Data Analysis Portfolio Project

Pratham Barot

Professional Background

I have completed Integrated M.Tech in AI & ML. I have several skills including Data Analysis, Python, Machine Learning, MySQL, Excel, Tableau, Power BI. I have workedon various personal projects related to Data Analysis, Machine Learning, Excel and Python.

Also, I have participated in various hackathons like in Neo Codeathon, Girl Script Summer of Code, and HacktoberFest where I have consistently contributed in Python, Data Analysis and ML.

Currently I am selected as a contributor in Omdena's collaborative project. My primary role involved data analysis, where I utilized my analytical skills to process and interpret data, contributing significantly to the project's objective of creating innovative and data-driven solutions.

I am very flexible and ready to adjust in any environment. I am adaptive to work under pressure. I have Leadership Quality with strong influencing skills. I am Team player and Motivator having group dynamics.

Table of Contents

Table of Contents

Hiring Process Analytics

IMDB Movies Analysis

Bank Loan Case Study

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 Life Scenario: Going to a mall to buy apparel

The following steps would be taken by the person while making the right decision:-

Step 1: Plan: -- First, I will decide a mall to go shopping that offers a variety of stores. Then I will identify the specific clothing items I need, such as shirts, jeans, and formal attire. Finally, I'll create a detailed list to ensure I purchase everything efficiently during the visit.

Step 2: Prepare:-- I will Set a budget for apparel ensures me stay within my financial limits while shopping. Considering style preferences and current trends helps me choose outfits that align with my taste and keep me fashionable. This approach will narrow the options, making the shopping process more efficient and enjoyable.

Data Analytics Process

Step 3: Process: -- I will research current fashion trends provides inspiration and helps me stay updated with what's in style. Then I will select stores based on my budget, opting for high-end options for formal wear and budget-friendly stores for casual attire. This strategy ensures a balanced shopping experience that meets both my style and financial goals.

Step 4: Analyze: -- While shopping, conduct in-store comparisons to assess the fit, comfort, and prices of different brands. Then I will analyze trends to choose versatile apparel that is both stylish and suitable for multiple occasions. I will consider buying matching outfits to create cohesive looks for various events, ensuring practicality and value.

Step 5: Share: -- Now I will discuss with friends, family who recently went shopping there.

Step 6: Act: -- I will execute my plan and enjoy buying apparel.

Data Analytics Process

Conclusions

Hence, we have seen how we can use the 6 steps of Data Analytics while making any decision in real life scenarios (finding the best place for solo travel) 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.

Instagram User Analytics

- **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.

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

- Created new Schema as ig_clone in MySql
- Copied the dataset in MySql and executed it
- all the tables and data loaded in database
- By using the 'select' command we can query the desired output

Software used for querying the results

--> MySQL Workbench 8.0.38 CE

Instagram User Analytics

Findings – I

To find the 5 oldest users of the Instagram from the database:-

- We will use select statement to select username and created_at column from users table.
- We will use order by clause with asc in created_at column to sort the output in ascending order.
- Using limit function, we can display the output for top 5 oldest users of the Instagram.

Output/Results: -

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

Instagram User Analytics

Findings – II

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

- We will use select statement to select username and user id column from users table then assign alias for users table as u and photos table as p
- we will do left join photos table on users table to retrieve photo id column from photos table on u.id= p.user_id because both columns have common content.
- Using where clause we will filter rows from users table where p.id is null.
- We will use order by clause in id column in users table to sort the output.

Output/Results:-

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 – III

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

- We will use **select** statement to select id, username column from users table, photo_id column from Likes table, image_url column from photos table.
- Then we will use **count** function in Likes table to count the likes and assign alias as no_of_likes.
- we will do **inner join** photos table on Likes table **on l.photo_id = p.id ,** Users table on photos table **on p.user_id = u.id** and assign alias for likes table as **l ,** photos table as **p ,** users table as **u.**
- We will use **group by** clause in photo_id column from likes table to get number of likes for each photo_id.
- We will use **order by** clause with **desc** in no_of_likes column to sort the output in descending order.
- Using **limit** function, we can display the output of user with the most likes on single photo on Instagram.

Output/Results:-

id	username	photo_id	image_url	no_of_likes
52	Zack_Kemmer93	145	https://jarret.name	48

Instagram User Analytics

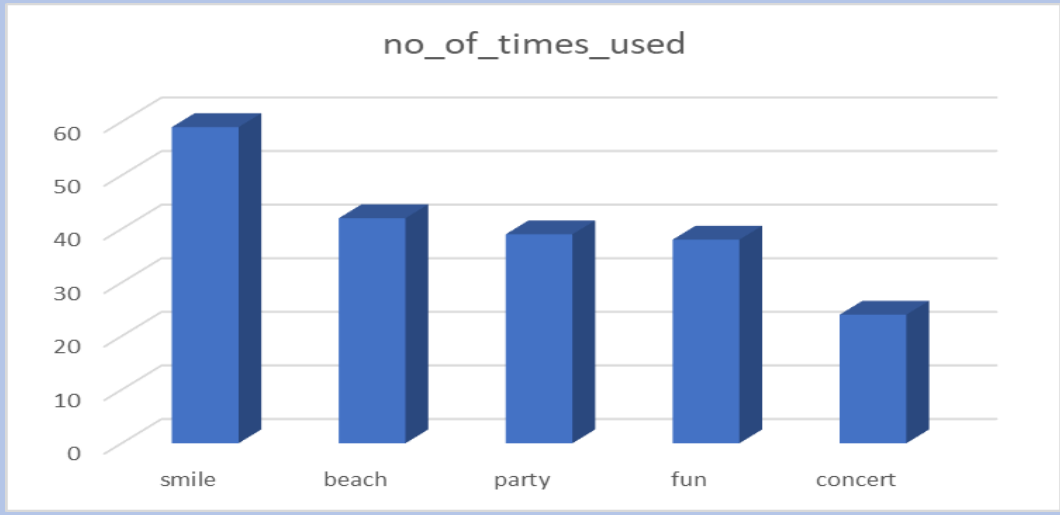
Findings – IV

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

- We will use **select** statement to select tag_name column from Tags table and use **count** function to count the number of times tags used then assign **alias** for Tags table as **t** and photo_tags table as **pt.**
- We will do **left join** tags table on photo_tags table **on pt.tag_id= t.id** because both columns have common content.
- Using **group by** clause we will get number of times tags used for each tag_id.
- We will use **order by** clause with **desc** in no_of_times_used column to sort the output in descending order.
- Using **limit,** we will get top five most commonly used hashtags.

Output/Result:-

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



Instagram User Analytics

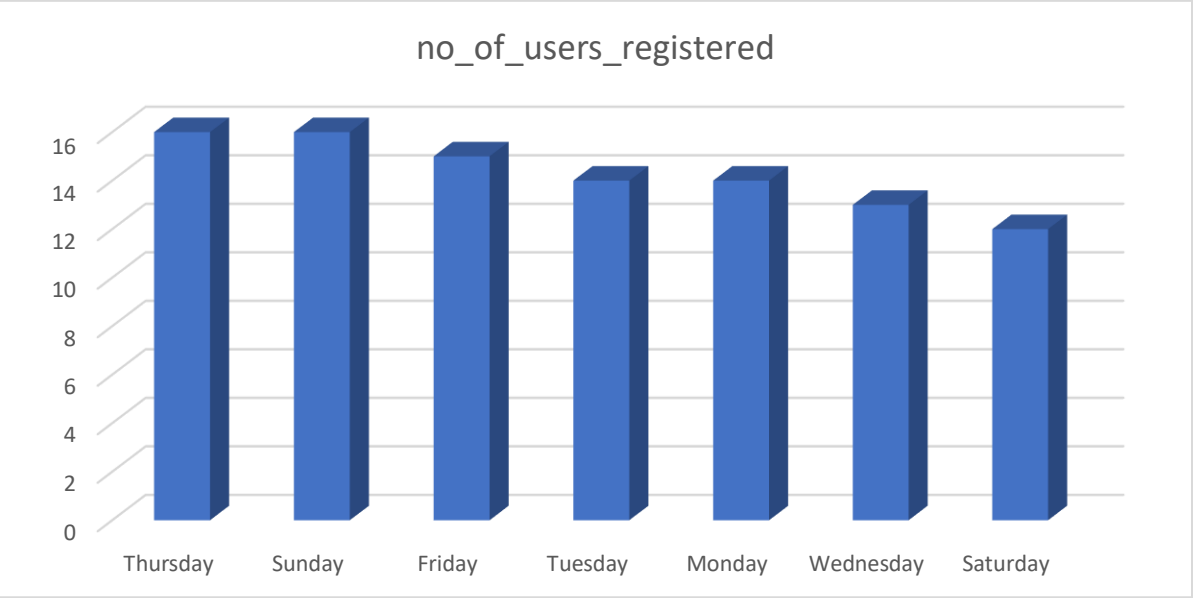
Findings – V

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

- We will create columns for desired output using select dayname(created_at) as day_of_week and count(*) as no_of_users_registered from users table.
- Using group by clause in day_of_week column to get number of users registered in each day of the week.
- Then We will use order by clause with desc in no_of_users_registered column to sort the output in descending order.

Output/Results:-

day_of_week	no_of_users_registered
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12



Instagram User Analytics

Findings - VI

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.

- We will use select statement and count function to count as total and to count distinct user_id from photos table and divide them then we will get Avg_num_of_post_per_user.
- We will use select statement for subquery and count function as total from photos table and as total from users table and divide them
- Then we will get the total number of photos on Instagram divided by the total number of users.

Output/Results: -

Avg_num_of_post_per_user
3.473

total_no_of_photos_divide_total_no_of_users
2.57

Instagram User Analytics

Findings – VII

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

- We will use select statement to select user_id from Likes table and username from users table.
- We will use count function in distinct photo_id from Likes table and assign alias for Likes table as l and users table u then we will left join Users table on Likes table
- Using group by clause we will get output for each user.
- Then we will use having function to search values count(*) from photos equal values to total_no_of_photos_liked.

Output/Results: -

user_id	username	total_no_of_photos_liked
5	Aniya_Hackett	257
14	Jaclyn81	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: -

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

- 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

Conclusion

In conclusion, I would like to conclude that not only Instagram but many other social media and commercial firms use such Analysis to find the insights from their customer data which in turn help the firms to find the customers who will be an Asset to the firm in the future and not some Liability.

Such Analysis and sorting of the customer base is done at weekly, monthly, quarterly or yearly basis as per the needs of the business firms so as to maximize their profits in future with minimal cost to the company.

Operation Analytics and Investigating Metric Spike

Description

Operational Analytics is a crucial process that involves analyzing a company's end-to-end 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:-

- Created new Schema as project3 in MySql
- Then add tables and column names with data types
- Then add the values into them using the 'insert into' function of MySQL
- By using the 'select' command we can query the desired output
- Steps taken to load the data into the data base

Software used for querying the results:-

--> MySQL Workbench 8.0.38 CE

Software used for analyzing using Bar plots:-

--> Microsoft Excel

Operation Analytics and Investigating Metric Spike

Findings – I

To calculate the number of jobs reviewed per hour for each day in November 2020.

- We will use select statement from job_data table.
- Then we will use count function in distinct job_id column and divide by (30 days * 24 hours) to get number of jobs reviewed per hour for each day.

Output/Results:-

no_of_jobs_reviewed
0.0083

Operation Analytics and Investigating Metric Spike

Findings - II

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.

- We will use select statement , count function on distinct job_id column and avg function in count(distinct job_id) from job_data table.
- By using ROWS function we will be considering the rows between 6 preceding and current row.
- Then we will get 7-day rolling average of throughput.
- We will use group by clause in ds column.
- By using order by clause in ds column we will sort the order.

Output /Result :-

ds	jobs_reviewed	throughput_7_rolling_avg
25-11-2020	1	1
26-11-2020	1	1
27-11-2020	1	1
28-11-2020	2	1.25
29-11-2020	1	1.2
30-11-2020	2	1.3333

Operation Analytics and Investigating Metric Spike

Findings – III

To calculate the percentage share of each language over the last 30 days:-

- We will use select statement to select job_id, language column from job_data table.
- Then we will use count function in language column and divide by total using sum(count(*)) over().
- Using group by in language column we will get percentage share of each language.

Output/Results:-

job_id	language	percentage_share_of_each_language
21	English	12.5
22	Arabic	12.5
23	Persian	37.5
25	Hindi	12.5
11	French	12.5
20	Italian	12.5

Operation Analytics and Investigating Metric Spike

Findings – IV

To display duplicate rows from the job_data table.

- First we will decide in which column we need to find duplicate rows.
- Then we will use row_number() function to find the row numbers which are having the same value.
- We will use partition on row_number function over the column which we decided i.e job_id.
- Then we will use where function to find the row_num having value greater than 1.

Output/Results:-

job_id	actor_id	event	Language	time_spent	org	Ds	no_of_rows
23	1005	transfer	Persian	22	D	28-11-2020	2
23	1004	skip	Persian	56	A	26-11-2020	3

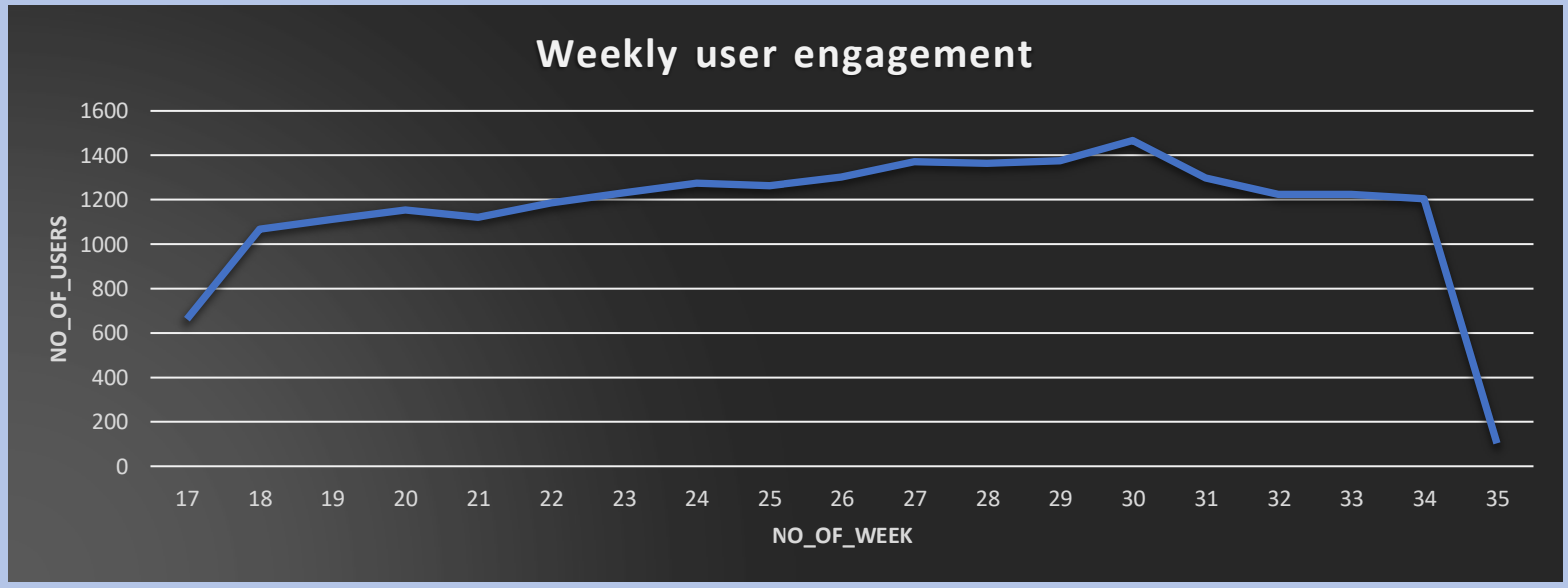
Operation Analytics and Investigating Metric Spike

Findings - V

To calculate the weekly user engagement:

- We will use select statement, week function in occurred_at column to extract number of weeks and count function in distinct user_id column to get number of users from events table.
- Using group by clause in no_of_week we will get weekly user engagement.

Output/Results:-



Operation Analytics and Investigating Metric Spike

Findings - V

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

Operation Analytics and Investigating Metric Spike

Findings - VI

To calculate the user growth for the product:

- We will use extract to extract year and week from activated_at column from users table.
- Using group by clause we will group extracted year and week on the basis of year and week number.
- Then we will use order by to sort the output based on extracted year and week
- We will use sum, over and row function between unbounded preceding and current row to find cumm_active_users.

Operation Analytics and Investigating Metric Spike

Findings - VI

Output/Results: -

year	no_of_weeks	no_of_active_users	cumm_active_users
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
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

year	no_of_weeks	no_of_active_users	cumm_active_users
2013	45	91	2607
2013	46	88	2695
2013	47	102	2797
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
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 – VII

To calculate the weekly retention of users based on their sign-up cohort.

- The weekly retention of users-sign up cohort can be calculated by two means i.e. either for the entire column of occurred_at of the events table or by specifying the week number (18 to 35)
- First we will use extract function to extract week from occurred_at column from events table.
- Then we will select the rows in which event_type = 'signup_flow' and event_name = 'complete_signup'.
- After that we will use left join on user_id to join the tables in which event_type = 'engagement'.
- Using group by clause in user_id we will get weekly retention for each user.
- Then we will use order by to sort the output on the basis of user_id.

Operation Analytics and Investigating Metric Spike

Findings - VII

Output/Results: - (entire column of occurred_at)

Link for the result
https://drive.google.com/file/d/1SkO5Rj-aiSkj0aPDehPnHWCrdF57p4By/view?usp=sharing

Output/Results: - (week number as 18)

Link for the result
https://drive.google.com/file/d/15d1pcbOQyTnZH1SIdyB9bAWd7tQvzVzz/view?usp=sharing

Operation Analytics and Investigating Metric Spike

Findings - VIII

To calculate the weekly engagement per device.

