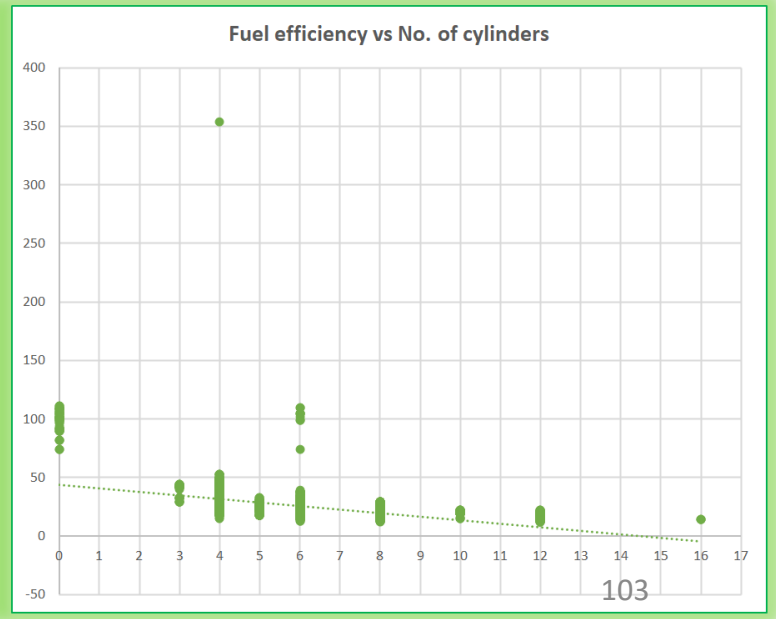
Analyzing the Impact of Car Features on Price and Profitability(Dashboard 1)



