Data Analysis Portfolio

A MADHAVA VARMA

Professional Background

I am currently pursuing B.Tech in Computer Science and Engineering and I am enrolled in BSc Data Science and Application Online degree of IIT MADRAS and I am dedicated to continuous learning and adapting to new technologies.

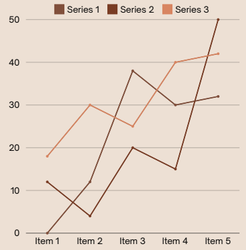
I’ve always been interested in Data Analysis and Data Science since my first year and I’ve done several personal projects based on it.
I am currently working on a Machine Learning project as an intern at Vizag Steel Plant.

As a fresher I am eager to use my skills for solving real-world problems and be a part of innovative projects and teams.

Table of Contents

Table of Contents

Table of Contents

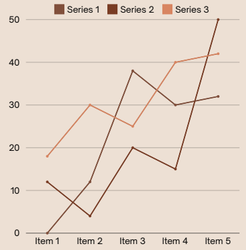


Data Analytics Process

Description

We use Data Analytics in everyday life without even knowing it.
For eg : Going to a market to buy something .

Your task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.



Data Analytics Process

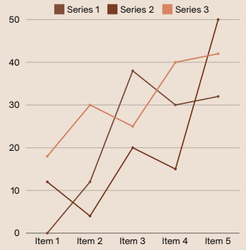
Design

Real World Application: Individual Finance Improvement

Step 1: Plan:
We set financial goals such as saving for retirement, paying off debt, or building an emergency fund. We plan the actions that are to be taken in order to achieve these goals, such as creating a budget, tracking expenses, and investing for the future.

Step 2: Prepare:
We gather relevant financial data including income statements, bank statements and bills. We may also utilize some tools to streamline data collection and to organize it like Excel or Googlesheets.

Step 3: Process:
We clean and organize the collected financial data, categorizing expenses, and income streams. We may use spreadsheets etc. to input and process the data, ensuring accuracy and consistency.



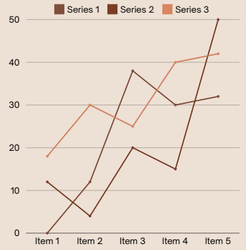
Data Analytics Process

Design

Step 4: Analyze:
We use the processed data and conduct a thorough analysis of our financial situation. We identify spending patterns, assess current financial status, and evaluate progress towards our goals. This analysis may involve calculating metrics such as savings rate, net worth, income to expenditure ratio etc.

Step 5: Share:
We now share this financial analysis and insights with trusted advisors, such as family members, close friends etc. Sharing insights can provide valuable feedback and support which helps us to make better financial strategies and reach our goals.

Step 6: Act:
Based on our analysis and feedback received, we take required actions to improve our financial situation and achieve our goals. This may include creating a budget, prioritize savings, investing with long-term objectives and so on.



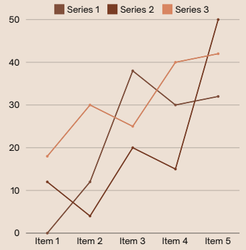
Data Analytics Process

Design

Step 4: Analyze:
We use the processed data and conduct a thorough analysis of our financial situation. We identify spending patterns, assess current financial status, and evaluate progress towards our goals. This analysis may involve calculating metrics such as savings rate, net worth, income to expenditure ratio etc.

Step 5: Share:
We now share this financial analysis and insights with trusted advisors, such as family members, close friends etc. Sharing insights can provide valuable feedback and support which helps us to make better financial strategies and reach our goals.

Step 6: Act:
Based on our analysis and feedback received, we take required actions to improve our financial situation and achieve our goals. This may include creating a budget, prioritize savings, investing with long-term objectives and so on.



Data Analytics Process

Conclusion

Hence, we have seen how we can use the 6 steps of Data Analytics while making any decision based on a real life application.

The 6 steps used to take decisions in real life scenarios are:-

- Plan
- Prepare
- Process
- Analyze
- Share
- Act

Instagram User Analytics

Description

User analysis involves tracking how users engage with a digital product (software application or mobile app) to provide valuable insights that can help the business grow.

These insights derived from this analysis can be used by various teams within the business, which might use these insights to launch a new campaign, decide on new features to build, and improve the overall user experience.

We are supposed to provide a detailed report for the Marketing and Investor metrics department. This analysis will help them make a decision based on different metrics and insights.

Instagram User Analytics

The Problem

A) Marketing Analysis:

• Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time.
Your Task: Identify the five oldest users on Instagram from the provided database.

• Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails.
Your Task: Identify users who have never posted a single photo on Instagram.

• Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo win.
Your Task: Determine the winner of the contest and provide their details to the team.

• Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.
Your Task: Identify and suggest the top five most commonly used hashtags on the platform.

• Ad Campaign Launch: The team wants to know the best day of the week to launch ads. Your Task: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

Instagram User Analytics

The Problem

B) Investor Metrics:

• User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.
Your Task: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.

• Bots & Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts.
Your Task: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

Instagram User Analytics

Design

Steps taken to load the data into the data base:

- Using the 'create db' function of MySQL create a data base.
- Then add tables and column names.
- Then add the values into them using the 'insert into' function of MySQL.
- By using the 'select' command we can query the desired output.

Software used for querying the results --› MySQL Workbench 8.0 CE

Instagram User Analytics

Findings - 1

To find the most loyal i.e. the top 5 oldest users of Instagram:

1. We will use the data from the users table by selecting the username and created_at columns.
2. Then using the order by function we will order the desired output by sorting with the created_at column in ascending order.
3. Then using the limit function, the output will be displayed for top 5 oldest Instagram users.

Result:

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

Instagram User Analytics

Findings - 2

To identify users who have never posted a single photo on Instagram:

1. We will first select username column from the users table.
2. Then we will left join photos table on the users table, on users.id = photos.user_id because, both the users.id and photos.user_id have common contents in them.
3. Then we will find rows from the users table where the photos.id IS NULL

Result:

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
Esmeralda.Mraz57	90
Bethany20	91

Instagram User Analytics

Findings - 3

To determine the winner of the contest and provide their details to the team:

- We need to select users.username, photos.id, photos.image_url and count(*) as total_likes .
- Next, we need to use inner join function to inner join users, photos and likes table on likes.photo_id = photos.id and photos.user_id = users.id.
- Now we need to use GROUP BY function to group the data based on photos.id.
- Then, we use ORDER BY function to sort the data based on total_likes in descending order using DESC and limit it to get the winner by using the limit function.

Result:

	id	username	id	image_url	total_likes
▶	52	Zack_Kemmer93	145	https://jarret.name	48

Instagram User Analytics

Findings - 4

To Identify and suggest the top five most commonly used hashtags on the platform:

- Select tags.tag_name, count(*) as tag_count to count the number of tags individually.
- Now we need to use JOIN function to join tags and photo_tags tables on tags.id = photo_tags.id as they contain common data.
- Next, we use GROUP BY function to group the data based on tags.tag_name.
- Then, we use ORDER BY function to sort the data based on tag_count in descending order using DESC and limit it to get the winner by using the limit function.

Result:

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

Instagram User Analytics

Findings - 5.

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

- Select dayname(created_at) as day_in_week, count(*) as user_count from the users table.
- Now, we use GROUP BY function to group the data based on day_in_week.
- Then, we use ORDER BY function to sort the data based on user_count in descending order using DESC.

Result:

day_in_week	user_count
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Instagram User Analytics

Findings - 6

To calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.

- First, we need to get the count of posts/photos present in photos.id column of photos table using count(*) from photos.
- Next, we need to get the count of users present in the users.id column of users table using count(*) from users.
- Now, we need to divide the total number of posts / total number of users to get the average number of posts per user.

Result:

	avg_posts_per_user
▶	2.5700

Instagram User Analytics

Findings - 7

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

- First, we need to select the user_id column in photos table, username column from users table and count(*) as total_likes to count the total number of likes.
- Next, we need to use the inner join function to inner join users and likes tables based on users.id = likes.user_id.
- Now, we need to use group by function to get the data based on likes.user_id. Finally, we need to get the data of users from photos table who are having count(*) equal to the value of total_likes.

Result:

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

Instagram User Analytics

Analysis

After performing the analysis, I have the following points: -

- Top 5 oldest users are: -

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

- There are 26 users who are inactive on Instagram. They have never posted any kind of stuff of Instagram may it be any photo, video or any type of text. So, the Marketing team of Instagram needs to remind such inactive users.

- Zack_Kemmer93 with user id 52 is the winner because he has most number of likes i.e 48 on his single photo with photo_id 145.

- The top 5 most commonly used #hashtags along with the total count are smile(59), beach(42), party(39), fun(38) and concert(24).

- Most of the users registered on Thursday and Sunday i.e 16. So best day of the week to launch ads are Thursday and Sunday.

- Average number of posts per user is 3. total number of photos on Instagram divided by the total number of users is 2.57.

- There are 30 users who have liked every single photo on the site. They can be identified as bots or fake accounts.

Instagram User Analytics

Analysis

Using the 5 Whys approach I am finding the root cause of the following: -

• Why did the Marketing team wanted to know the most inactive users?
➔ So, they can reach out to those users via mail and ask them the reasons which keeping them away from using the Instagram.

• Why did the Marketing team wanted to know the top 5 #hashtags used?
➔ May be the Marketing team wanted to add some filter features for photos and videos posted using the top 5 mentioned #hashtags.

• Why did the Marketing team wanted to know on which day of the week the platform had the newest users registered?
➔ So, that they can run more Ads of various brands during such days and also get profit from it.

• Why did the Investors wanted to know about the average posts per user has on Instagram?
➔ It is a fact that every brand or social platform is determined by the user engagement on such platforms, also investors wanted to know whether the platform has the right and authenticated user base. It also helps the tech team determine how to handle such traffic on the platform with the latest tech without disrupting the smooth and efficient functioning of the platform

• Why did the Investors wanted to know the count of BOTS and Fake accounts if any?
➔ So that the Investors are assured that they are investing into an Asset and not a Future Liability.

Instagram User Analytics

Conclusions

In conclusion, not only Instagram but also various other social media and commercial companies utilize data analysis to extract insights from customer data.

This process helps identify valuable customers who will contribute positively to the company's future success.

By regularly analyzing and categorizing the customer base, businesses can optimize profits while minimizing costs.

Operation Analytics and Investigating Metric Spike

Description

Operational Analytics is a crucial process that involves analyzing a company's end-toend operations. This analysis helps identify areas for improvement within the company.
You work closely with various teams, such as operations, support, and marketing, etc and help them derive valuable insights from the data they collect. One of the key aspects of Operational Analytics is investigating metric spikes.

This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales. You are working as a Lead Data Analyst at a company like Microsoft.
You'll be provided with various datasets and tables, and your task will be to derive insights from this data to answer questions posed by different departments within the company.

These insights will can help improve the company's operations and understand sudden changes in key metrics.

Operation Analytics and Investigating Metric Spike

The Problem

Case Study 1: Job Data Analysis

• Jobs Reviewed Over Time: Calculate the number of jobs reviewed per hour for each day in November 2020.
Your Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

• Throughput Analysis: Calculate the 7-day rolling average of throughput (number of events per second).
Your Task: Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.

• Language Share Analysis: Calculate the percentage share of each language in the last 30 days.
Your Task: Write an SQL query to calculate the percentage share of each language over the last 30 days.

• Duplicate Rows Detection: Identify duplicate rows in the data.
Your Task: Write an SQL query to display duplicate rows from the job_data table.

Operation Analytics and Investigating Metric Spike

The Problem

Case Study 2: Investigating Metric Spike

• Weekly User Engagement: Measure the activeness of users on a weekly basis.
Your Task: Write an SQL query to calculate the weekly user engagement.

• User Growth Analysis: Analyze the growth of users over time for a product.
Your Task: Write an SQL query to calculate the user growth for the product.

• Weekly Retention Analysis: Analyze the retention of users on a weekly basis after signing up for a product.
Your Task: Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.

• Weekly Engagement Per Device: Measure the activeness of users on a weekly basis per device.
Your Task: Write an SQL query to calculate the weekly engagement per device.

• Email Engagement Analysis: Analyze how users are engaging with the email service.
Your Task: Write an SQL query to calculate the email engagement metrics.

Operation Analytics and Investigating Metric Spike

Design

Steps taken to load the data into the data base:

- Using the 'create db' function of MySQL create a data base.
- Then add tables and column names.
- Then add the values into them using the 'insert into' function of MySQL.
- By using the 'select' command we can query the desired output.

Software used for querying the results --› MySQL Workbench 8.0 CE

Operation Analytics and Investigating Metric Spike

Findings-1

To find the number of jobs reviewed per hour per day of November 2020:

1. We will use the data from job_id columns of the job_data table.
2. Then we will divide the total count of job_id by (30 days * 24 hours) for finding the number of jobs reviewed per day.

Result:

	no_of_jobs_reviewed_per_day
▶	0.0083

Operation Analytics and Investigating Metric Spike

Findings-2

For calculating the 7-day rolling daily metric average of throughput:-

1. We will be first taking the count of job_idand ordering them w.r.t ds (date of interview)
2. Then by using the ROW function we will be considering the rows between 6 preceding rows and the current row
3. Then we will be taking the average of the jobs_reviewed

Result:

review_date	jobs_reviewed	throughput
2020-11-25 00:00:00	1	1.0000
2020-11-26 00:00:00	1	1.0000
2020-11-27 00:00:00	1	1.0000
2020-11-28 00:00:00	2	1.2500
2020-11-29 00:00:00	1	1.2000
2020-11-30 00:00:00	2	1.3333

Operation Analytics and Investigating Metric Spike

Findings-3

To Calculate the percentage share of each language :-

1. We will first divide the total number of languages by the total number of rows presents in the table
2. Then we will do the grouping based on the languages.

Result:

language	no_of_jobs	Percentage
English	1	12.50
Arabic	1	12.50
Persian	3	37.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50

Operation Analytics and Investigating Metric Spike

Findings-4

To view the duplicate rows having the same value we will:-
1. First decide in which do we need to find the duplicate row values
2. After deciding the column(parameter) we will use the ROW_NUMBER function to find the row numbers having the same value
3. Then we will portioning the ROW_NUMBER function over the column (parameter) that we decided i.e. job_id
4. Then using the WHERE function we will find the row_num having value greater than 1 i.e. row_num based on the occurrence of the job_id in the table

Result:

job_id	actor_id	evnt	language	time_spent	org	ds	row_num
23	1005	transfer	Persian	22	D	2020-11-28 00:00:00	2
23	1004	skip	Persian	56	A	2020-11-26 00:00:00	3

Operation Analytics and Investigating Metric Spike

Findings-5

To find the weekly user engagement:-

1. We will extract the week from the occurred_at column of the events table using the EXTRACT function and WEEK function
2. Then we will be counting the number of distinct user_id from the events table
3. Then we will use the GROUP BY function to group the output w.r.t week from occurred_at.

Result:

nweek	users_count
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104

Operation Analytics and Investigating Metric Spike

Findings-6

To find the user growth (number of active users per week):-

1. First we will the extract the year and week for the occurred_at column of the users table using the extract, year and week functions
2. Then we will group the extracted week and year on the basis of year and week number 3. Then we ordered the result on the basis of year and week number
4. Then we will find the cumm_active_users using the SUM, OVER and ROW function between unbounded preceding and current row

Result:

year	nweeks	active_users	total_ausers
2013	0	23	23
2013	1	30	53
2013	2	48	101
2013	3	36	137
2013	4	30	167
2013	5	48	215
2013	6	38	253
2013	7	42	295
2013	8	34	329
2013	9	43	372
2013	10	32	404
2013	11	31	435
2013	12	33	468
2013	13	39	507
2013	14	35	542
2013	15	43	585
2013	16	46	631
2013	17	49	680
2013	18	44	724
2013	19	57	781
2013	20	39	820
2013	21	49	869
2013	22	54	923
2013	23	50	973

year	nweeks	active_users	total_ausers
2013	48	97	2894
2013	49	116	3010
2013	50	124	3134
2013	51	102	3236
2013	52	47	3283
2014	0	83	3366
2014	1	126	3492
2014	2	109	3601
2014	3	113	3714
2014	4	130	3844
2014	5	133	3977
2014	6	135	4112
2014	7	125	4237
2014	8	129	4366
2014	9	133	4499
2014	10	154	4653
2014	11	130	4783
2014	12	148	4931
2014	13	167	5098
2014	14	162	5260
2014	15	164	5424
2014	16	179	5603
2014	17	170	5773
2014	18	163	5936

year	nweeks	active_users	total_ausers
2013	24	45	1018
2013	25	57	1075
2013	26	56	1131
2013	27	52	1183
2013	28	72	1255
2013	29	67	1322
2013	30	67	1389
2013	31	67	1456
2013	32	71	1527
2013	33	73	1600
2013	34	78	1678
2013	35	63	1741
2013	36	72	1813
2013	37	85	1898
2013	38	90	1988
2013	39	84	2072
2013	40	87	2159
2013	41	73	2232
2013	42	99	2331
2013	43	89	2420
2013	44	96	2516
2013	45	91	2607
2013	46	88	2695
2013	47	102	2797

2014	19	185	6121
2014	20	176	6297
2014	21	183	6480
2014	22	196	6676
2014	23	196	6872
2014	24	229	7101
2014	25	207	7308
2014	26	201	7509
2014	27	222	7731
2014	28	215	7946
2014	29	221	8167
2014	30	238	8405
2014	31	193	8598
2014	32	245	8843
2014	33	261	9104
2014	34	259	9363
2014	35	18	9381

Operation Analytics and Investigating Metric Spike

Findings-7

The weekly retention of users-sign up cohort can be calculated by two means i.e. either by specifying the week number (18 to 35) or for the entire column of occurred_at of the events table.

1. Firstly we will extract the week from occurred_at column using the extract, week functions
2. Then, we will select out those rows in which event_type = 'signup_flow' and event_name = 'complete_signup'
3. If finding for a specific week we will specify the week number using the extract function
4. Then using the left join we will join the two tables on the basis of user_id where event_type = 'engagement'
5. Then we will use the Group By function to group the output table on the basis of user_id
6. Then we will use the Order By function to order the result table on the basis of user_id

Result: **CLICK_HERE!!!**

Weekly_Retention_Analysis_Result_drive_link

Operation Analytics and Investigating Metric Spike

Findings-8

To find the weekly user engagement per device:-
1. Firstly we will extract the year_num and week_num from the occurred_at column of the events table using the extract, year and week function
2. Then we will select those rows where event_type = 'engagement' using the WHERE clause
3. Then by using the Group By and Order By function we will group and order the result on the basis of year_num, week_num and device

Result:

Click_HERE!!

Weekly_Engagement_Result_drive_link

Operation Analytics and Investigating Metric Spike

Findings-9

To find the email engagement metrics(rate) of users:-

1. We will first categorize the action on the basis of email_sent, email_opened and email_clicked using the CASE, WHEN, THEN functions
2. Then we select the sum of category of email_opened divide by the sum of the category of email_sent and multiply the result by 100.0 and name is as email_opening_rate
3. Then we select the sum of category of email_clicked divide by the sum of the category of email_sent and multiply the result by 100.0 and name is as email_clicking_rate
4. email_sent = ('sent_weekly_digest','sent_reengagement_email')
5. email_opened = 'email_open'
6. email_clicked = 'email_clickthrough'

Result:

	opening_rate	clicking_rate
▶	33.5834	14.7899

Operation Analytics and Investigating Metric Spike

Analysis

From the tables we can infer the following:-

- number of distinct job reviewed per day is 0.0083 number of non-distinct jobs reviewed per day is 0.0111 7 day rolling average throughput for 25, 26, 27, 28, 29 and 30 Nov 2020 are 1, 1, 1, 1.25, 1.2 and 1.3333 respectively.
- Percentage Share of each language i.e. Arabic, English, French, Hindi, Italian and Persian are 12.5, 12.5, 12.5, 12.5, 12.5 and 37.5 respectively.
- There are 2 duplicates values/rows having job_id = 23 and language = Persian in both the rows
- The weekly user engagement is the highest for week 31 i.e. 1685
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014
- The email_opening_rate is 33.5833 and email_clicking_rate is 14.78988

Operation Analytics and Investigating Metric Spike

Analysis

Using the Why's approach I am trying to find more insights:

- Why there is a difference of values between the number of distinct jobs reviewed per day and number of non-distinct jobs reviewed per day?

----› May be due to repeated values in two or more rows or the dataset consisted of duplicate rows

- Why one shall use 7 day rolling average for calculating throughput and not daily metric average?

----› For calculating the throughput we will be using the 7-day rolling because 7-day rolling gives us the average for all the days right from day 1 to day 7 Whereas daily metric gives us average for only that particular day itself.

- Why is it that percentage share of all other languages is 12.5% but that of language = 'Persian' is 37.5?

-----› In such cases there are two chances i.e. either there were duplicate rows having language as 'Persian' or there were really two or more unique people who were speaking in Persian language

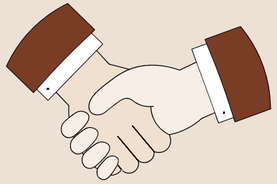
- Why do we need to look for duplicate rows in an dataset?

----› Duplicates have a direct influence of the Analysis going wrong and may led to wrong Business Decision leading to loss to the company or any entity; so to avoid these one must look for duplicates and remove them where necessary

Operation Analytics and Investigating Metric Spike

Conclusion

In conclusion, operational analytics and investigating metric spikes are crucial for any business.
These analyses should occur regularly—daily, weekly, monthly, quarterly, or yearly—based on the firm's specific needs. Additionally, firms should prioritize email engagement with customers.
Catchy subject lines, reasonable discounts, and coupons can help boost the existing customer base. Furthermore, having a dedicated department to address the concerns of visitors who abandoned the sign-up process is essential.
Guiding these visitors can potentially convert them from mere visitors to loyal customers.

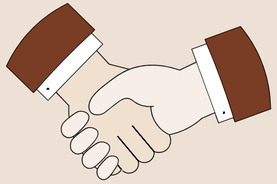


Hiring Process Analytics

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 analyse before hiring freshers or any other individual.

Thus, making an opportunity for a Data Analyst job here too!
Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.
You are 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 you to answer certain questions making sense out of that data.



Hiring Process Analytics

The Problem

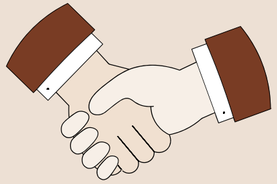
- Hiring: Process of intaking of people into an organization for different kinds of positions.
  Your task: How many males and females are Hired ?

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

- Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.
  Your task: Draw the class intervals for salary in the company ?

- Charts and Plots: This is one of the most important part of analysis to visualize the data.
  Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?

- Charts: Use different charts and graphs to perform the task representing the data. Your task: Represent different post tiers using chart/graph?



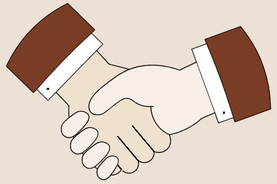
Hiring Process Analytics

Design

I performed the following actions before the data analysis:

- Created a copy of the raw data to perform analysis without affecting the original dataset.
- Removed irrelevant columns that were unnecessary for the analysis.
- Checked for blank spaces and NULL values in the dataset.
- Imputed missing values in numeric columns using the mean and median.
- Identified and handled outliers by replacing them with the median of the respective column.
- Imputed missing values in categorical columns using the mode (most frequent value).
- Removed duplicate rows from the dataset.

Software Used: Microsoft Excel



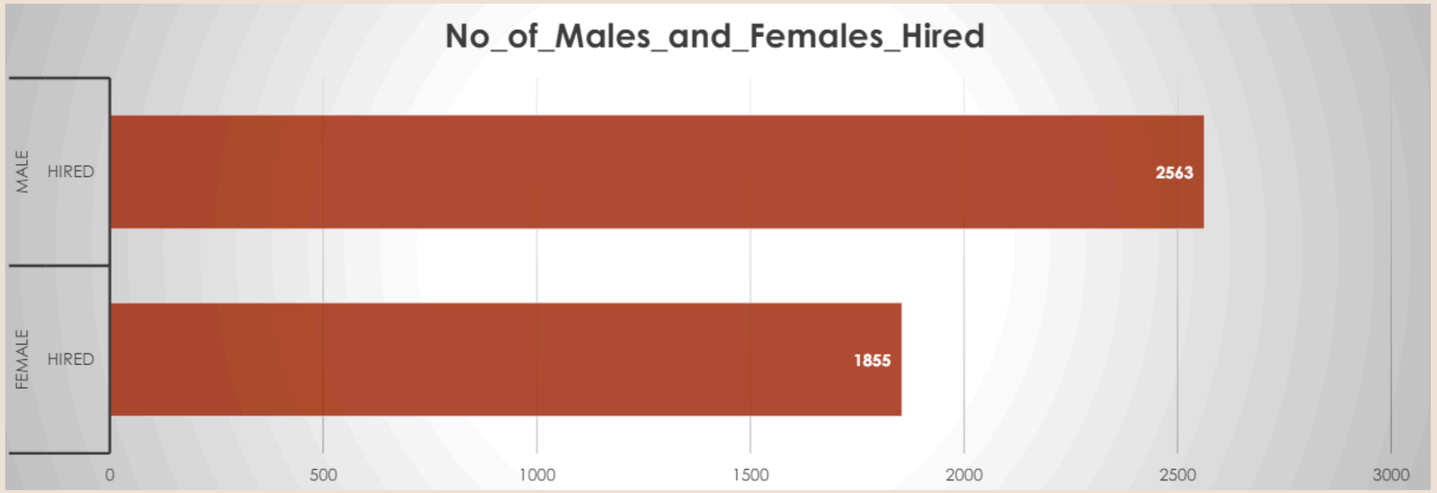
Hiring Process Analytics

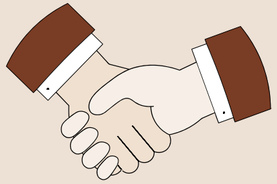
Findings-1

To determine the gender distribution of hires and check how many males and females have been hired by the company:

Result:

event_name	Status	no_of_male_and_female
Male	Hired	2563
Female	Hired	1856





Hiring Process Analytics

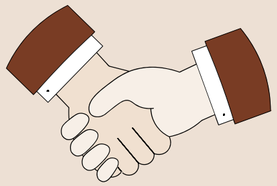
Findings-2

To find the average salary offered by this company:

Using the Formulae to calculate average salary offered by this company.
=AVERAGE(G:G)

Result:

Average	49983.02902

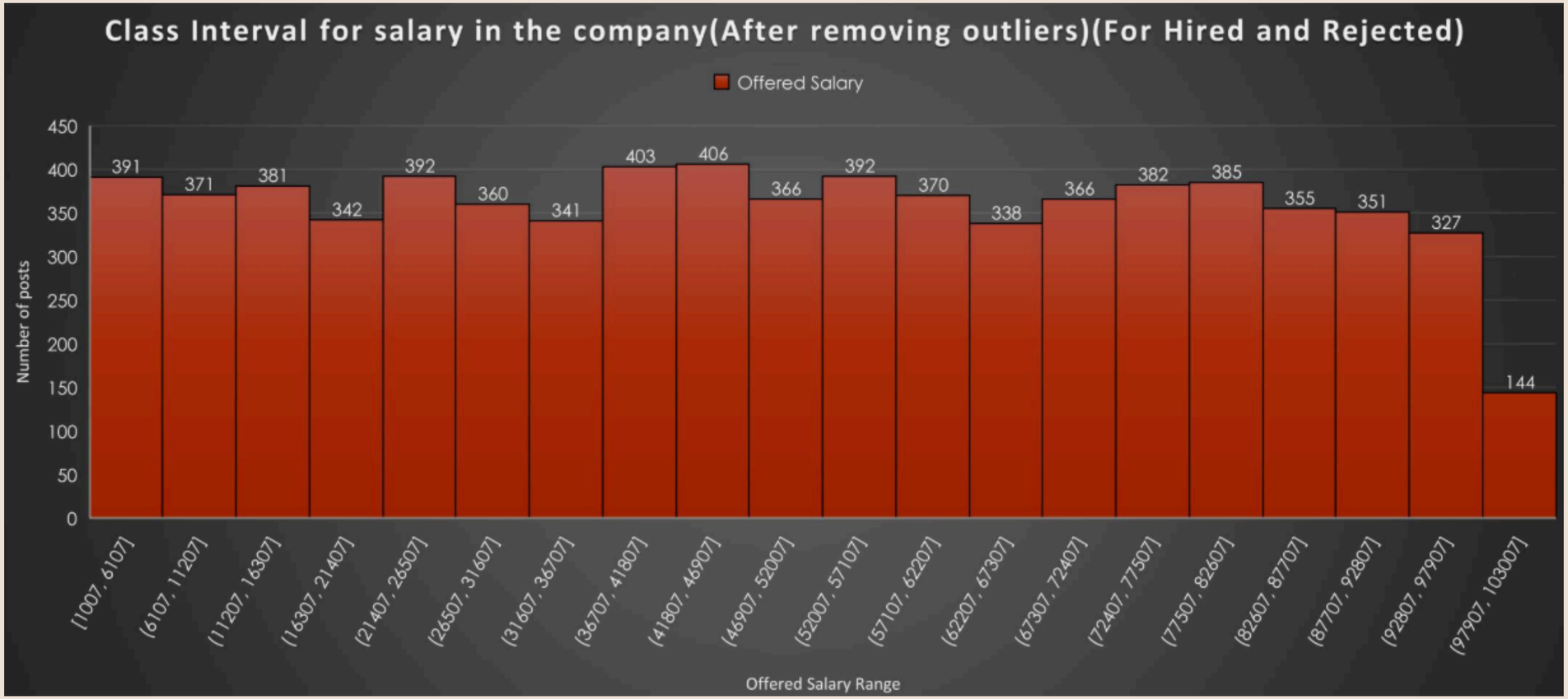


Hiring Process Analytics

Findings-3

To create class intervals for the salaries in the company:

Result:

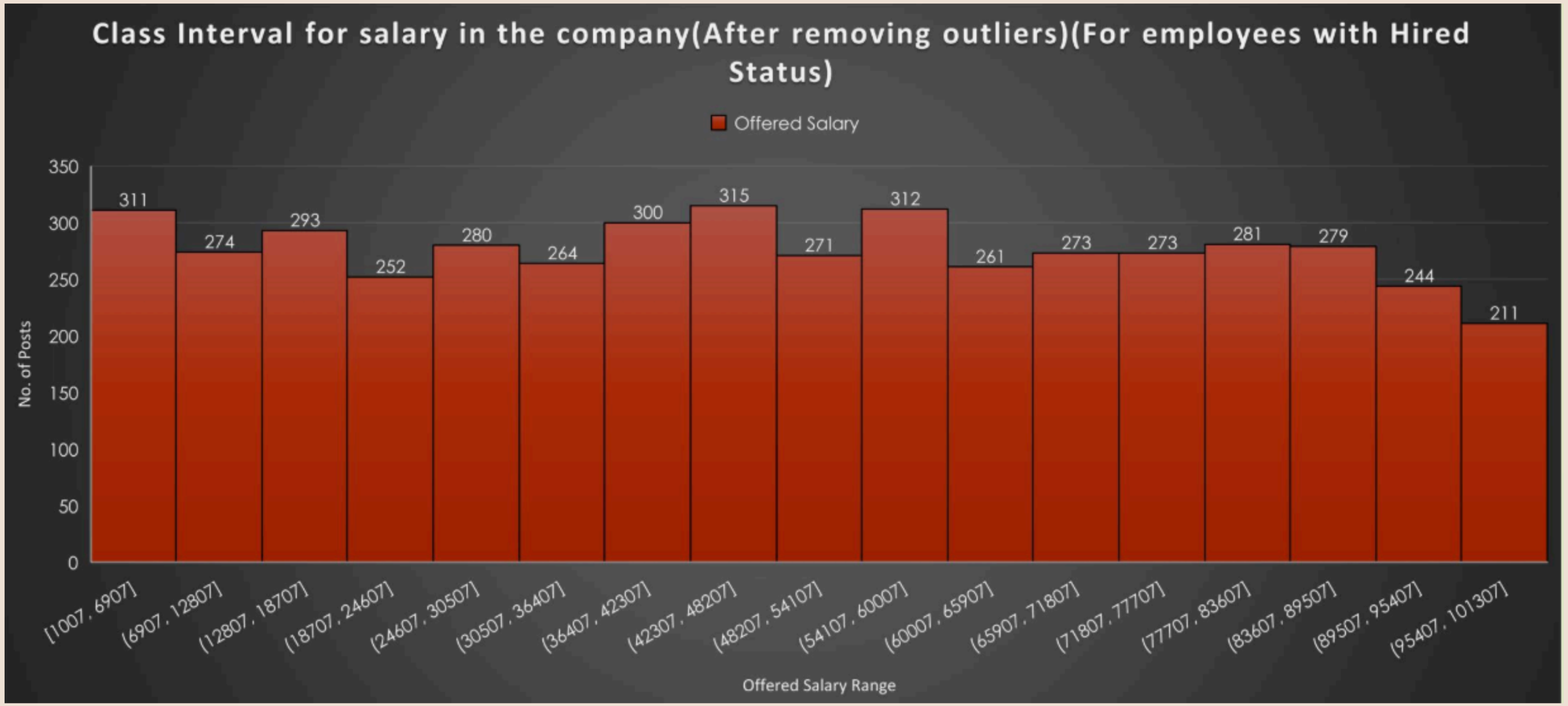


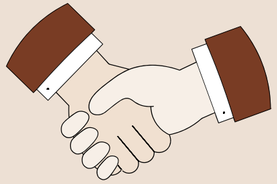
Hiring Process Analytics

Findings-3(Contd.)