- We will extract year and week from occurred_at column from events table.
- Then we will select device column and use count function to get number of users.
- Using **where** clause we will select rows where **event_type='engagement'.**
- We will use **group by** and **order by** function to group and order the output based on year, no_of_weeks, device.

Output/Results: -

Link for the result
https://drive.google.com/file/d/19nLJVLlXLF68b_rKIsQCwWfna5r1TQR0/view?usp=sharing

Operation Analytics and Investigating Metric Spike

Findings – IX

To calculate the email engagement metrics: -

- First, we will categorize the action into 'email_opened', 'email_sent', 'email_clicked' using when, case, then functions.
- We will divide sum of category 'email_opened' and sum of category 'email_sent' and multiply by 100 and put the name as email_opening_rate.
- Then we will divide sum of category 'email_clicked' and sum of category 'email_sent' and multiply by 100 and put the name as email_clicking_rate.

Categorizing of action: -

email_opened = ('email_open')

email_sent = ('sent_weekly_digest','sent_reengagement_email')

email_clicked = ('email_clickthrough')

Operation Analytics and Investigating Metric Spike

Findings – IX

Output/Results:-

email_opening_rate	email_clicking_rate
33.5834	14.7899

Operation Analytics and Investigating Metric Spike

Analysis

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

- The number of jobs reviewed per hour for each day in November 2020 is 0.0083
- 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.
- Persian has the highest percentage share of each language over the last 30 days.
- 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 30 i.e. 1467.
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014.
- The email_opening_rate is 33.5834 and email_clicking_rate is 14.7899.

Operation Analytics and Investigating Metric Spike

Analysis

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

- Why is there number of jobs reviewed per hour for each day?
  ----> So they won't miss any jobs.
- 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
- Why is the weekly user engagement so less in the beginning and then got increased?
  -----> It is a fact that for any new product or service launched, during its initial period in the market it is less known to all people only some people use the product and based on their experience the product/service engagement increases or decreases depending on whether the consumer experience

Operation Analytics and Investigating Metric Spike

Analysis

was good or bad. In this case since the user engagement increased after 2-3 weeks of the launch means that the consumer had a good experience with the product/service.

- Why is weekly retention so important?
  ---> Weekly retention helps the firms to convince and help those visitors who just complete the sign-up or leave the sign-up process in between, such visitors may become customers in future if they are guided and convinced properly.
- Why is weekly engagement per device plays an important role?
  ----> Based on the reviews from users weekly engagement per device helps the firms on which devices they must focus more and which devices need more improvements so they also get a good review in users weekly engagement per device.
- Why is Email Engagement plays an important role?
  ----> Email Engagement helps the firms to decide the discounts and offers on specific products. In this case the email_opening_rate is 33.58 i.e., out of the 100 mails send only 34 mails were opened and the email_clicking_rate is 14.789 i.e., out of 100 mails opened only 15 mails were clicked for more details regarding the discount/product details. This means that the current firm needs to have some more catchy line for mails also the firm needs to do rigorous planning and deciding content before sending the mails.

Operation Analytics and Investigating Metric Spike

Conclusion

In Conclusion, I would like to conclude that Operation Analytics and Investigating Metric Spike are very necessary and they must be done on daily, weekly, Monthly, Quarterly or Yearly basis based on the Business needs of the firm.

Also, any firm/entity must focus on the Email Engagement with the customers; the firm must use catchy headings along with reasonable discounts and coupons so as to increase their existing customer base.

Also any firm must have a separate department (if possible) so as to hear out to the problems of those Visitors who had left the Sign-up Process in between, the firm must guide them so as to convert them from Visitors to Customers.

Hiring Process Analytics

Description

Hiring Process is most important thing that can affect the ultimate growth of a company. The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department.

Being a Data Analyst, your job is to analyze the company's hiring process data and draw meaningful insights from it.

You are working as a Data Analyst for a multinational company such as Google and the company has provided with the data records of their previous hirings and have asked certain questions making sense out of that data.

Hiring Process Analytics

The Problem

- **Hiring Analysis:** The hiring process involves bringing new individuals into the organization for various roles. Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company?
- **Salary Analysis:** The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees. Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.
- **Salary Distribution:** Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class. Your Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.
- **Departmental Analysis:** Visualizing data through charts and plots is a crucial part of data analysis. Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.
- **Position Tier Analysis:** Different positions within a company often have different tiers or levels. Your Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

Hiring Process Analytics

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.
- Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.
- I looked for blank spaces and NULL values if there is any.
- I did imputations on Blank and NULL cells of the numeric columns with Mean and Median.
- Then I checked for outliers and replaced them with the Median of that particular column where the outlier existed.
- I did imputations on blank and NULL cells of the categorical columns with Mode.
- Then I removed duplicate rows from the datasets.

Software used for doing the overall Analysis: -

----> Microsoft Excel

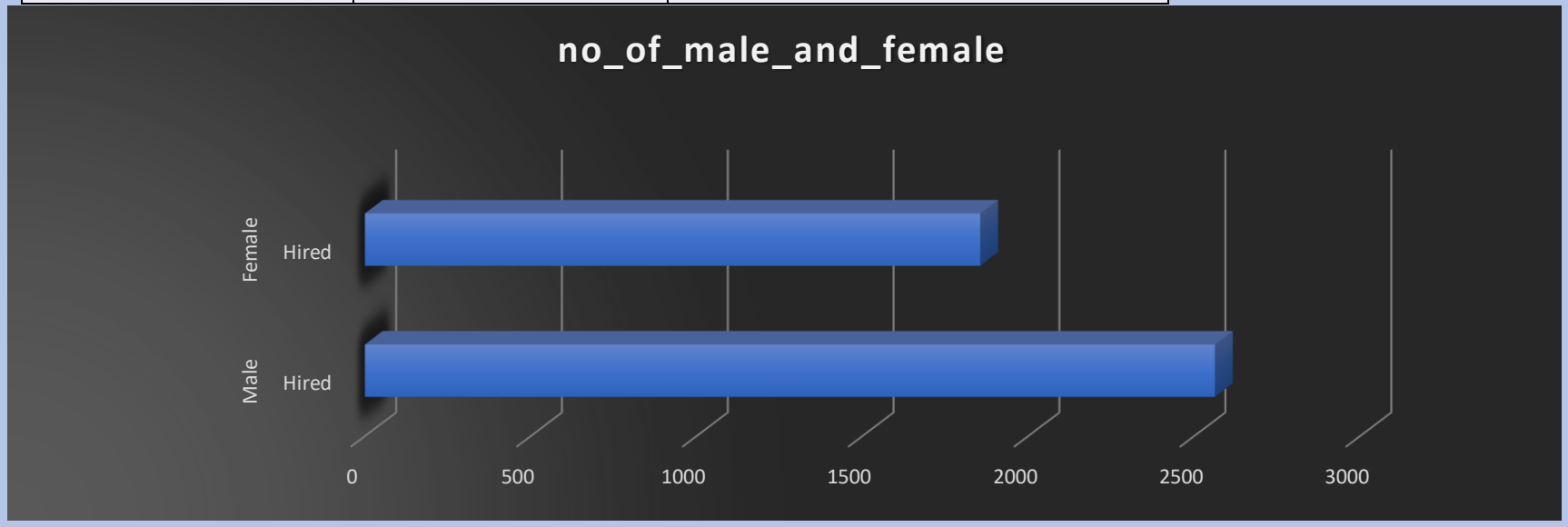
Hiring Process Analytics

Findings – I

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

Output/Results: -

event_name	Status	no_of_male_and_female
Male	Hired	2563
Female	Hired	1856



Hiring Process Analytics

Findings - II

To find the average salary offered by this company:

- Using the Formulae to calculate average salary offered by this company.
  =AVERAGE

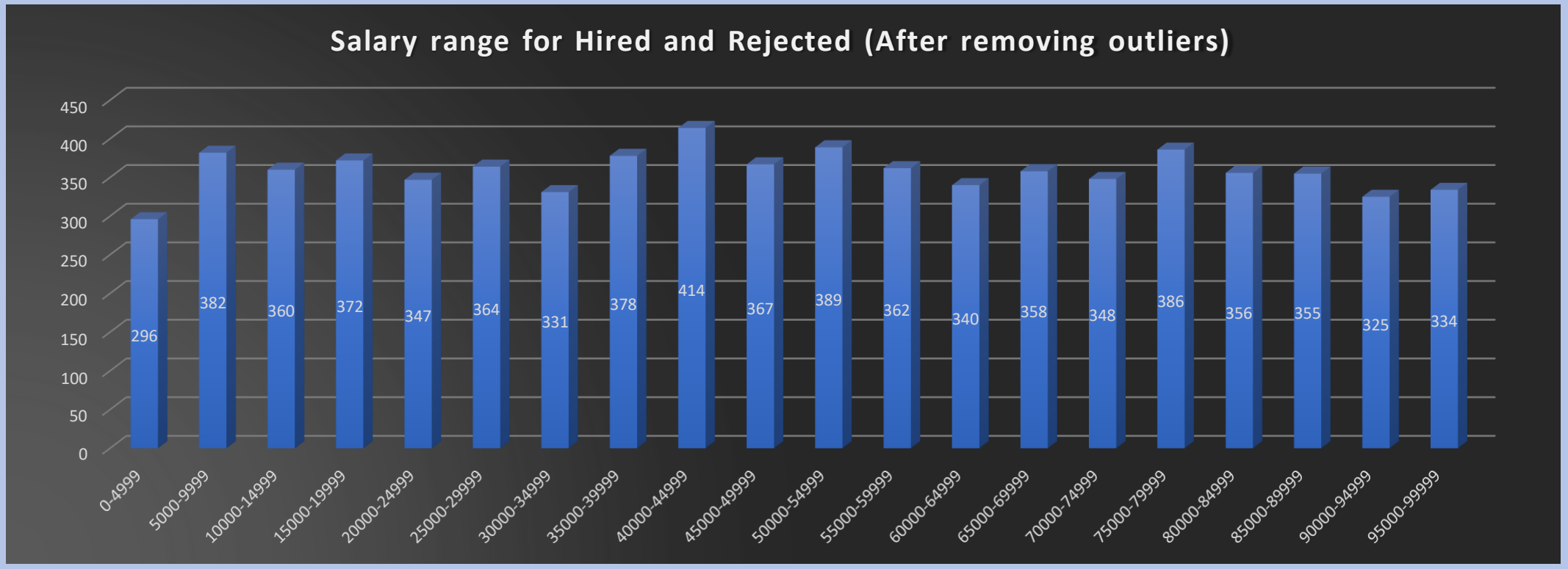
Output/Results: -

Average	49983.02902

Hiring Process Analytics

Findings - III

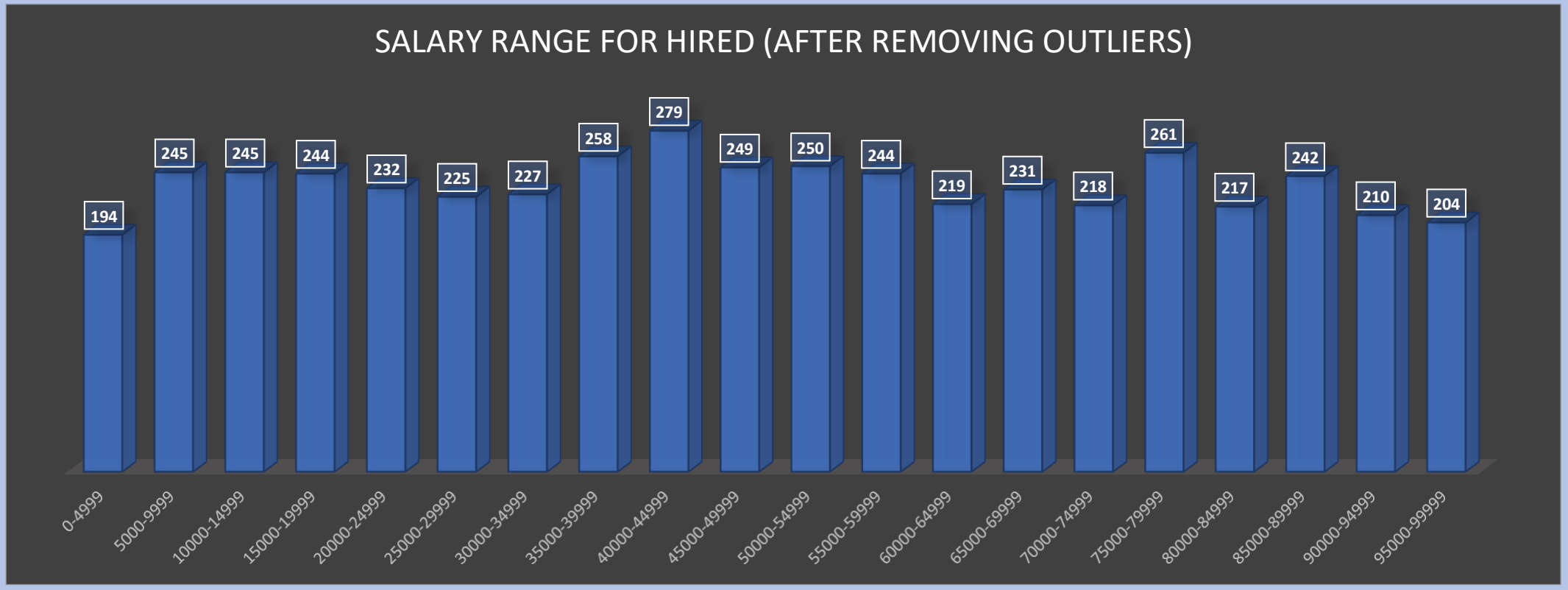
To create class intervals for the salaries in the company:



From the above Bar plot, I can see that the highest number of posts (both hired and rejected) is 414 for the salary range 40000 to 44999.

Hiring Process Analytics

Findings - III



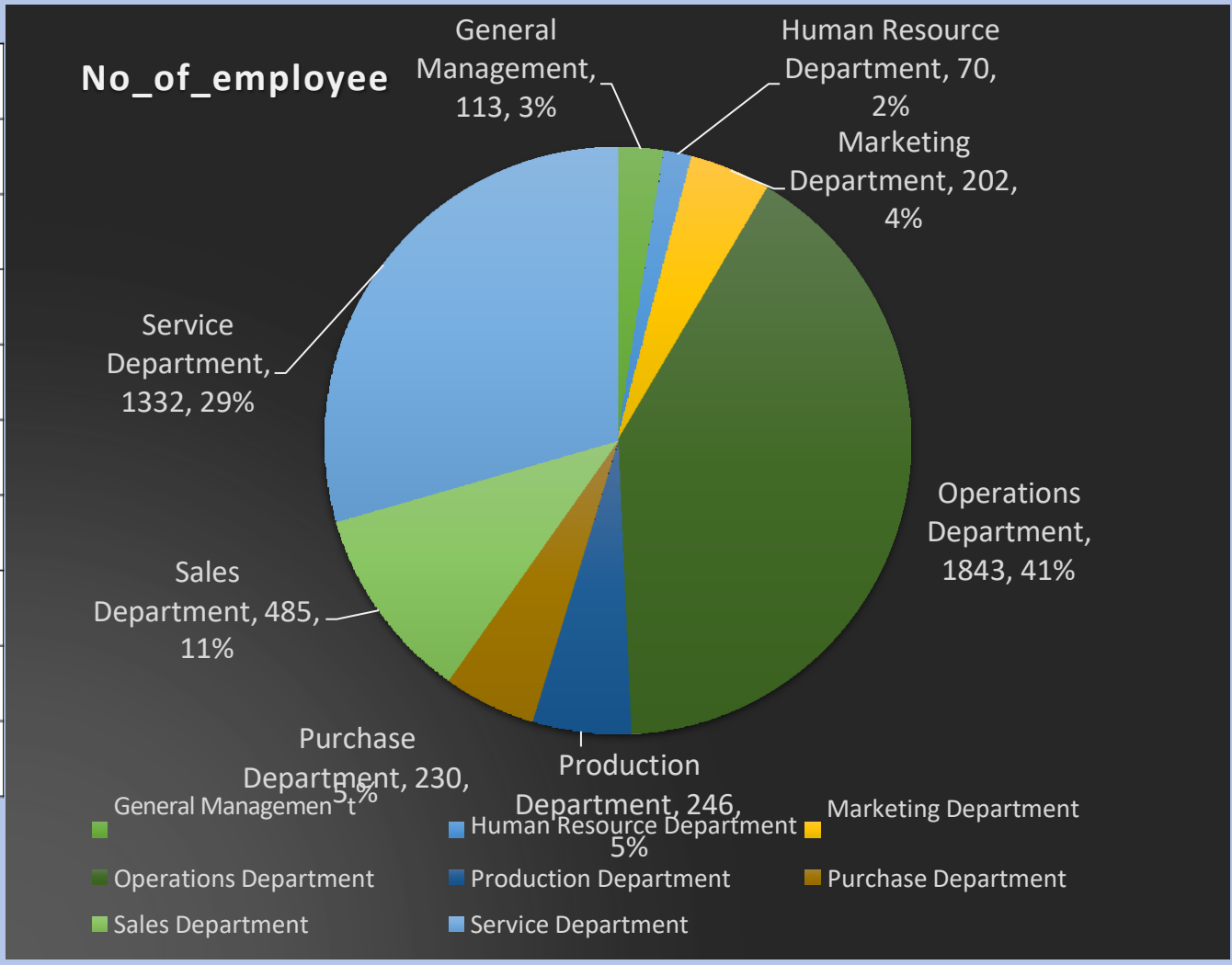
From the above Bar plot, I can see that the highest number of posts (hired) is 279 for the salary range 40000 to 44999.