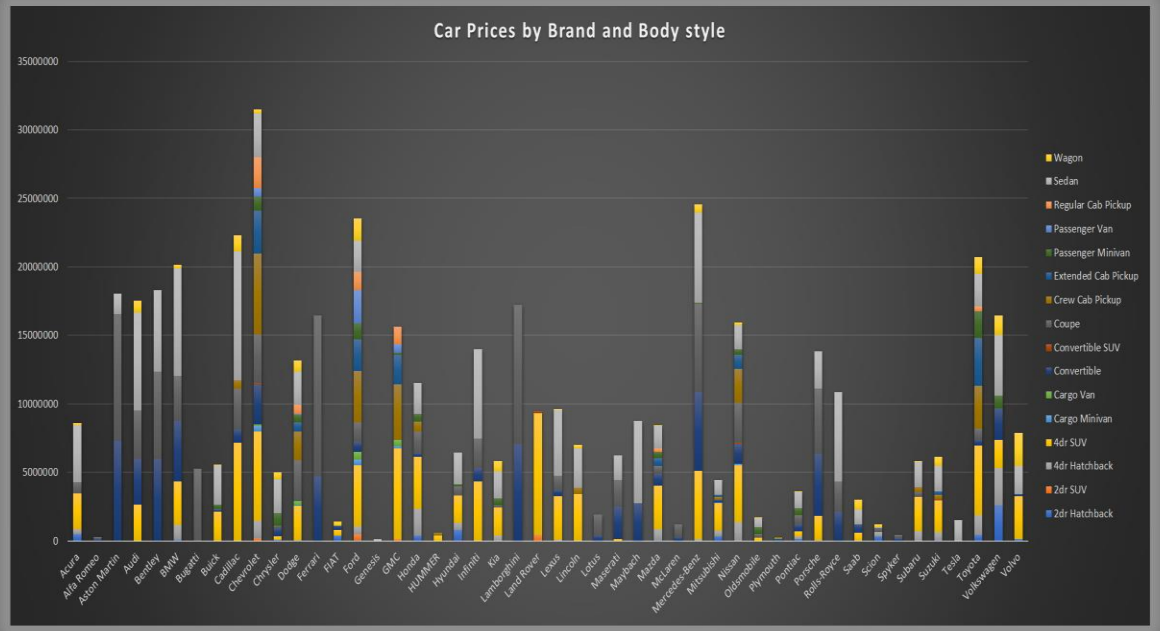


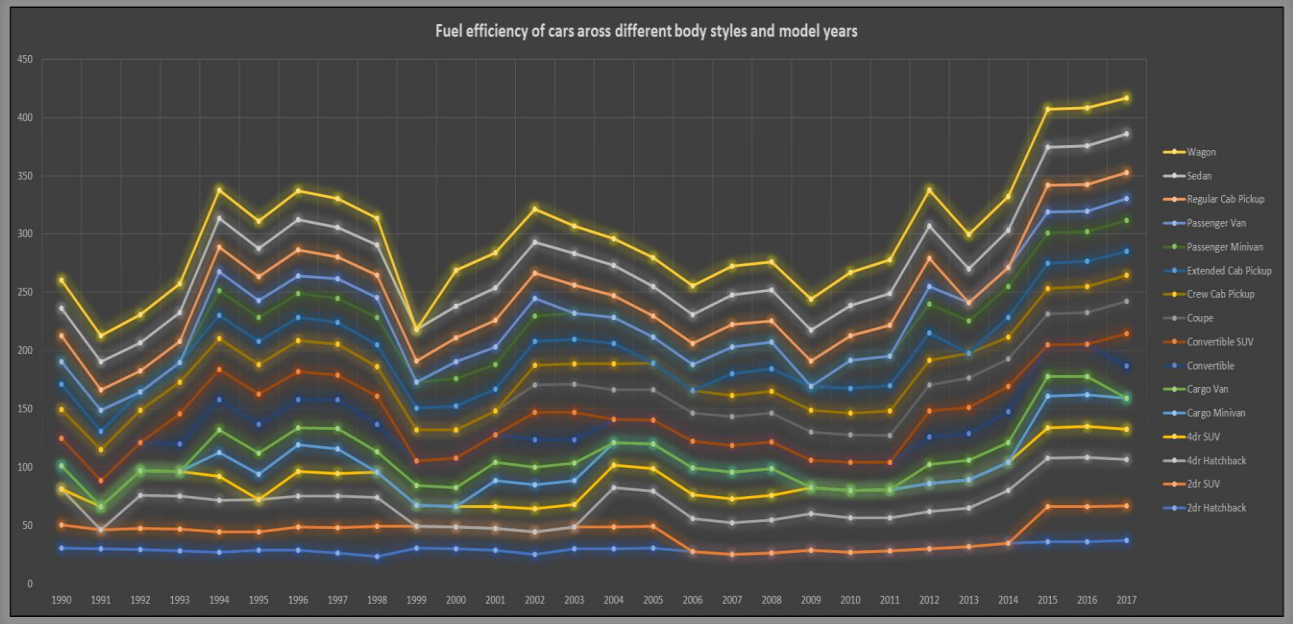


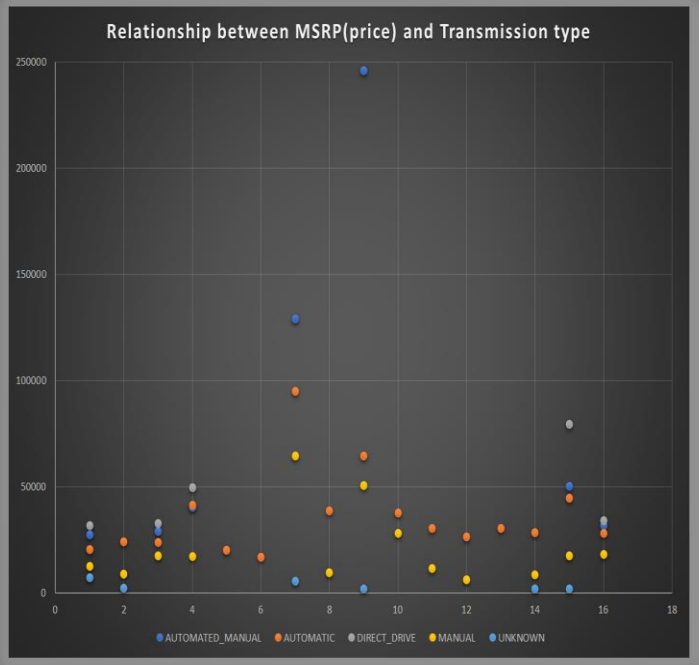


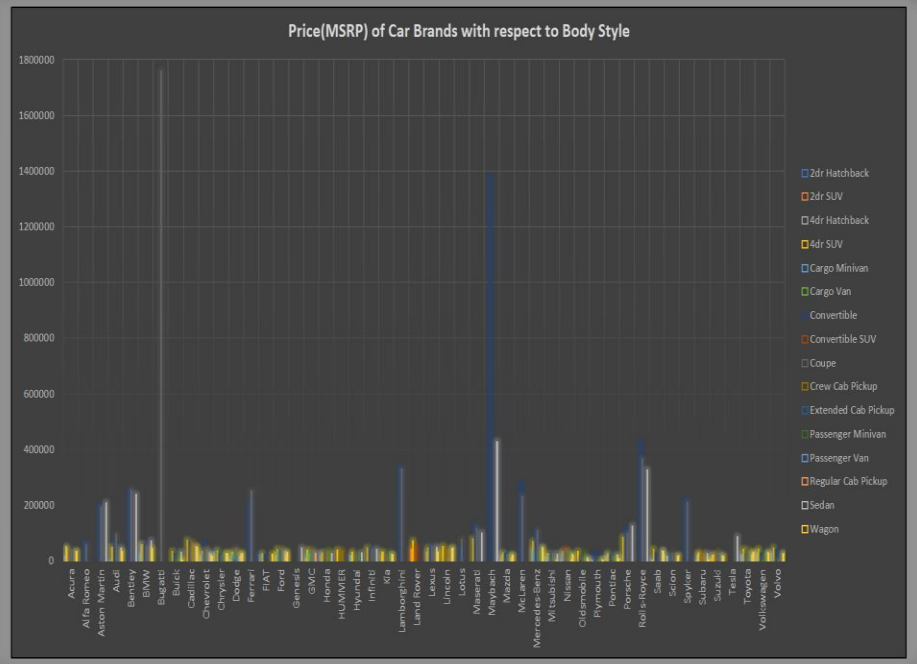


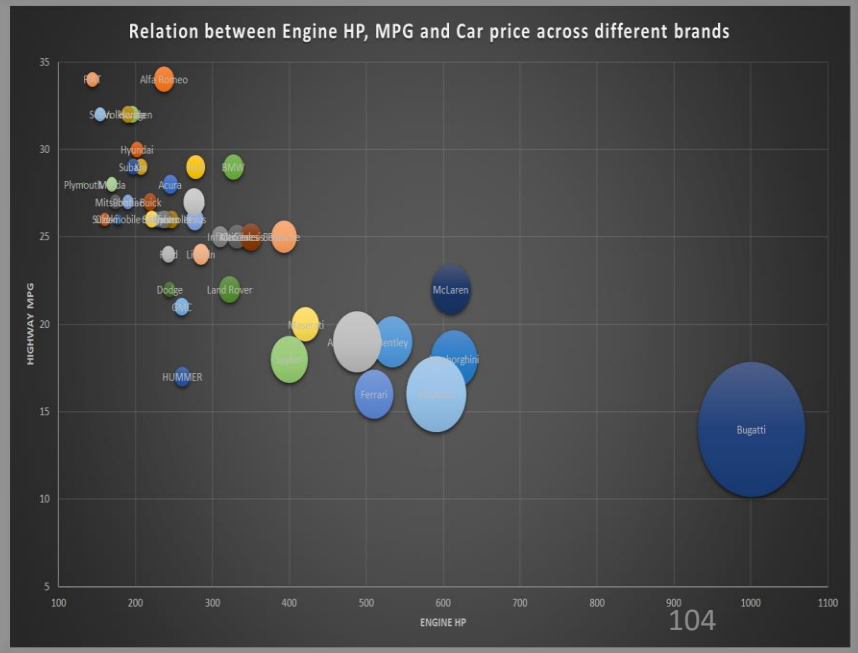
Analyzing the Impact of Car Features on Price and Profitability(Dashboard 2)







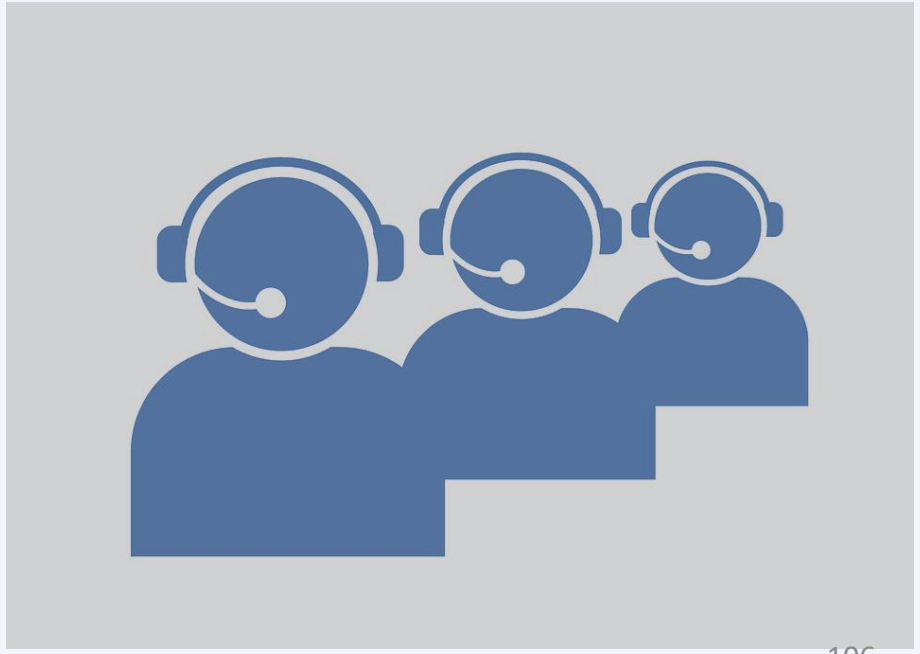




Conclusion

- Hatchback, Flex Fuel, crossover are the highest populated and have the highest number of cars
- The cars that have high engine power have higher prices. Thus, as the car's power increases car's price will also increase.
- Engine cylinders have the positive relationship with the car's price stating that it is the most important feature determining the car's price. Number of doors have negative relationship with the car's price.
- Bugatti has the highest average price followed by Maybatch. Chevrolet and Mercedes-Benz have the highest contribution to the car's price.
- The couple style Bugatti and the Convertible style of Maybach have the highest average car prices. The automated convertible and automatic are highly contributing in MSRP.
- If Engine HP increases, Highway MPG will decrease and the price(MSRP) will also increase.
- In transmission type automated_manual creates high impact because in a single car having both automated and manual gear systems will be more beneficial rather than a single gear system.
- Companies need to produce high or at least good fuel efficiency of cars by which the majority of the class can afford a car.

Project 8 - ABC Call Volume Trend Analysis



Project Description

Advertising is a crucial aspect of any business. It helps increase sales and makes the audience aware of the company's products or services. The first impressions of a business are often formed through its advertising efforts.

The target audience for businesses can be local, regional, national, or international. Various types of advertising are used to reach these audiences, including online directories, trade and technical press, radio, cinema, outdoor advertising, and national papers, magazines, and TV.

The advertising business is highly competitive, with many players bidding large amounts of money to target the same audience segment. This is where the company's analytical skills come into play. The goal is to identify those media platforms that can convert audiences into customers at a low cost. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business.