To create class intervals for the salaries in the company:

Result:





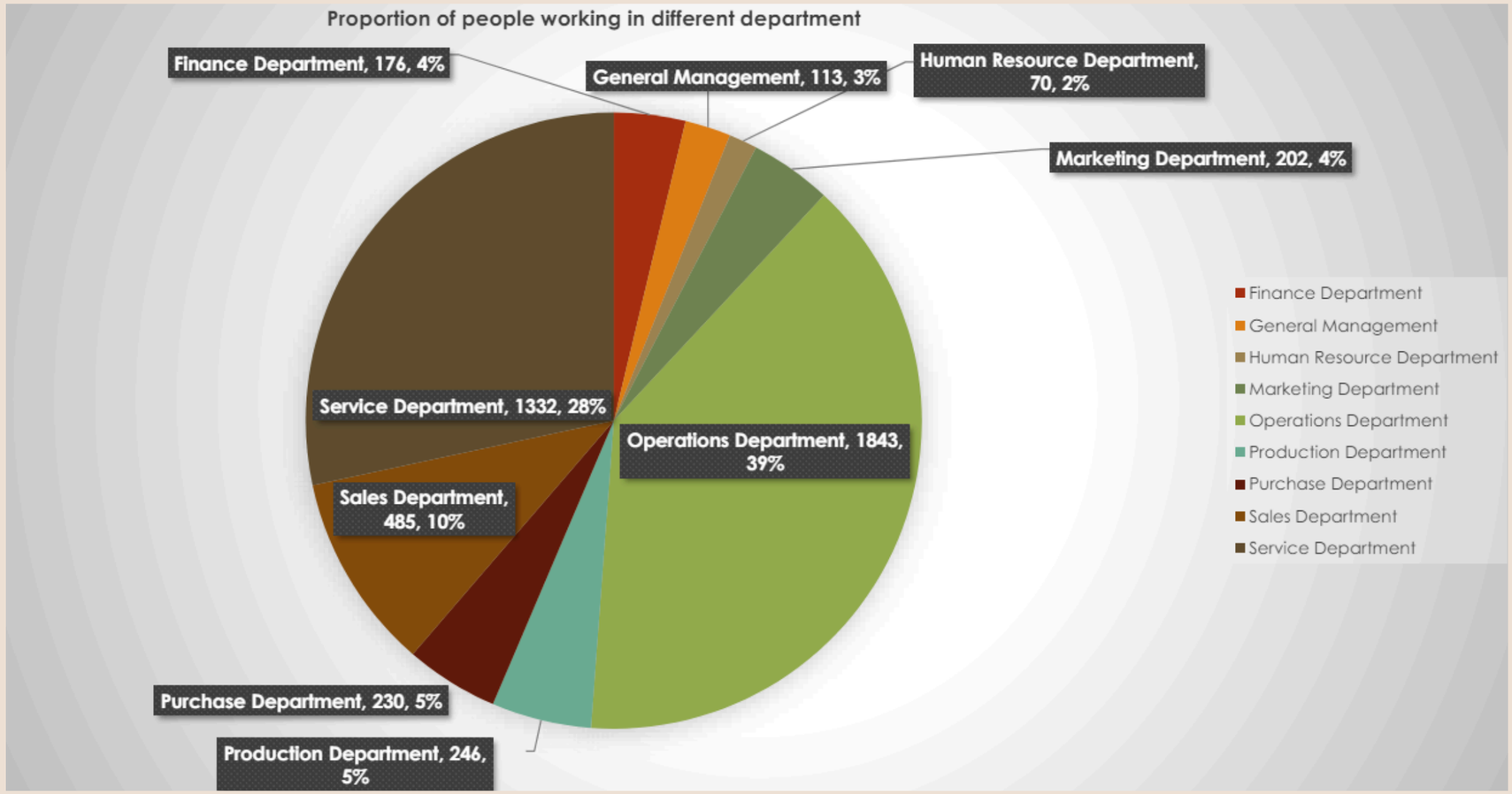
Hiring Process Analytics

Findings-4

To show the proportion of people working in different departments:

Result:

Department	no_of_people
Finance Department	176
General Management	113
Human Resource Department	70
Marketing Department	202
Operations Department	1843
Production Department	246
Purchase Department	230
Sales Department	485
Service Department	1332

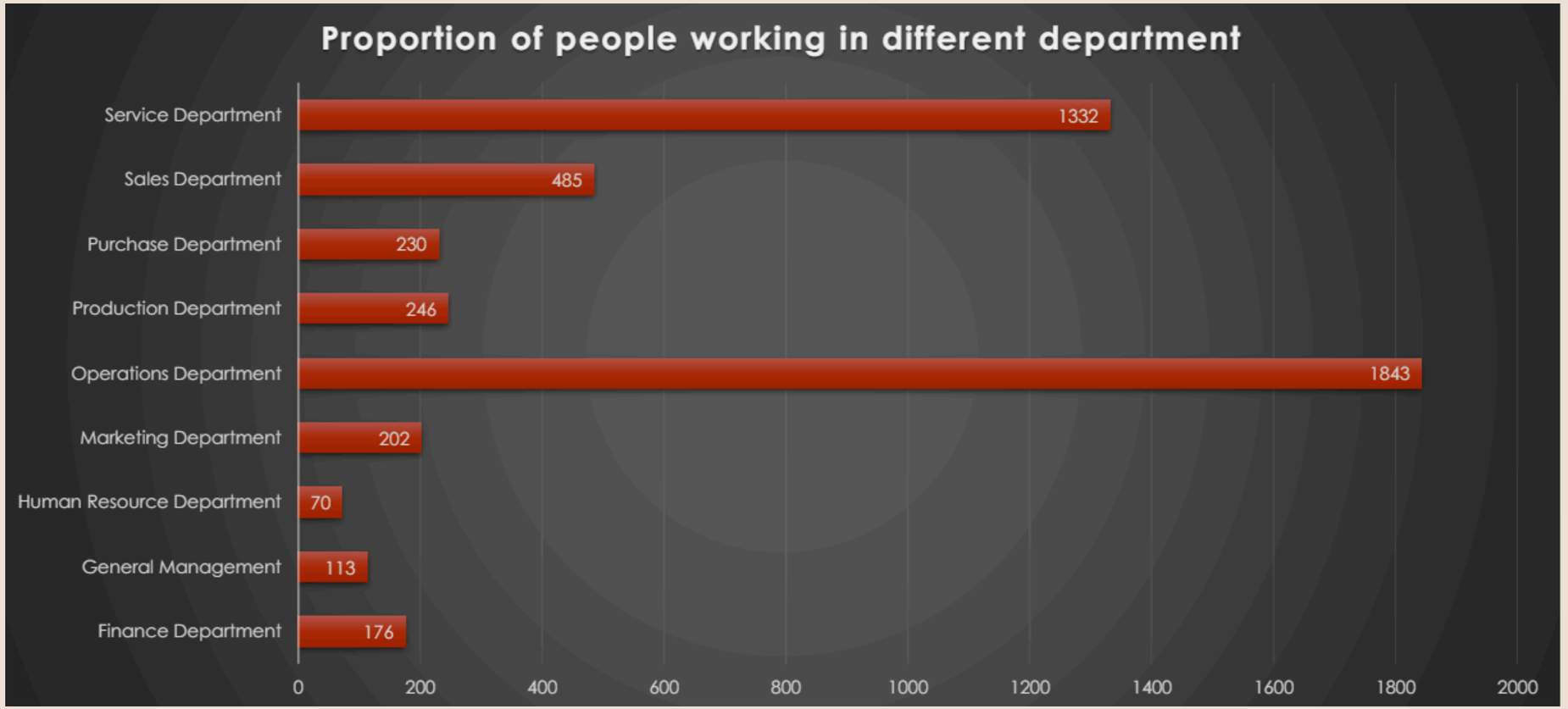


Hiring Process Analytics

Findings-4(Contd.)

To show the proportion of people working in different departments:

Result:

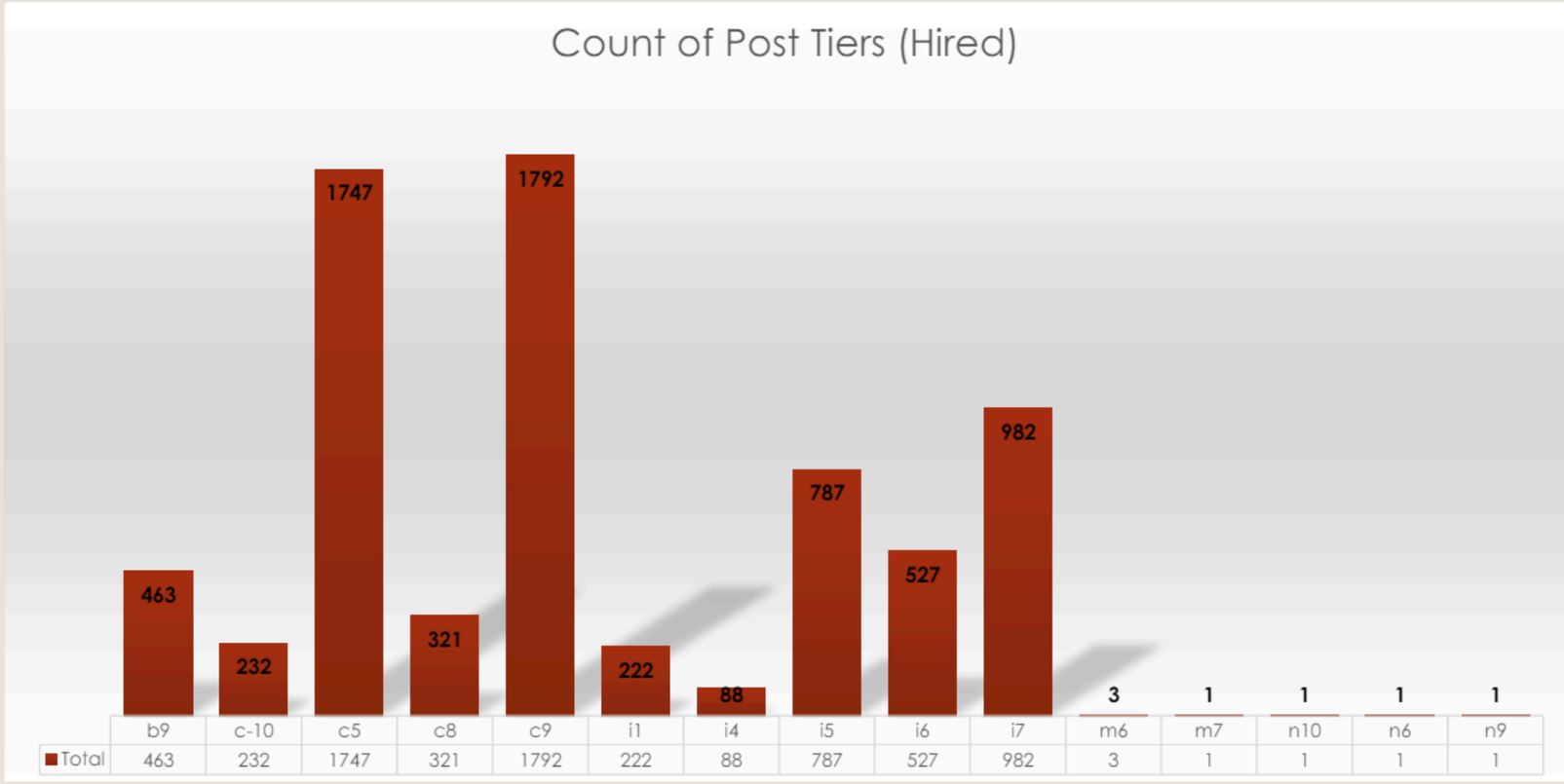


Hiring Process Analytics

Findings-5

To find different position tiers within the company:
Result:

Post Name	no_of_people_hired
b9	308
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
n6	1

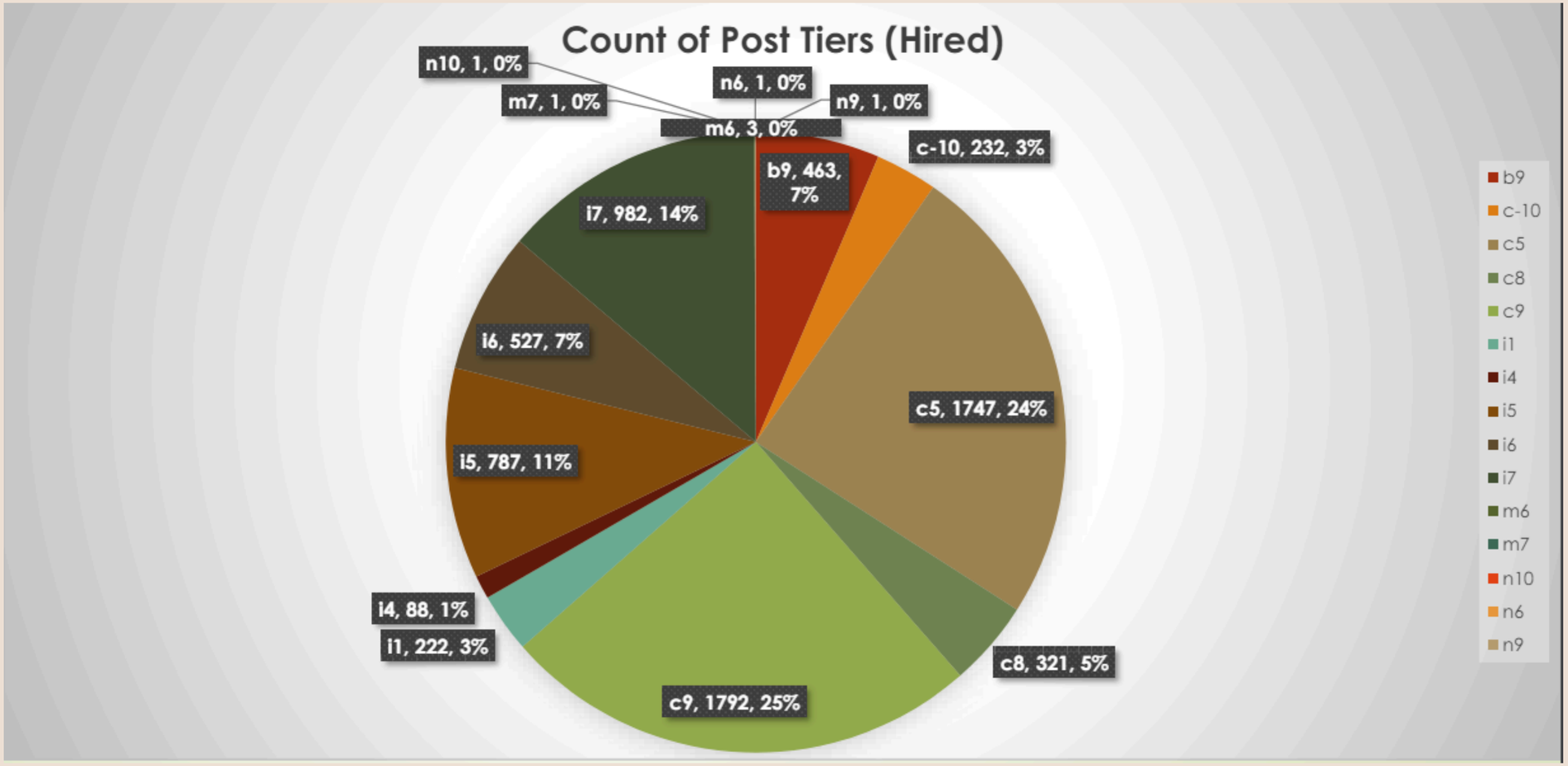


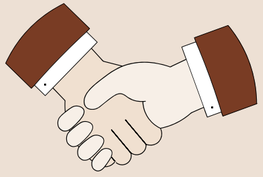
Hiring Process Analytics

Findings-5

To find different position tiers within the company:

Result:





Hiring Process Analytics

Analysis

Using the why approach I am finding the root cause of the following: -

• Why is there so much difference in the total number of Males and Females hired?

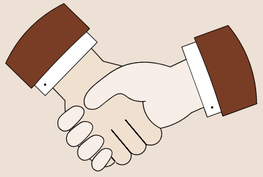
---› Since, the Company is an MNC and people from all around the world work here; such difference exists due to the fact that the equality has not yet reached to each and every part of the world. Some regions in the Gulf countries and in African continents along with some Asian countries face this problem.

• Why are there a few numbers of people whose salaries more than 85000 and a greater number of people whose salaries between 35000 to 60000?

----› It is a fact that there are some positions in company who require a specialist person with years of experience in that particular field of work and hence company looks for such people and offer them higher salary packages also such people regularly prove themselves an asset to the company. For any company there are more people having the salary in the range 35000 to 60000; such people have spent 3-4 years in the company and their salary and increments are decided based on their monthly, quarterly and yearly performance.

• Why is that the Operations department has the highest number of people working?

----› Operations Department works like a central hub for all other departments, all the execution tasks are carried out by this department. Operations department has the highest work load when compared to all other departments.



Hiring Process Analytics

Conclusion

I conclude by saying that Hiring Process Analytics plays a vital role for companies and firms in determining job openings for the near future. This analysis is conducted on a monthly, quarterly, or yearly basis, aligning with company needs and policies.

The Operations Department typically has the highest workforce due to its central role in executing various tasks. Within any company, certain employees receive higher salary packages compared to others—often due to specialized skills and years of experience.

Hiring Process Analytics helps companies decide salaries for new hires, assess workforce requirements by department, and guide appraisals and increments for existing employees. By leveraging these insights, companies can make informed decisions and optimize their workforce management effectively.

IMDB Movie Analysis

Description

The dataset provided having various columns of different IMDB Movies.
A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success of the Movies can be defined by high IMDB ratings.

The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.
First you need to clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

IMDB Movie Analysis

The Problem

• Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score. Your Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

• Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.
Your Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

• Language Analysis: Situation: Examine the distribution of movies based on their language.
Your Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

• Director Analysis: Influence of directors on movie ratings.
Your Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

• Budget Analysis: Explore the relationship between movie budgets and their financial success.
Your Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

IMDB Movie Analysis

Design

I performed the following actions before the data analysis:

• First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.

• Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.

• Columns like color, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes are irrelevant data. It needs to be dropped.

• We need to remove the rows which contains null values. Then we need to remove duplicates from dataset.

Software Used: Microsoft Excel

IMDB Movie Analysis

Findings-1

To find the most common genres of movies in the dataset: -

- First, we need to separate multiple genres and use COUNTIF function to count the number of movies for each genre.
- Then we will use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics.

Result:

Most common genres are:-								
genres	Count	Mean	Median	Mode	Max	Min	Variance	Standard Deviation
Drama	153	7.041830065	7.2	7.3	8.8	3.4	0.687054524	0.828887522
Comedy|Drama|Romance	151	6.494701987	6.5	6.5	8	4.3	0.562771744	0.750181141
Comedy|Drama	147	6.583673469	6.7	6.7	8.8	3.3	0.734800112	0.857204825
Comedy	145	5.840689655	6	6.5	8	1.9	1.481874521	1.217322686
Comedy|Romance	135	5.896296296	6	6.1	8.4	2.7	0.768269762	0.87650999

IMDB Movie Analysis

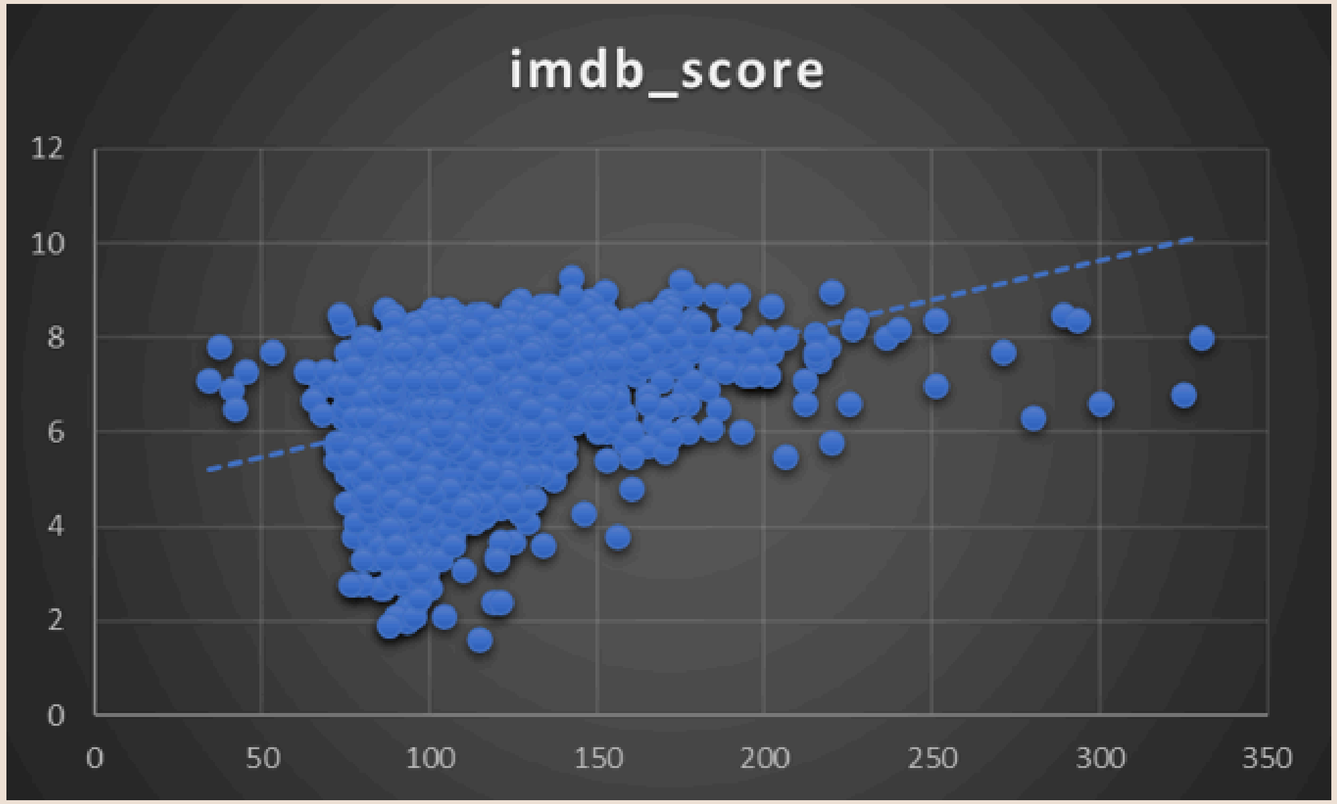
Findings-2

To Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score: -

- First, we will select column duration and imdb_score.
- Then we will use Excel's functions like AVERAGE, MEDIAN, and STDEV to calculate descriptive statistics.

Result:

Average	109.9241164
Median	106
Standard Deviation	22.75364979



IMDB Movie Analysis

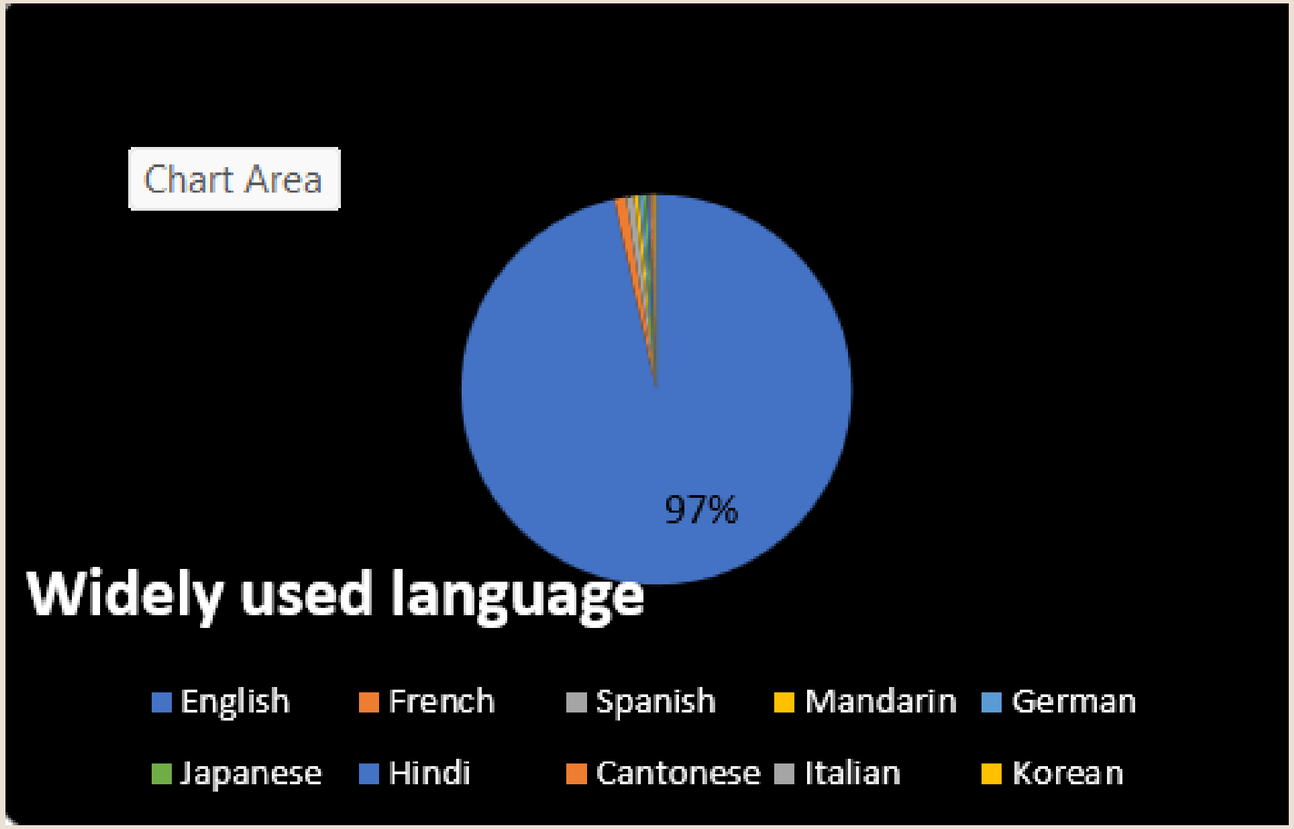
Findings-3

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

- First, we will select Column language and imdb_score.
- Then we will use COUNTIF function to count the number of movies for each language.
- Using AVERAGE, MEDIAN, and STDEV function we will calculate Mean, Median and Standard Deviation of IMDB Scores for each language.