Hiring Process Analytics

Findings – IV

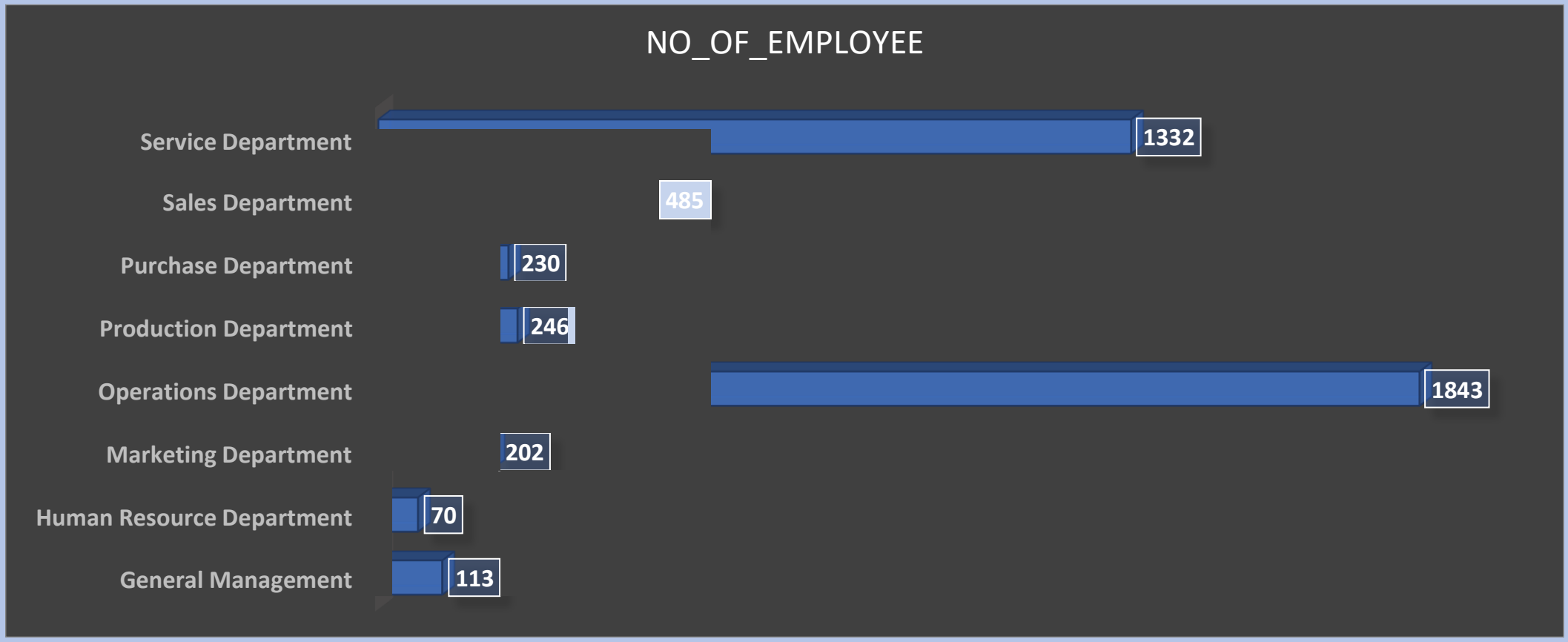
To show the proportion of people working in different departments:

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 – IV



From the above table, pie chart and Bar Plot I have inferred that the Highest number of people are working in the Operations Department i.e., 1843 which accounts for almost 41% of the total workforce of the company.

Hiring Process Analytics

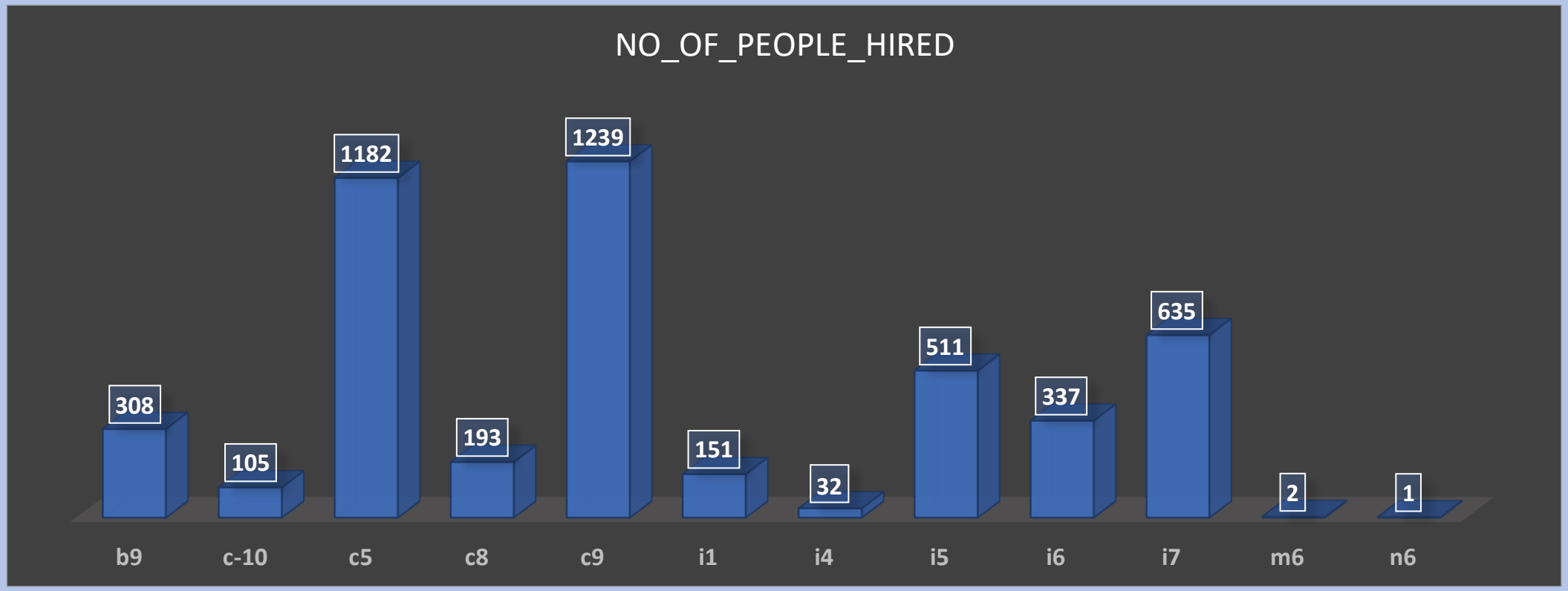
Findings – V

To find different position tiers within the company:

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 – V



From the above Bar plot, I can see that the c9 post has the highest number of openings i.e. 1239.

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

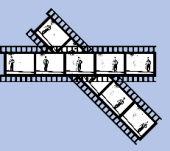
Conclusion

I can conclude that Hiring Process Analytics plays vital role for all the companies and firms to decide the job openings for the near future. Hiring Process Analytics is done on monthly, quarterly or yearly basis as per the needs and policies of the companies.

For any company the Operations Department has the highest number of workforce due to the workload on this department as this department acts as a central hub for all the executive tasks carried out.

For any company there will some employees who have high salary packages compared to other employees, and this is due to the fact that they have some special skills and years of experience in their particular field of work.

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



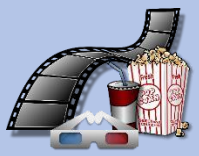
IMDB Movie Analysis

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.

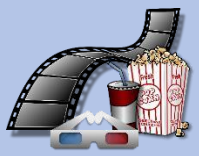
First you need to clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.



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

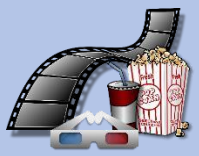
Design

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.
- 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 for doing the overall Analysis: -

----> Microsoft Excel



IMDB Movie Analysis

Findings – I

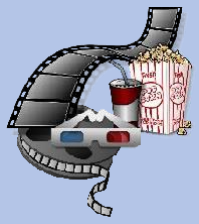
To find the most common genres of movies in the dataset: -

- 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.

Output/Results: -

Most common genres are: -								
Genres	**Count**	**Mean**	**Median**	**Mode**	**Max**	**Min**	**Variance**	**Standard Deviation**
Drama	153	7.04183	7.2	7.3	8.8	3.4	0.687055	0.828887522
Comedy|Drama|Romance	151	6.494702	6.5	6.5	8	4.3	0.562772	0.750181141
Comedy|Drama	147	6.583673	6.7	6.7	8.8	3.3	0.7348	0.857204825
Comedy	145	5.84069	6	6.5	8	1.9	1.481875	1.217322686
Comedy|Romance	135	5.896296	6	6.1	8.4	2.7	0.76827	0.87650999

From the above table it shows that the most common Genre is Drama.



IMDB Movie Analysis

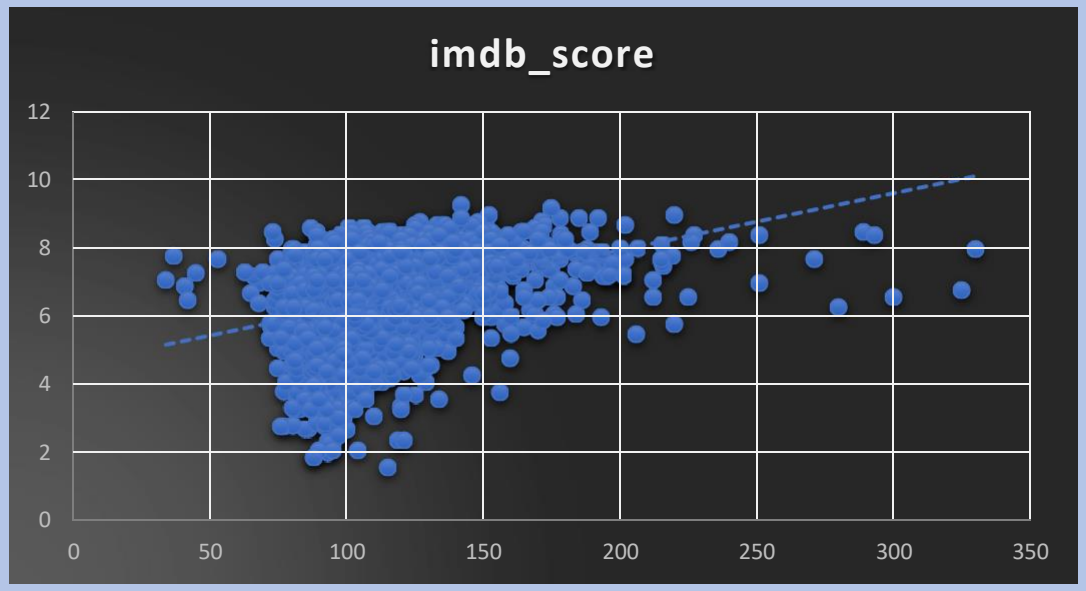
Findings – II

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.

Output/Result: -

Average	109.9241164
Median	106
Standard Deviation	22.75364979

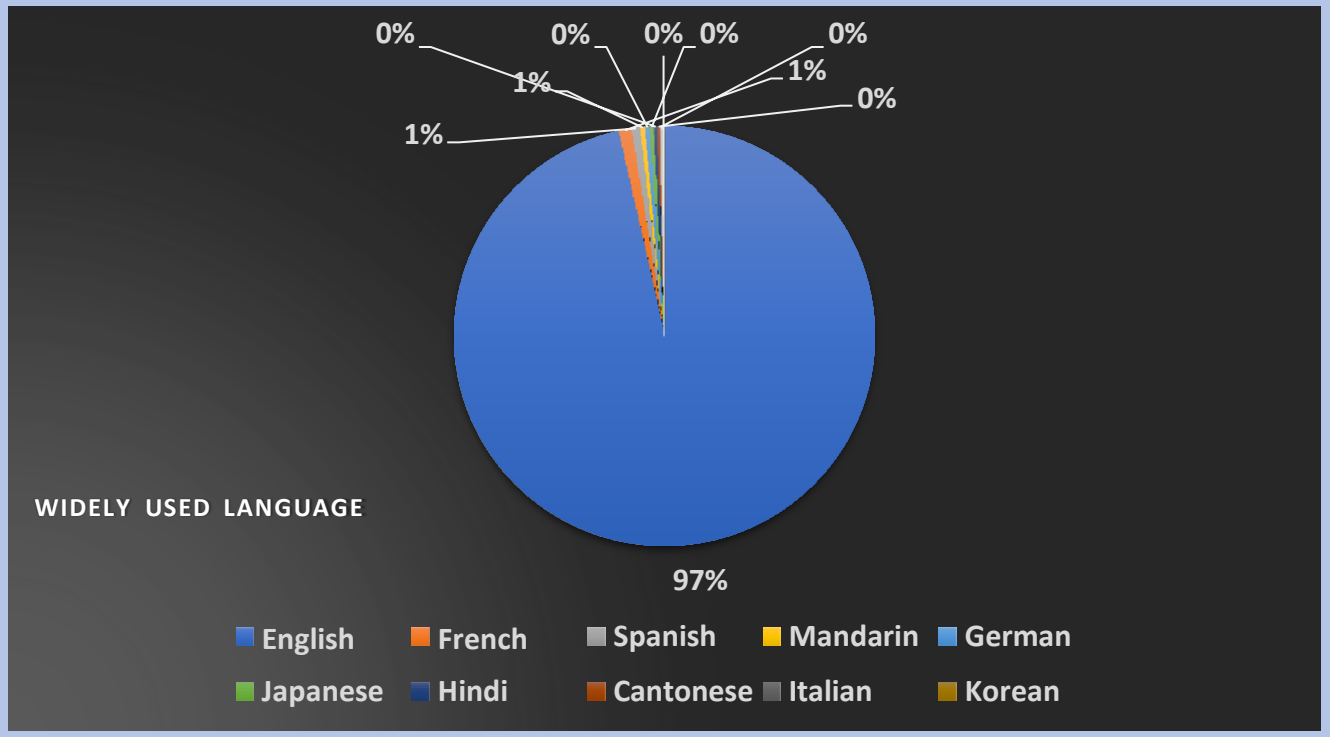


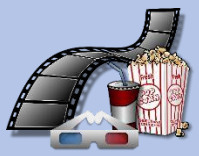
IMDB Movie Analysis

Findings - III

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.



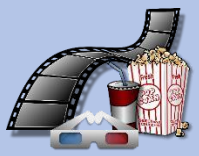


IMDB Movie Analysis

Findings – III

Most common Languages are: -				
Language	**Count**	**Mean**	**Median**	**Standard Deviation**
English	3668	6.423909	6.5	1.048750752
French	37	7.286486	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.021429	7.25	0.765786244
German	13	7.692308	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	7.185714	7	1.155318962
Korean	5	7.7	7.7	0.570087713

The most common languages used in movies is English.



IMDB Movie Analysis

Findings - IV

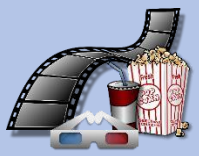
To Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations: -

- 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.

Output/Results: -

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

From above table it shows that top directors are Charles Chaplin, Tony Kaye.



IMDB Movie Analysis

Findings – V

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.