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

Tech stack used : Microsoft Excel

Average Call Duration:

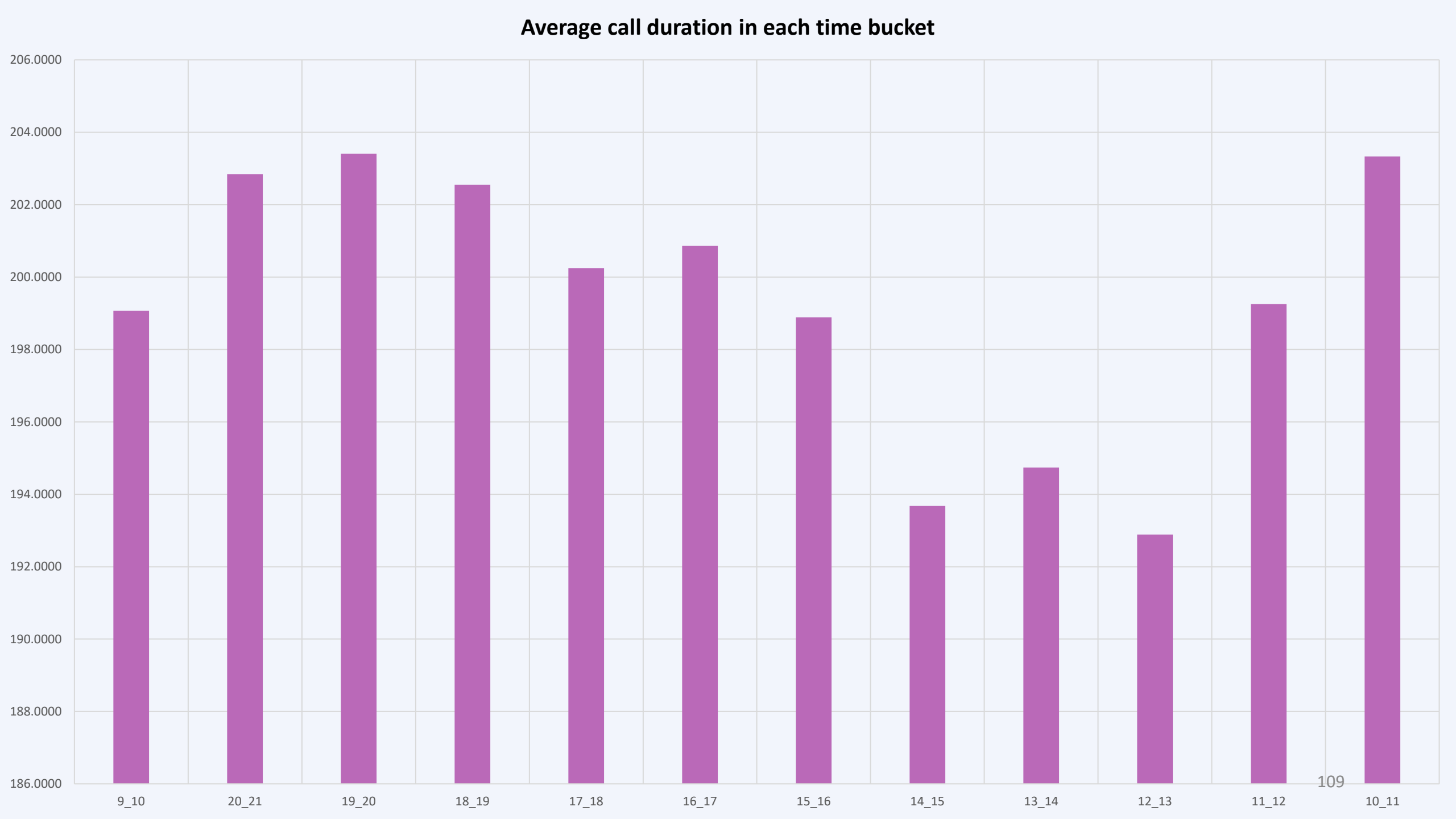
What is the average duration of calls for each time bucket?

Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
9_10	199.0691
20_21	202.8460
19_20	203.4061
18_19	202.5510
17_18	200.2488
16_17	200.8682
15_16	198.8889
14_15	193.6771
13_14	194.7402
12_13	192.8888
11_12	199.2550
10_11	203.3310
Grand Total	**198.6227745**

I have used pivot table to find out the average call_seconds which are “Answered” and plotted a graph to visualize it.

From the graph we can say that,

Most of the calls are answered during 10am -11am and 7pm – 8pm.