Result:

Most common Languages are:-				
Language	**Count**	**Mean**	**Median**	**Standard Deviation**
English	3668	6.42391	6.5	1.048750752
French	37	7.28649	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.02143	7.25	0.765786244
German	13	7.69231	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.2375	7.3	0.440575922
Italian	7	7.18571	7	1.155318962
Korean	5	7.7	7.7	0.570087713



IMDB Movie Analysis

Findings-4

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

- First, we need to select column director_name and imdb_score.
- Then we will use AVERAGE function to Calculate the average IMDB score for each director.
- Then we will calculate percentrank and use PERCENTILE function to identify the directors with the highest scores.

Result:

director_name	Average
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

IMDB Movie Analysis

Findings-5

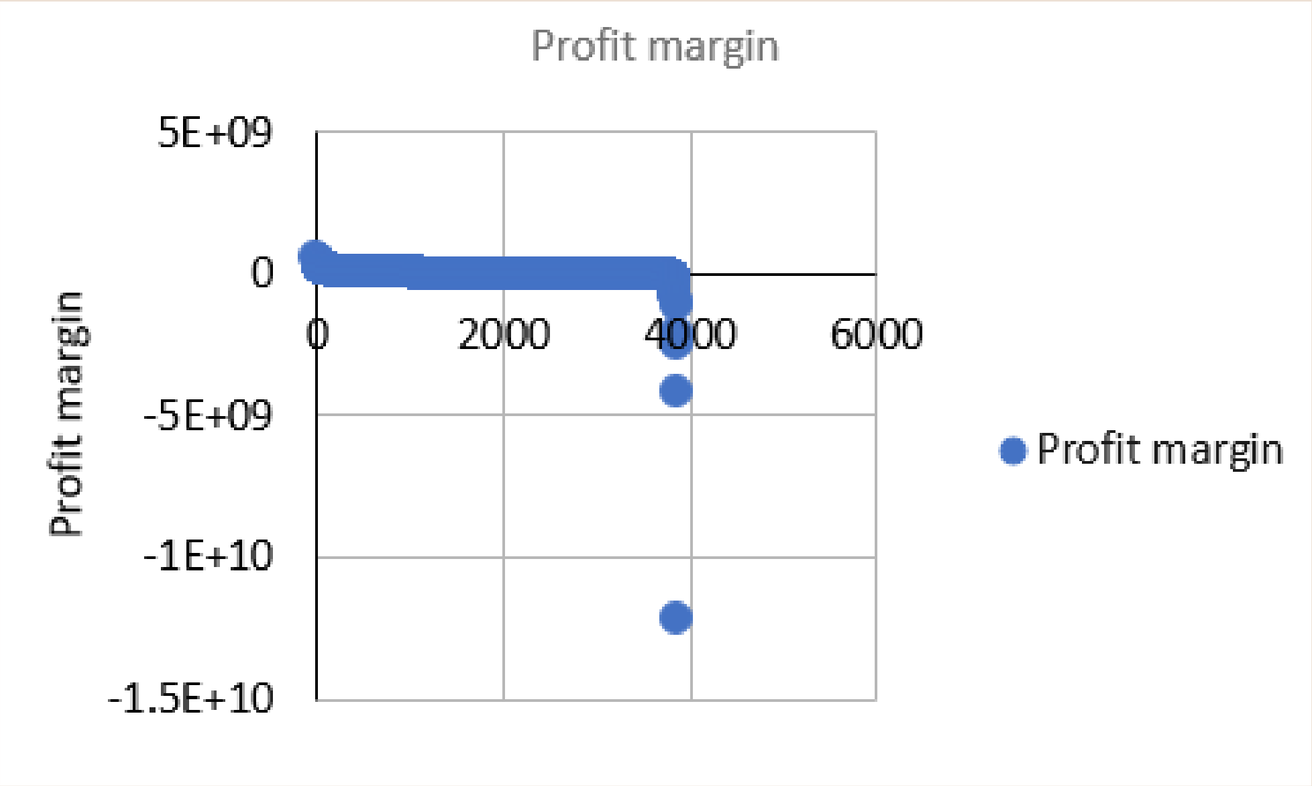
To Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

- First, we will calculate profit margin for each movie by subtracting budget value from gross value.
- We will use CORREL function to calculate correlation coefficients between movie budgets and gross earnings.
- Using MAX function, we will get highest profit margin then we will use INDEX function to get the title of the movie.

Result:

CORRELATION
0.100850218

MAX PROFIT	MOVIE TITLE
523505847	AvatarÂ



IMDB Movie Analysis

Findings-5

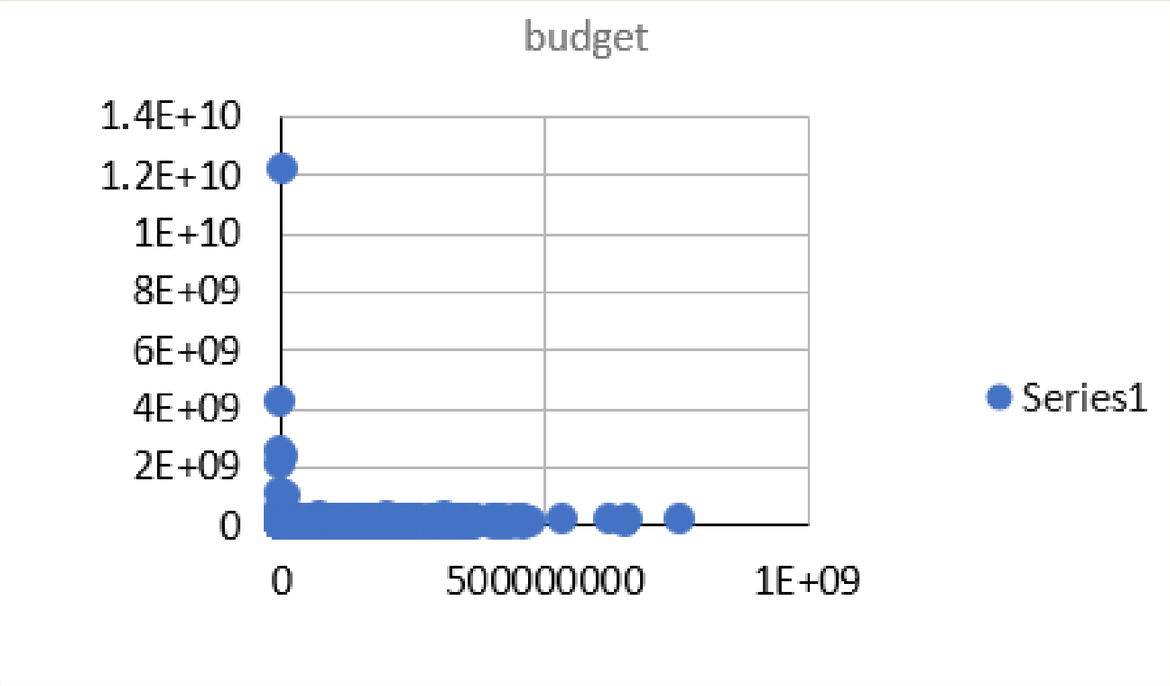
To Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

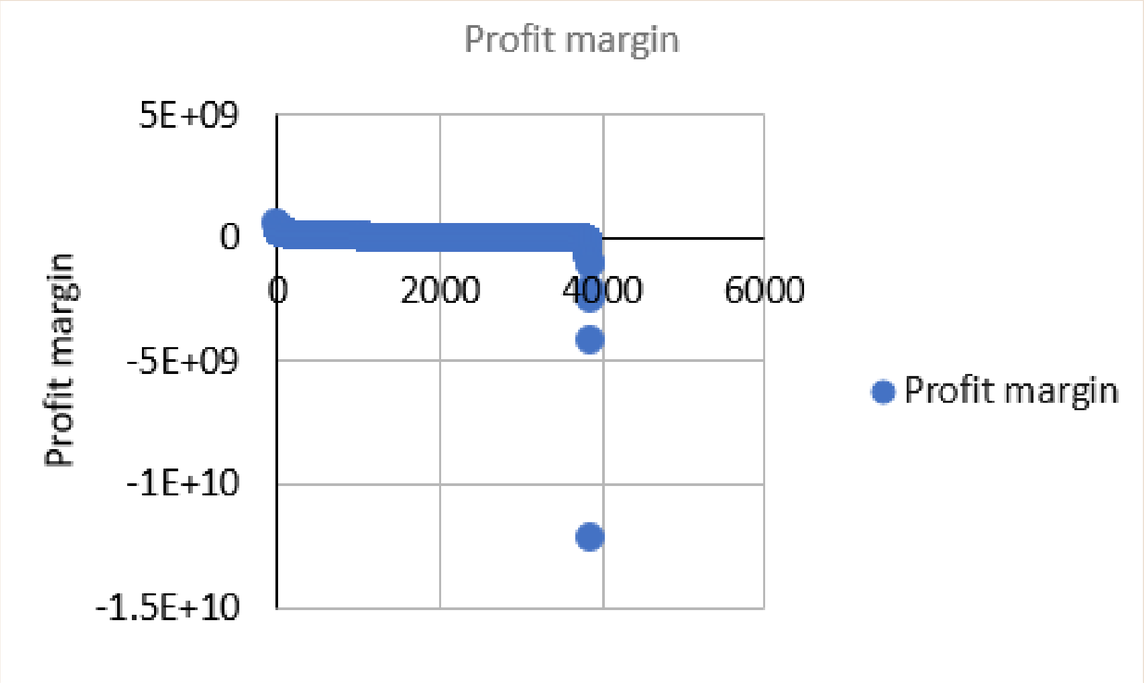
- First, we will calculate profit margin for each movie by subtracting budget value from gross value.
- We will use CORREL function to calculate correlation coefficients between movie budgets and gross earnings.
- Using MAX function, we will get highest profit margin then we will use INDEX function to get the title of the movie.

Result:

CORRELATION
0.100850218

MAX PROFIT	MOVIE TITLE
523505847	AvatarÂ





IMDB Movie Analysis

Analysis

Using the Why's approach I am trying to uncover the root cause: -

• Why is it that only Drama and Comedy had the highest popularity?

----› Most of people all over the world are stressed with their work life so they need a relaxing refreshment and not some action or horror type thing. So, people prefer watching movies that were of Drama or Comedy genre or both.

• Why is the highest duration of movies being not the Most rated IMDB Movie?

-----› Maybe because generally people like whole movie and they vote on IMDB portal based on whole movie. So, duration of the movie does not affect on IMDB rating.

• Why is 'English' the most common languages used in movies?

-----› Movies having language as English were having country of origin as USA. Also, it is a well-known fact that USA economy was robust during those days. So, the social media investors looked for directors made movies in English so as to gain some financial gains.

• Why is it that the Most rated IMDB movie and the highest profit movie not the same?

-----› Maybe, due to fact that during the IMDB rating only recognized and people who know how to vote on IMDB have the access to the IMDB portal. On the other hand, the profit is calculated on the basis of the tickets sold in theatres worldwide.

IMDB Movie Analysis

Conclusion

In conclusion, IMDb movie analysis isn’t solely the domain of filmmakers—it extends to investors, stakeholders, and theater owners.

While normal viewers might not delve into such analysis, it significantly impacts both pre-production and post-production phases. Contrary to popular belief, a movie’s IMDb rating doesn’t always correlate directly with its profit. True profitability hinges on global ticket sales.

Interestingly, tired audiences often gravitate toward Comedy/Drama genres rather than Action/Horror. Filmmakers should consider these insights during pre-production planning to create engaging movies that resonate with viewers worldwide.

BANK LOAN CASE STUDY

Description

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some customers who don't have a sufficient credit history take advantage of this and default on their loans.
Suppose you work for a consumer finance company which specializes in lending various types of loans to customers. You have to use EDA to analyze the patterns in the data and ensure that capable applicants are not rejected.

When the company receives a loan application, company faces two risks:
• If the applicant can repay the loan but is not approved, the company loses business.
• If the applicant cannot repay the loan and is approved, the company faces a financial loss.

When a customer applies for a loan, there are four possible outcomes:
• Approved: The company has approved the loan application.
• Cancelled: The customer cancelled the application during the approval process.
• Refused: The company rejected the loan.
• Unused Offer: The loan was approved but the customer did not use it.

BANK LOAN CASE STUDY

The Problem

• Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset.
Your Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

• Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
Your Task: Detect and identify outliers using Excel statistical functions, focusing on numerical variables.

• Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.
Your Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

• Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
Your Task: Perform univariate analysis to understand the distribution of individual variables, to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

• Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
Your Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

BANK LOAN CASE STUDY

Design

Before starting the actual analysis, I have: -

• First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.
• The columns which have null values more than or equal to 50%.

These columns need to be dropped.

Software used for doing the overall Analysis: - ----› Microsoft Excel

BANK LOAN CASE STUDY

Design

The following columns
needs to be dropped
as they have
more than 50% of NULL values
in application datasets: -

Column name	no_of_null_values	Percentage_of_null_values
COMMONAREA_AVG	34960	70%
COMMONAREA_MODE	34960	70%
COMMONAREA_MEDI	34960	70%
NONLIVINGAPARTMENTS_AVG	34714	69%
NONLIVINGAPARTMENTS_MODE	34714	69%
NONLIVINGAPARTMENTS_MEDI	34714	69%
LIVINGAPARTMENTS_AVG	34226	68%
LIVINGAPARTMENTS_MODE	34226	68%
LIVINGAPARTMENTS_MEDI	34226	68%
FONDKAPREMONT_MODE	34191	68%
FLOORSMIN_AVG	33894	68%
FLOORSMIN_MODE	33894	68%
FLOORSMIN_MEDI	33894	68%
YEARS_BUILD_AVG	33239	66%
YEARS_BUILD_MODE	33239	66%
YEARS_BUILD_MEDI	33239	66%
OWN_CAR_AGE	32949	66%
LANDAREA_AVG	29721	59%
LANDAREA_MODE	29721	59%
LANDAREA_MEDI	29721	59%
BASEMENTAREA_AVG	29199	58%
BASEMENTAREA_MODE	29199	58%
BASEMENTAREA_MEDI	29199	58%
EXT_SOURCE_1	28172	56%
NONLIVINGAREA_AVG	27572	55%
NONLIVINGAREA_MODE	27572	55%
NONLIVINGAREA_MEDI	27572	55%
ELEVATORS_AVG	26651	53%
ELEVATORS_MODE	26651	53%
ELEVATORS_MEDI	26651	53%
WALLSMATERIAL_MODE	25459	51%
APARTMENTS_AVG	25385	51%
APARTMENTS_MODE	25385	51%
APARTMENTS_MEDI	25385	51%
ENTRANCES_AVG	25195	50%
ENTRANCES_MODE	25195	50%
ENTRANCES_MEDI	25195	50%
LIVINGAREA_AVG	25137	50%
LIVINGAREA_MODE	25137	50%
LIVINGAREA_MEDI	25137	50%
HOUSETYPE_MODE	25075	50%
FLOORSMAX_AVG	24875	50%
FLOORSMAX_MODE	24875	50%
FLOORSMAX_MEDI	24875	50%

BANK LOAN CASE STUDY

Design

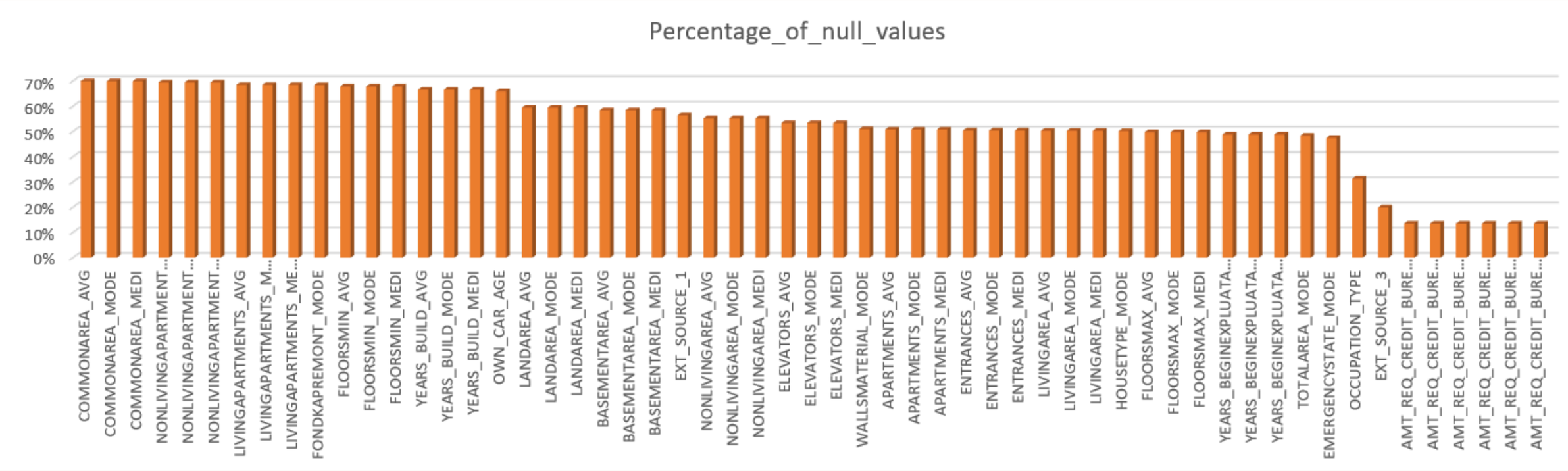
Columns having irrelevant data which is not required for analysis.

Column name	no_of_null_values	Percentage_of_null_values
FLAG_MOBIL	0	0%
FLAG_EMP_PHONE	0	0%
FLAG_WORK_PHONE	0	0%
FLAG_CONT_MOBILE	0	0%
FLAG_PHONE	0	0%
FLAG_EMAIL	0	0%
CNT_FAM_MEMBERS	1	0%
REGION_RATING_CLIENT	0	0%
REGION_RATING_CLIENT_W_CITY	0	0%
EXT_SOURCE_2	126	0%
EXT_SOURCE_3	9944	20%
YEARS_BEGINEXPLUATATION_AVG	24394	49%
YEARS_BEGINEXPLUATATION_MODE	24394	49%
YEARS_BEGINEXPLUATATION_MEDI	24394	49%
TOTALAREA_MODE	24148	48%
EMERGENCYSTATE_MODE	23698	47%
DAYS_LAST_PHONE_CHANGE	1	0%
FLAG_DOCUMENT_2	0	0%
FLAG_DOCUMENT_3	0	0%
FLAG_DOCUMENT_4	0	0%
FLAG_DOCUMENT_5	0	0%
FLAG_DOCUMENT_6	0	0%
FLAG_DOCUMENT_7	0	0%
FLAG_DOCUMENT_8	0	0%
FLAG_DOCUMENT_9	0	0%
FLAG_DOCUMENT_10	0	0%
FLAG_DOCUMENT_11	0	0%
FLAG_DOCUMENT_12	0	0%
FLAG_DOCUMENT_13	0	0%
FLAG_DOCUMENT_14	0	0%
FLAG_DOCUMENT_15	0	0%
FLAG_DOCUMENT_16	0	0%
FLAG_DOCUMENT_17	0	0%
FLAG_DOCUMENT_18	0	0%
FLAG_DOCUMENT_19	0	0%
FLAG_DOCUMENT_20	0	0%
FLAG_DOCUMENT_21	0	0%

BANK LOAN CASE STUDY

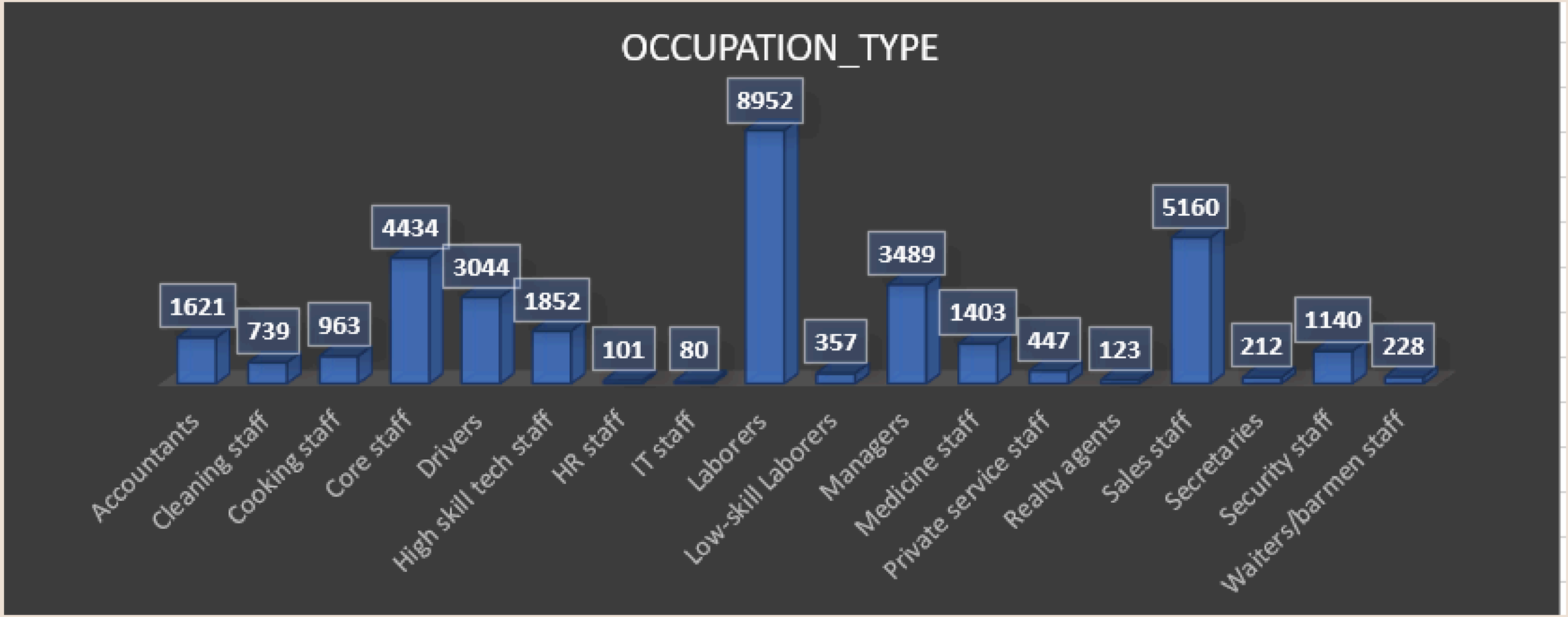
Design

Bar graph showing the percentage of null values that are dropped.



BANK LOAN CASE STUDY

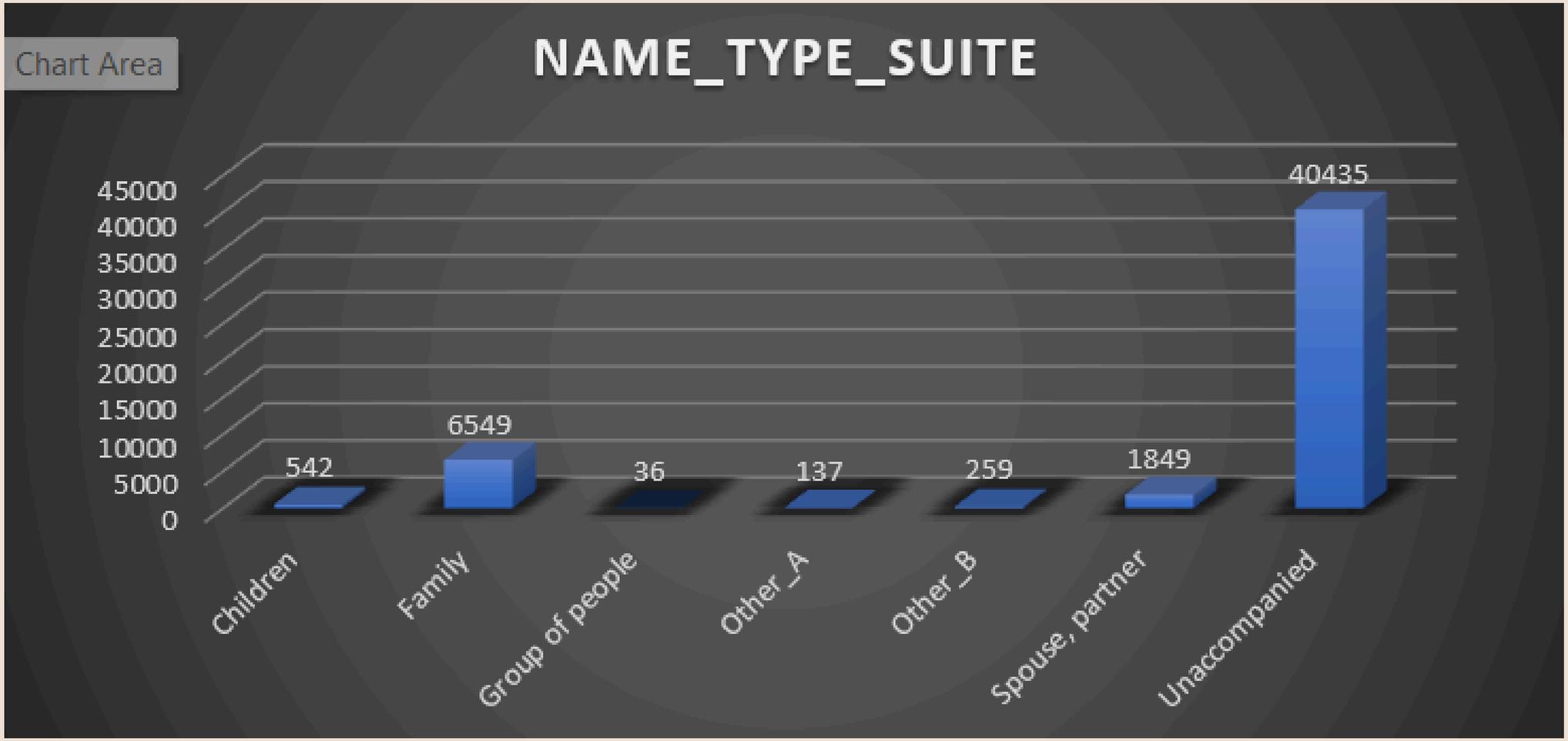
Findings-1



Highest Occurring Variable is **Laborers.**

BANK LOAN CASE STUDY

Findings-2

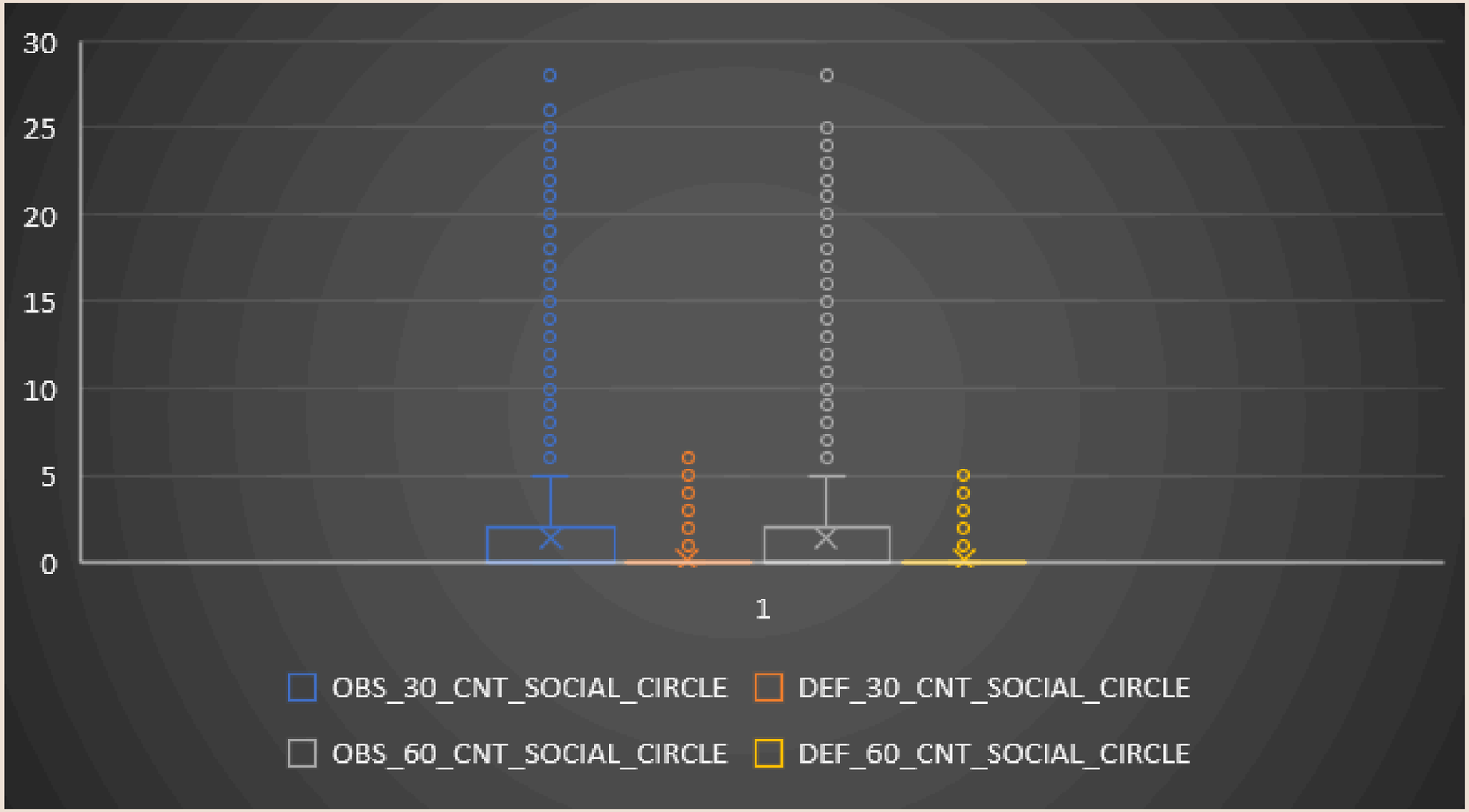


Highest Occurring Variable is **Unacccompanied.**

BANK LOAN CASE STUDY

Findings-3

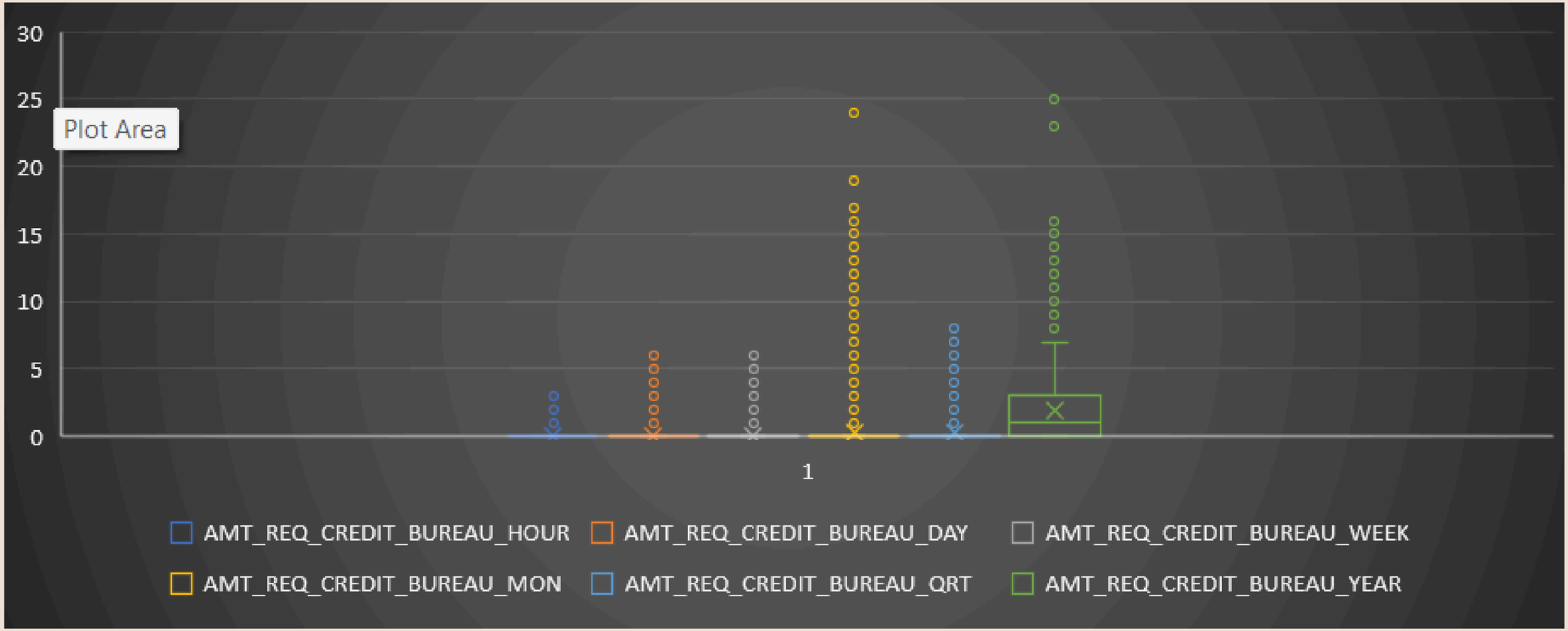
RESULT: Replacing Blanks in other columns of the Application Dataset using MEDIAN and MODE