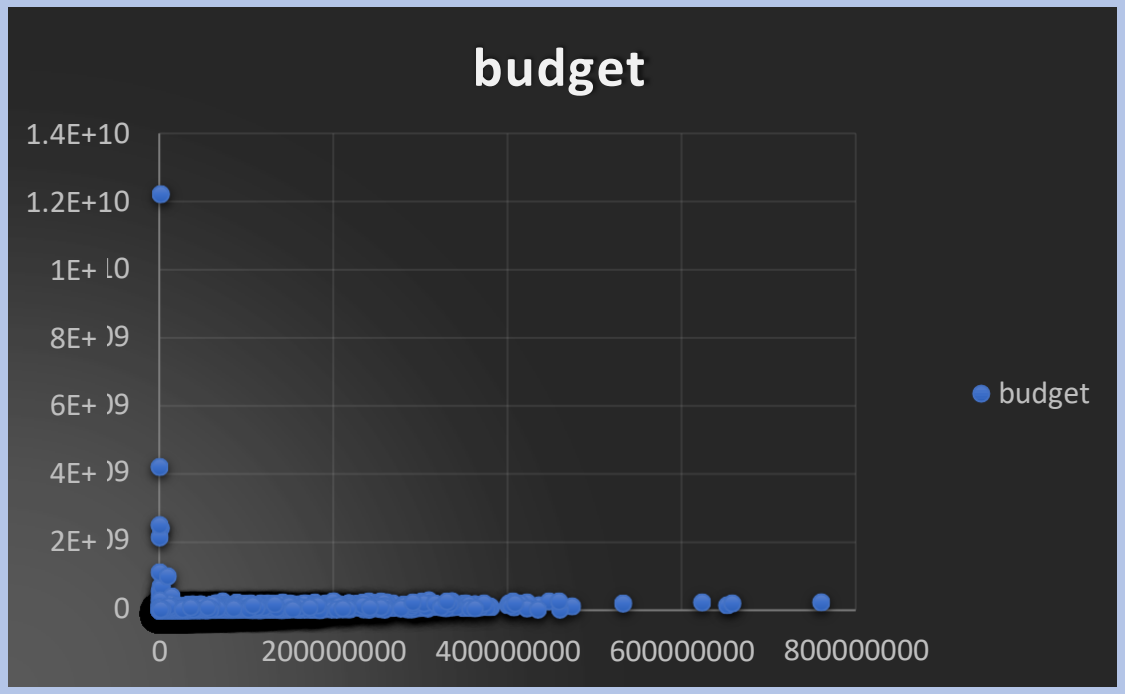
Output/Results: -

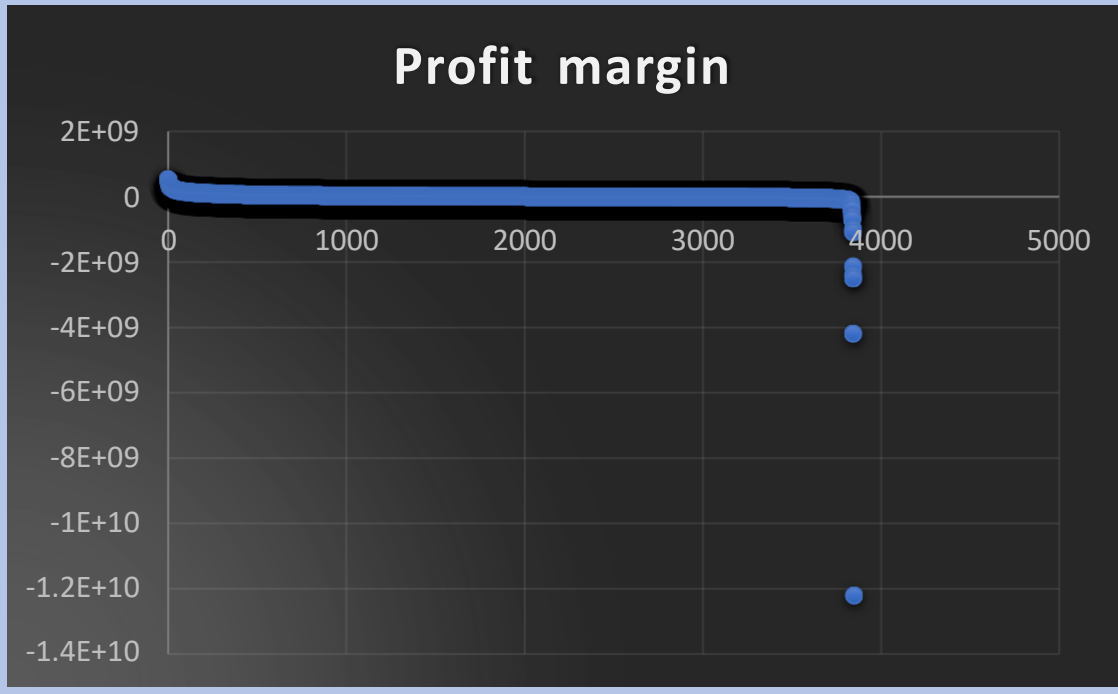
CORRELATION
0.100850218

MAX PROFIT	MOVIE TITLE
523505847	AvatarÂ

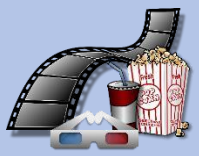
IMDB Movie Analysis

Findings – V





Above table shows that the movies with the highest profit margin is **AvatarÂ .**

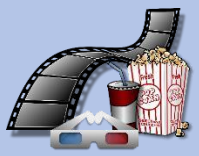


IMDB Movie Analysis

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

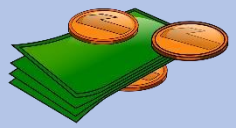
In Conclusion, I would like to conclude that IMDB Movie Analysis or any such analysis is done not only by Movie makers before movie production, but it is also done by various investors, stakeholders, theatre outlet owners.

Normal people would not mind to do such analysis but such analysis plays an crucial part during the pre-production phase of the movies and also during the post-production phase. Also, it is not necessary that the movie with the highest IMDB rating will have the highest profit.

Profit is calculated truly on the basis on the number of tickets sold by theatres all over the world. Most of the people are tired with their daily lives and they prefer movies with Comedy/ Drama genre or both, and they would not go for movies with Action/Horror genre.

So, directors and production team must keep in mind the above points and shall do the pre-production analysis before the commencement of filming.

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.

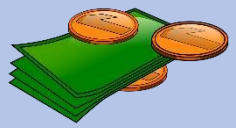
When the company receives a loan application, company faces two risks:

- 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.

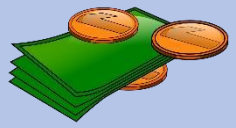
Bank Loan Case Study



The Problem

- **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.
  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 in the dataset using Excel statistical functions and features, 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.

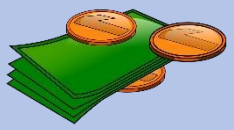
Bank Loan Case Study



The Problem

- **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, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.
- **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%. Those columns need to be dropped.

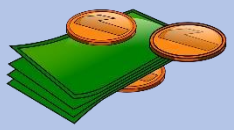
The following columns needs to be dropped as they have more than 50% of NULL values in application datasets: -

COMMONAREA_AVG
COMMONAREA_MODE
COMMONAREA_MEDI
NONLIVINGAPARTMENTS_AVG
NONLIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MEDI
LIVINGAPARTMENTS_AVG
LIVINGAPARTMENTS_MODE
LIVINGAPARTMENTS_MEDI
FONDKAPREMONT_MODE
FLOORSMIN_AVG
FLOORSMIN_MODE
FLOORSMIN_MEDI
YEARS_BUILD_AVG
YEARS_BUILD_MODE
YEARS_BUILD_MEDI
OWN_CAR_AGE

LANDAREA_AVG
LANDAREA_MODE
LANDAREA_MEDI
BASEMENTAREA_AVG
BASEMENTAREA_MODE
BASEMENTAREA_MEDI
EXT_SOURCE_1
NONLIVINGAREA_AVG
NONLIVINGAREA_MODE
NONLIVINGAREA_MEDI
ELEVATORS_AVG
ELEVATORS_MODE
ELEVATORS_MEDI
WALLSMATERIAL_MODE
APARTMENTS_AVG
APARTMENTS_MODE
APARTMENTS_MEDI

ENTRANCES_AVG
ENTRANCES_MODE
ENTRANCES_MEDI
LIVINGAREA_AVG
LIVINGAREA_MODE
LIVINGAREA_MEDI
HOUSETYPE_MODE
FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI

Bank Loan Case Study



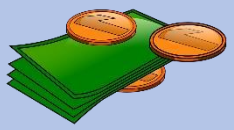
Design

The following columns needs to be dropped in application datasets because these are having irrelevant data for analysis: -

FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EXT_SOURCE_2
EXT_SOURCE_3
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EMERGENCYSTATE_MODE
DAYS_LAST_PHONE_CHANGE

FLAG_DOCUMENT_2
FLAG_DOCUMENT_3
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
FLAG_DOCUMENT_7
FLAG_DOCUMENT_8
FLAG_DOCUMENT_9
FLAG_DOCUMENT_10
FLAG_DOCUMENT_11
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21

Bank Loan Case Study



Design

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

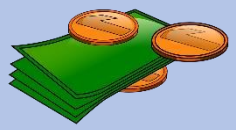
RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED
AMT_DOWN_PAYMENT
RATE_DOWN_PAYMENT

The following columns needs to be dropped in previous_application datasets because these are having irrelevant data for analysis: -

NAME_TYPE_SUITE
PRODUCT_COMBINATION
WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START
FLAG_LAST_APPL_PER_CONTRACT
NFLAG_LAST_APPL_IN_DAY

- I did imputations on Blank and NULL cells of the numeric columns with Mean and Median.

Bank Loan Case Study



Design

- Then I checked for outliers and replaced them with the Median of that particular column where the outlier existed.
- On CNT_PAYMENT I did custom imputation. Most of Blank cells of cnt_payments have contract_status as cancelled, refused, unused offer. So, it makes more sense replacing them with 0 rather than Mean or Median.
- I did imputations on blank and NULL cells of the categorical columns with Mode.
- Then I removed duplicate rows from the datasets.

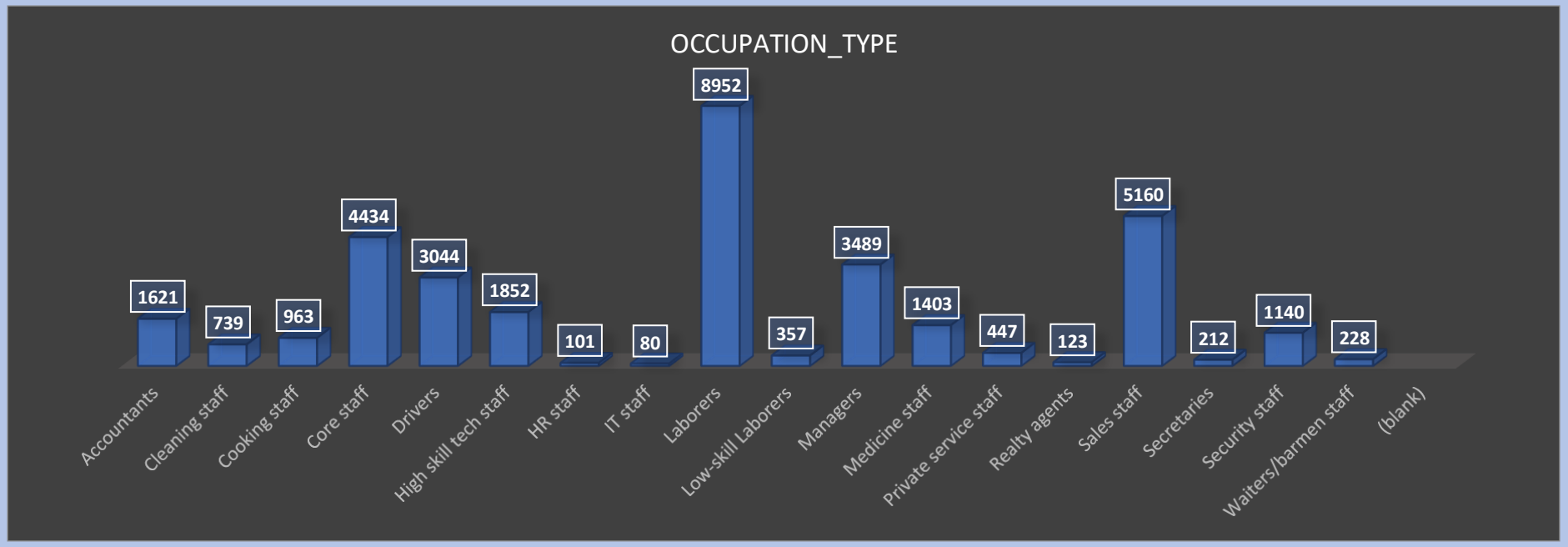
Software used for doing the overall Analysis: -

----> Microsoft Excel

Bank Loan Case Study

Findings - I

OCCUPATION_TYPE

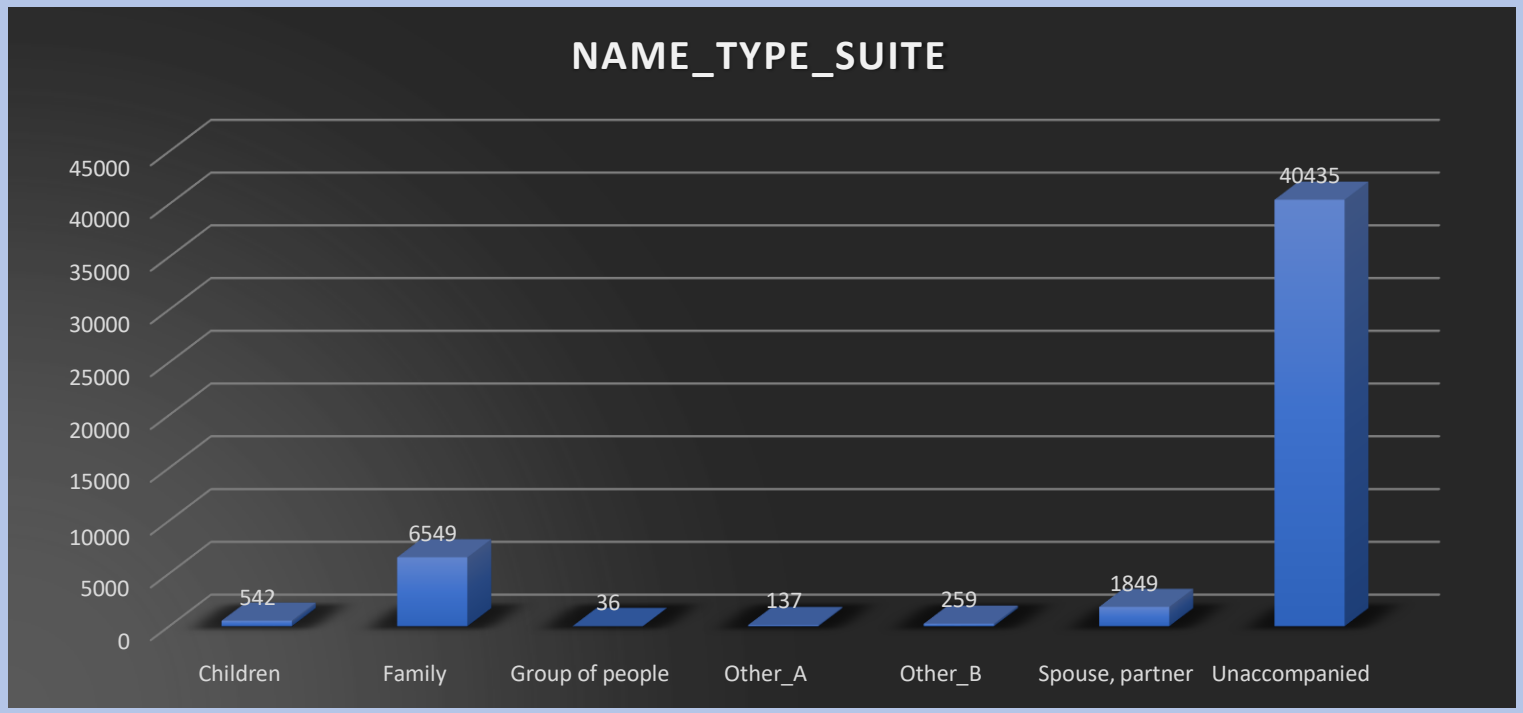


Most Occurring Variable is Laborers. We will replace blanks with 8952.

Bank Loan Case Study

Findings - II

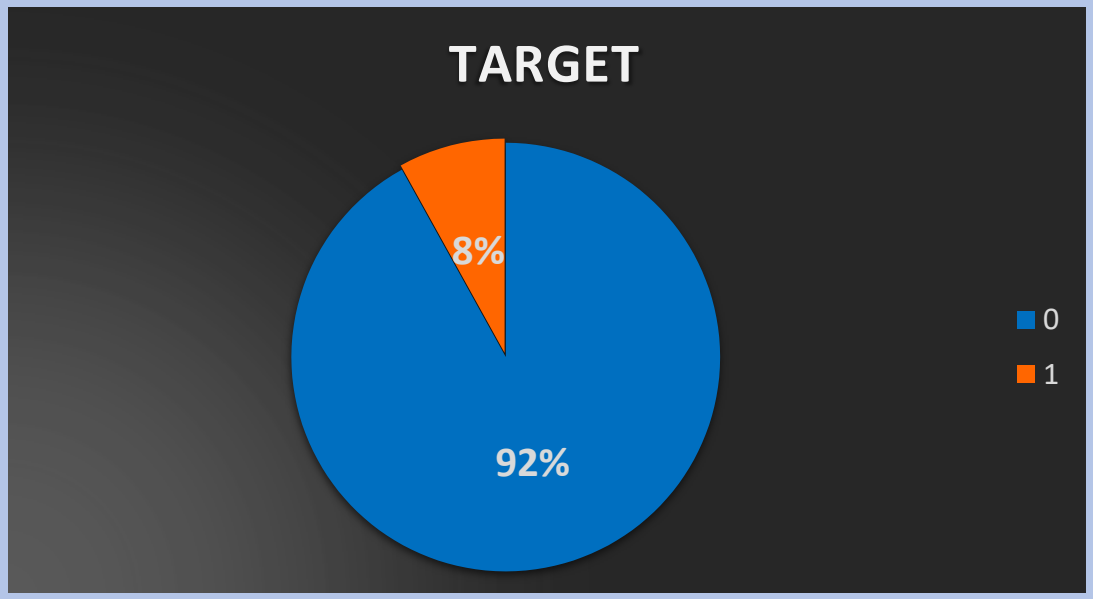
NAME_TYPE_SUITE



Most Occurring Variable is Unaccompanied.