Average call duration in each time bucket
206.0000
204.0000
202.0000
200.0000
198.0000
196.0000
194.0000
192.0000
190.0000
188.0000
186.0000
9_10
20_21
19_20
18_19
17_18
16_17
15_16
14_15
13_14
12_13
11_12
10_11

Call Volume Analysis:

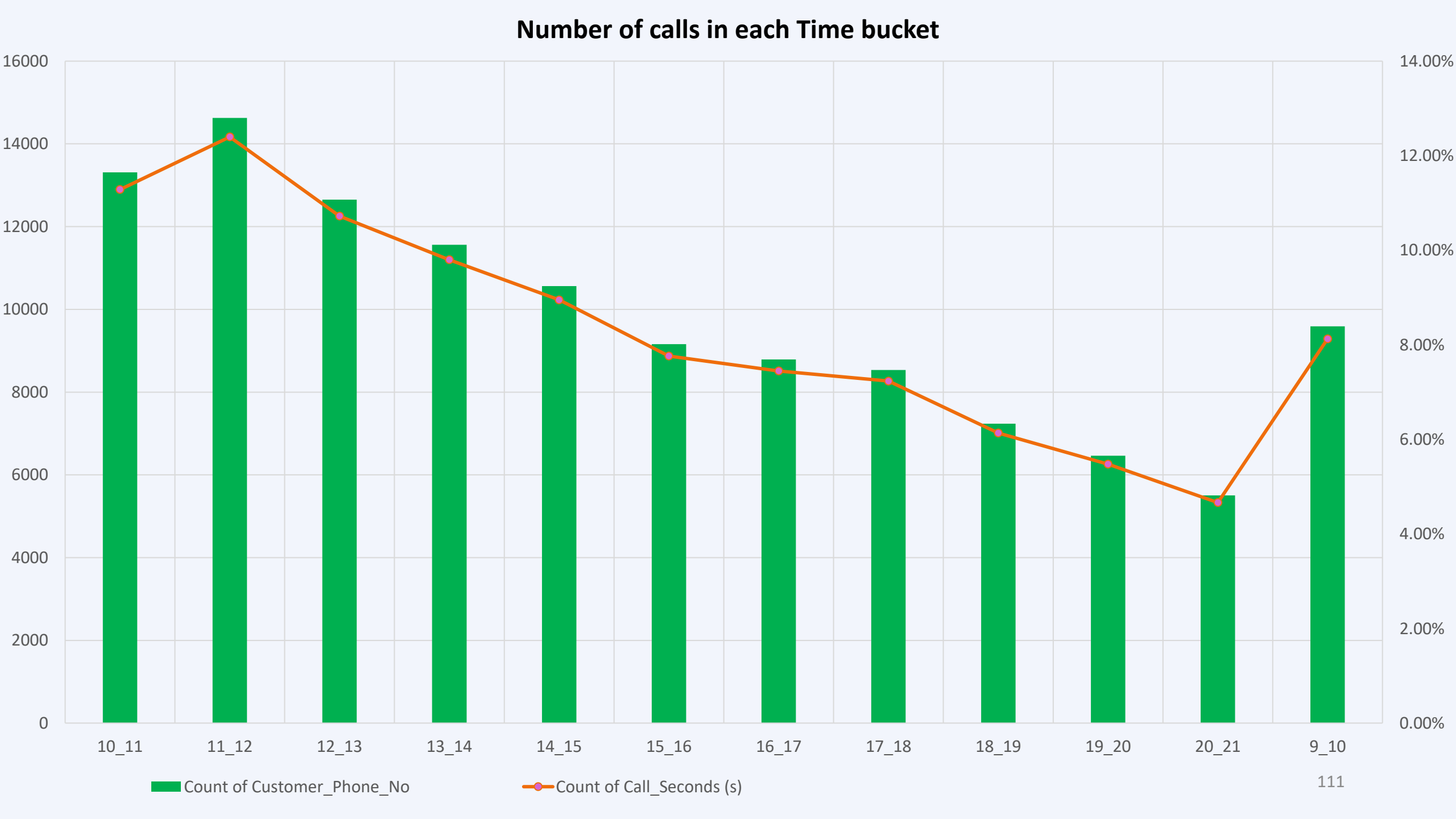
Create a chart or graph that shows the number of calls received in each time bucket?

Time bucket	Count of Customer_Phone_No	Count of Call_Seconds (s)
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%
9_10	9588	8.13%
Grand Total	**117988**	**100.00%**

I have used pivot table to find out the number of calls received and plotted a graph to visualize it.

From the graph we can say that,

Most of the calls are received during 11am-12pm.