BANK LOAN CASE STUDY

Findings-4

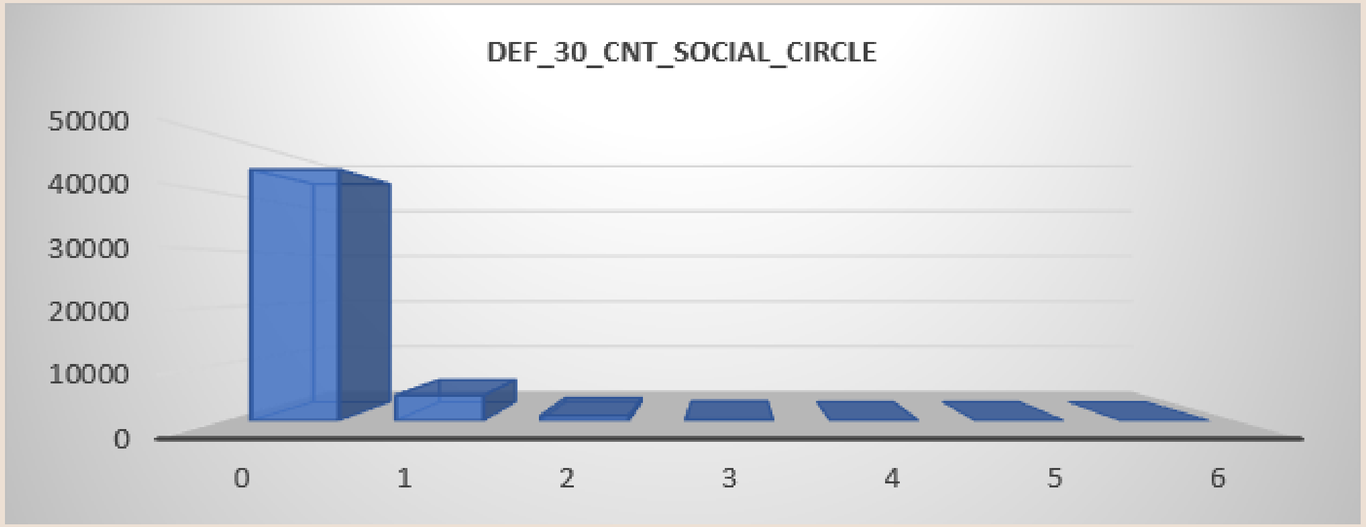
RESULT: Replacing Blanks in other columns of the Application Dataset using MEDIAN and MODE

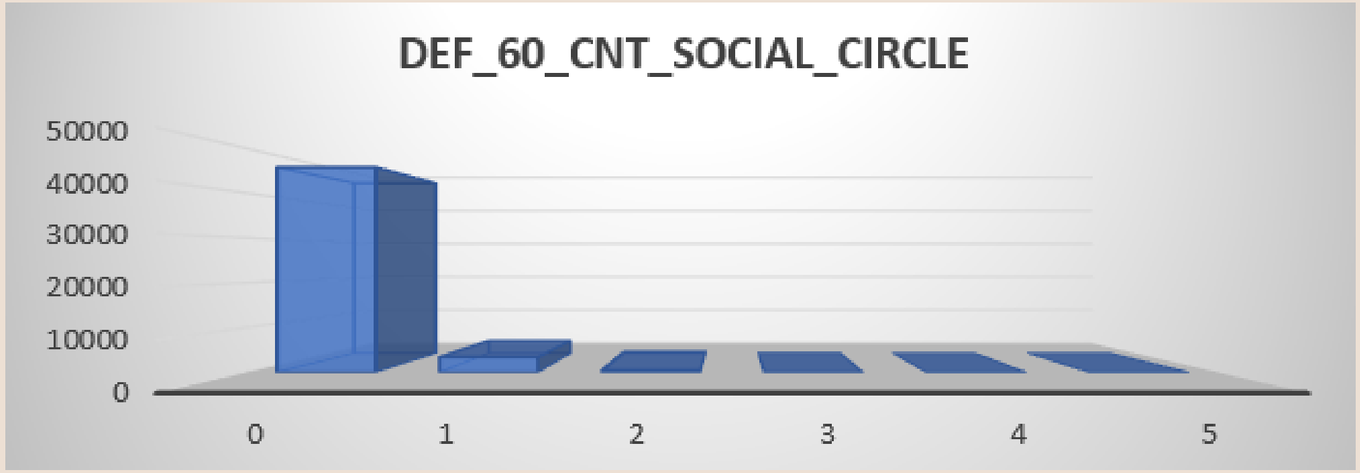


BANK LOAN CASE STUDY

Findings-5

RESULT: Replacing Blanks in other columns of the Application Dataset using MEDIAN and MODE

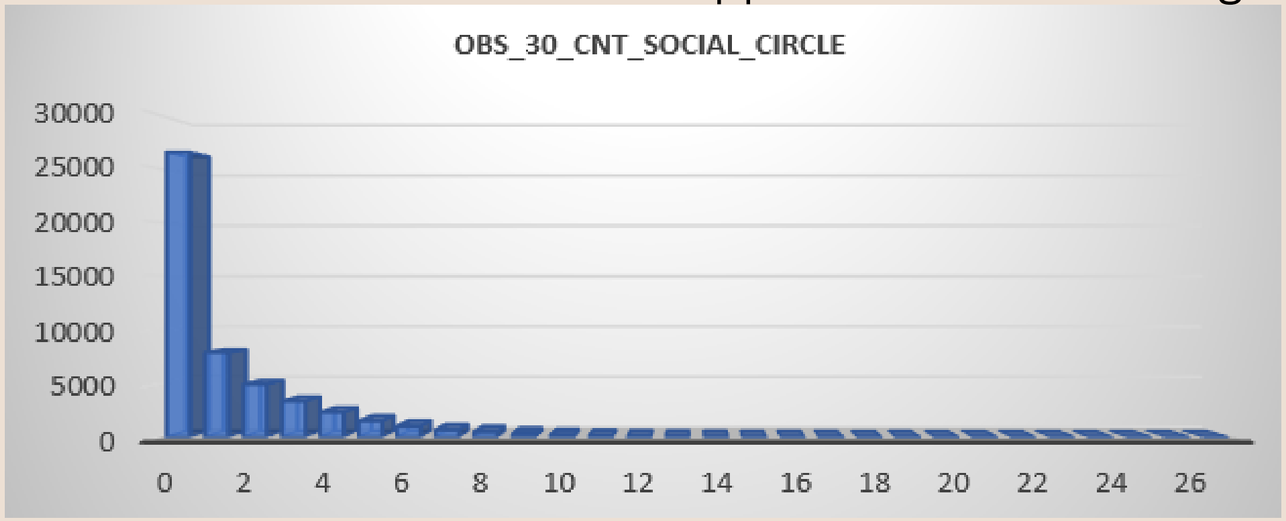


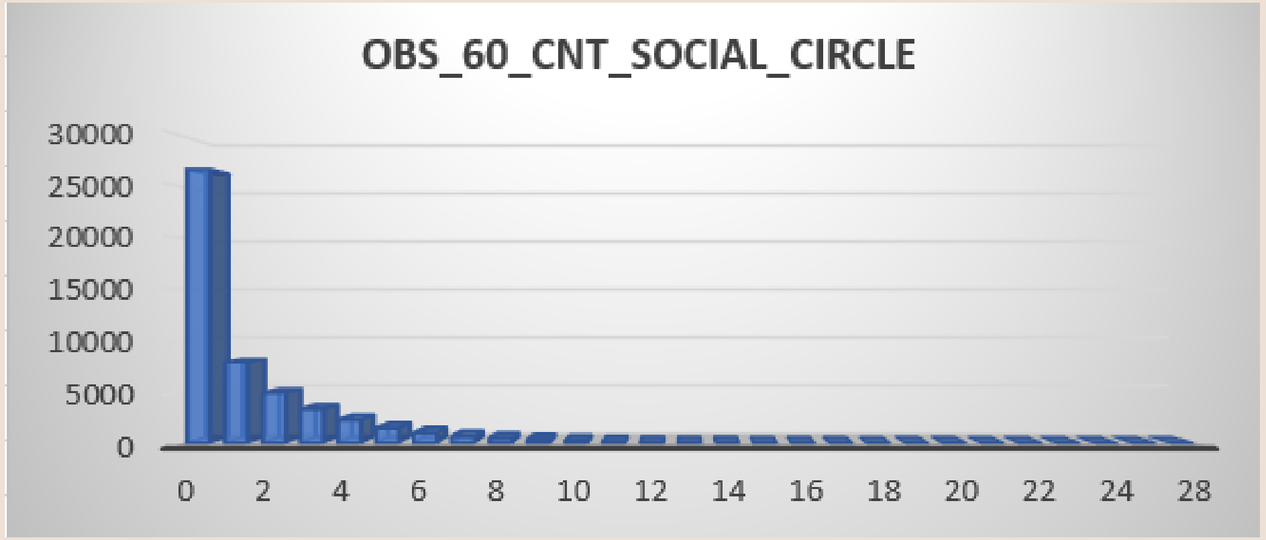


BANK LOAN CASE STUDY

Findings-6

RESULT: Replacing Blanks in other columns of the Application Dataset using MEDIAN and MODE

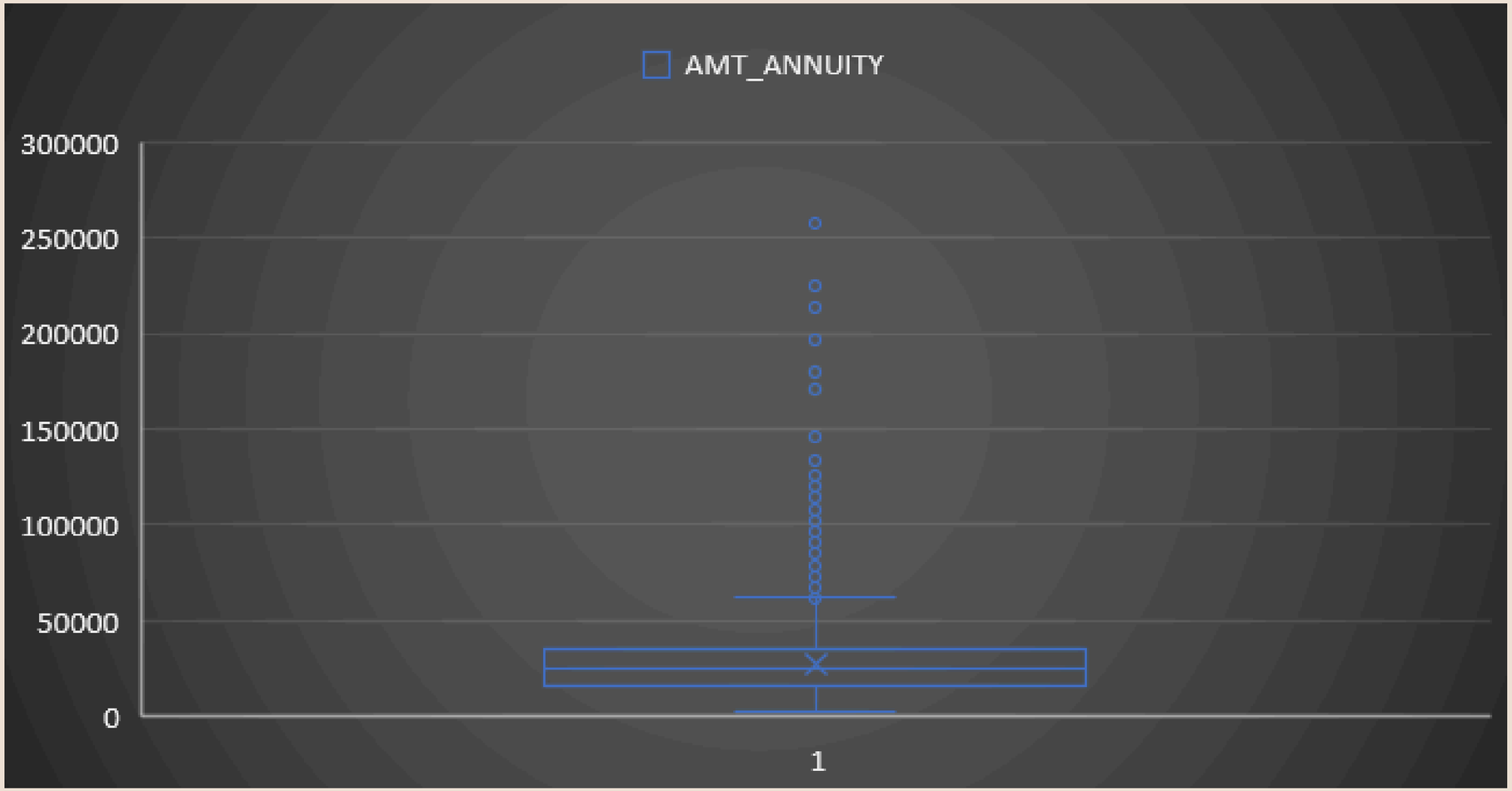




BANK LOAN CASE STUDY

Findings-7

RESULT: Replacing Blanks in AMT_ANNUTIY column of the Application Dataset with the median



BANK LOAN CASE STUDY

Findings-8

RESULT: Replacing Blanks in AMT_GOODS_PRICE column of the Application Dataset with the median

BANK LOAN CASE STUDY

Findings-9

To detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Formulae:

Quartile 1 : =QUARTILE(A:A,1)

Quartile 3 : =QUARTILE(A:A,3)

IQR = Quartile 3 - Quartile 1

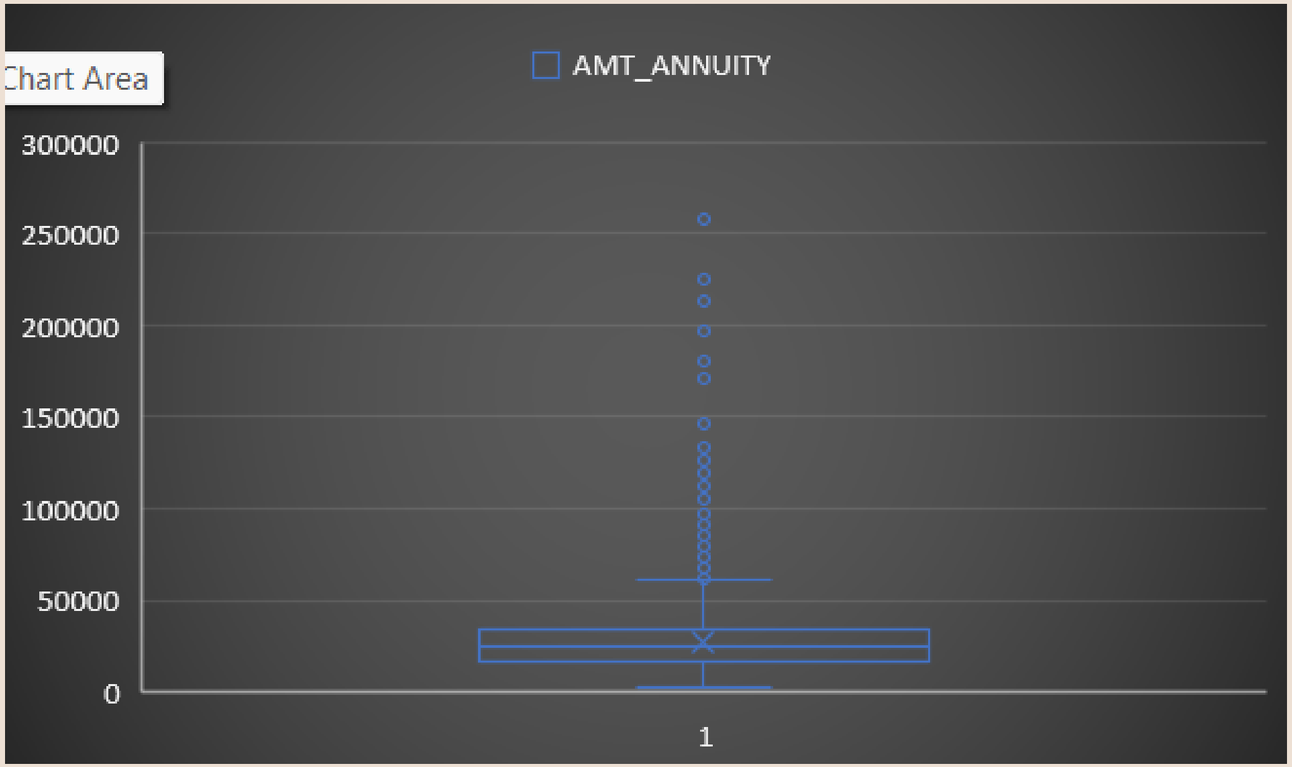
Upper Limit = Quartile 3 + 1.5*IQR

Lower Limit = Quartile 1 – 1.5IQR

AMT_INCOME_TOTAL	
Quartile 1	112500
Quartile 3	202500
IQR	90000
Upper Limit	337500
Lower Limit	-22500

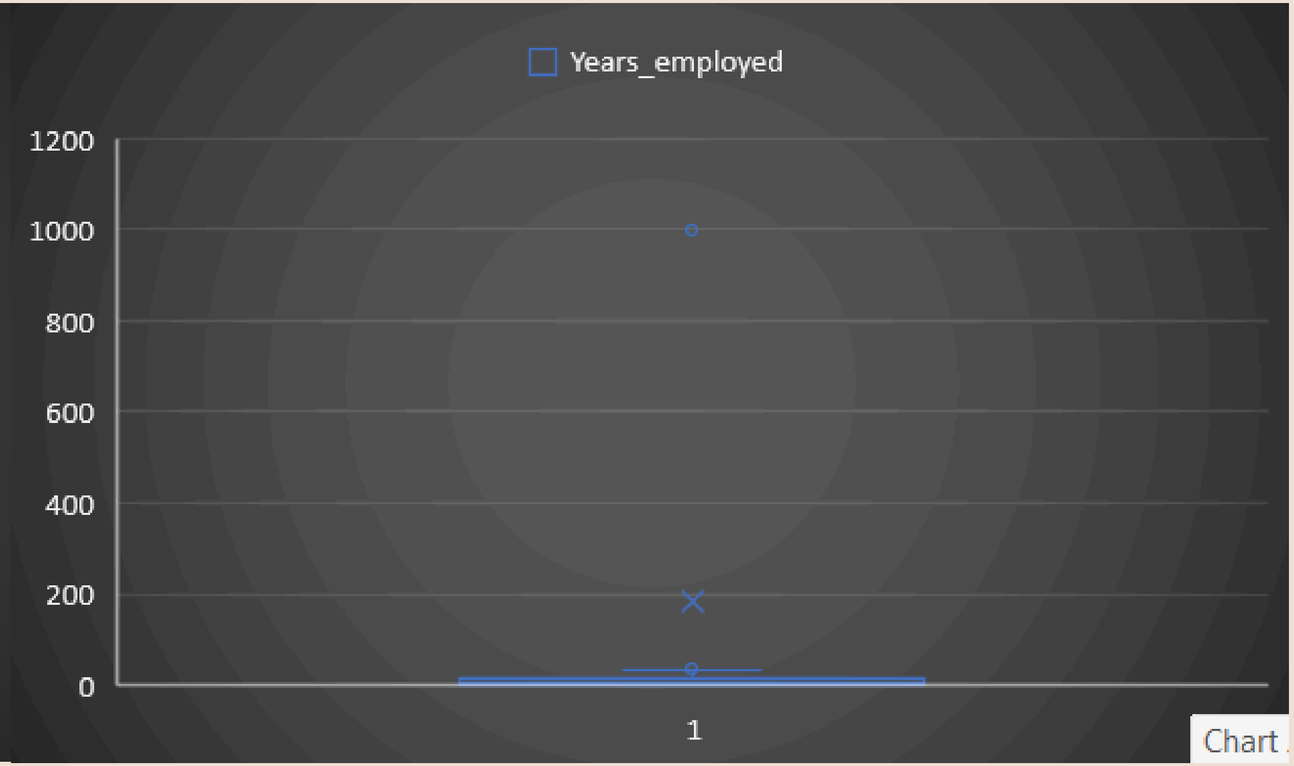
BANK LOAN CASE STUDY

Findings-9.



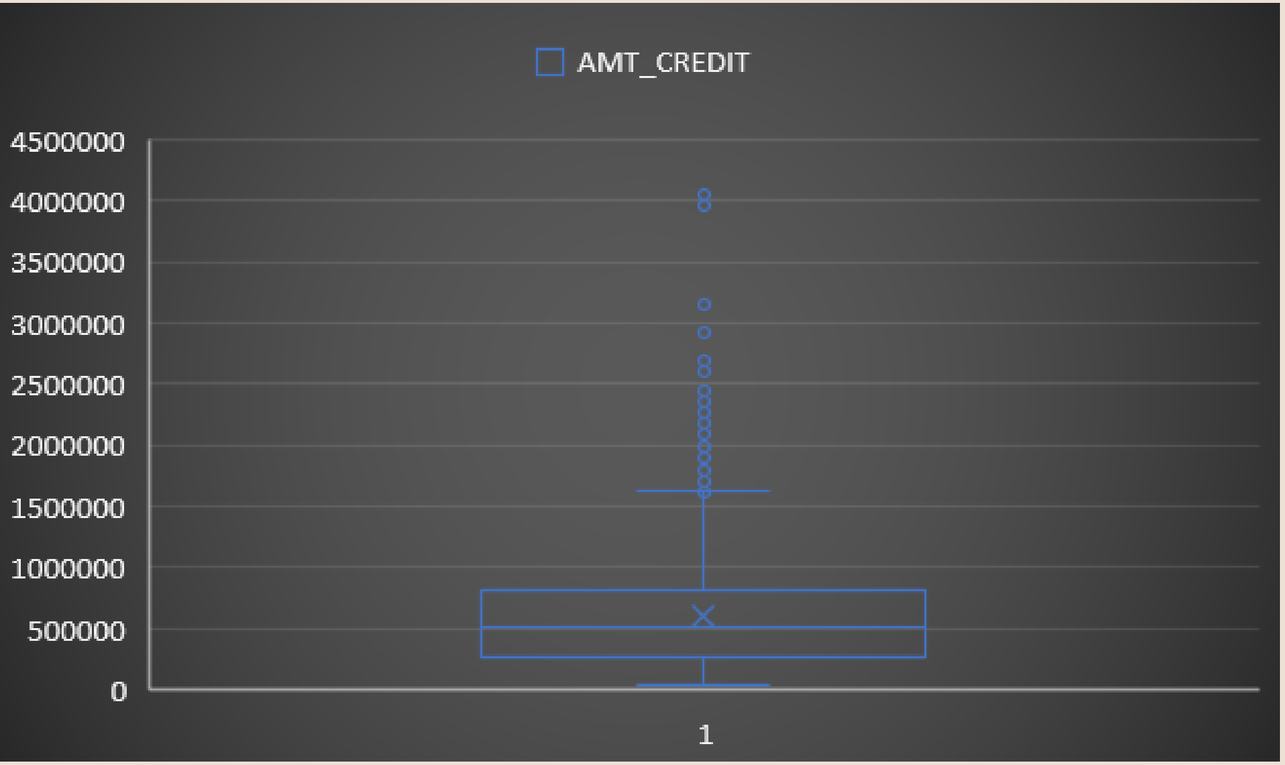
BANK LOAN CASE STUDY

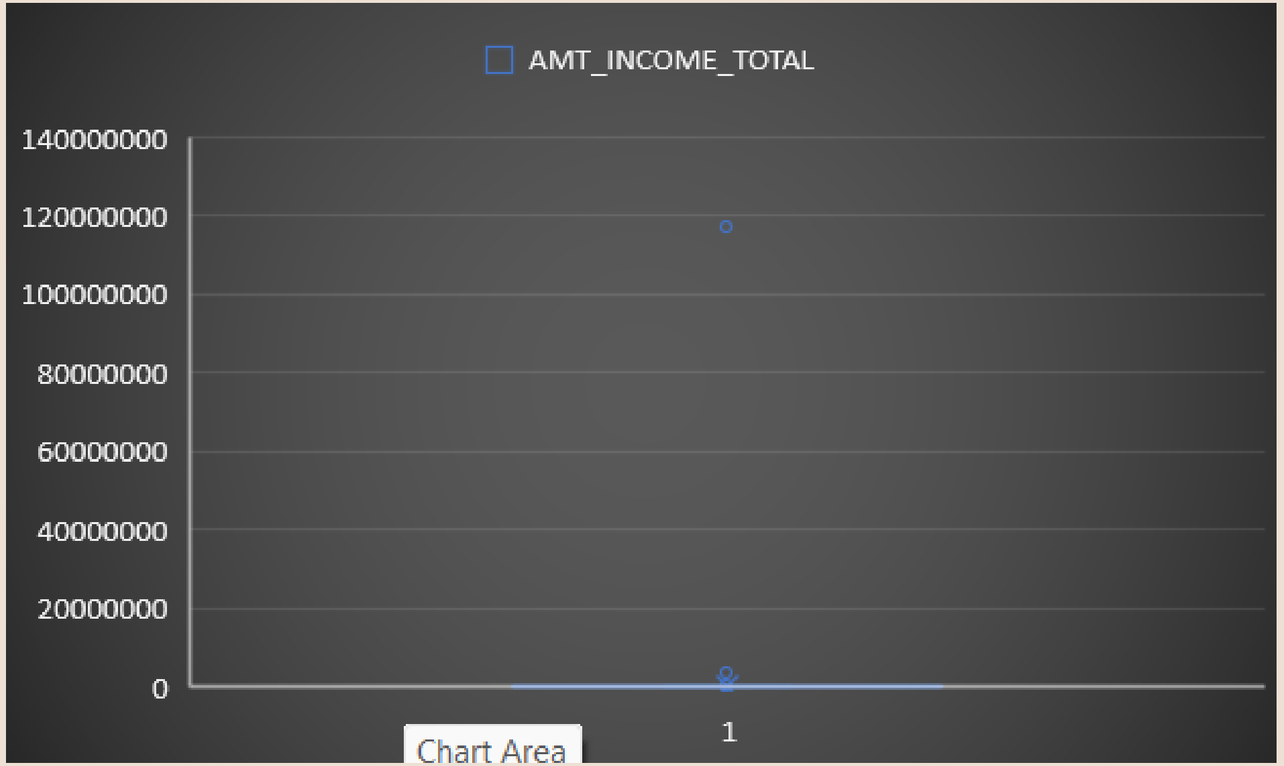
Findings-9.



BANK LOAN CASE STUDY

Findings-9.