Bank Loan Case Study

Findings - III



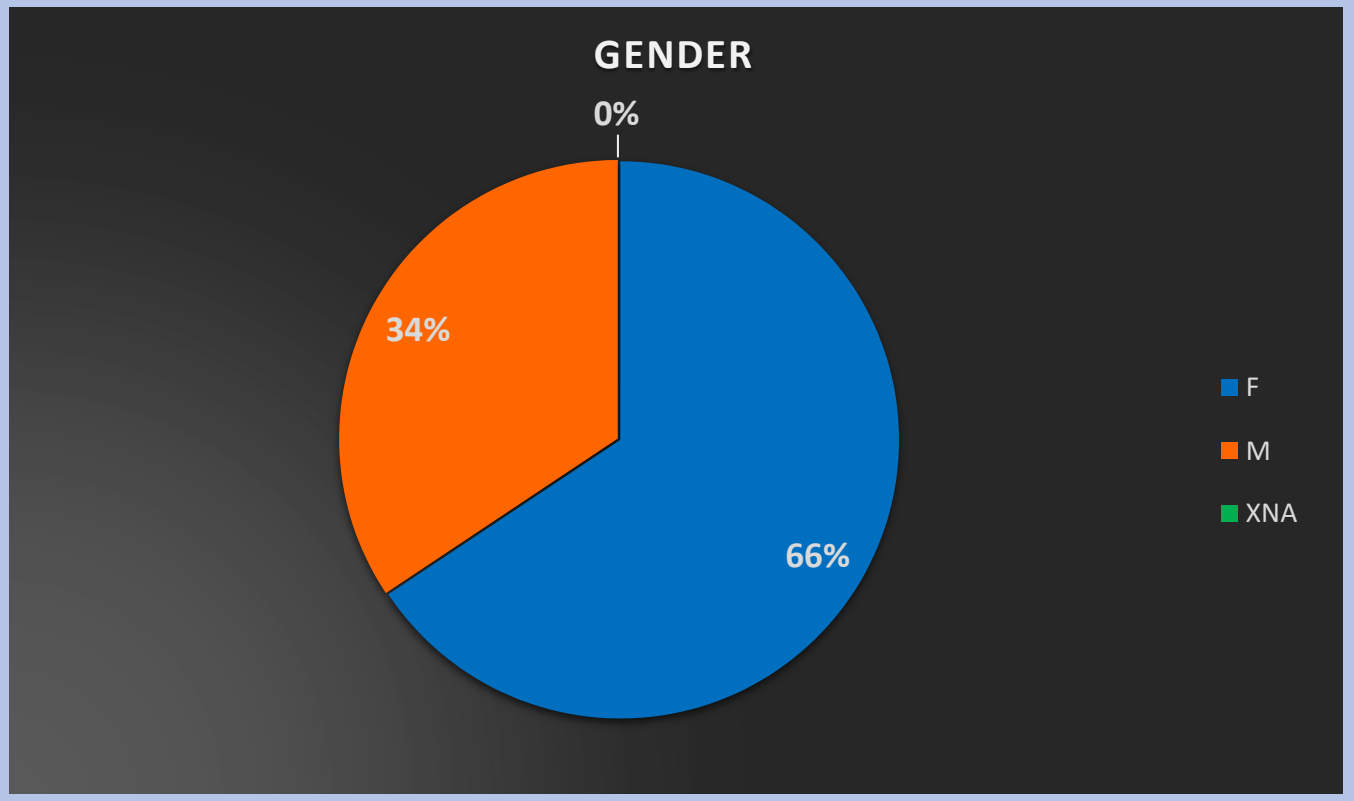
0 – Payment on time

1 – Late Payment

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

Bank Loan Case Study

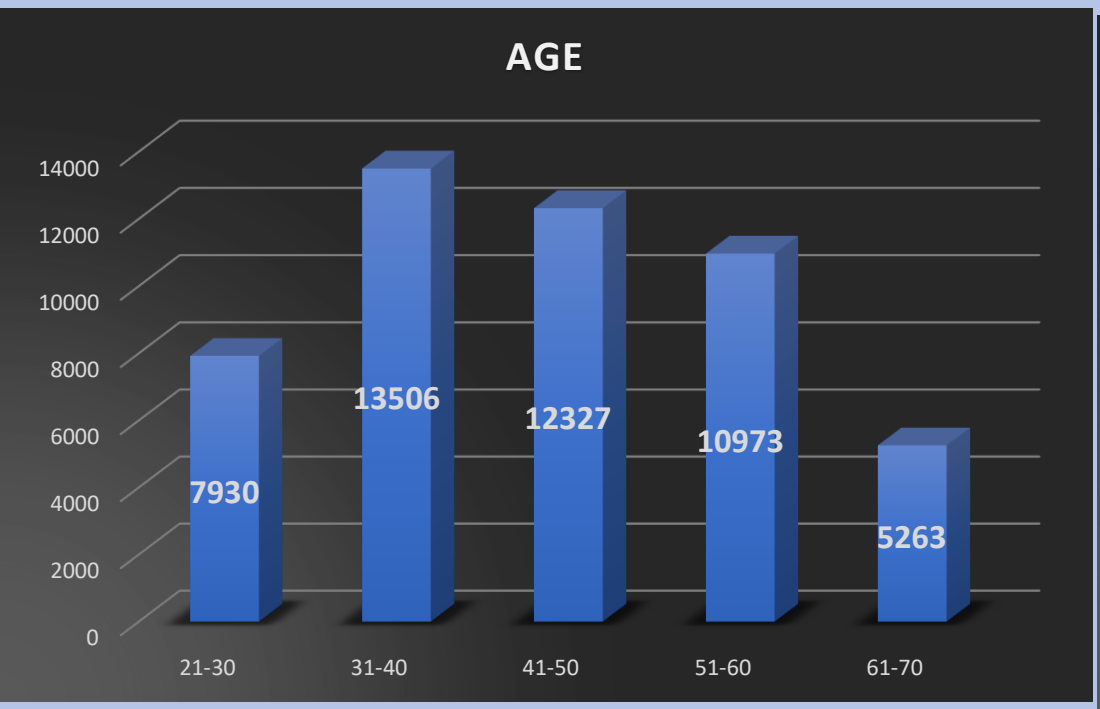
Findings - IV

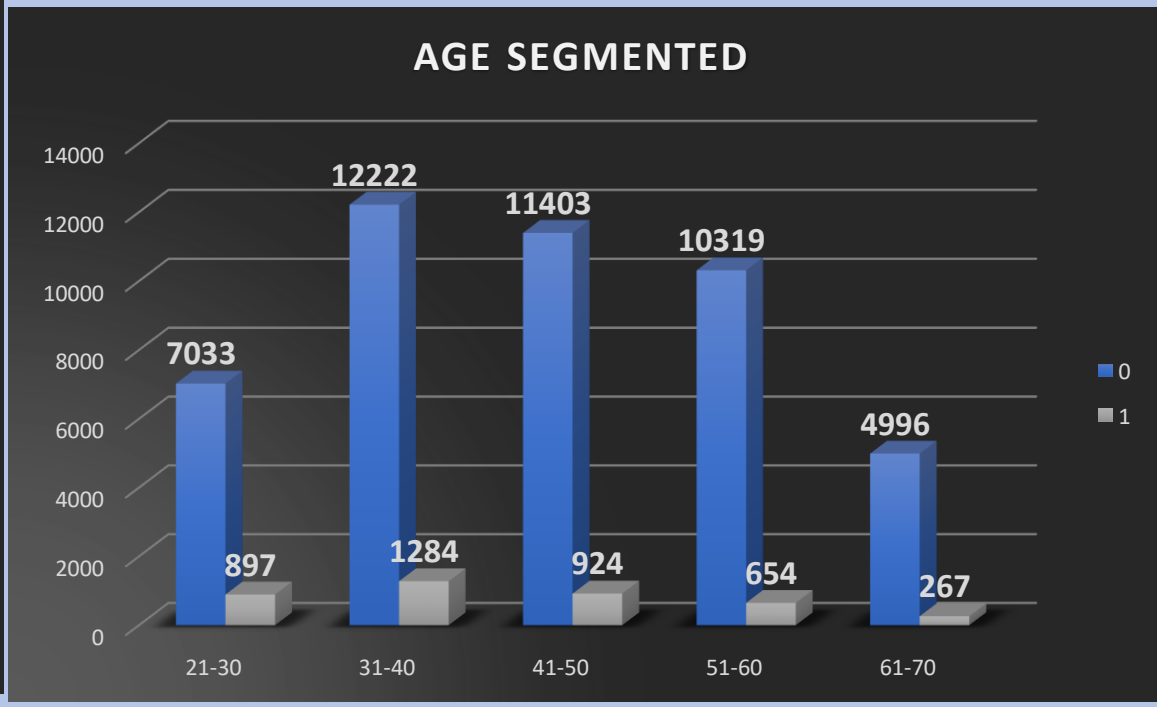


Almost 66% client are Female and 34% clients are Male.

Bank Loan Case Study

Findings - V



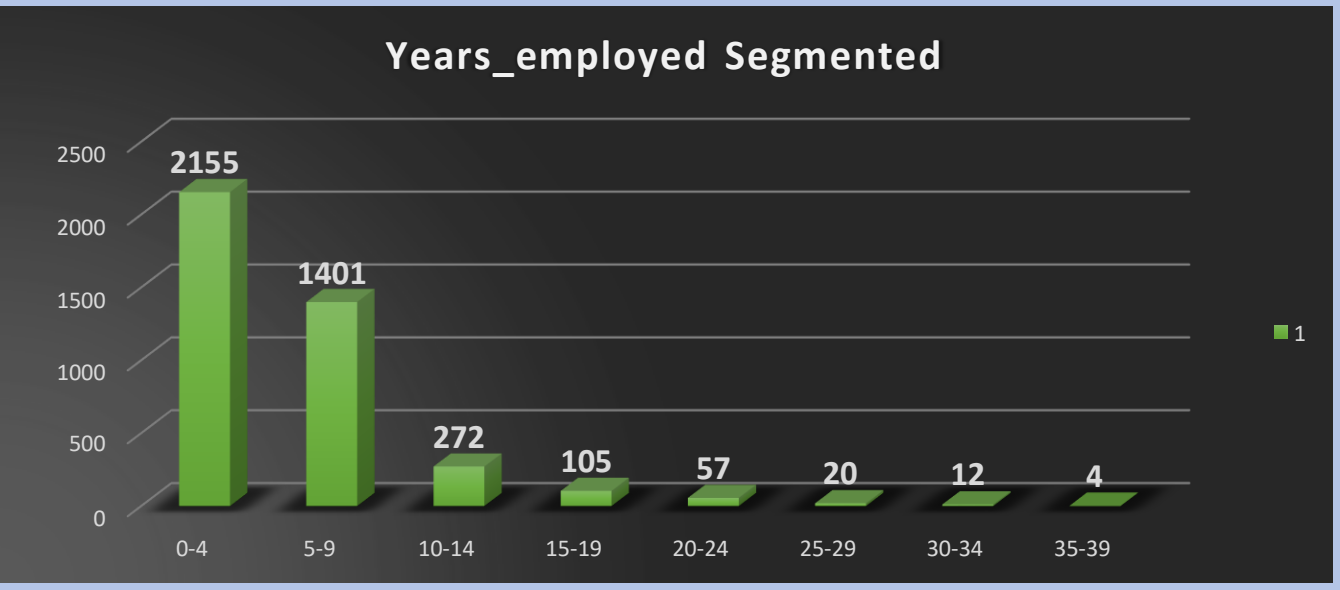


Majority of the Clients are in the age group 31-40.

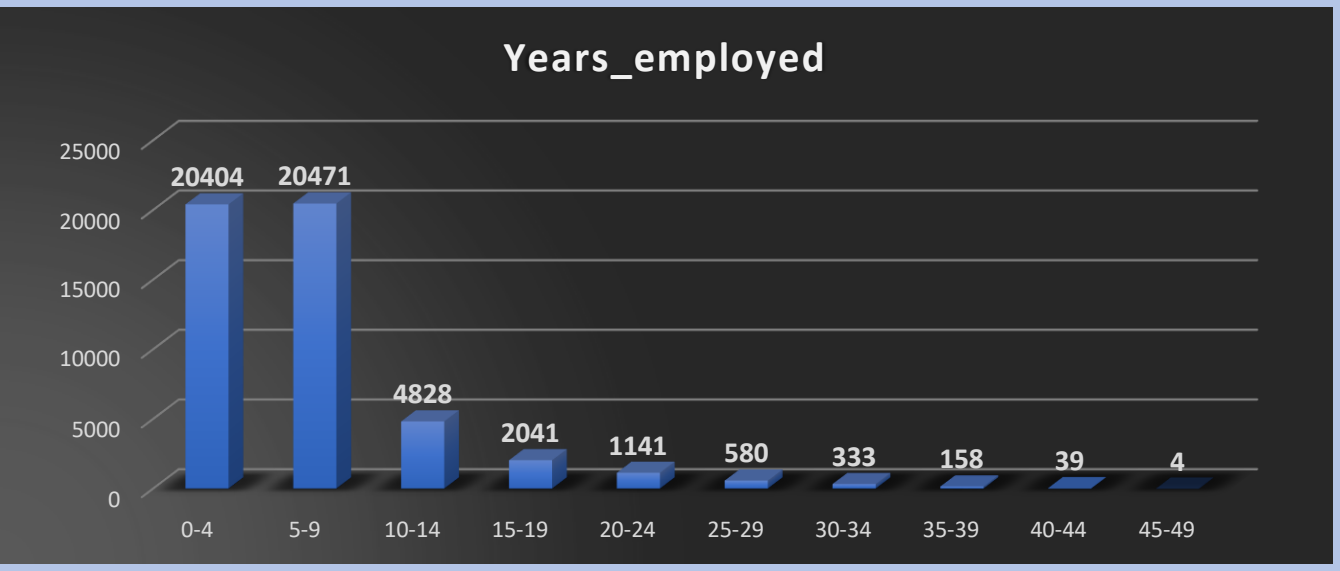
We can see in AGE SEGMENTED, as age increases , chances of defaulter decreases.

Bank Loan Case Study

Findings - VI



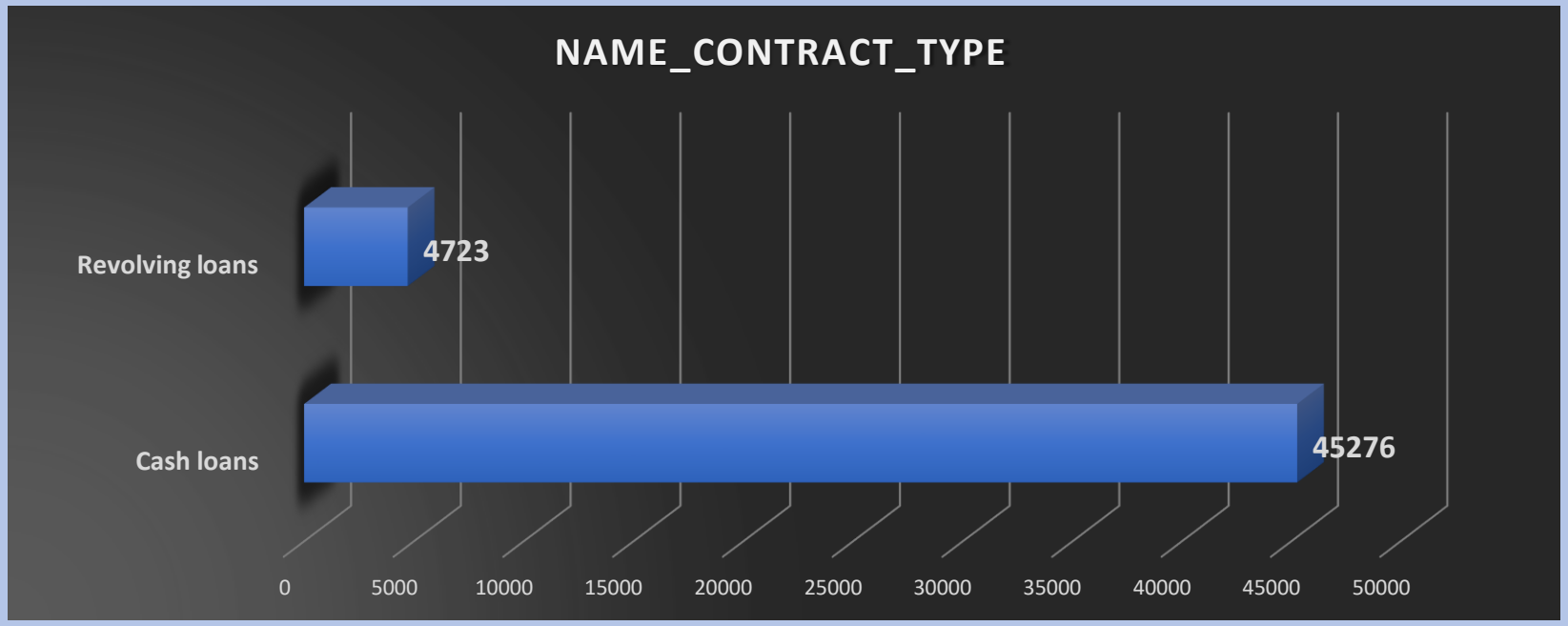
Majority of the Clients are having 0-9 years of experience.



we can see as experience increases, chances of defaulting decreases.