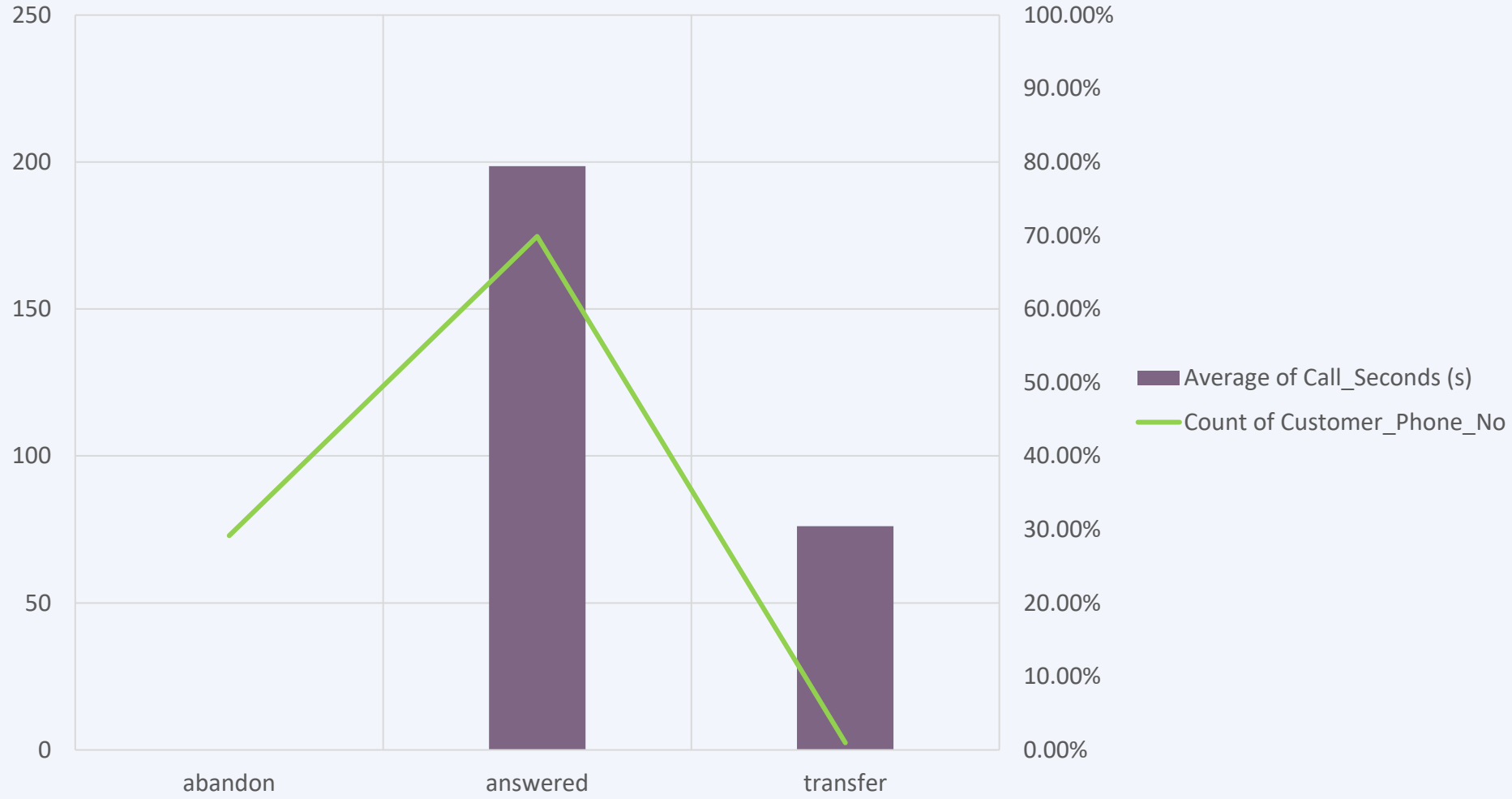

Number of calls in each Time bucket
16000
14000
12000
10000
8000
6000
4000
2000
0
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
10_11
11_12
12_13
13_14
14_15
15_16
16_17
17_18
18_19
19_20
20_21
9_10
Count of Customer_Phone_No
Count of Call_Seconds (s)

Manpower Planning:

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_No
abandon	0	29.16%
answered	198.6227745	69.88%
transfer	76.14651368	0.96%
Grand Total	**139.5321473**	**100.00%**

Firstly I have found out the average call_seconds and count of customer_phone_no based on the call status and plotted a graph to visualize it and get a better understanding to move further.



Assumptions

- An agent works 6 days a week
- On average each agent takes 4 unplanned leaves

Total working hours	9 hrs
Break(lunch & snacks)	1.5 hrs
60% of their total actual working hours	4.5
Total no. of days in a month	30

Date_&_Time	(Multiple Items)	
Row Labels	**Sum of Call_Seconds (s)**	**Sum of hours**
01-Jan	**676664**	187.9622
Grand Total	**676664**	