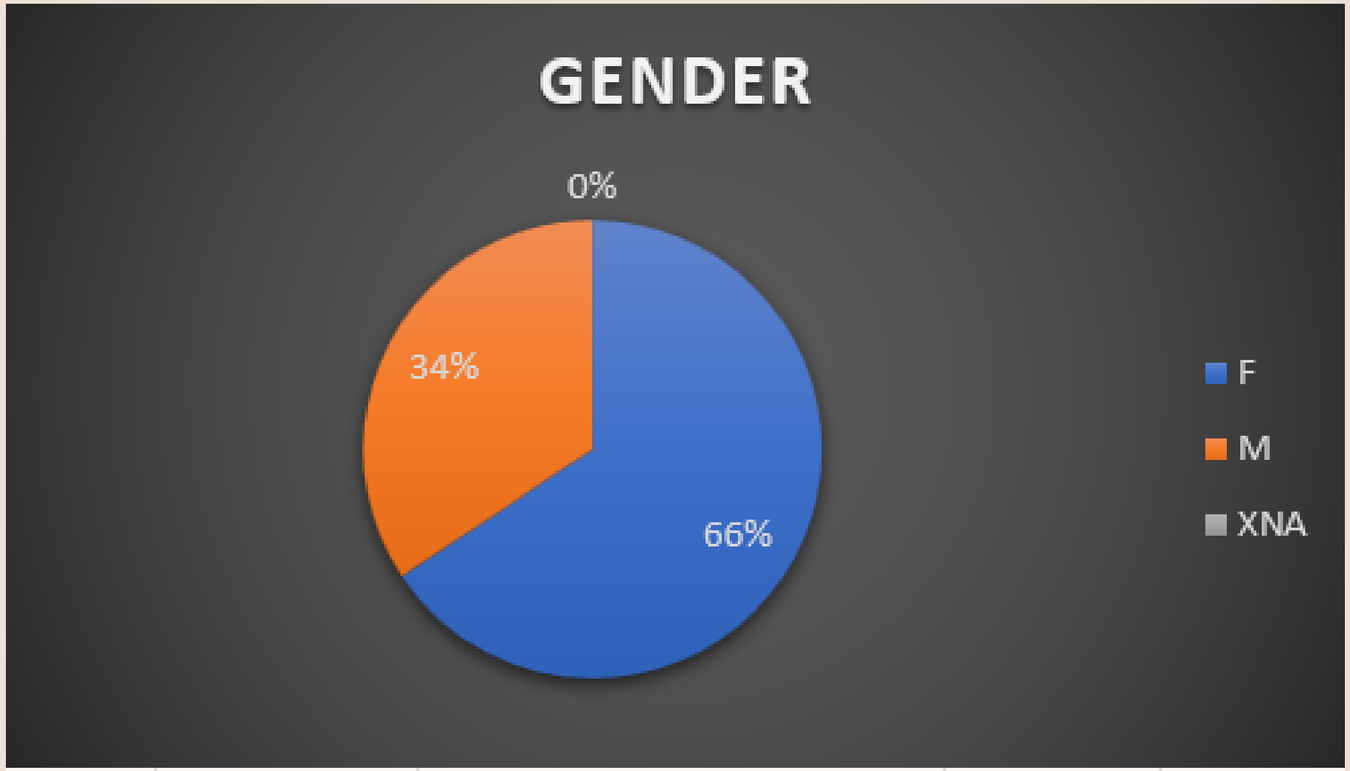




BANK LOAN CASE STUDY

Findings-10

To determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

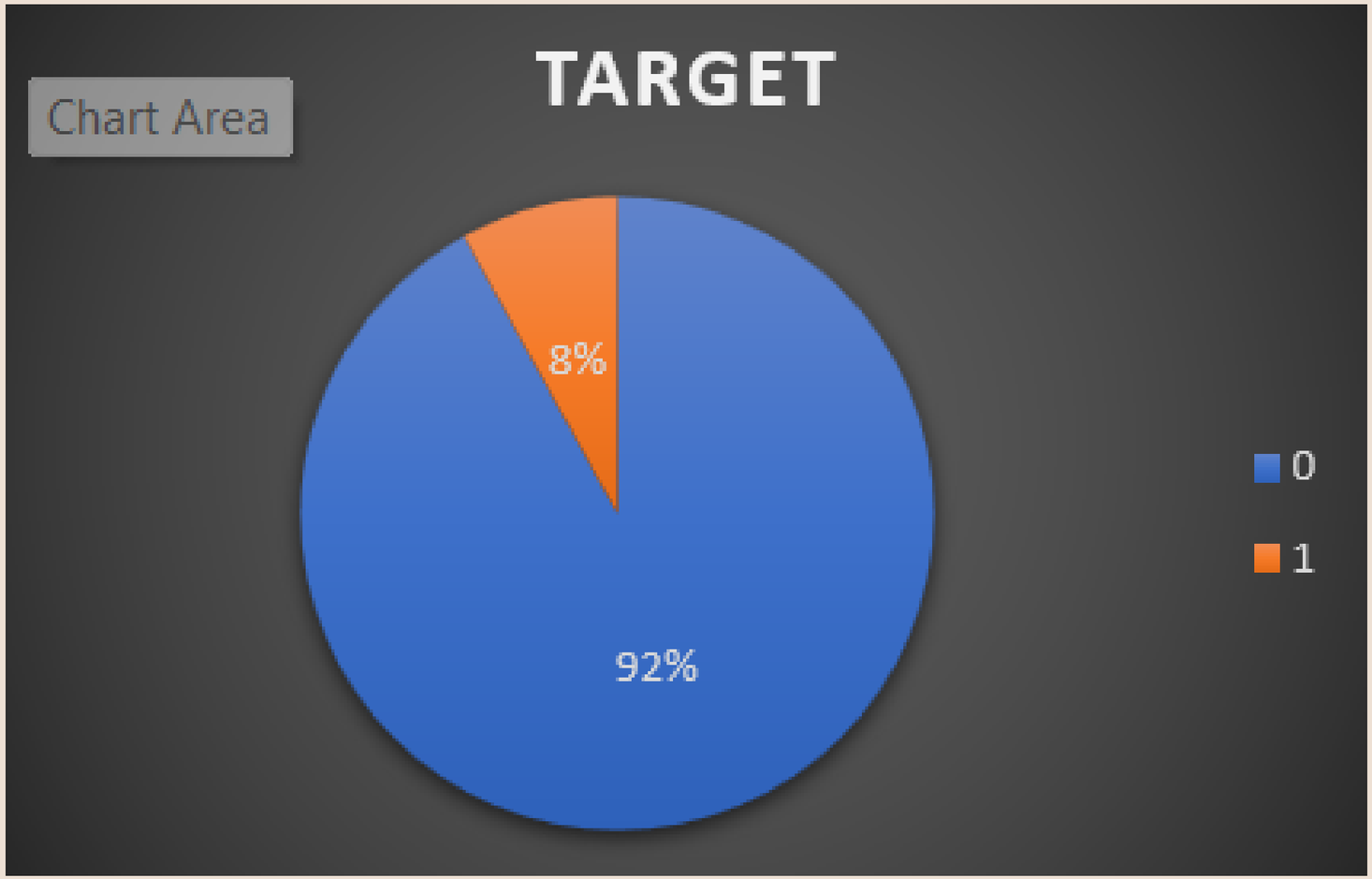


GENDER	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**

BANK LOAN CASE STUDY

Findings-10

To determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.



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

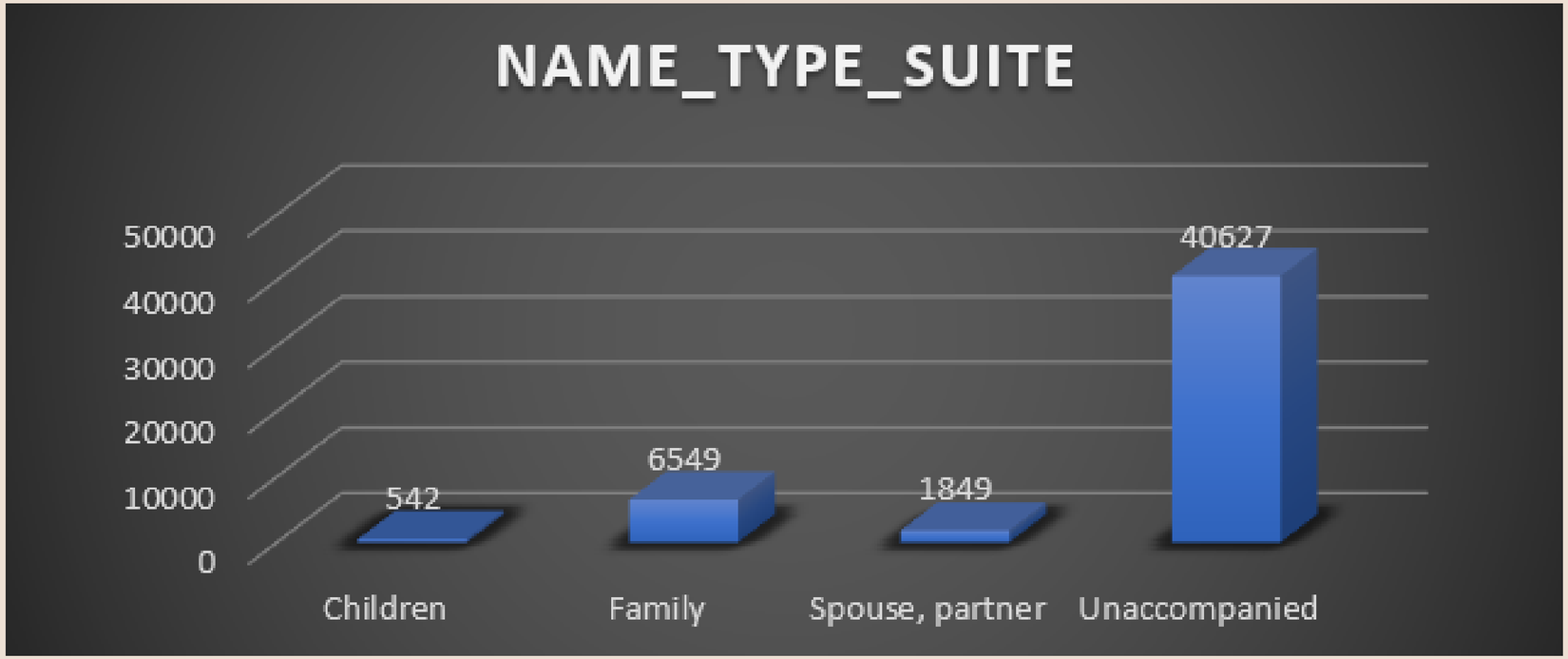
Almost 92% clients are loan re-payers.
8% client are Defaulters.

BANK LOAN CASE STUDY

Findings-II

To perform univariate analysis to understand the distribution of individual variables, segmented univariat analysis to compare variable distributions for different scenarios, and bivariate analysis to explor relationships between variables and the target variable using Excel functions and features.

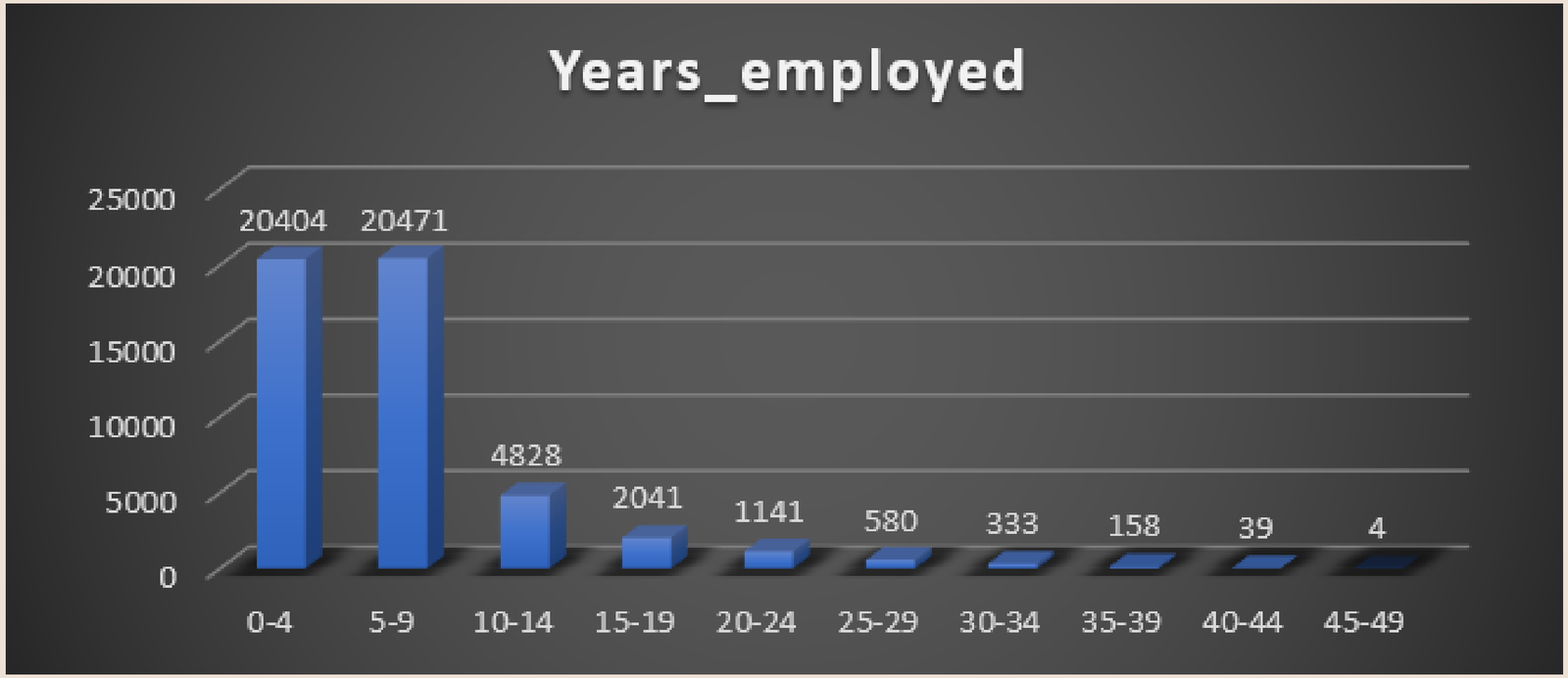
Row Labels	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Spouse, partner	1849
Unaccompanied	40627
Grand Total	**49567**



BANK LOAN CASE STUDY

Findings-12

Row Labels	Count of Years_employed
0-4	20404
5-9	20471
10-14	4828
15-19	2041
20-24	1141
25-29	580
30-34	333
35-39	158
40-44	39
45-49	4
Grand Total	**49999**

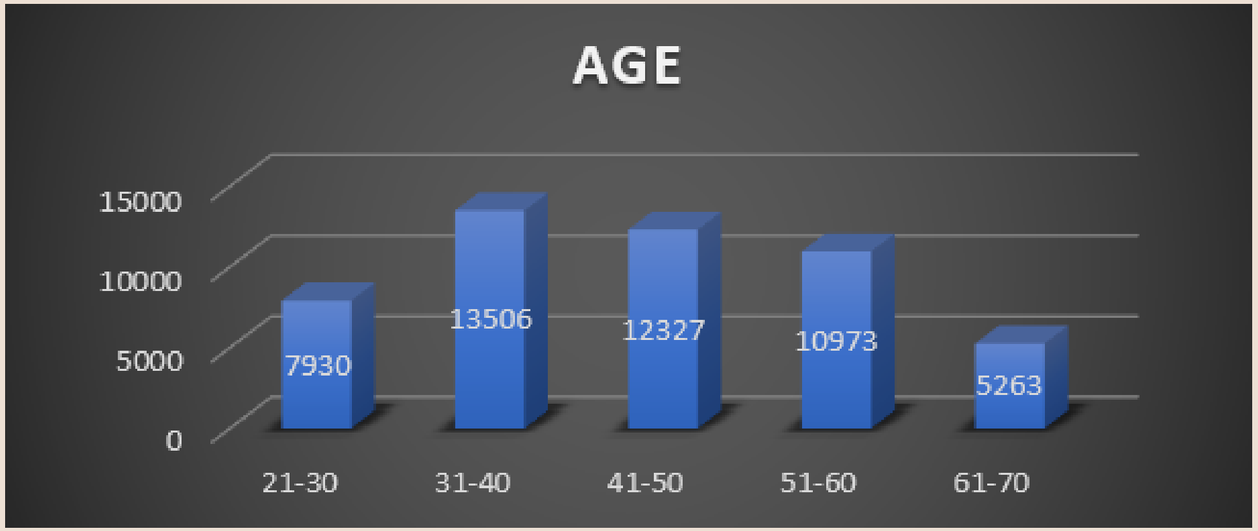


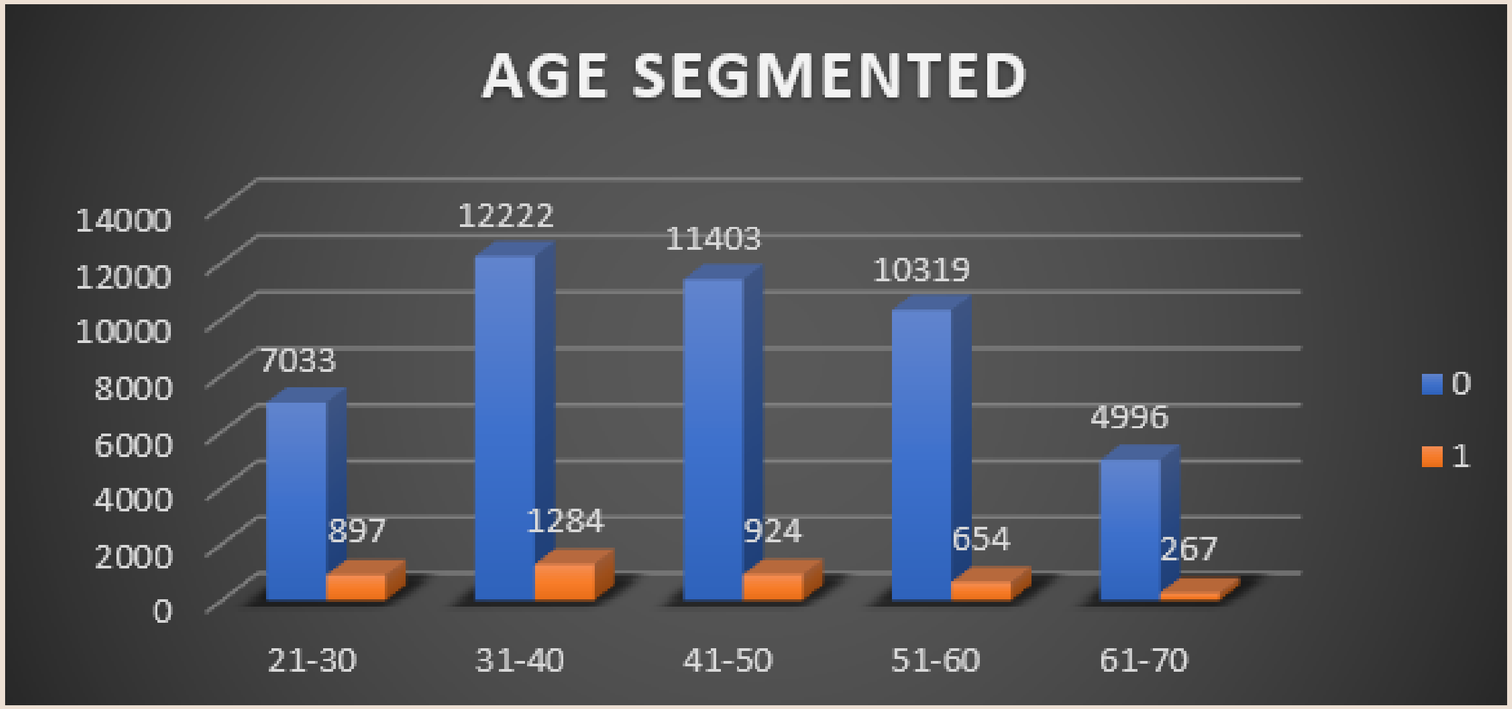
Majority of the Clients are having 0-9 years of experience.
We can see as experience increases, chances of defaultingdecreases.

BANK LOAN CASE STUDY

Findings-14

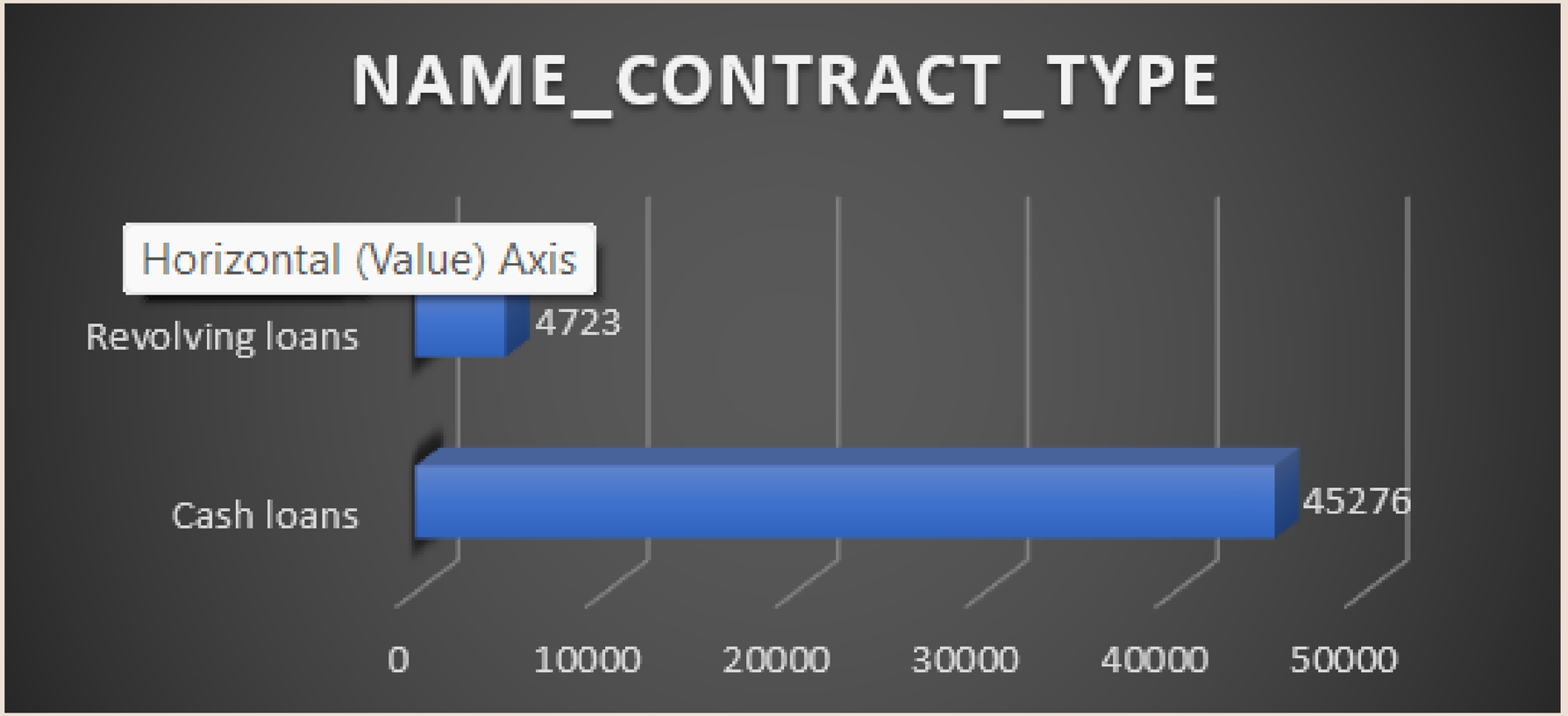
Majority of the Clients are in the age group 31-40.
We can see in AGE SEGMENTED, as age increases , chances of defaulter decreases.





BANK LOAN CASE STUDY

Findings-15

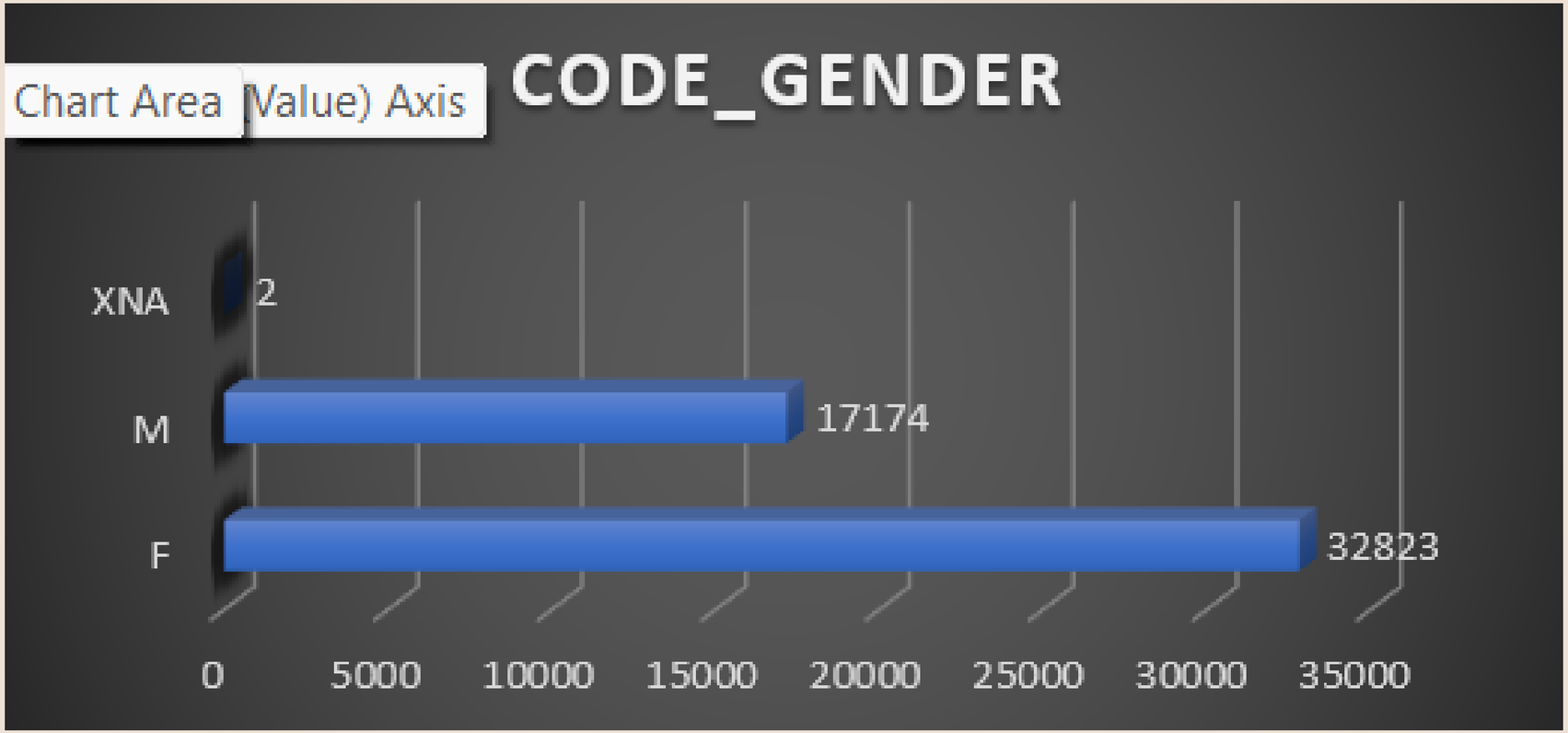


Majority of the Clients are taking Cash loans.

BANK LOAN CASE STUDY

Findings-16

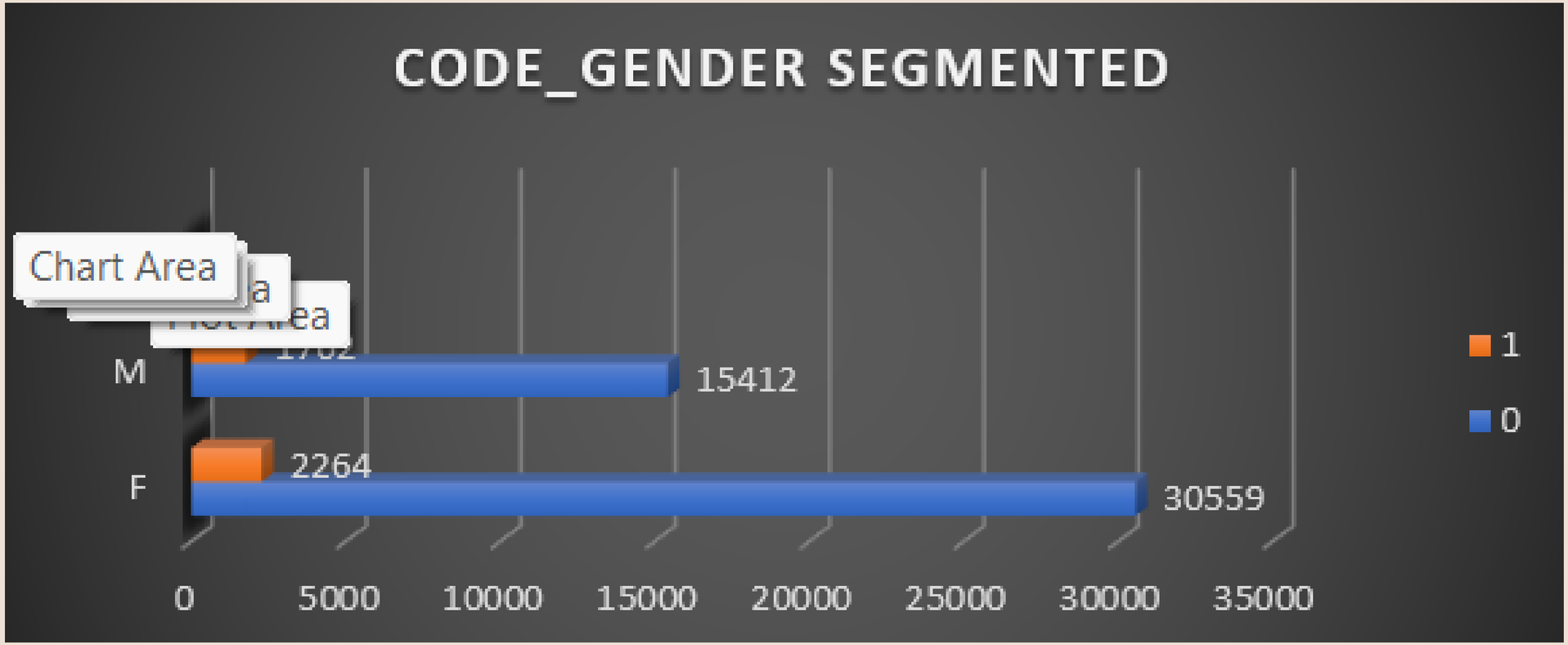
Row Labels	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	**49999**



Male are less defaulters compared to Female.