Bank Loan Case Study

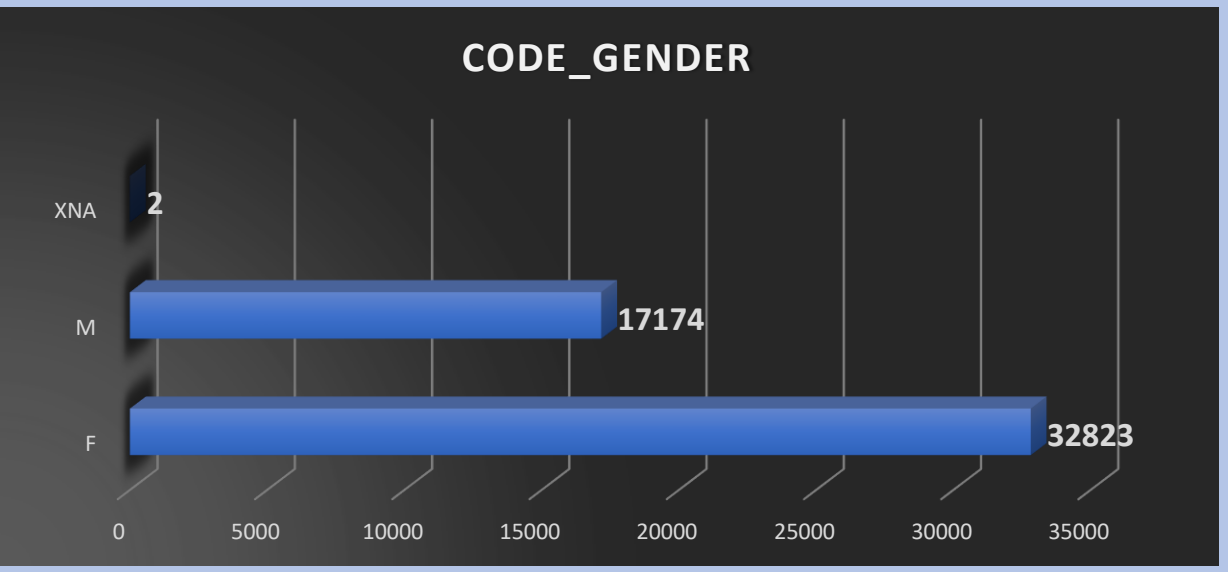
Findings - VII

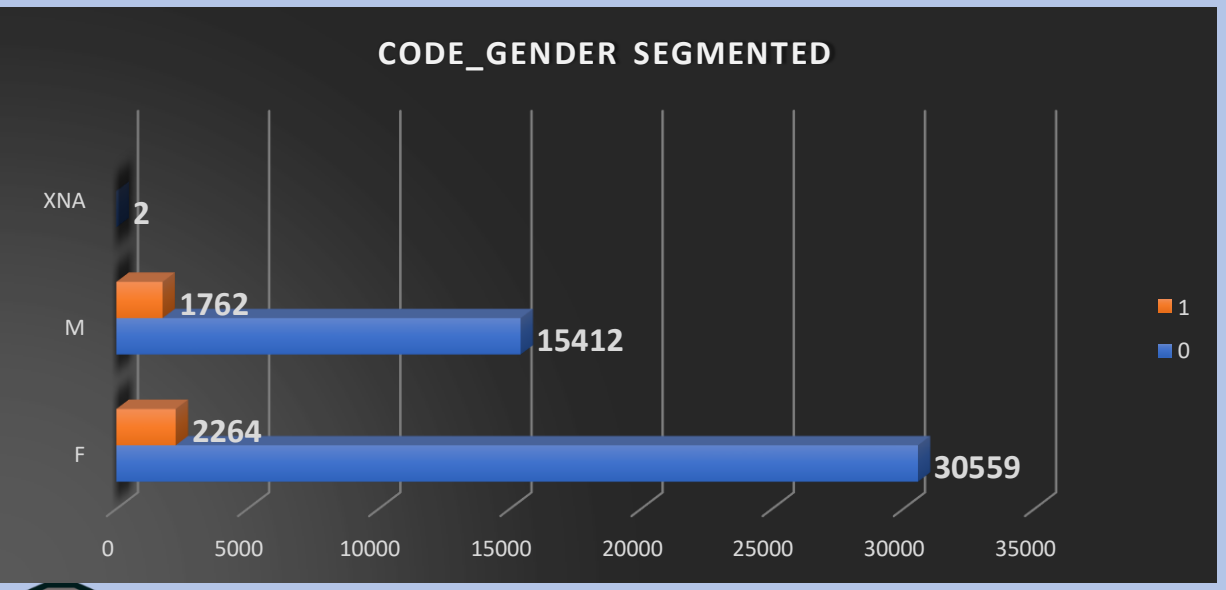


Majority of the Clients are taking Cash loans.

Bank Loan Case Study

Findings - VIII

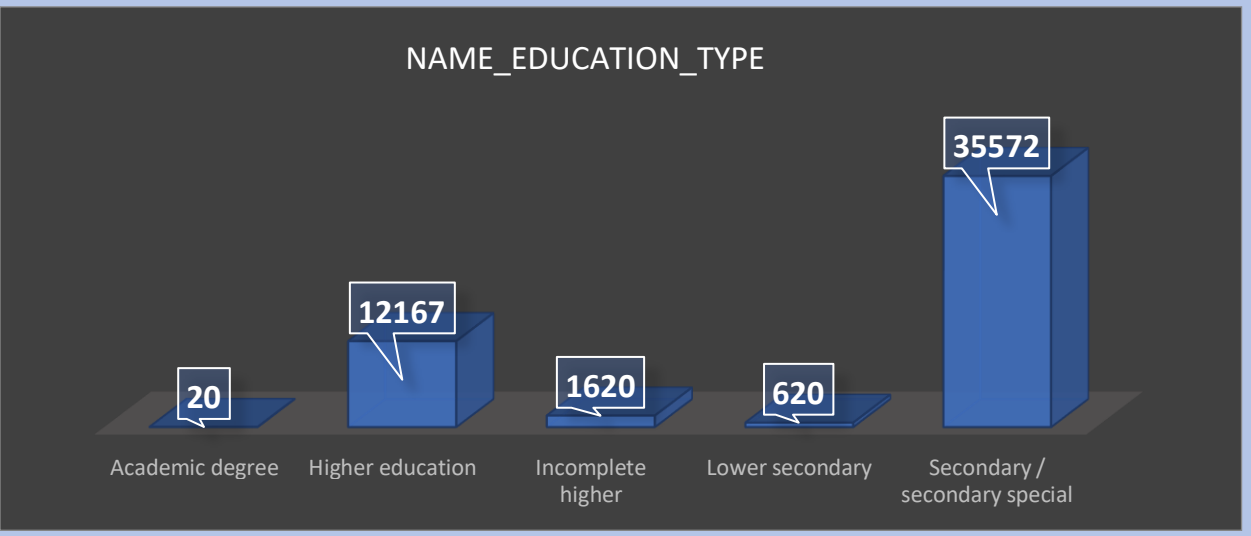




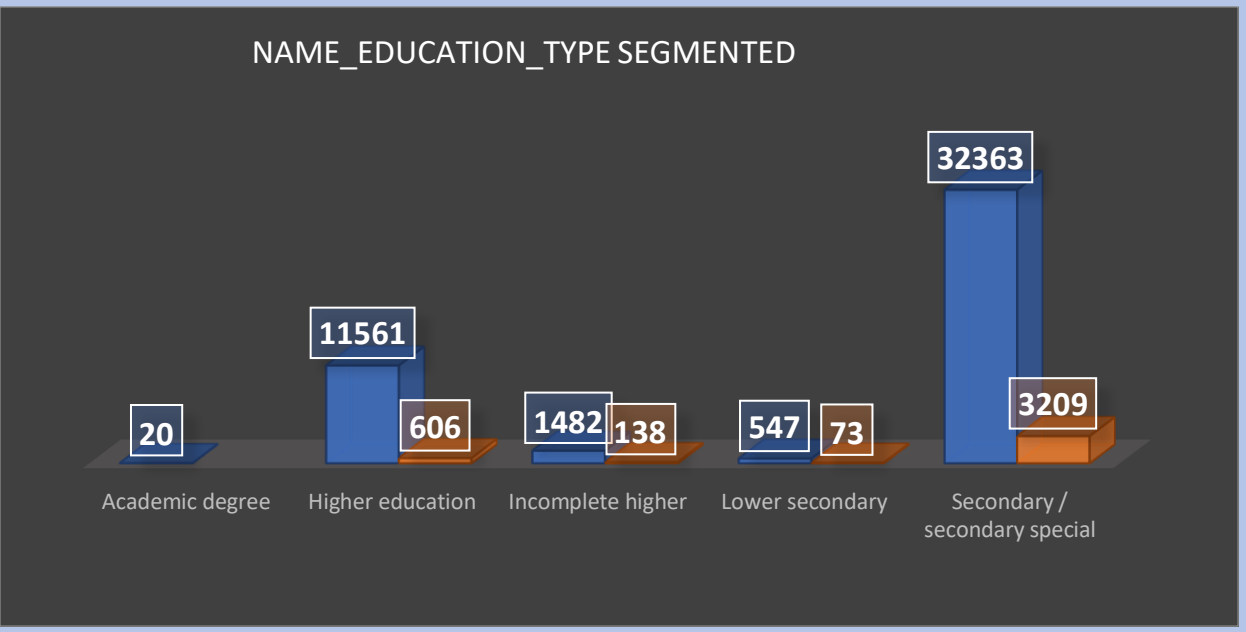
Male are less defaulters compared to Female.

Bank Loan Case Study

Findings - IX



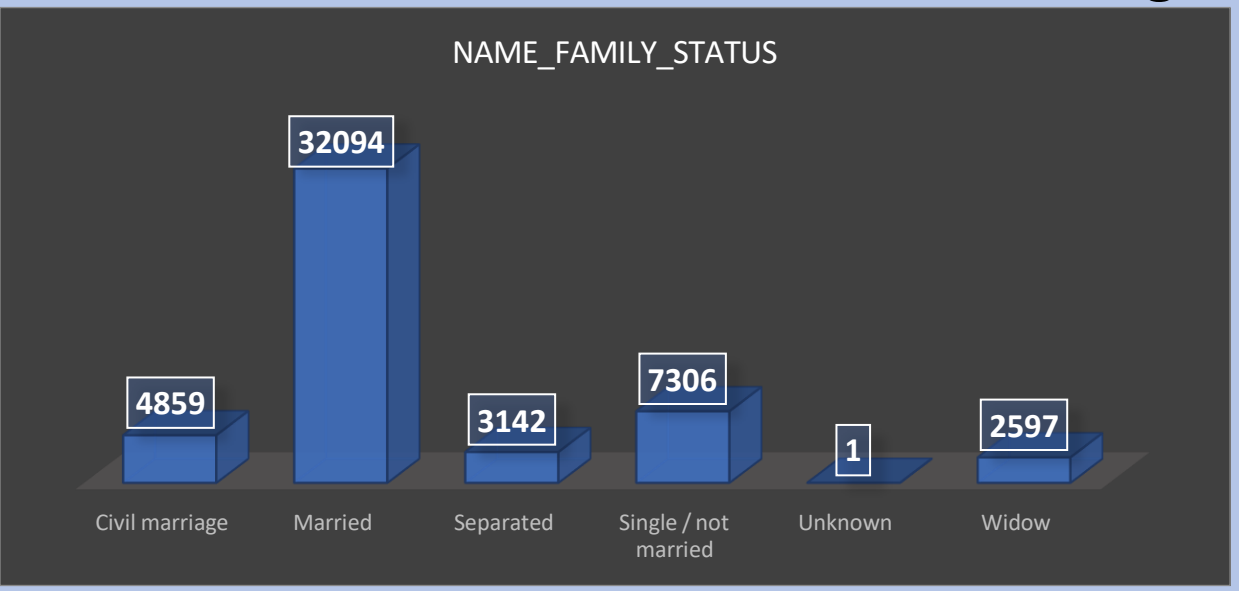
The numbers of loans taken by Clients with Secondary special Education is the highest and Academic degree is the lowest.

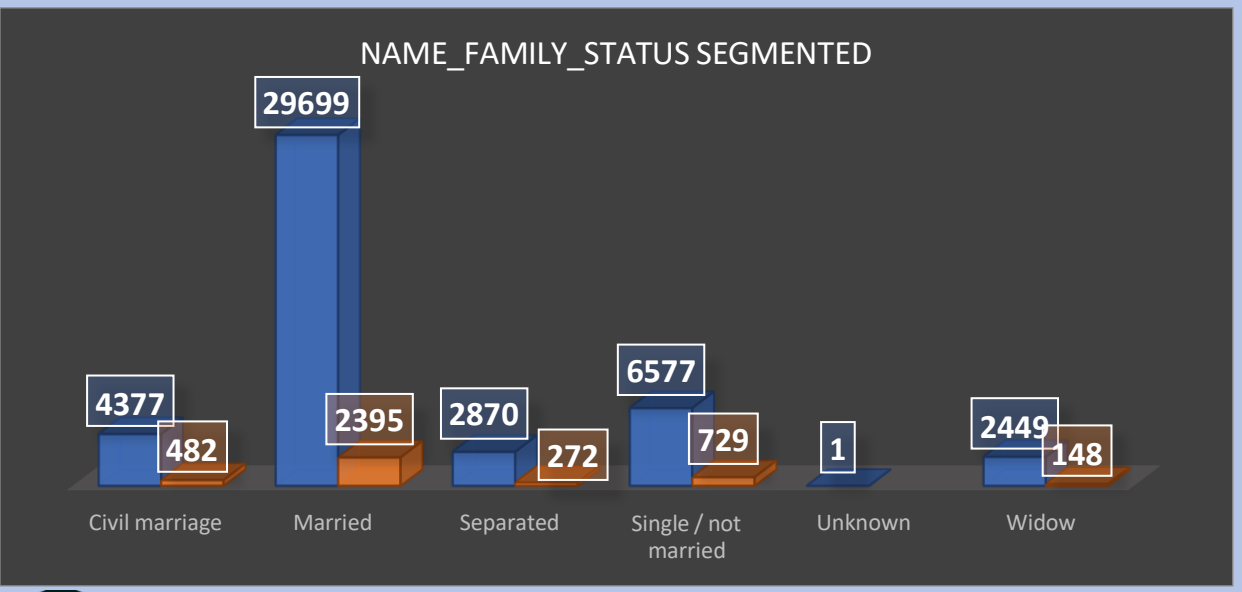


Least default: Academic degree
Highest default: Secondary special

Bank Loan Case Study

Findings - X





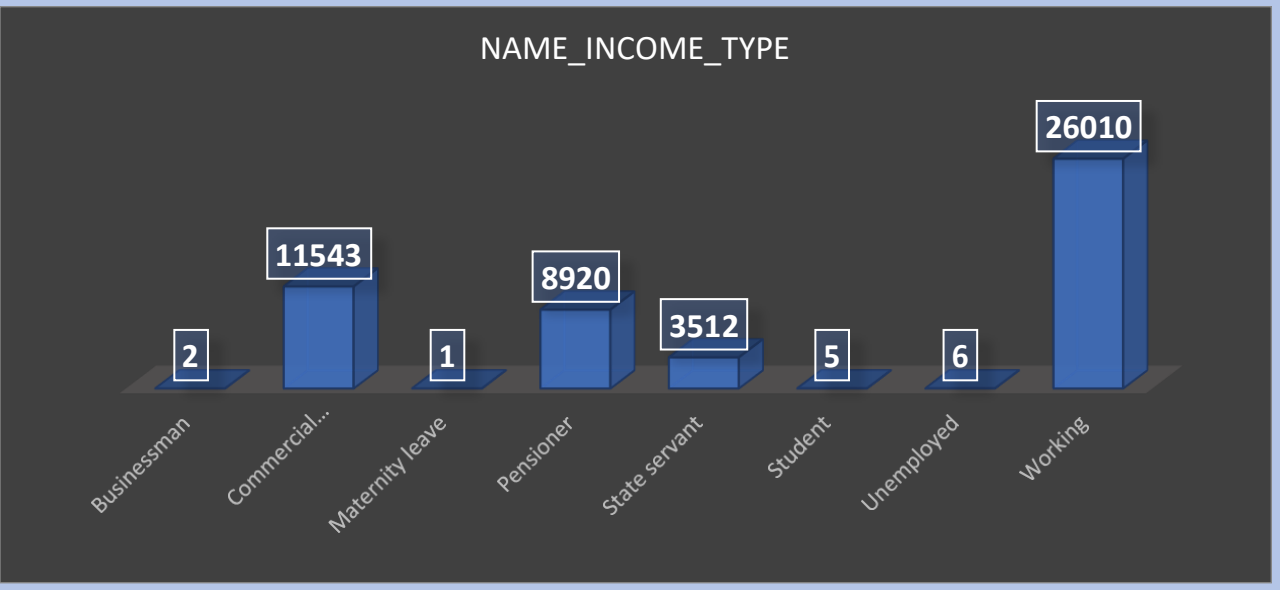
The number of loans taken by Married clients are the highest and clients who are widows are the least if we ignore unknown.

Least Defaulter: Widow

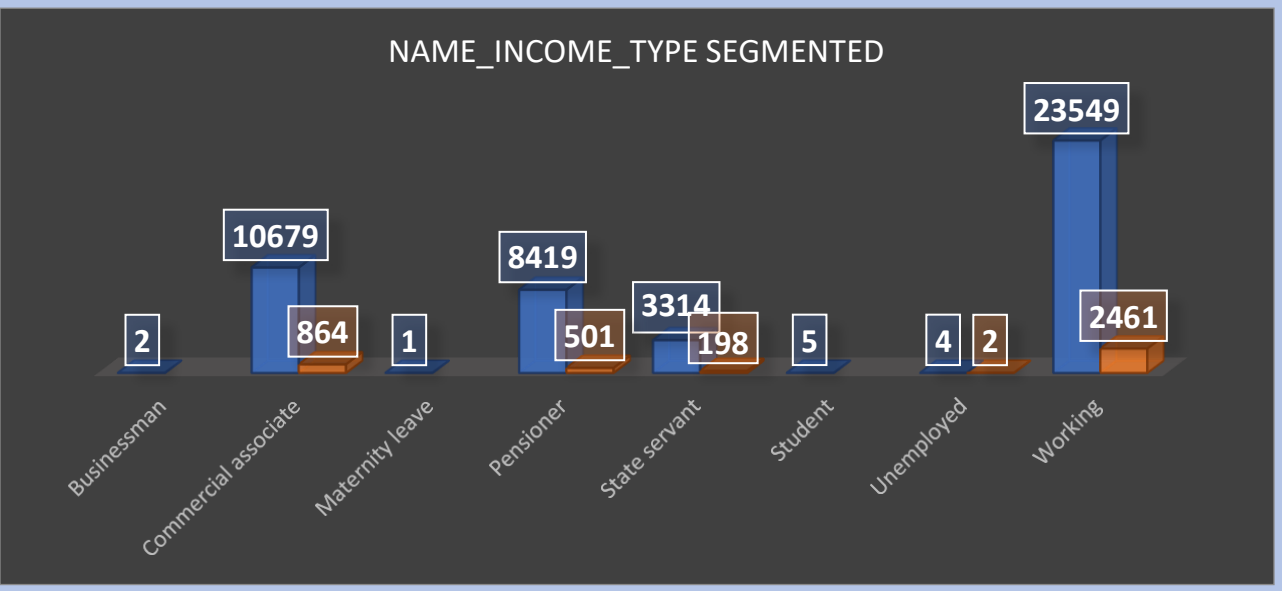
Highest Defaulter: Married

Bank Loan Case Study

Findings - XI



Bank target those groups whose income type is working.