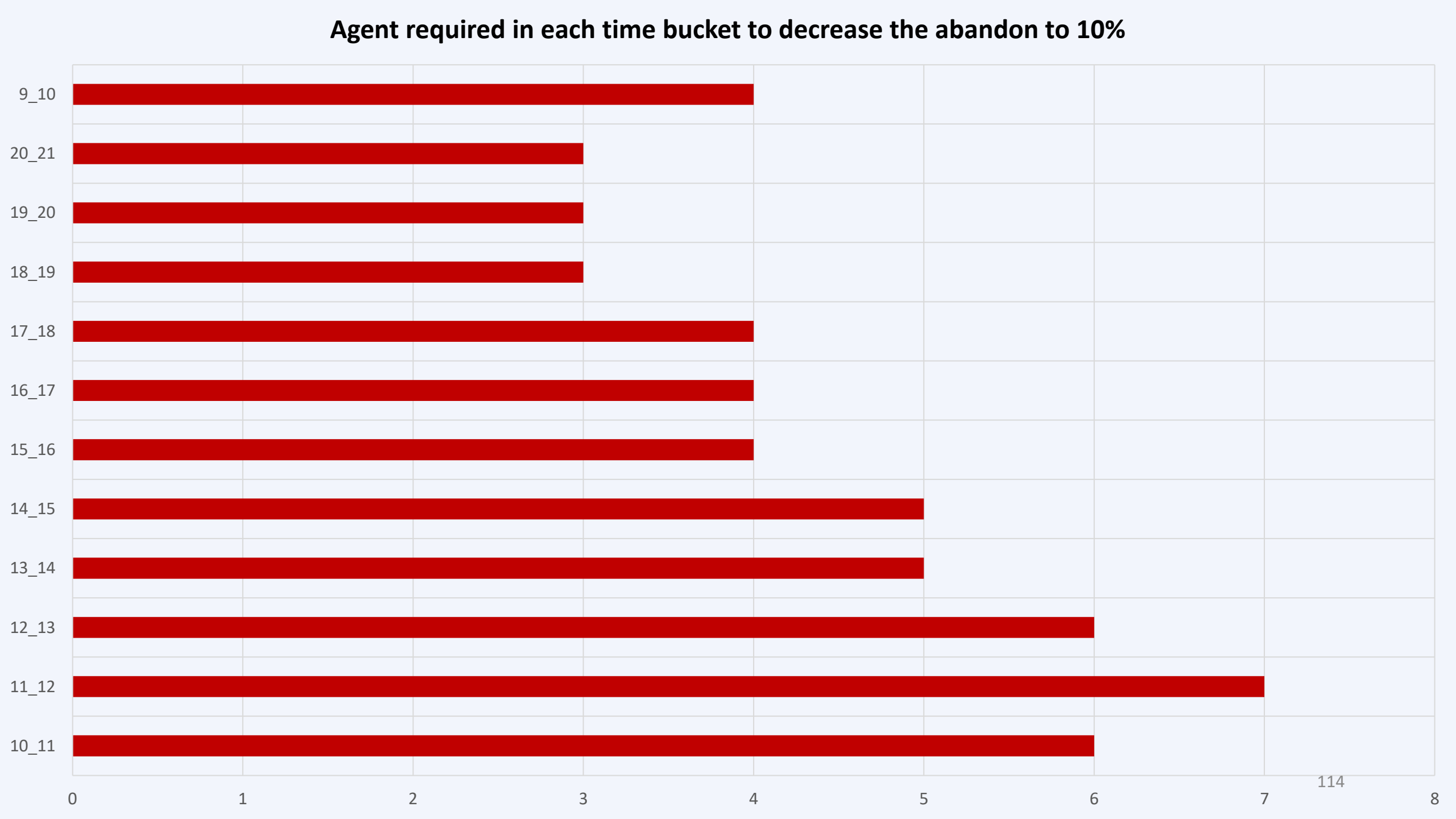
Then I have found out the sum of call_seconds in one day. After extracting it, I have found out the sum of hours by dividing it by 3600

Total no.of hours worked by the agent in a day	187.9622
60% of the agent's work(i.e 30% abandon rate)	41.76938
In order to avoid the abandon rate to 10 %90% of the agent's work should be utilized	
In order to decrease the abondon rate to 10%12 agents(54-42) are required.	

Unitary Method	
If	Then
70	41.76938
90	x
x	53.70351

After getting the value of x, I have distributed the no.of agents required I each time bucket based on the percentage of call_seconds to avoid the rejection of calls and decrease the abandon rate to 10%

Time_bucket	Percent Count of Call_Se	Agent required
10_11	0.11	6
11_12	0.12	7
12_13	0.11	6
13_14	0.10	5
14_15	0.09	5
15_16	0.08	4
16_17	0.07	4
17_18	0.07	4
18_19	0.06	3
19_20	0.05	3
20_21	0.05	3
9_10	0.08	4
Grand Total	**1.00**	**54**


Agent required in each time bucket to decrease the abandon to 10%
9_10
20_21
19_20
18_19
17_18
16_17
15_16
14_15
13_14
12_13
11_12
10_11
0
1
2
3
4
5
6
7
8
114

Night Shift Manpower Planning:

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

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

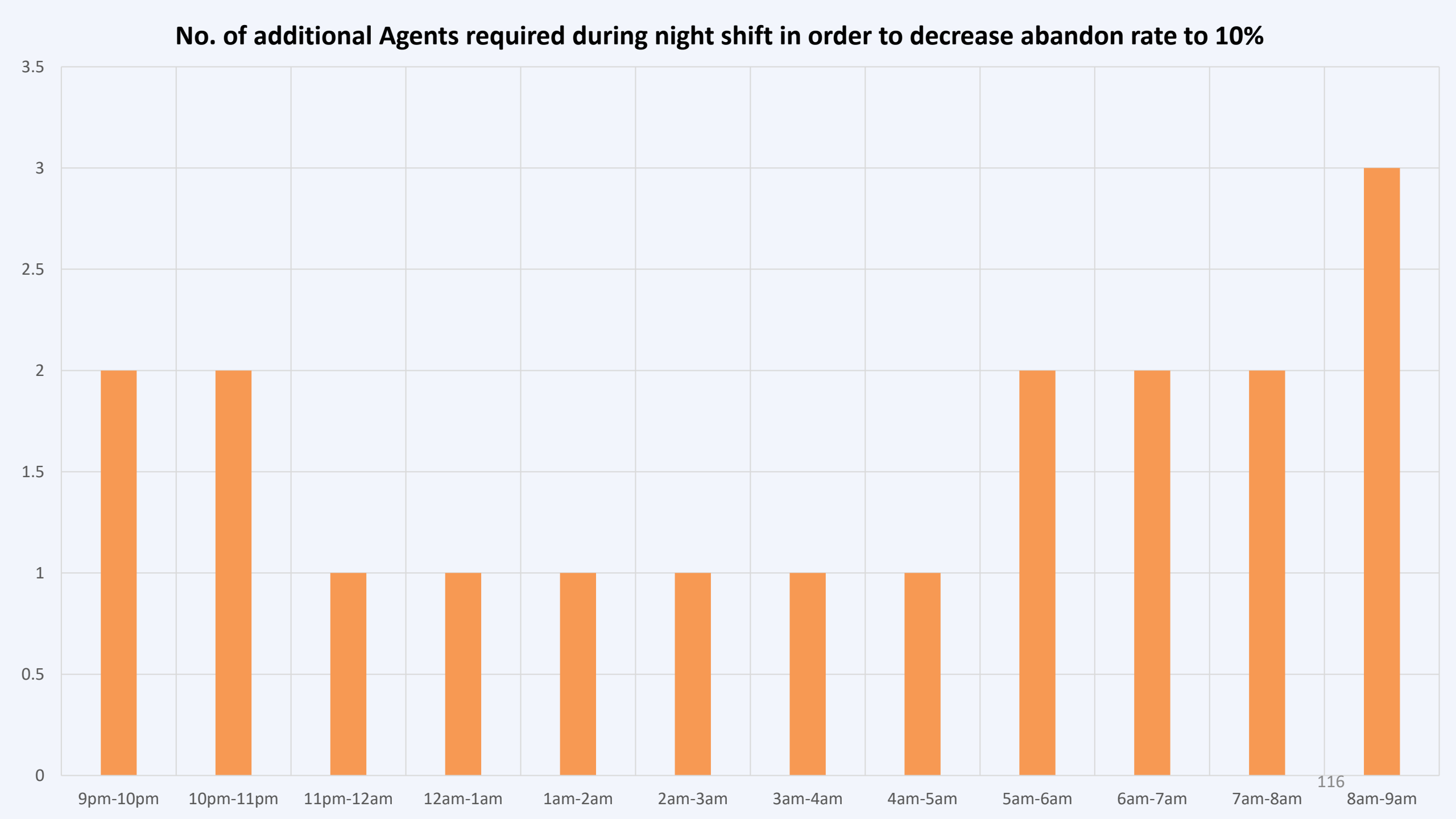
Daily average calls	5129.913
30% of daily calls(at night)	1539
Average call_seconds	198.623
No.of hours required	76.42011
No.of agents required	16.88889

I have taken the total of calls which are answered, transferred and abandoned.

Then I have found out the daily average calls based on the grand total.

Based on the calculation done until now , I have calculated the no.of hours required to decrease the abandon rate. Since the average working hours of an agent is 4.5, I have divided the no.of hours required by 4.5 to get the additional no.of agents required to decrease the abandon rate to 10%. Then I have distributed the agents based on time.

Time bucket	Call duration	Time distribution	Agents required	Agents required(rounded)
9pm-10pm	3	10.00	1.7	2
10pm-11pm	3	10.00	1.7	2
11pm-12am	2	15.00	1.133333333	1
12am-1am	2	15.00	1.133333333	1
1am-2am	1	30.00	0.566666667	1
2am-3am	1	30.00	0.566666667	1
3am-4am	1	30.00	0.566666667	1
4am-5am	1	30.00	0.566666667	1
5am-6am	3	10.00	1.7	2
6am-7am	4	7.50	2.266666667	2
7am-8am	4	7.50	2.266666667	2
8am-9am	5	6.00	2.833333333	3
Total	30	1.00	17	19


No. of additional Agents required during night shift in order to decrease abandon rate to 10%
3.5
3
2.5
2
1.5
1
0.5
0
9pm-10pm
10pm-11pm
11pm-12am
12am-1am
1am-2am
2am-3am
3am-4am
4am-5am
5am-6am
6am-7am
7am-8am
8am-9am

Conclusion

- Most of the calls are answered during 10am -11am and 7pm – 8pm.
- Most of the calls are received during 11am-12pm.
- During the day shift or general shift, in order to decrease the abandon rate to 10% , 12 agents are additionally required to answer the call , so that the customer satisfaction is not neglected and there is a less chance of receiving customer complaints on abandoning of calls.
- During the night shift, 17 agents are additionally required in order to decrease the abandon rate to 10%. And the agents must be very attentive during 9pm-11pm and 5am-9am, as most of the calls are received during this time period.

My learnings through these projects

- I leveraged Excel's advanced functions and features, including pivot tables, data validation, conditional formatting, and scenario analysis. I also utilized Excel's built-in statistical functions to analyze the data.
- To address data quality concerns, I implemented data validation rules and applied data cleaning techniques. For effective visualization, I employed pivot tables to summarize and aggregate data.
- Dealing with missing data and outliers proved challenging. I applied data imputation techniques to handle missing values and implemented outlier detection methods to clean the dataset.
- These projects enhanced my proficiency in data manipulation and analysis within a familiar tool. I gained a deeper understanding of financial metrics and how they contribute to business success. I also improved my ability to organize and present data in a clear and concise manner.

Link to my projects:
https://drive.google.com/drive/folders/1R0uihZJc2lLp07Vcv2-DOATBE6KhDpny?usp=sharing

Thank you