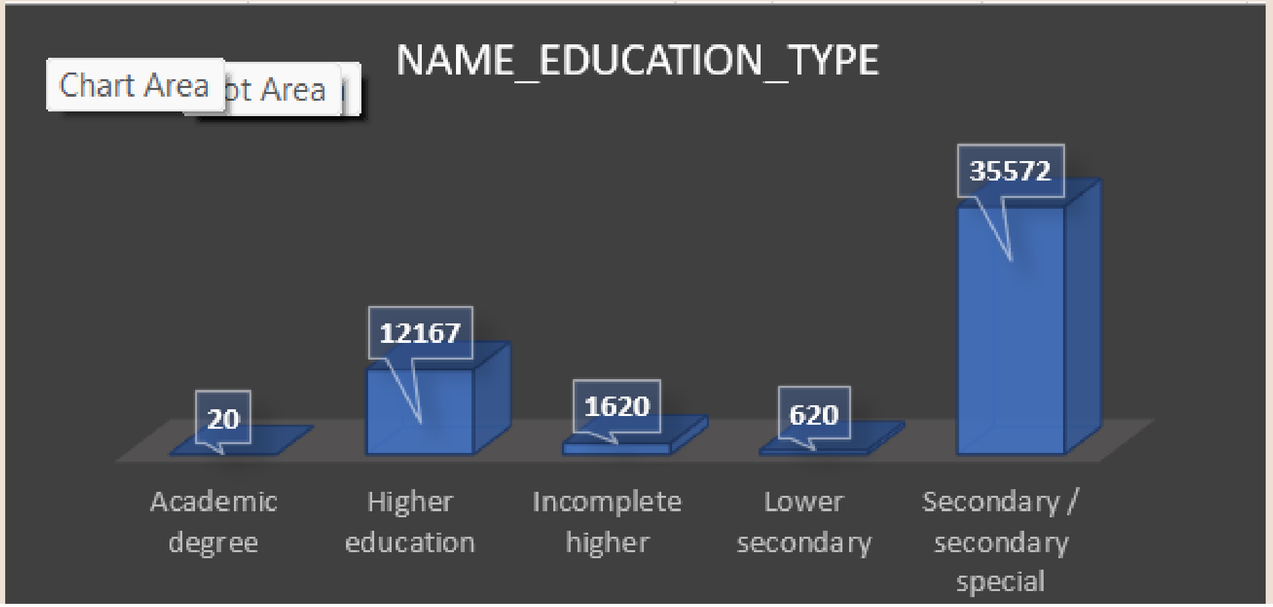
BANK LOAN CASE STUDY

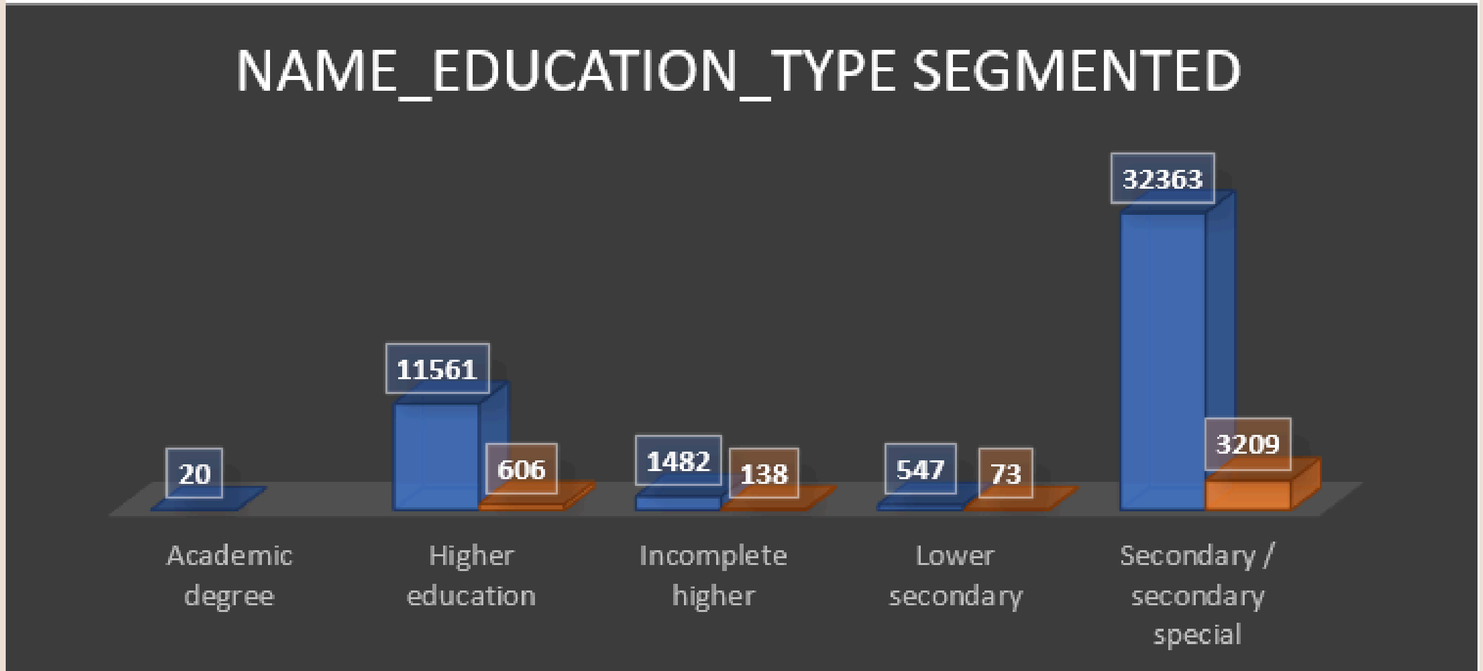
Findings-17.



Male are less defaulters compared to Female.

BANK LOAN CASE STUDY





The numbers of loans taken by Clients with Secondary special Education is the highest and Academic degree is the lowest.

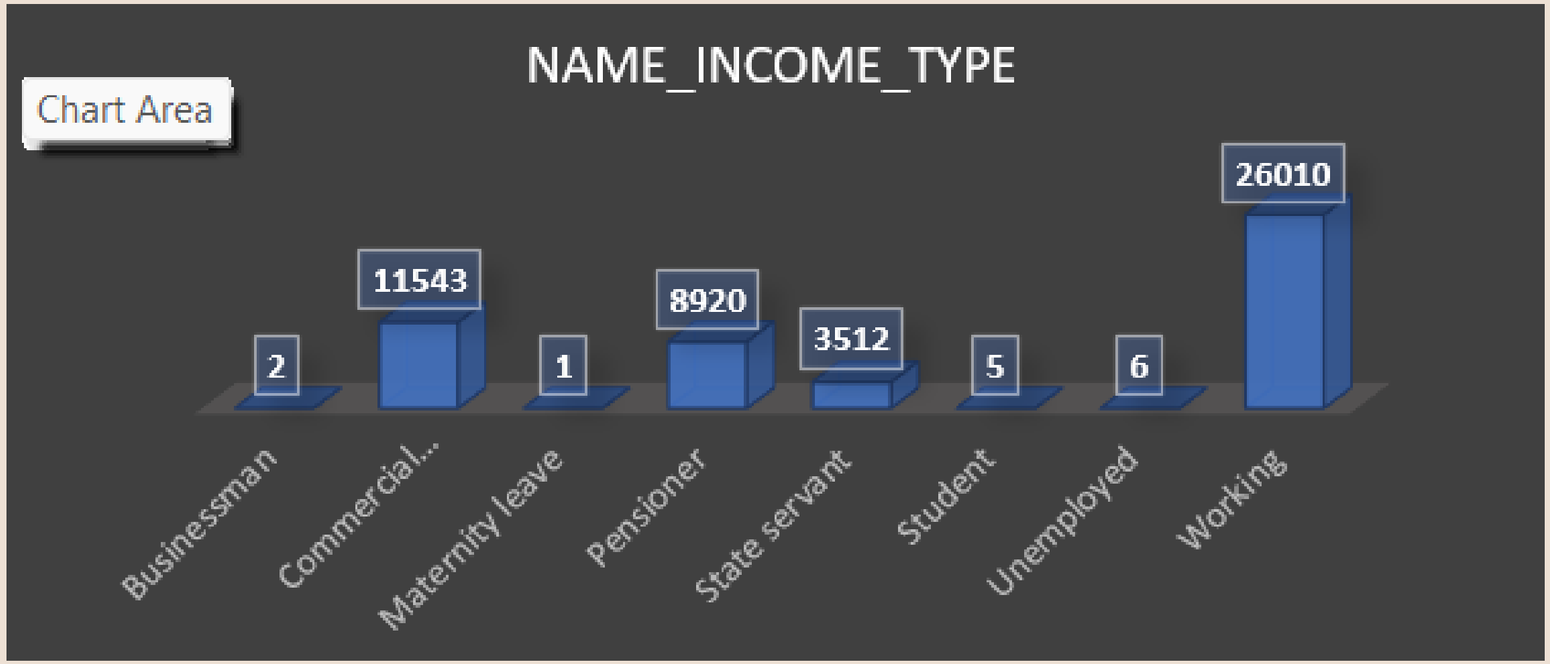
Least default: Academic degree

Highest default: Secondary special

BANK LOAN CASE STUDY

Findings-19

Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010
Grand Total	**49999**



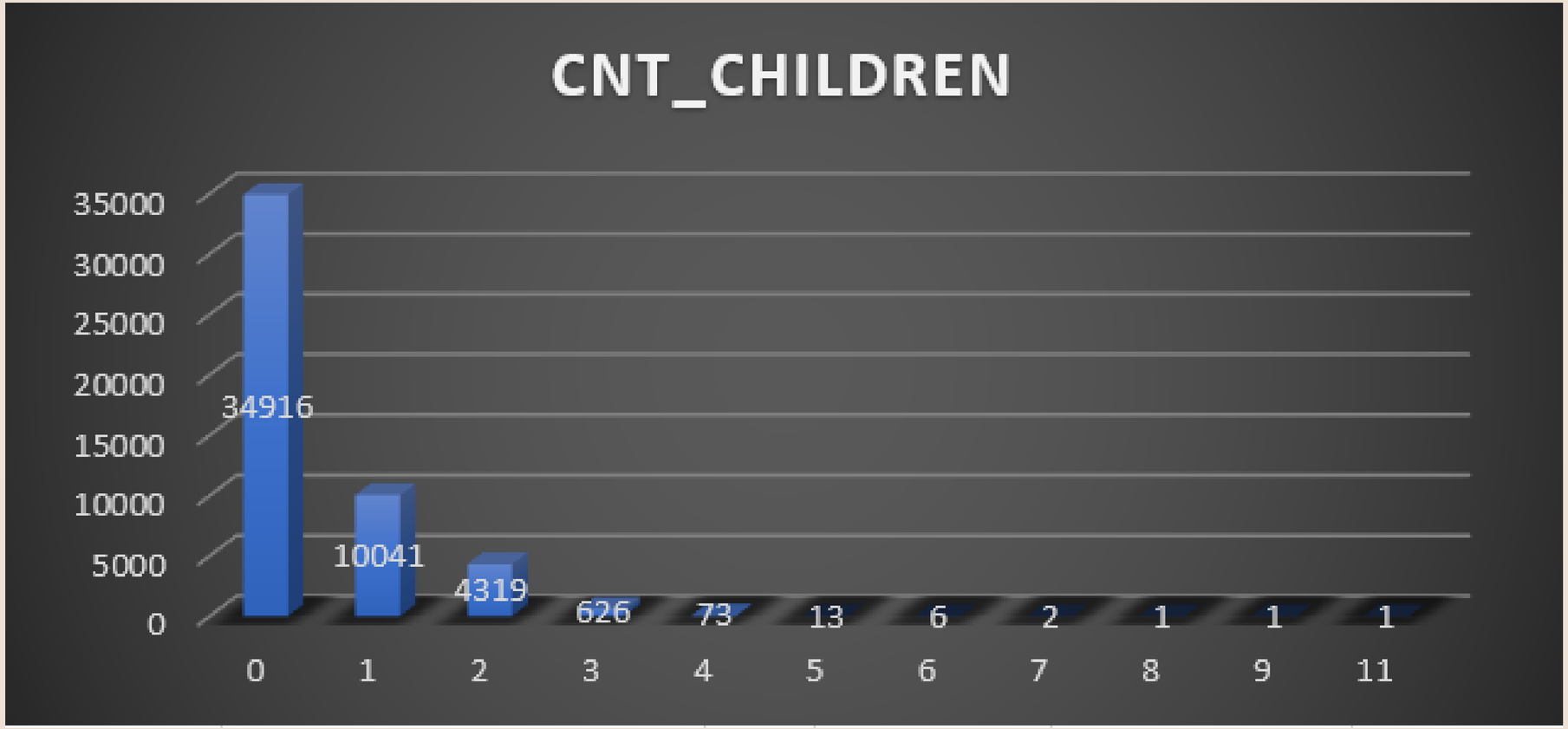
Bank target those groups whose income type is working.

Least default: Client who is Businessman or student or at Maternity leave.

Highest default: Client who is working

BANK LOAN CASE STUDY

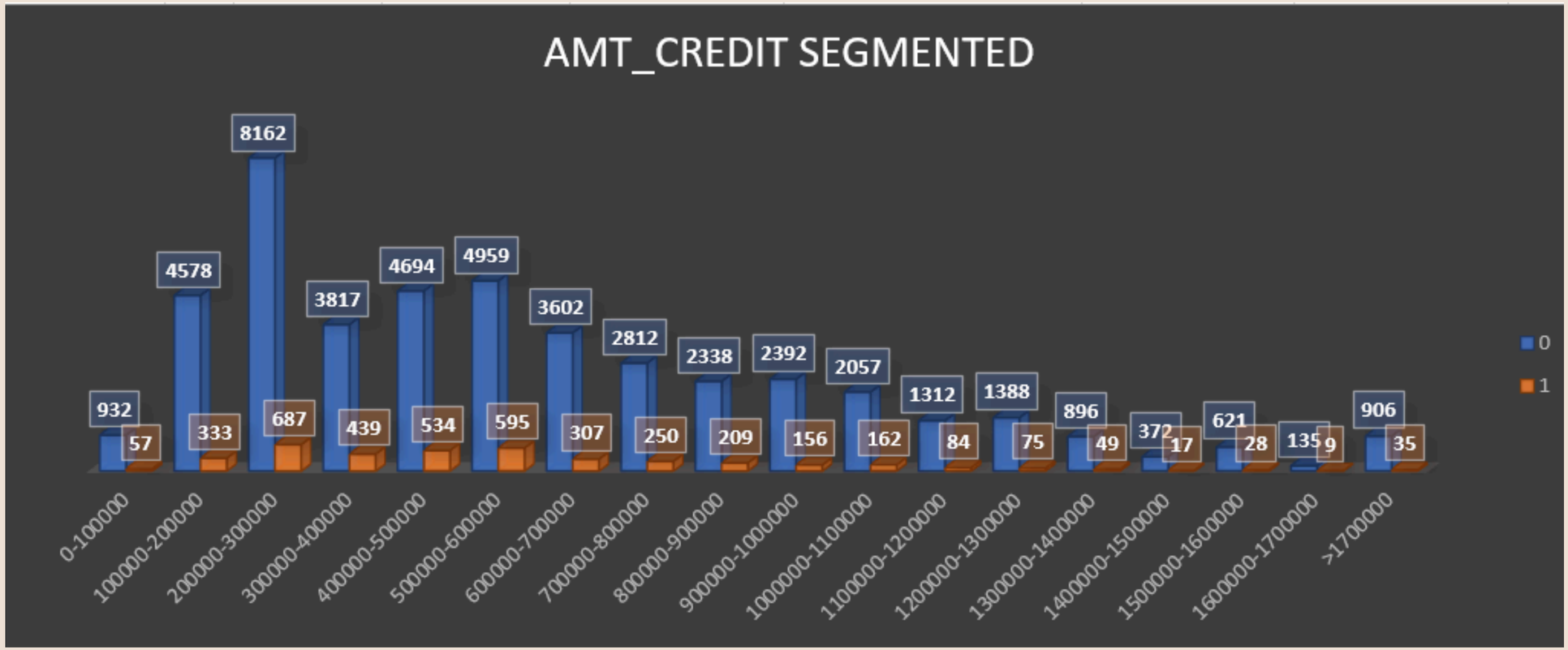
Findings-20



The highest number of loans are taken by Clients who does not have a child.

BANK LOAN CASE STUDY

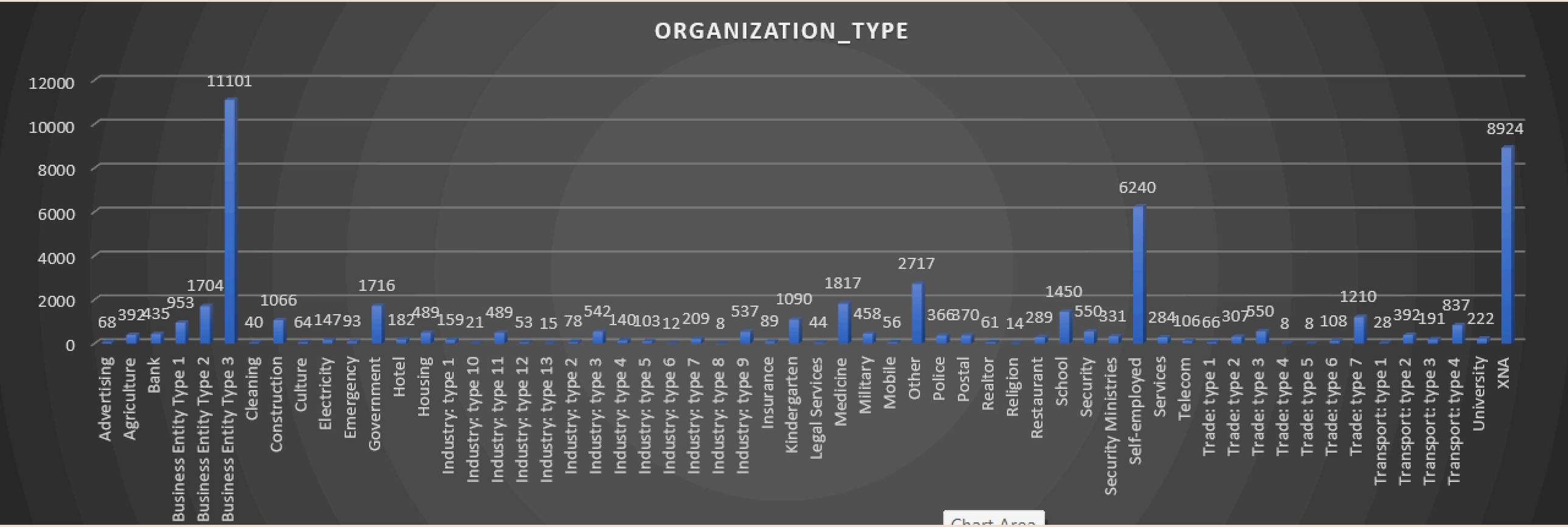
Findings-20



Majority of the Clients took the loan between 2L – 3L.

BANK LOAN CASE STUDY

Findings-21



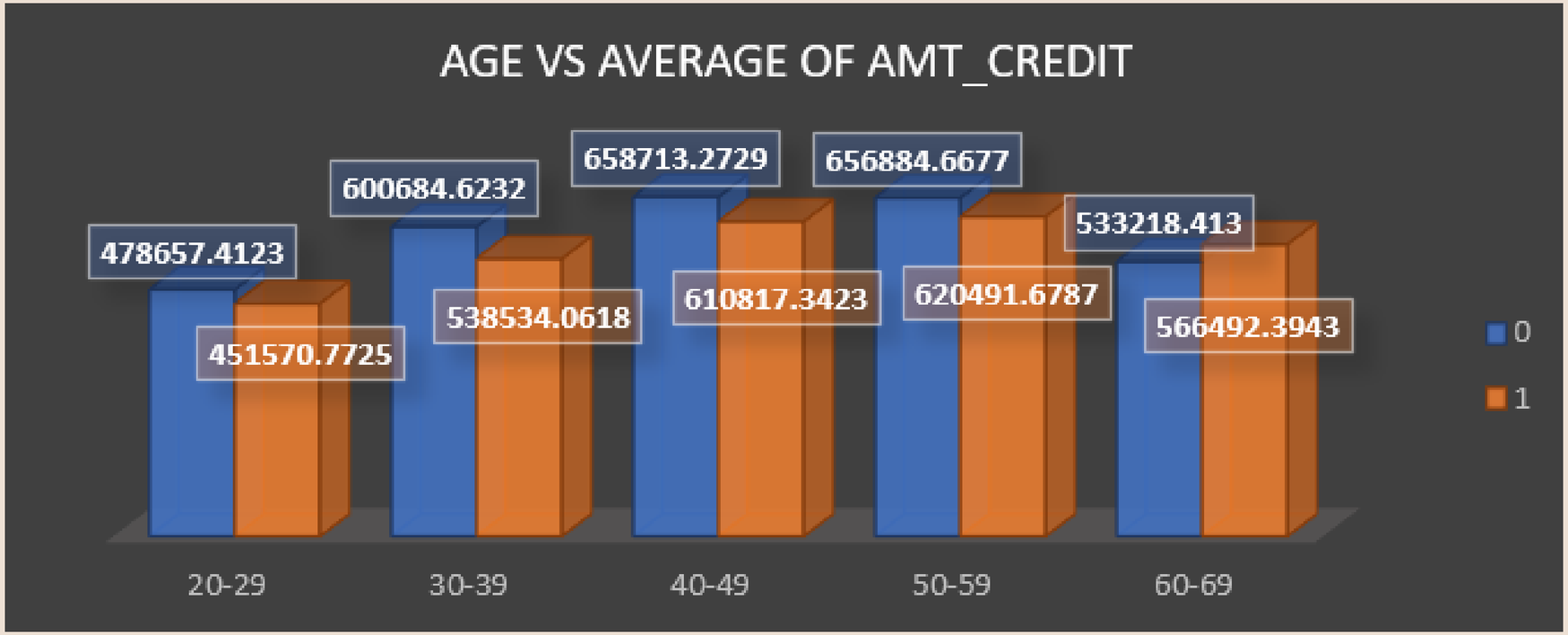
Clients who are working in business Entity type of Organization took the highest number of loans.

BANK LOAN CASE STUDY

Findings-22

Average of AMT_CREDIT	Column Labels	
Row Labels	0	1
20-29	478657.4123	451570.7725
30-39	600684.6232	538534.0618
40-49	658713.2729	610817.3423
50-59	656884.6677	620491.6787
60-69	533218.413	566492.3943
Grand Total	**603562.2995**	**555603.522**

Age group 40-49 took the highest amount of loan but age group 50-59 are defaulter
with highest amount of loan.

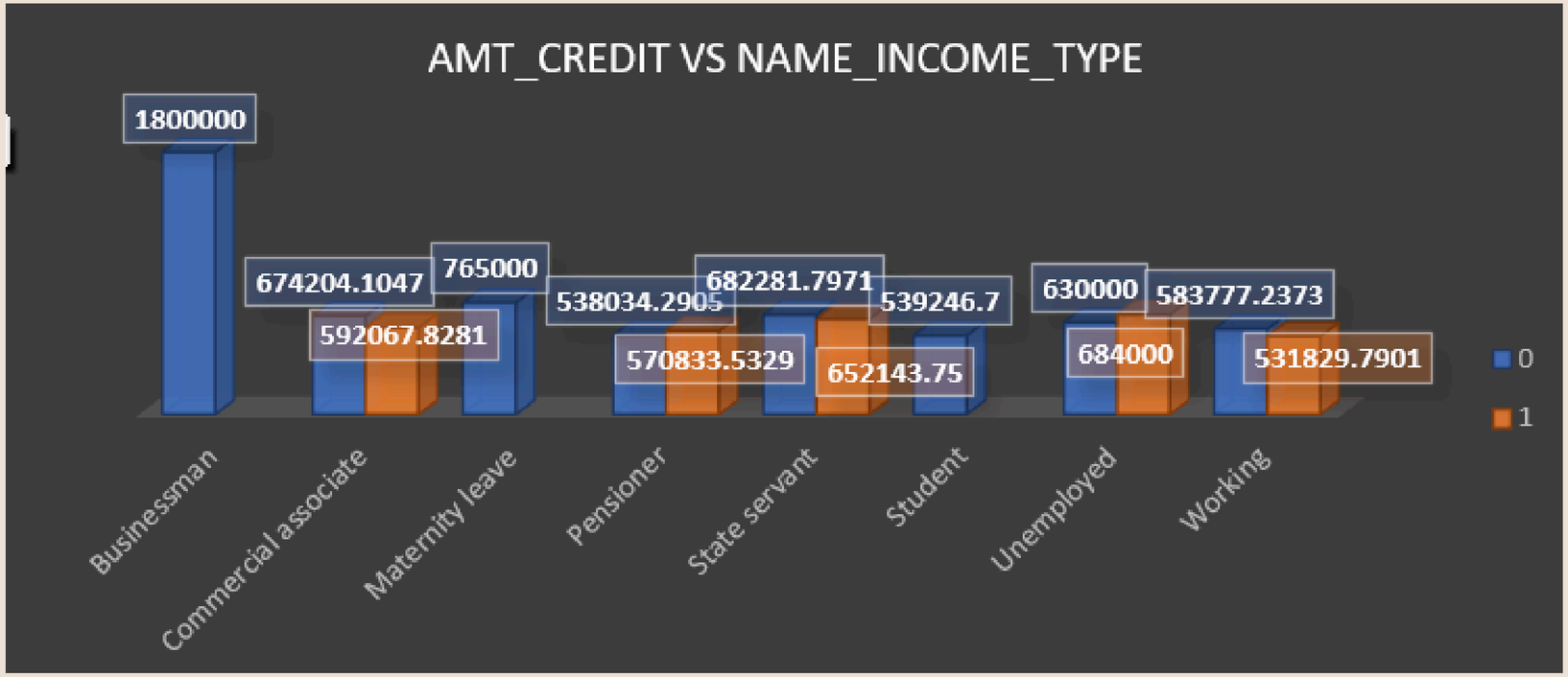


BANK LOAN CASE STUDY

Findings-23

Average of AMT_CREDIT	Column Labels	
NAME_INCOME_TYPE	0	1
Businessman	1800000	
Commercial associate	674204.1047	592067.8281
Maternity leave	765000	
Pensioner	538034.2905	570833.5329
State servant	682281.7971	652143.75
Student	539246.7	
Unemployed	630000	684000
Working	583777.2373	531829.7901
Grand Total	**603562.2995**	**555603.522**

Businessman took the highest amount of loan and did the payment on time.
Clients who are unemployed have highest amount of loan which they didn't repay on time.



BANK LOAN CASE STUDY

Findings-24

Top Correlation Coefficients for Payment difficulties are: -

Correlation between Columns	Values
OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SO	0.998357563
AMT_GOODS_PRICE - AMT_CREDIT	0.986051701
LIVE_REGION_NOT_WORK_REGION - REG_REGIO	0.861374946
DEF_60_CNT_SOCIAL_CIRCLE - DEF_30_CNT_SOC	0.850995792
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_W	0.825358079
AMT_ANNUITY - AMT_GOODS_PRICE	0.774006842
AMT_ANNUITY - AMT_CREDIT	0.770772818

BANK LOAN CASE STUDY

Findings-25

Top Correlation Coefficients for Re-payers are: -

Correlation between Columns	Values
AMT_CREDIT - AMT_GOODS_PRICE	0.982267963
OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE	0.998065853
DEF_60_CNT_SOCIAL_CIRCLE - DEF_30_CNT_SOCIAL_CIRCLE	0.89051161
REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION	0.806743886
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY	0.783754676
AMT_CREDIT - AMT_ANNUITY	0.749665201
AMT_GOODS_PRICE - AMT_ANNUITY	0.74950403

BANK LOAN CASE STUDY

Findings-25

Top Correlation Coefficients for Re-payers are: -

Correlation between Columns	Values
AMT_CREDIT - AMT_GOODS_PRICE	0.982267963
OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE	0.998065853
DEF_60_CNT_SOCIAL_CIRCLE - DEF_30_CNT_SOCIAL_CIRCLE	0.89051161
REG_REGION_NOT_WORK_REGION - LIVE_REGION_NOT_WORK_REGION	0.806743886
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY	0.783754676
AMT_CREDIT - AMT_ANNUITY	0.749665201
AMT_GOODS_PRICE - AMT_ANNUITY	0.74950403

BANK LOAN CASE STUDY

Analysis

Using the Why's approach I am trying to uncover root cause: -

• Why is it that the target_variable is of so much importance?

---› In this dataset target_variable represents whether the client had some payment difficulties (1) or the client didn't had some payment difficulties (0); It is important because the target_variable decides whether the bank should increase/decrease its interest rates on various loans given by the bank; Also in this case almost 92% of the clients didn't had any payment issues and only 8% of them had payment issues, this tells that bank's credit score is good.

• Why is it that proportion of Female clients more than that of the Male clients?

---› In countries like India especially there have been laws made by the Government for Women who want to establish their own Start-up, Business or their own classes, catering services, etc. These laws offer loans to women clients at a relatively low interest rate. Also, in some cases people purposely use their retired/household mother or household wife so that they can get some sort of concession i.e., low interest rates while applying for home loans.

BANK LOAN CASE STUDY

Analysis

• Why should bank prefer other Housing type clients though House/Apartments Housing type clients have the highest proportion of non-defaulters?