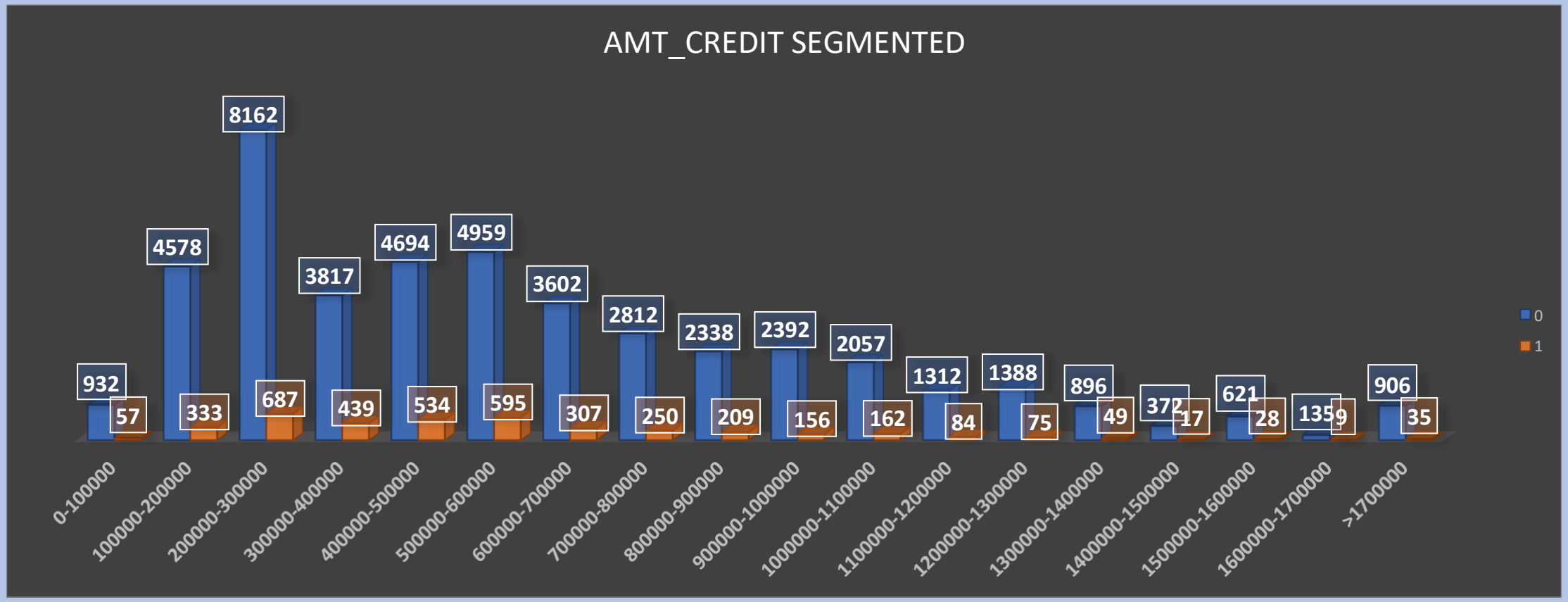


Least default: Client who is Businessman or student or at Maternity leave.

Highest default: Client who is working

Bank Loan Case Study

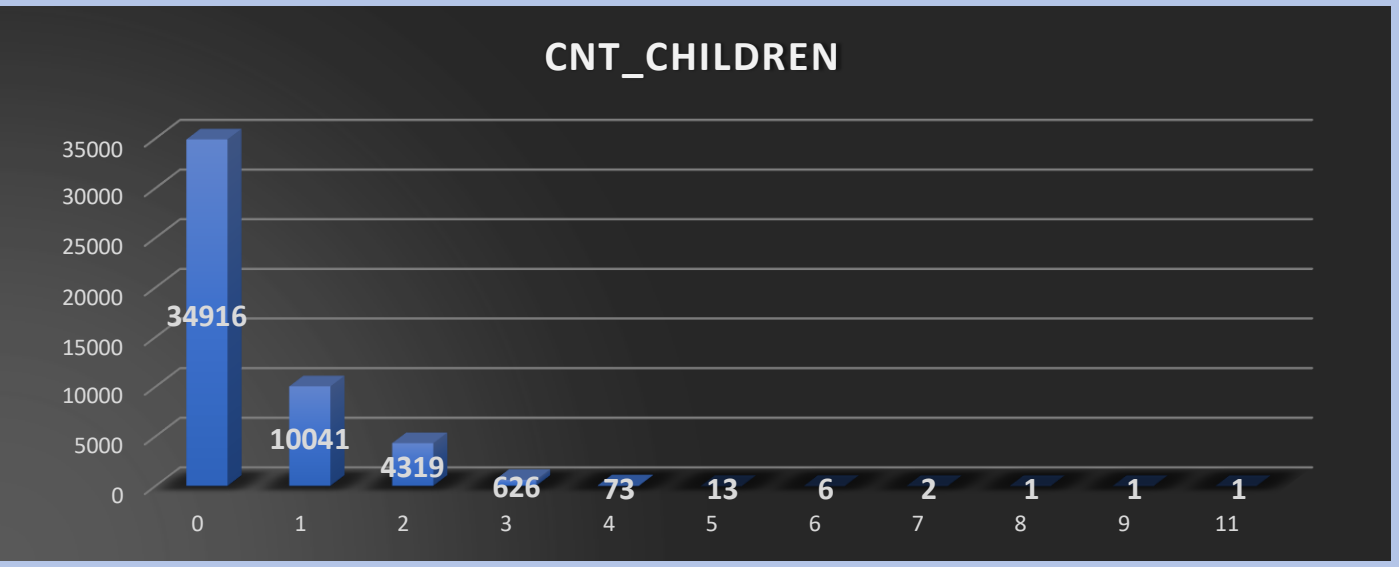
Findings - XII



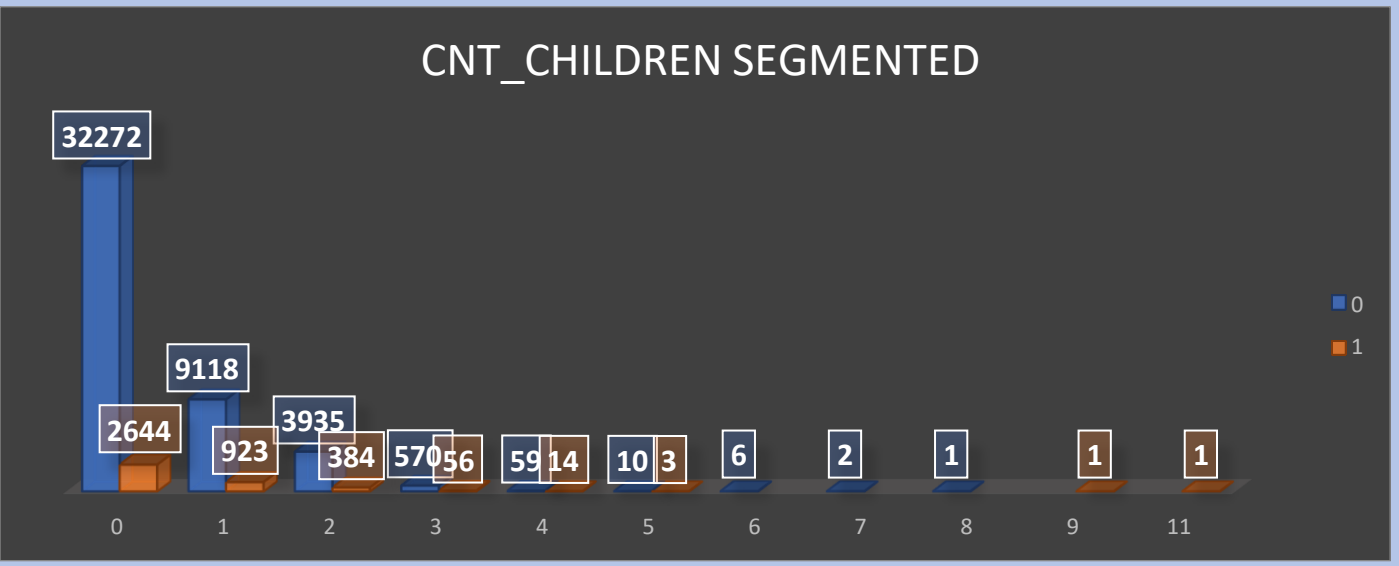
Majority of the Clients took the loan between 2L – 3L.

Bank Loan Case Study

Findings - XIII



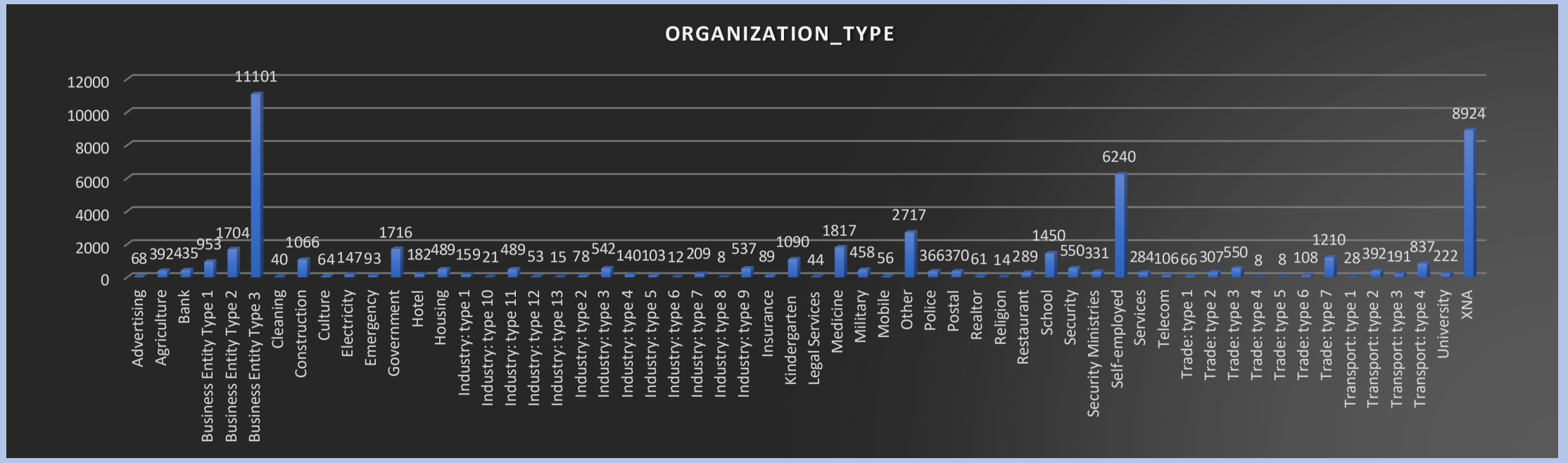
The highest number of loans are taken by Clients who does not have a child.



As number of children increases, number of clients who took loan decreases.

Bank Loan Case Study

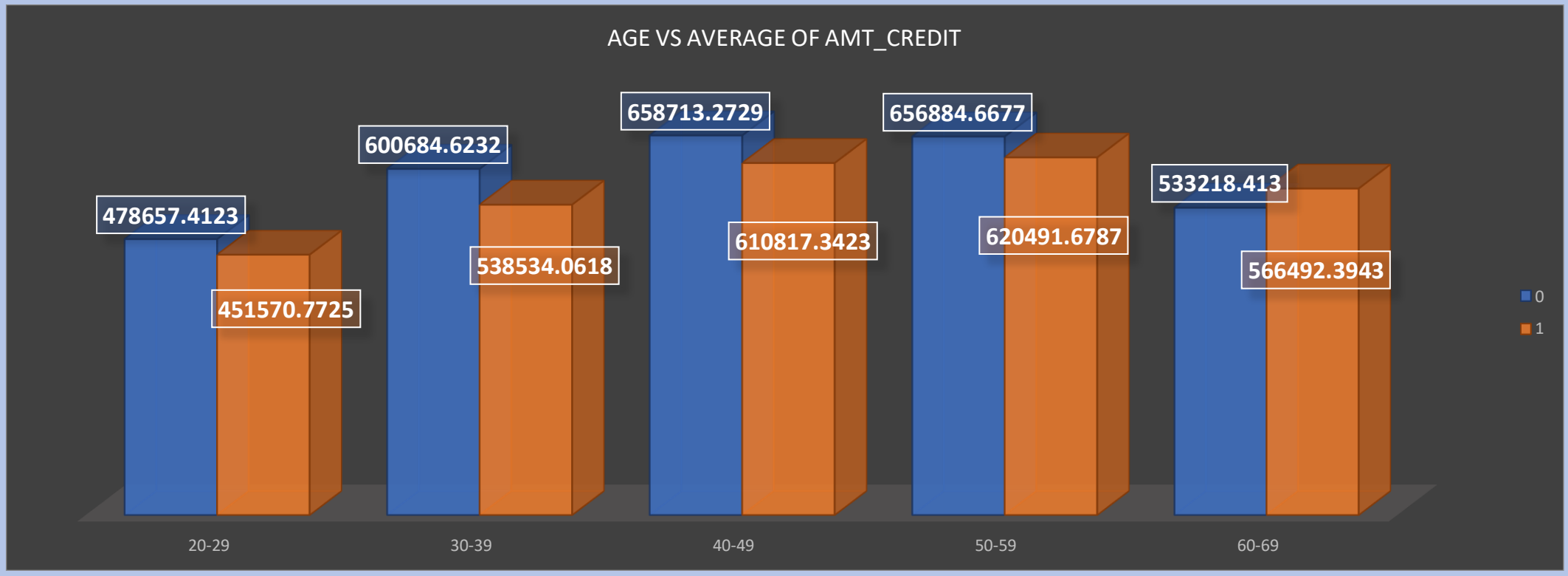
Findings - XIV



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

Bank Loan Case Study

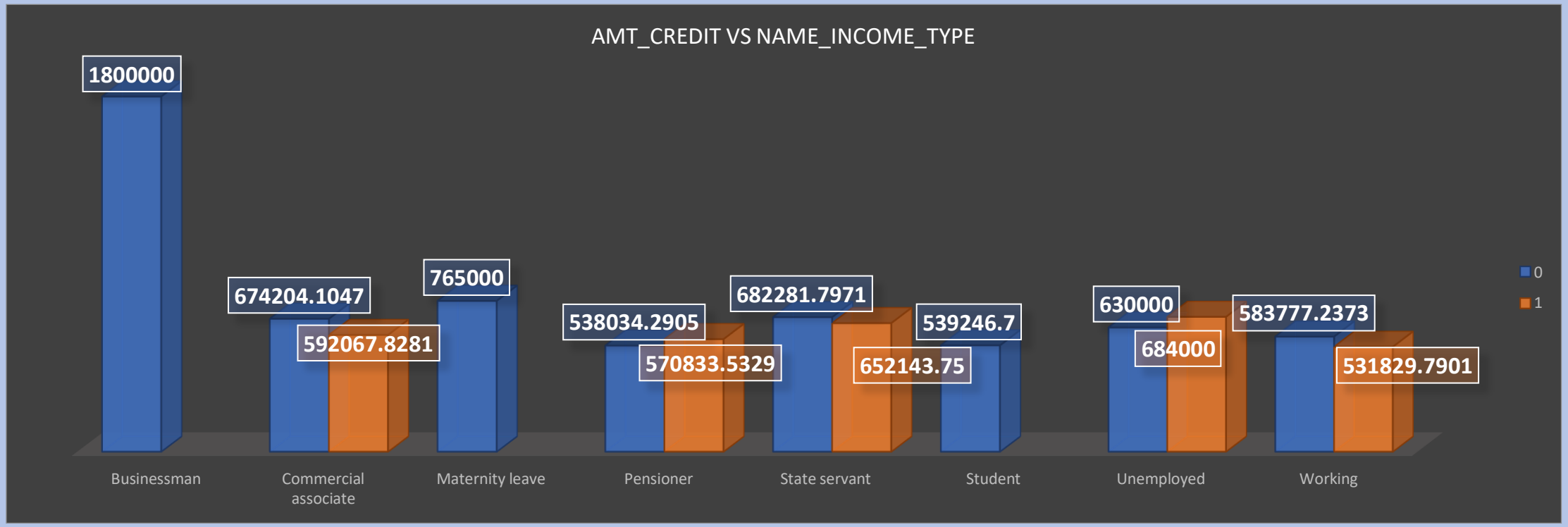
Findings - XV



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

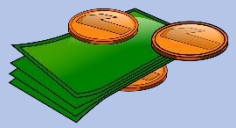
Findings - XVI



As we see in the graph, 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

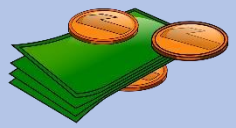
Findings - XVII



Top Correlation Coefficients for Payment difficulties are: -

Correlation between Columns	Value
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
	0.74950403