----› Cause people in other groups like Municipal Apartment, Rented Apartment, with Parents are in the search of their own house of their own name plate. Also, now a day in India the joint family system is declining and the future generations opt to live in their own 1/2 BHK's rather than living together will all family members in big Family Apartments Using the Why's approach I am trying to find some more useful insights

• Why should bank opt for working class clients more than the state-government class clients though state-government employees enjoy a lot of benefits and regular salary?

----› It is true that state government employee enjoy a lot of benefits but they also get housing allowances greater than that of working class and in some cases they even get an apartment to live with their families as long as they work for the state government. On the other hand, the working class don't enjoy such housing allowances or get very less of it, also the working class don't get an apartment to live in for their entire professional life (i.e. , until retirement) and so working class opt for purchasing their own house by taking house loan.

BANK LOAN CASE STUDY

Analysis

• Why should Bank not go for approving loans to 'Laborers' occupation_type clients though they have the highest non- defaulters count?

-----› Laborers take only personal loans for marriage or house repair purpose and their loan amount is also less and the interest on such loans is also less as compared to home loan, car loan, etc. which in turn will cause less profits to the bank.

• Why is it that females with low-income group have the lowest count of defaulters?

-----› Females belonging to such groups take loan of small amounts just for starting their own start-ups, business or catering/ parlour services and they usually enjoy benefit from government schemes for such purpose.

BANK LOAN CASE STUDY

Conclusion

In conclusion, I would like to conclude the following: -

- Most of the clients are loan re-payers.
- The Bank generally lends more loan to Female as compared to Male but Male are less defaulters compared to Female.
- As age and experience increases, chances of defaulter decreases.
- Most of the clients are taking cash loans.
- Educated clients tend to less defaulter compared to clients with lower education such as secondary special education so Bank should prefer clients with having such education status.
- As number of children increases, number of clients who take loan decreases.
- The Bank should be more cautious when lending money to clients who are unemployed because they are the most defaulters with highest amount of credit.
- As age increases amount taken by Clients are considerably high but with higher age defaulter percentage is lower. These are least risky and more profitable for Bank.

IMPACT OF CAR FEATURES

Description

The dataset includes variables such as car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP).

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. It is important to know the impact of car features on price and profitability in the automotive industry.

The purpose is to analyze 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.

IMPACT OF CAR FEATURES

The Problem

Tasks: Analysis
• Insight Required: How does the popularity of a car model vary across different market categories?
Task 1.A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores. T
ask 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

• Insight Required: What is the relationship between a car's engine power and its price?
Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

• Insight Required: Which car features are most important in determining a car's price?
Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

IMPACT OF CAR FEATURES

The Problem

Tasks: Analysis

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

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

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

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

Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.

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

IMPACT OF CAR FEATURES

The Problem

Building the Dashboard:

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

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

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

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

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

IMPACT OF CAR FEATURES

Design

I performed the following before the actual data analysis: -

• First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.

• Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.

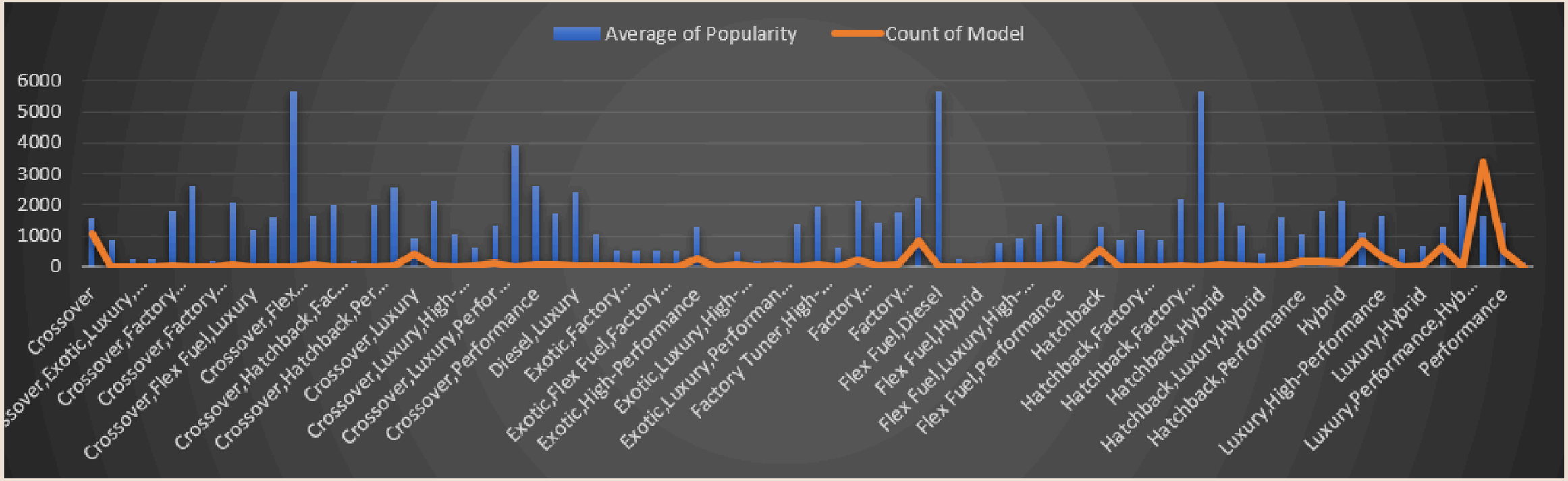
• I removed rows having blank spaces and NULL values.

• Then I removed duplicate rows from the datasets.

Software used for doing the overall Analysis: - ----› Microsoft Excel

IMPACT OF CAR FEATURES

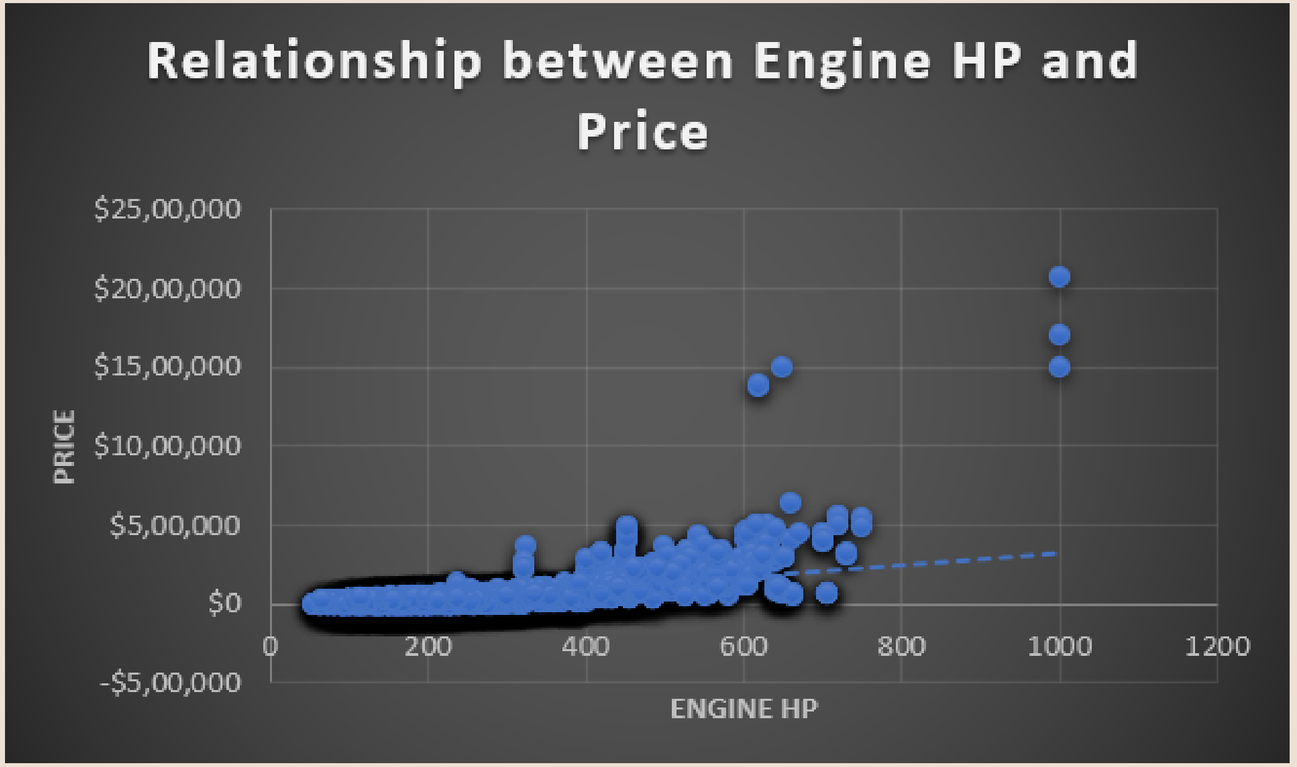
Findings-1



Insight: The popularity of Flex Fuel, Diesel, Hatchback, Crossover, and Performance market categories reflects diverse consumer preferences.

IMPACT OF CAR FEATURES

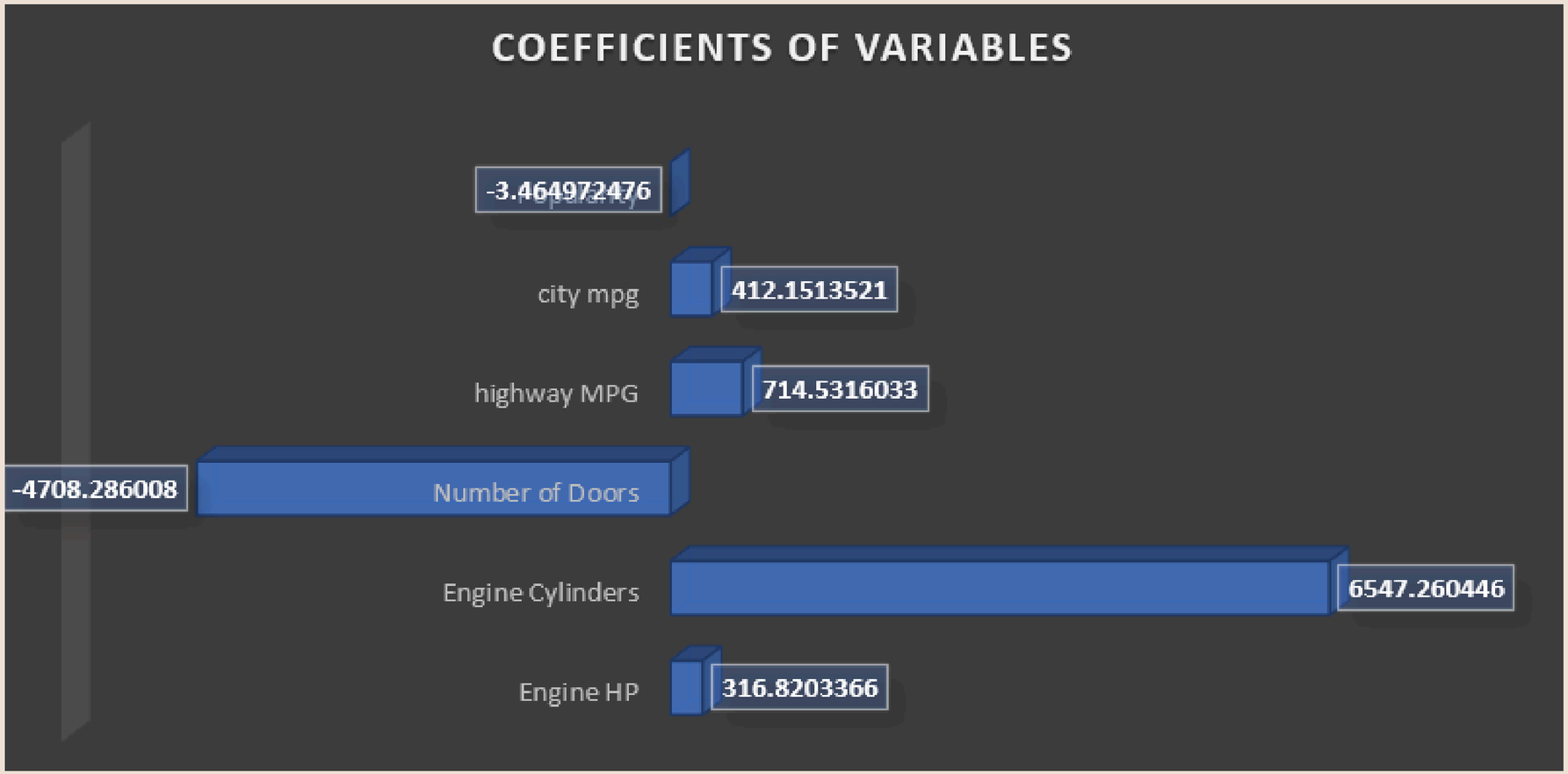
Findings-2



Insight: When the Engine power increases Price also increase. Hence, they are directly proportional to each other and have a positive relationship.

IMPACT OF CAR FEATURES

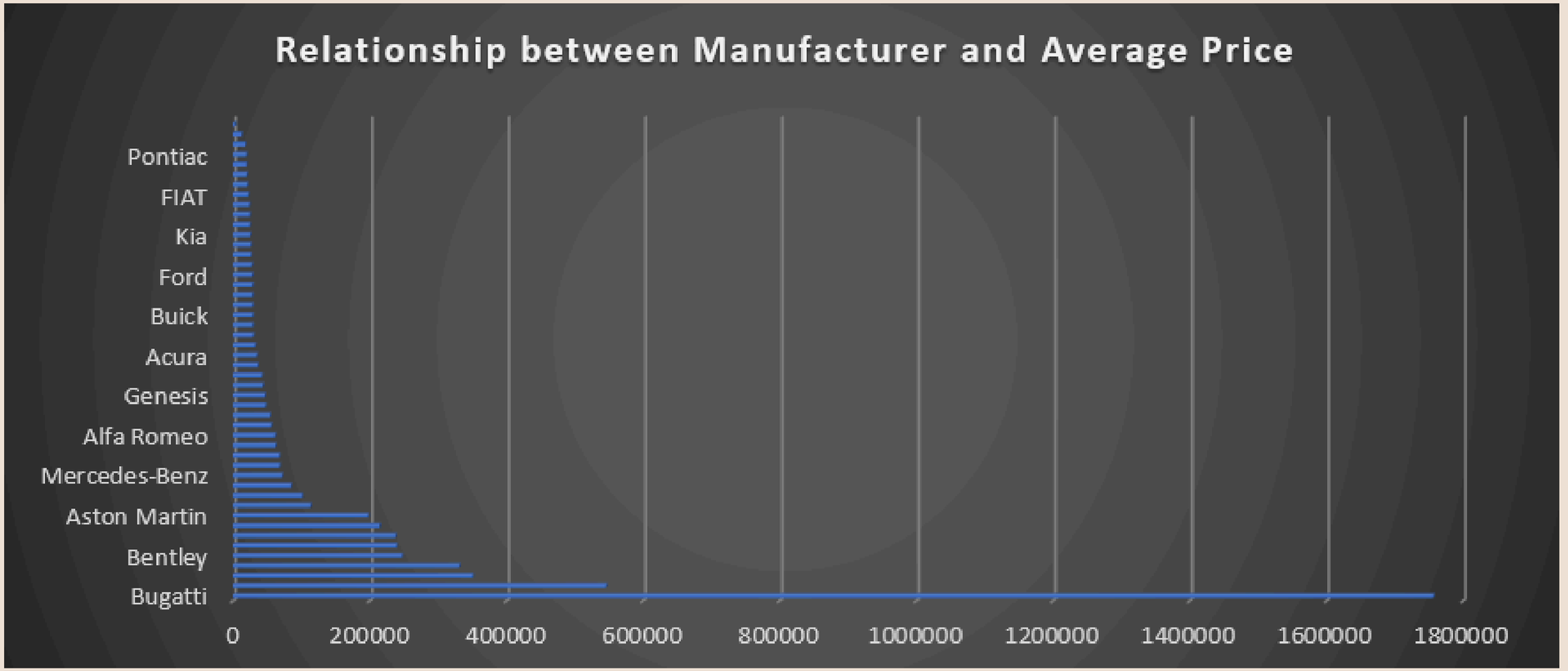
Findings-3



Insight: - It can be observed that Engine Cylinders are the most important features in determining a car's price**.**

IMPACT OF CAR FEATURES

Findings-4

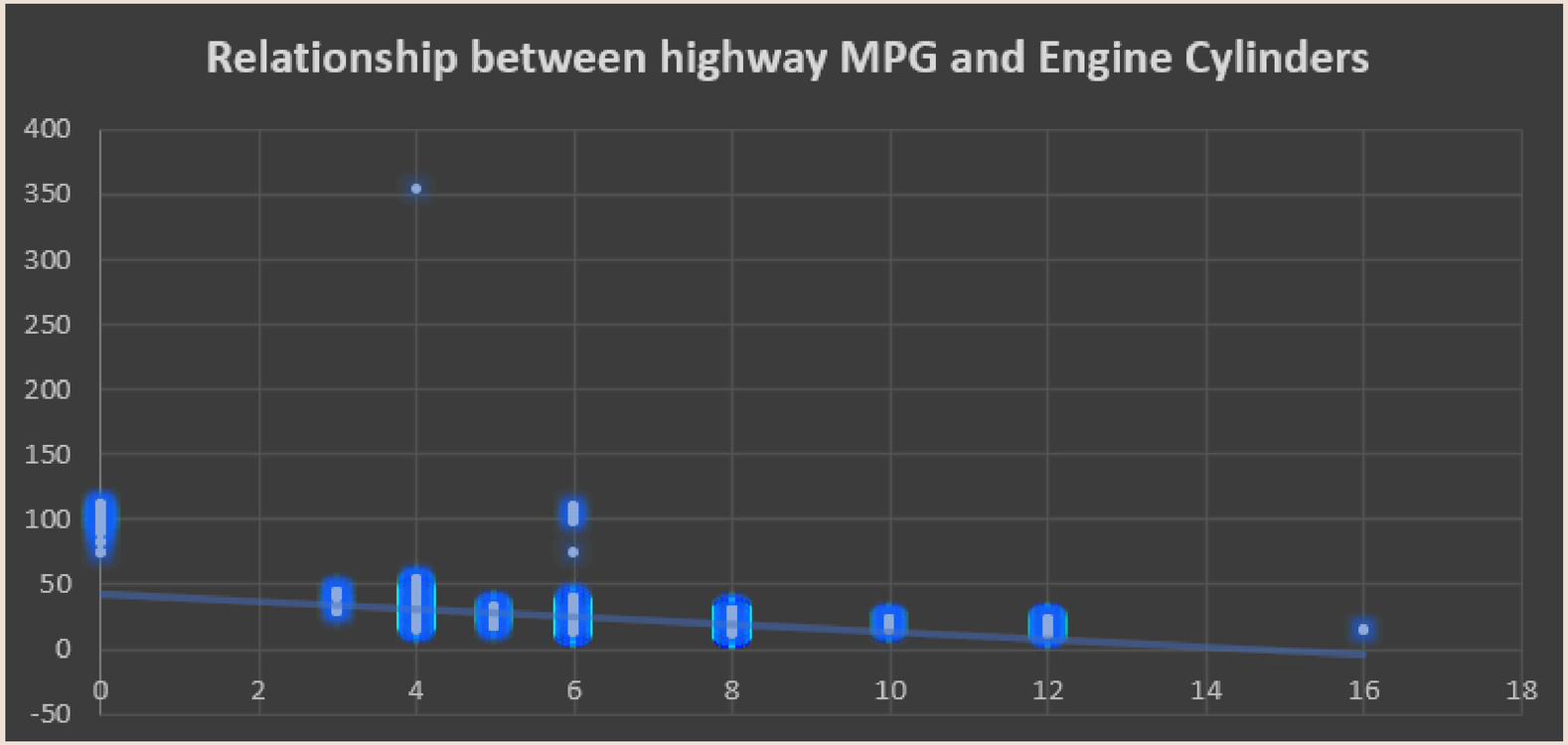


Insight: - The Average price is the highest for Bugatti and Plymouth has the lowest average price.

IMPACT OF CAR FEATURES

Findings-5

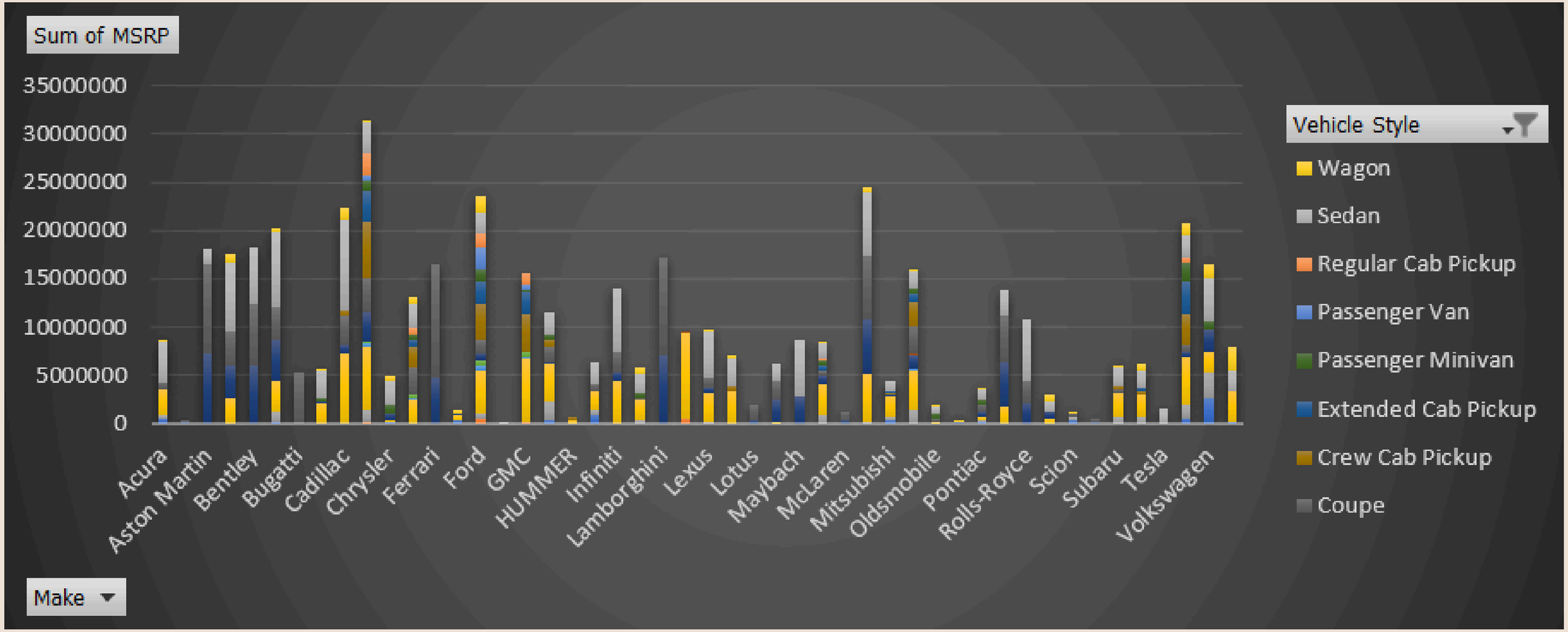
Correlation Coefficients	-0.596246019



Insight: - Number of Cylinders will increase then highway MPG will decrease. It's negative relationship between both of them.

IMPACT OF CAR FEATURES

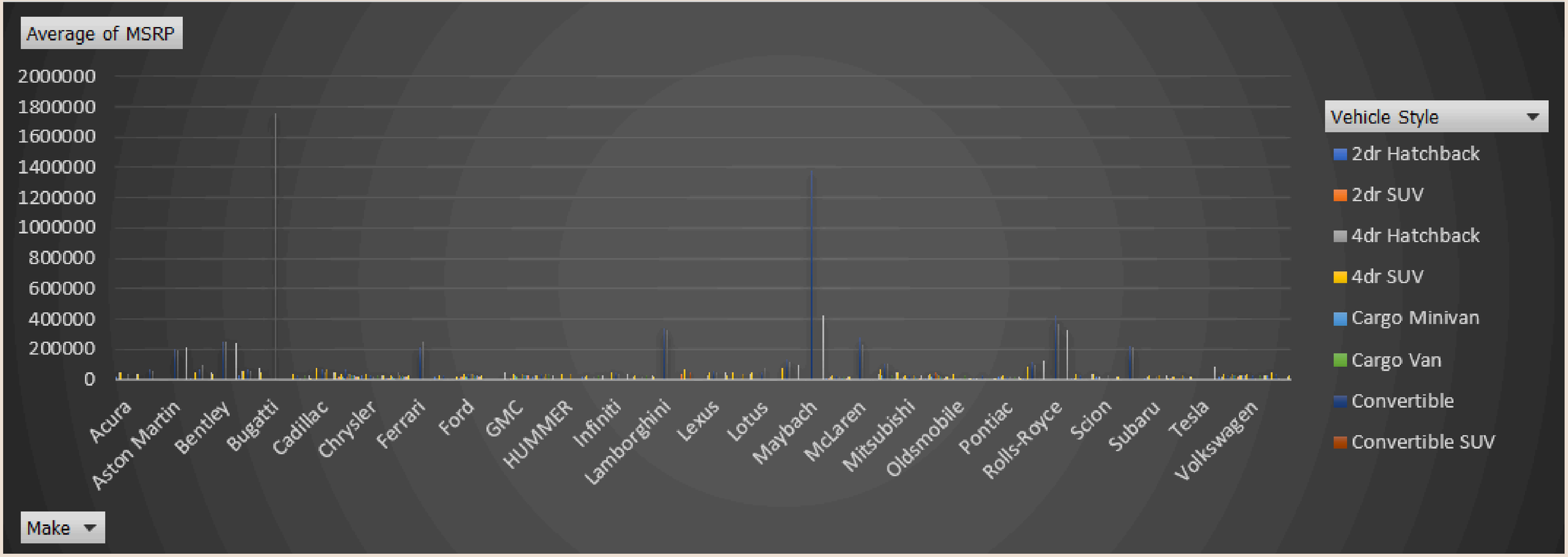
Findings-6



Insight: Chevrolet has the highest price distribution by body style

IMPACT OF CAR FEATURES

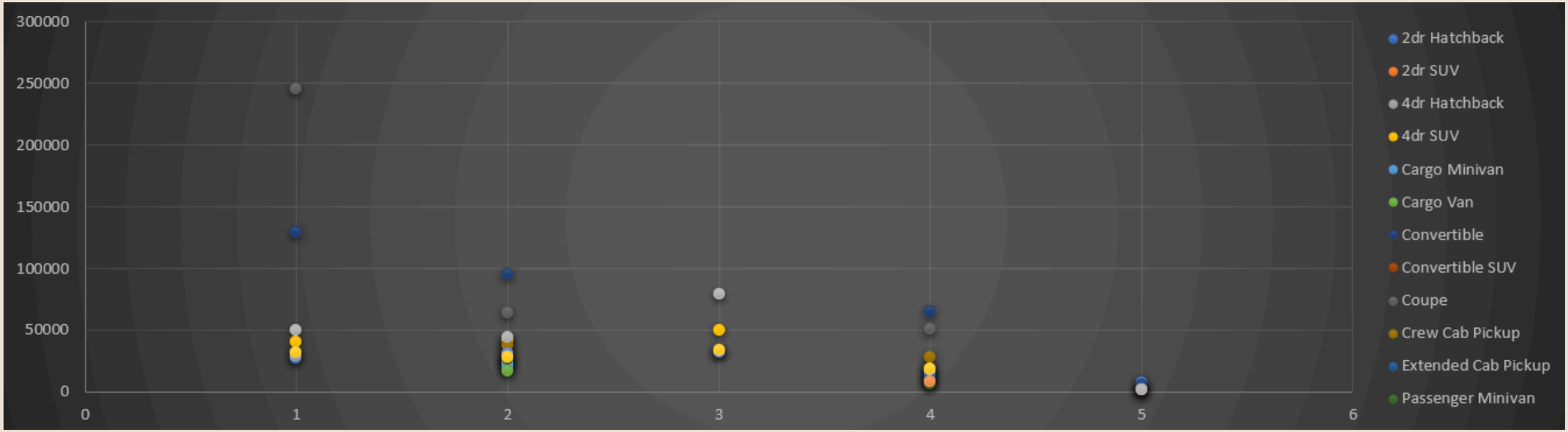
Findings-7.



Insight: Bugatti has the highest average MSRPs and Plymouth has the lowest average MSRPs by body style.

IMPACT OF CAR FEATURES

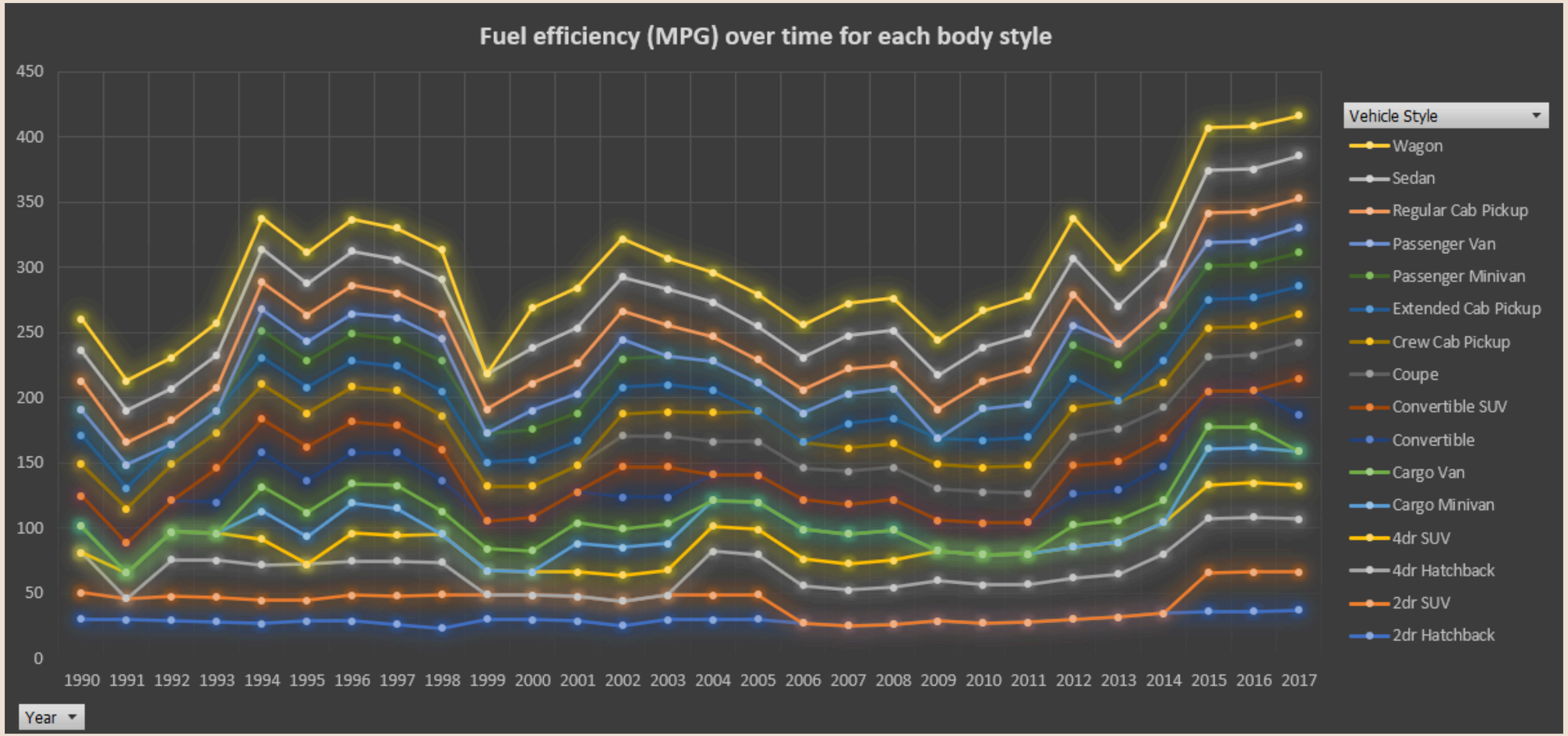
Findings-8



Insight: AUTOMATED_MANUAL with Coupe body style is the most expensive transmission.

IMPACT OF CAR FEATURES

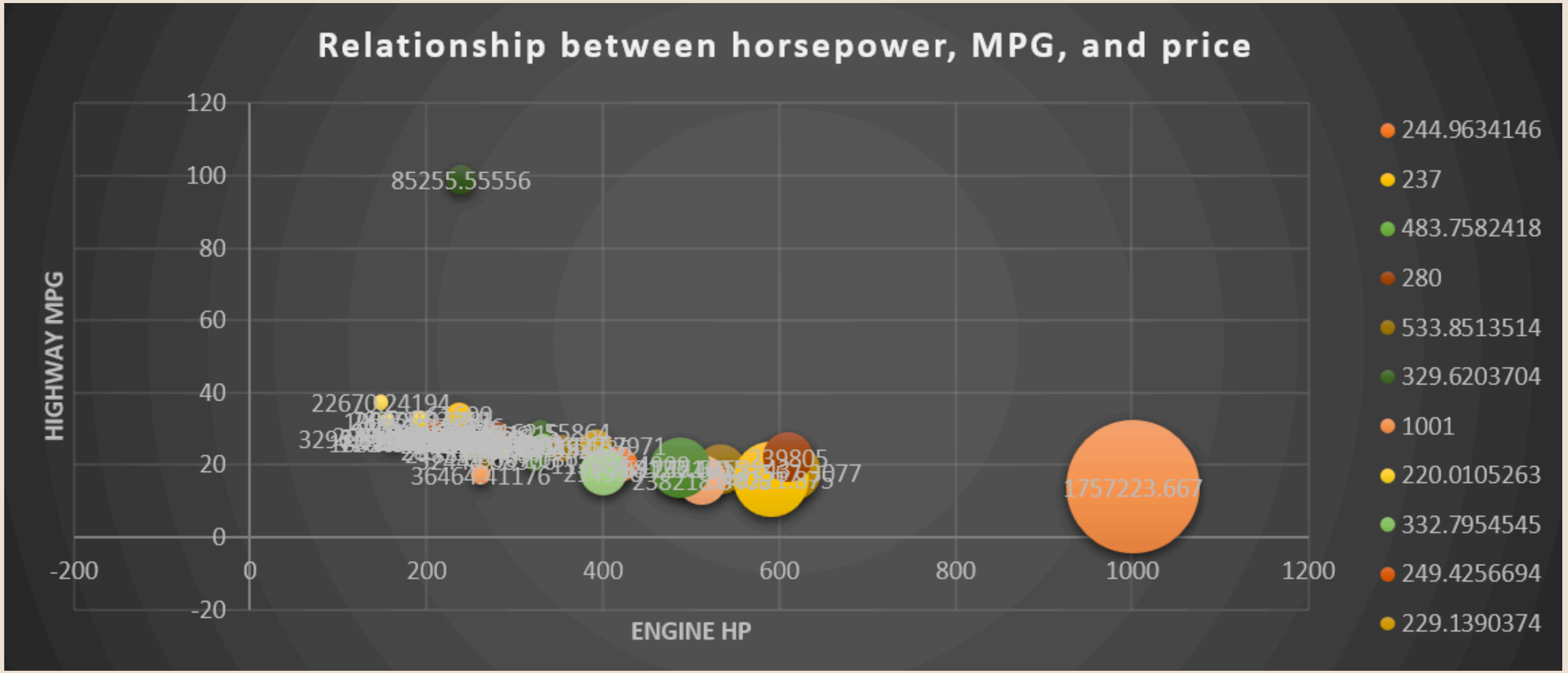
Findings-9



Insight: Wagon body style has the highest fuel efficiency in 2017.
Fuel efficiency of cars increased across different body styles and model years.

IMPACT OF CAR FEATURES

Findings-10



Insight: The Engine HP goes up when Highway MPG goes down but theprice increases.

IMPACT OF CAR FEATURES

Conclusion

In conclusion, the analysis of car data reveals valuable insights for manufacturers, buyers, and enthusiasts.

Popular market categories include Flex Fuel, Diesel, Hatchback, Crossover, and Performance.

Engine power impacts price, while engine cylinders play a significant role.

Bugatti commands the highest average price, and Plymouth the lowest.

As cylinder count increases, highway MPG decreases.

Chevrolet leads in price distribution by body style.

AUTOMATED_MANUAL transmission with Coupe body style is the most expensive.
Wagon body style excels in fuel efficiency.

Lastly, increasing engine horsepower affects both highway MPG and price. These findings inform decisions across the automotive industry.

ABC CALL VOLUME TREND ANALYSIS

Description

A Customer Experience (CX) team analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, various types of support, including email, inbound, outbound, and social media support, among others.

There are several AI-powered tools like include Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing are being used to enhance customer experience. Inbound customer support, which is the focus of this project, involves handling incoming calls from existing or prospective customers.

The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business. We have dataset that contains information about the inbound calls received by a company named ABC that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

ABC CALL VOLUME TREND ANALYSIS

The Problem

• Average Call Duration: Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.
Your Task: What is the average duration of calls for each time bucket?

• Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets. Your Task: Can you create a chart or graph that shows the number of calls received in each time bucket?

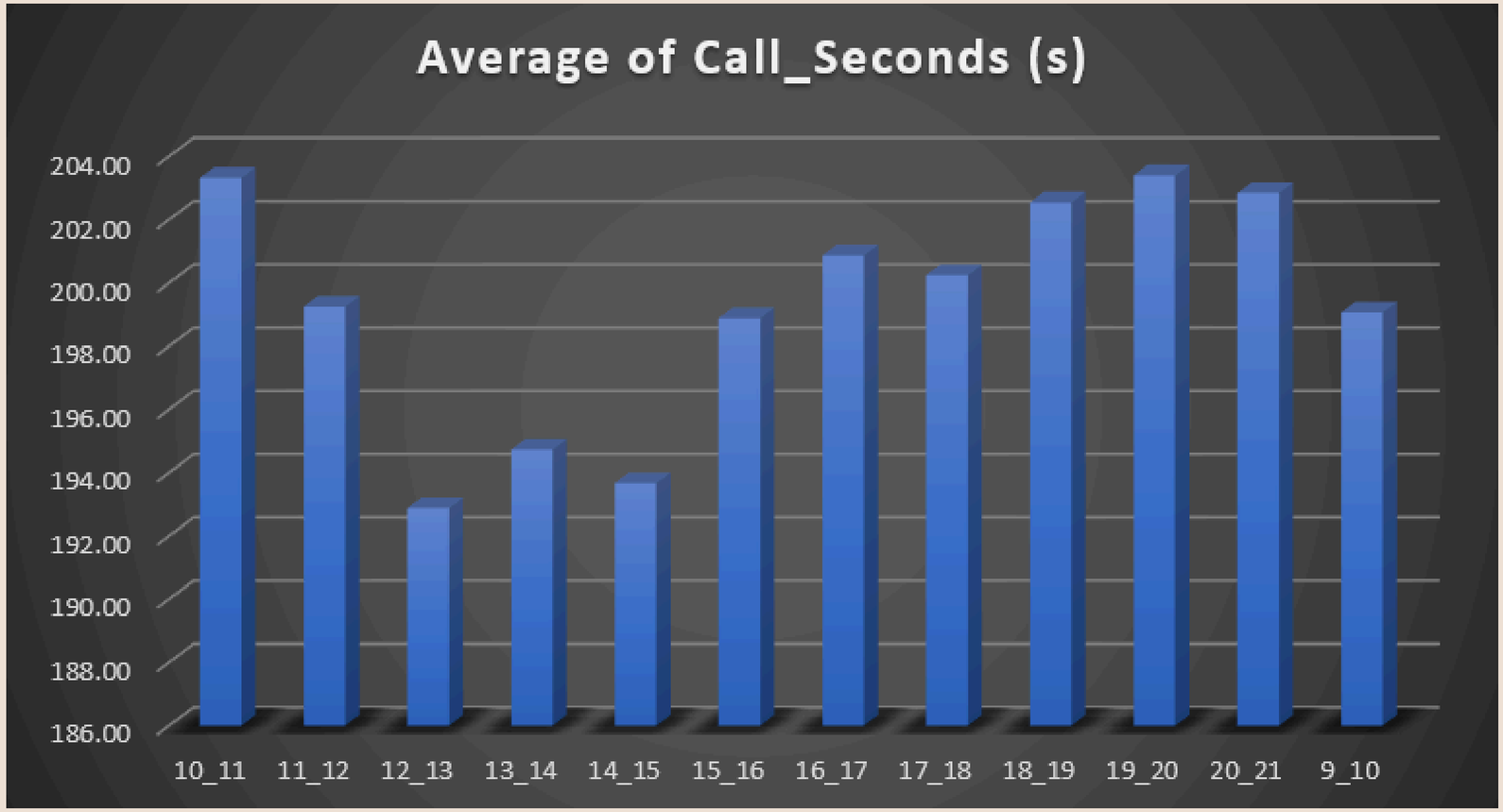
• Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.
Your Task: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

• Night Shift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience.
Your Task: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%. duration of the call, and the call status (whether it was abandoned, answered, or transferred).

ABC CALL VOLUME TREND ANALYSIS

Findings-1

Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
10_11	203.33
11_12	199.26
12_13	192.89
13_14	194.74
14_15	193.68
15_16	198.89
16_17	200.87
17_18	200.25
18_19	202.55
19_20	203.41
20_21	202.85
9_10	199.07
Grand Total	**198.62**

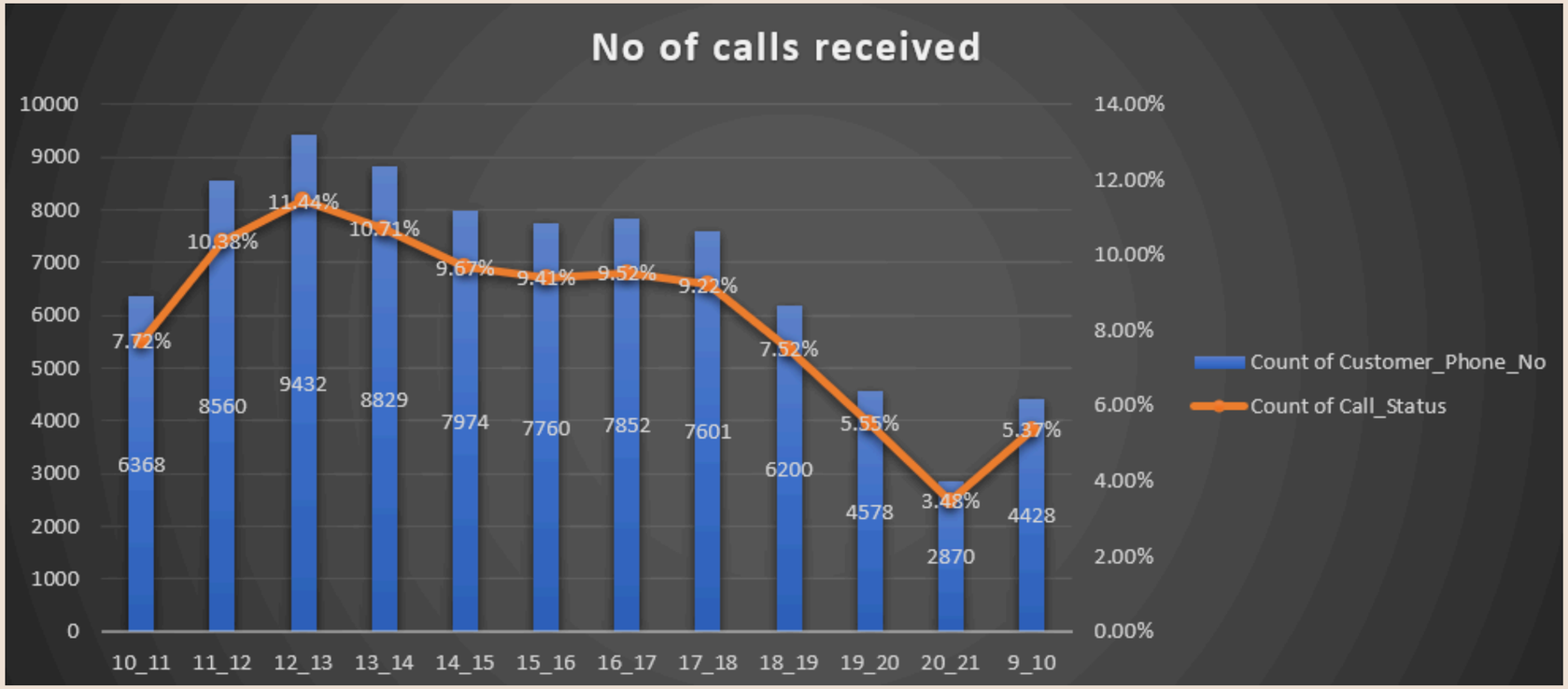


Based on Analysis maximum average duration of calls for incoming calls are at 10_11 AM and 7_8 PM i.e.203.33 and 203.41.

ABC CALL VOLUME TREND ANALYSIS

Findings-2

Call_Status	answered	
Row Labels	**Count of Customer_Phone_No**	**Count of Call_Status**
10_11	6368	7.72%
11_12	8560	10.38%
12_13	9432	11.44%
13_14	8829	10.71%
14_15	7974	9.67%
15_16	7760	9.41%
16_17	7852	9.52%
17_18	7601	9.22%
18_19	6200	7.52%
19_20	4578	5.55%
20_21	2870	3.48%
9_10	4428	5.37%
Grand Total	**82452**	**100.00%**

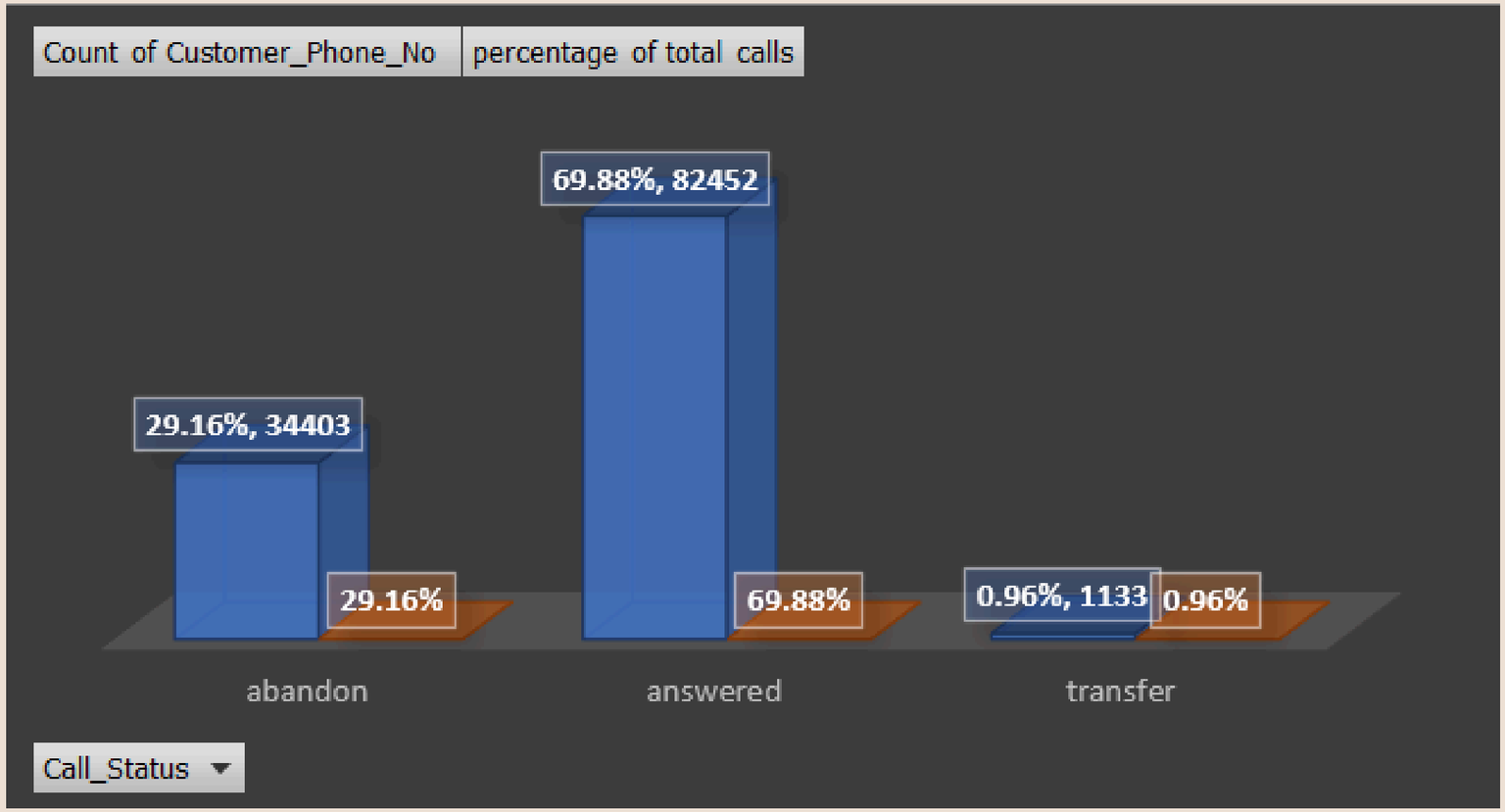


Based on chart highest number of calls received in between 12 PM and 1 PM which is 9432.

ABC CALL VOLUME TREND ANALYSIS

Findings-3

Row Labels	Count of Customer_Phone_No	percentage of total calls
abandon	34403	29.16%
answered	82452	69.88%
transfer	1133	0.96%
Grand Total	**117988**	**100.00%**



Total no. of agents required to reduce the abandon rate to 10% is 54. Maximum number of agents are required at 11_12 AM i.e., 7.

ABC CALL VOLUME TREND ANALYSIS

Findings-3

Row Labels	Sum of Call_Seconds (s)	Sum of Call_hours
1-Jan	676664	187.96

Assumption: An agent's total working hours are 9 hours. Out of this, 1.5 hours are allocated for lunch and snacks. Therefore, the actual working hours available for tasks (excluding breaks) is 7.5 hours. On average, an agent spends 60% of these actual working hours on calls with customers or users.

Total Working hrs	9 hrs
Lunch & snacks	1.5 hrs
Total actual working hrs	7.5 hrs
Actual working hrs	4.5 hrs

ABC CALL VOLUME TREND ANALYSIS

Findings-3

Row Labels	Count of Call_Seconds (s)	Percentage of Call_Seconds(s)	Time distribution	No of agents required for answered rate 90%
10_11	13313	11.28%	0.11	6
11_12	14626	12.40%	0.12	7
12_13	12652	10.72%	0.11	6
13_14	11561	9.80%	0.10	5
14_15	10561	8.95%	0.09	5
15_16	9159	7.76%	0.08	4
16_17	8788	7.45%	0.07	4
17_18	8534	7.23%	0.07	4
18_19	7238	6.13%	0.06	3
19_20	6463	5.48%	0.05	3
20_21	5505	4.67%	0.05	3
9_10	9588	8.13%	0.08	4
Grand Total	**117988**	**100.00%**	**1.00**	**54**

Total no. of agents required to reduce the abandon rate to 10% is 54. Maximum number of agents are required at 11_12 AM i.e., 7.

ABC CALL VOLUME TREND ANALYSIS

Findings-4

Average incoming calls	5130
Average incoming calls at night between 9 pm - 9 am (30% of 5130)	1539
Average seconds required to answer the calls (Avg incoming calls at night * Avg calls answered)	305680.4499
Average hours required to answer the calls	84.91123608
keeping the maximum abandon rate at 10%	
Actual average hours required to answer the calls	76.42011247
We know from the previous task that Actual working hrs is 4.5 hrs	
No. of agents required to answer the call	16.98224722

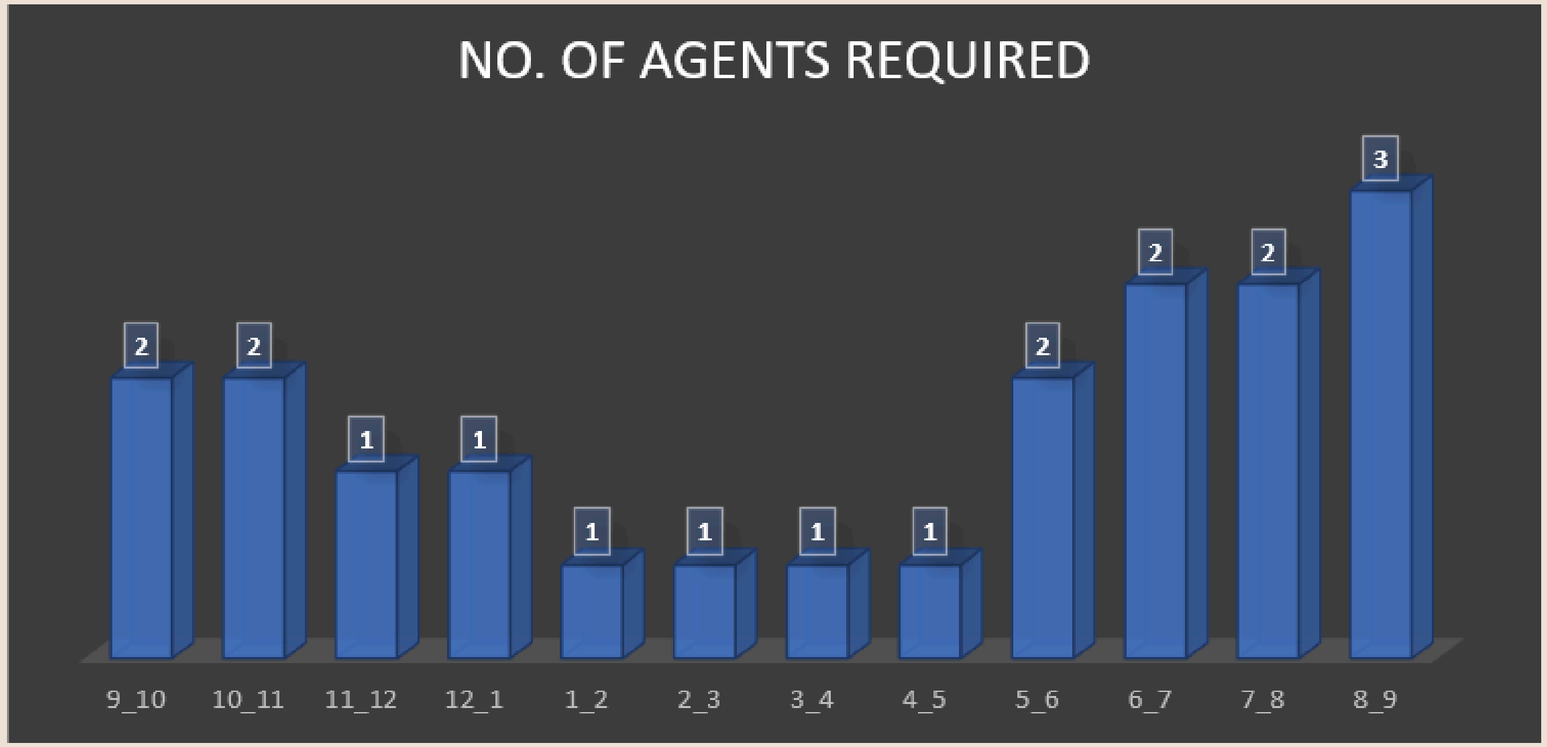
Total number of agents required is 17

Total number of agents required to answer the call at night 9 PM to 9 AM is 17.
Maximum agents are required at 8_9 AM i.e., 3

ABC CALL VOLUME TREND ANALYSIS

Findings-4

Time_bucket	Call distribution	Time distribution	No. of agents required
9_10	3	0.10	2
10_11	3	0.10	2
11_12	2	0.07	1
12_1	2	0.07	1
1_2	1	0.03	1
2_3	1	0.03	1
3_4	1	0.03	1
4_5	1	0.03	1
5_6	3	0.10	2
6_7	4	0.13	2
7_8	4	0.13	2
8_9	5	0.17	3
Total	**30**	**1.00**	**17**



Total number of agents required to answer the call at night 9 PM to 9 AM is 17.
Maximum agents are required at 8_9 AM i.e., 3

ABC CALL VOLUME TREND ANALYSIS

Analysis

Using the Why's approach I am trying to find root cause: -

- Why is that the average call answered were more in count in the time bucket of 10_11, 18_19, 19_20 and 20_21 as compared to other time buckets?
  ---› Most of the customers are office people and they need to reach office by 10 AM or 11 AM, so these customers call during 10_11 time bucket i.e. while they in transit to office or have reached office and have some free time before they start their work; During the time bucket 18_19, 19_20 and 20_21 the customers have either left their office and reached home or they are in the transit to reach home and during these time period i.e. 6 Pm to 9 Pm people have free time where they can share their concern to the customer service. During these time buckets most of the calls are from individual people with small problems which can be resolved quickly.

- Why is it that the time bucket 11_12 has the highest number of incoming calls but it does not have the highest number of average answered calls?
  ---› Maybe there were more number of incoming calls in the time bucket 11_12 and there were not enough personnel to handle most of the queries of the customers during the 11_12 time bucket.

ABC CALL VOLUME TREND ANALYSIS

Analysis

• Why is that one cannot provide the exact distribution of agents during the night time i.e. from 9 PM to 9 AM if the number of agents available during the night shift are already defined, so as to keep the abandon rate 10%?

---› For this particular case, since we have only 17 agents during night, we need to distribute in non-analytical way i.e. the agents who work in 19_20, 20_21 time bucket to wait and work in 21_22 and 22_23 time buckets as well. Also, agents who work during 9_10, 10_11 time bucket can be asked to work for 7_8 and 8_9 time bucket as well. he agents who work in the time bucket 1_2, 2_3, 3_4 and 4_5 can be asked to work in time buckets 6_7, 7_8 and 8_9 so as to keep the abandon rate at 10%. Also, the company needs to consider various factors like how far is the home of the agent if he/she is made to do night shift, Is the transport facility available during the night hours from the agent's home to company and many other factors and hence the exact distribution cannot be given using an analytical approach.

ABC CALL VOLUME TREND ANALYSIS

Conclusion

I hereby conclude that:

- The company can divide its workforce into three shifts to ensure round-the-clock availability for addressing customer queries and concerns.
- Agents handle calls with an average duration of 198.62 seconds.
- Incoming calls have their maximum average duration during two time periods:
  --- 10 AM to 11 AM
  --- 7 PM to 8 PM
- The minimum average call duration for incoming calls received by agents occurs during the time slot:
  ---12 PM to 1 PM
- The highest number of calls is received between 12 PM and 1 PM.
- The least number of calls answered occurs between 8 PM and 9 PM.
- To achieve a 10% abandon rate, the company requires a total of 54 agents.
- The company can hire 17 agents available during night hours (from 9 PM to 9 AM) to handle calls or consider shifting some day workers to the night shift.

APPENDIX

Data Analytics Process: -
---› Link for the shared PDF on Google Drive:

Data Analytics Trainee Task -1 .pdf - Google Drive

Instagram User Analytics: -
----› Link for the shared file on Google Drive:

Data Analytics Trainee Task - 2.pdf - Google Drive

Operation Analytics and Investigating Metric Spike Analysis: -
-----› Link for the shared file on Google Drive:

Data Analytics Trainee Task - 3.pdf - Google Drive

Hiring Process Analytics: -
----› Link for shared PDF on google drive:

Data Analytics Trainee Task - 4.pdf - Google Drive

APPENDIX

IMDB Movie Analysis:

---› Link for the shared PDF on Google Drive:

Data Analytics Trainee Task - 5.pdf - Google Drive

Bank Loan Case Study:

-----› Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task 6.pdf - Google Drive

Analyzing the Impact of Car Features on Price and Profitability:

------› Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task - 7.pdf - Google Drive

ABC Call Volume Trend Analysis:

-----› Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task 8.pdf - Google Drive