Bank Loan Case Study



Findings - XVIII

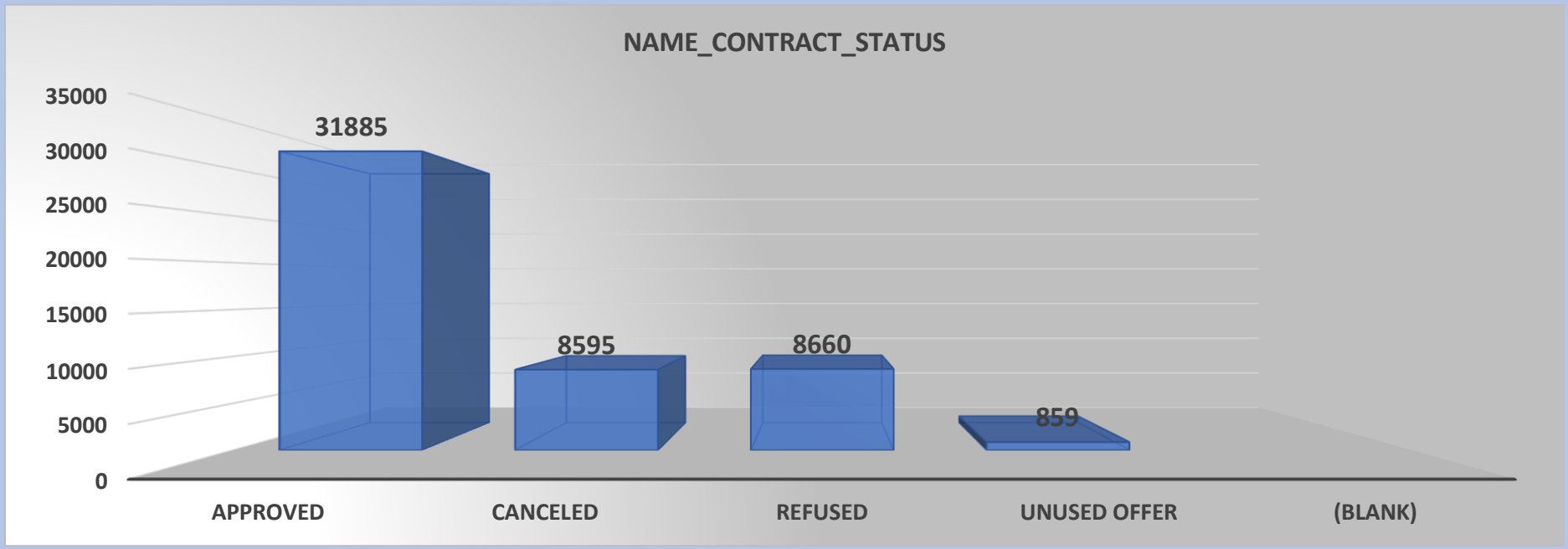
Top Correlation Coefficients for Re-payers are: -

Correlation between Columns	Value
OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE	0.998357563
AMT_GOODS_PRICE - AMT_CREDIT	0.986051701
LIVE_REGION_NOT_WORK_REGION - REG_REGION_NOT_WORK_REGION	0.861374946
DEF_60_CNT_SOCIAL_CIRCLE - DEF_30_CNT_SOCIAL_CIRCLE	0.850995792
REG_CITY_NOT_WORK_CITY - LIVE_CITY_NOT_WORK_CITY	0.825358079
AMT_ANNUITY - AMT_GOODS_PRICE	0.774006842
AMT_ANNUITY - AMT_CREDIT	0.770772818

Bank Loan Case Study

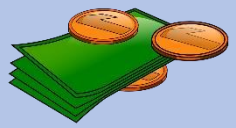
Findings - XIX

Previous_application datasets



Most number of Clients were approved for loans previously.

Bank Loan Case Study



Analysis

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

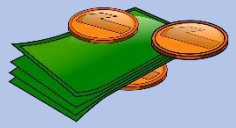
- Why is it that the target_variable is of so much importance?
  ---> In this dataset target_variable represents whether the client had some payment difficulties (1) or the client didn't had some payment difficulties (0); It is important because the target_variable decides whether the bank should increase/decrease 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.

Analyzing the Impact of Car Features on Price and Profitability

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.

Analyzing the Impact of Car Features on Price and Profitability

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.
  Task 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.

Analyzing the Impact of Car Features on Price and Profitability

The Problem

- 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.

Analyzing the Impact of Car Features on Price and Profitability

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?

Analyzing the Impact of Car Features on Price and Profitability

Design

Before starting the actual analysis, I have: -

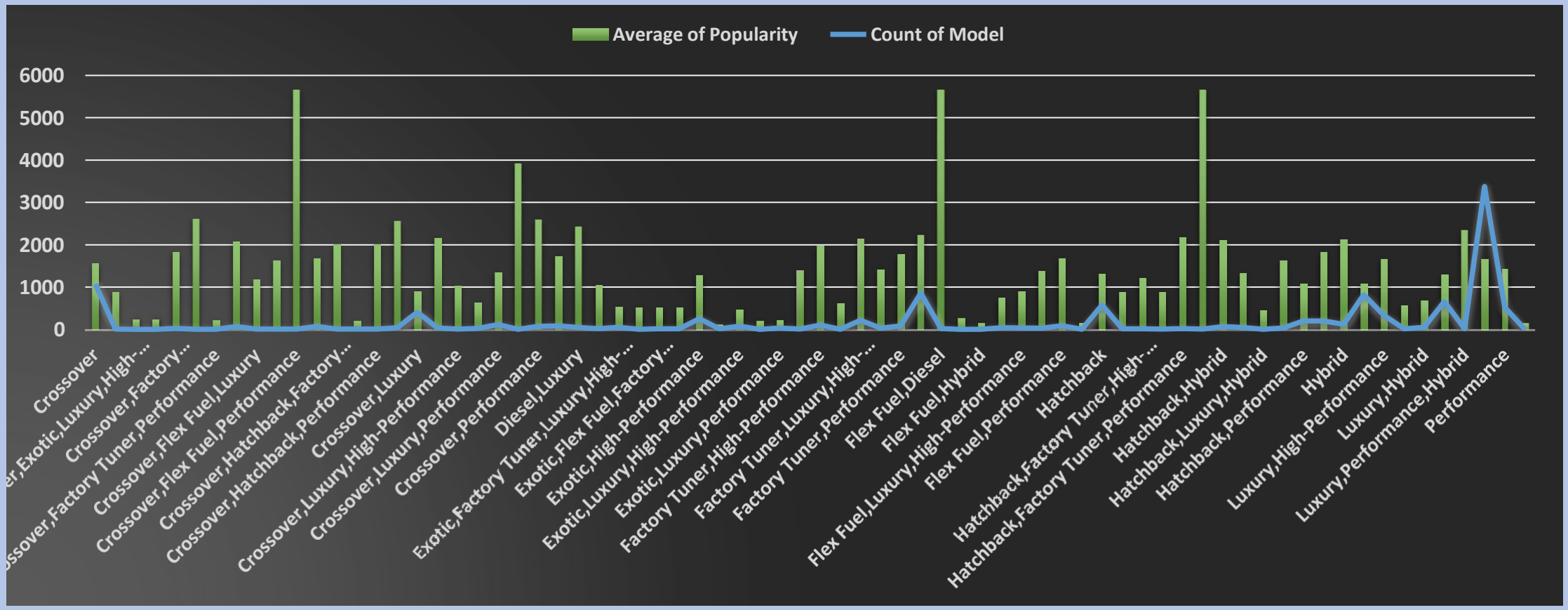
- First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.
- Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.
- I removed rows having blank spaces and NULL values.
- Then I removed duplicate rows from the datasets.

Software used for doing the overall Analysis: -

----> Microsoft Excel

Analyzing the Impact of Car Features on Price and Profitability

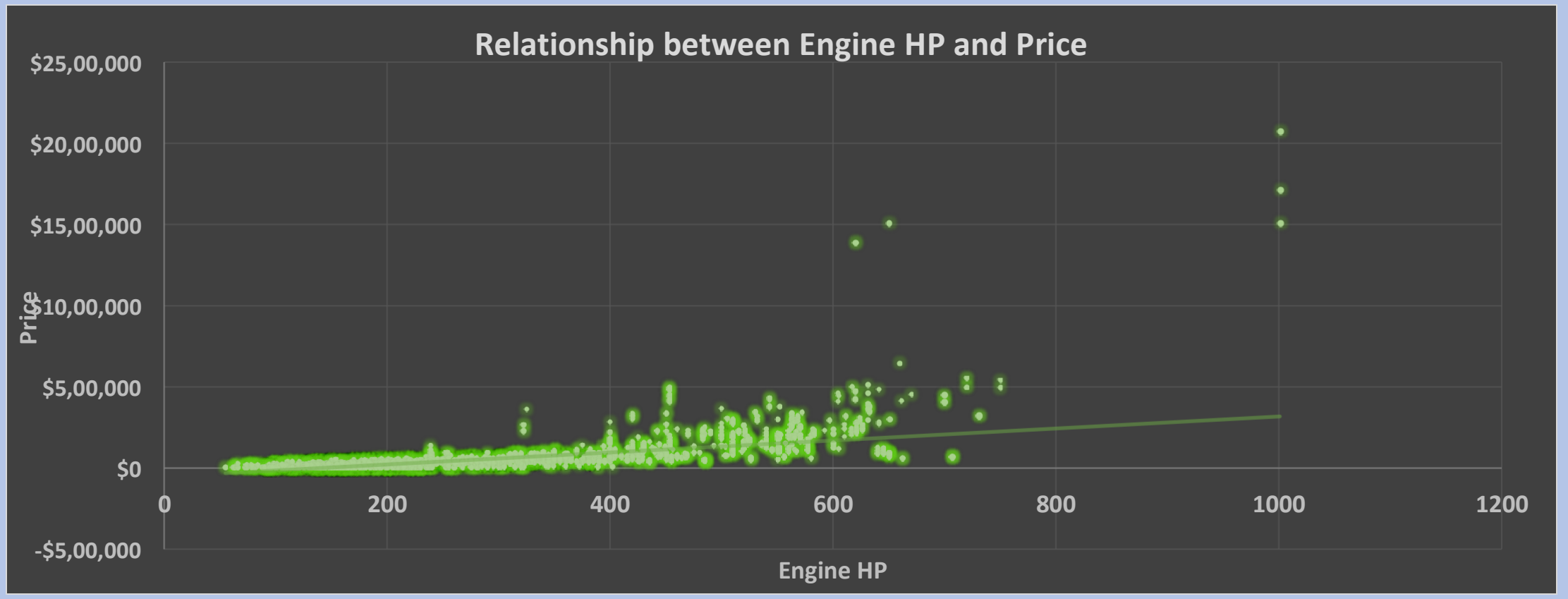
Findings – I



Insights: - Flex Fuel, Diesel, Hatchback, Crossover, Performance are the most popular market category for car models.

Analyzing the Impact of Car Features on Price and Profitability

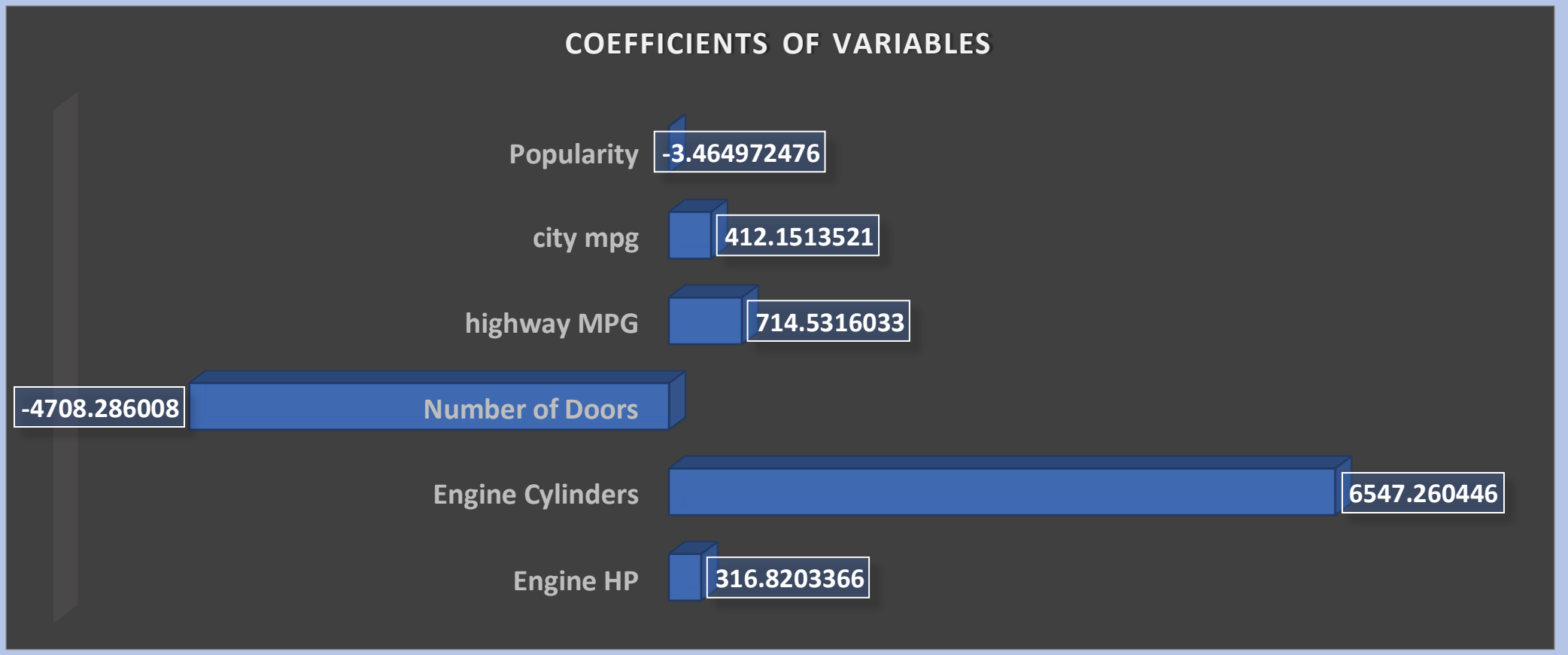
Findings - II



Insights: - If Engine power increases Price will also increase. So, it's positive relationship between both of them.

Analyzing the Impact of Car Features on Price and Profitability

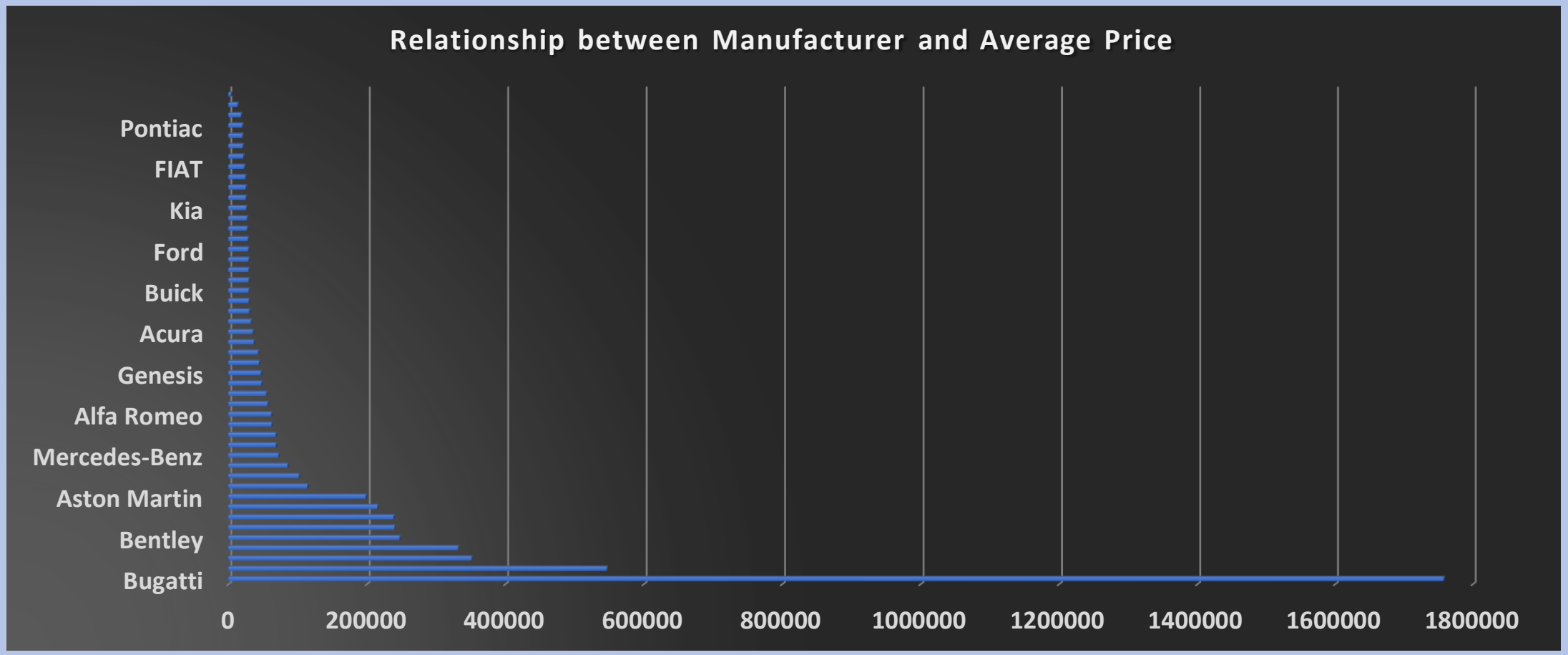
Findings - III



Insights: - Engine Cylinders are the most important features in determining a car's price.

Analyzing the Impact of Car Features on Price and Profitability

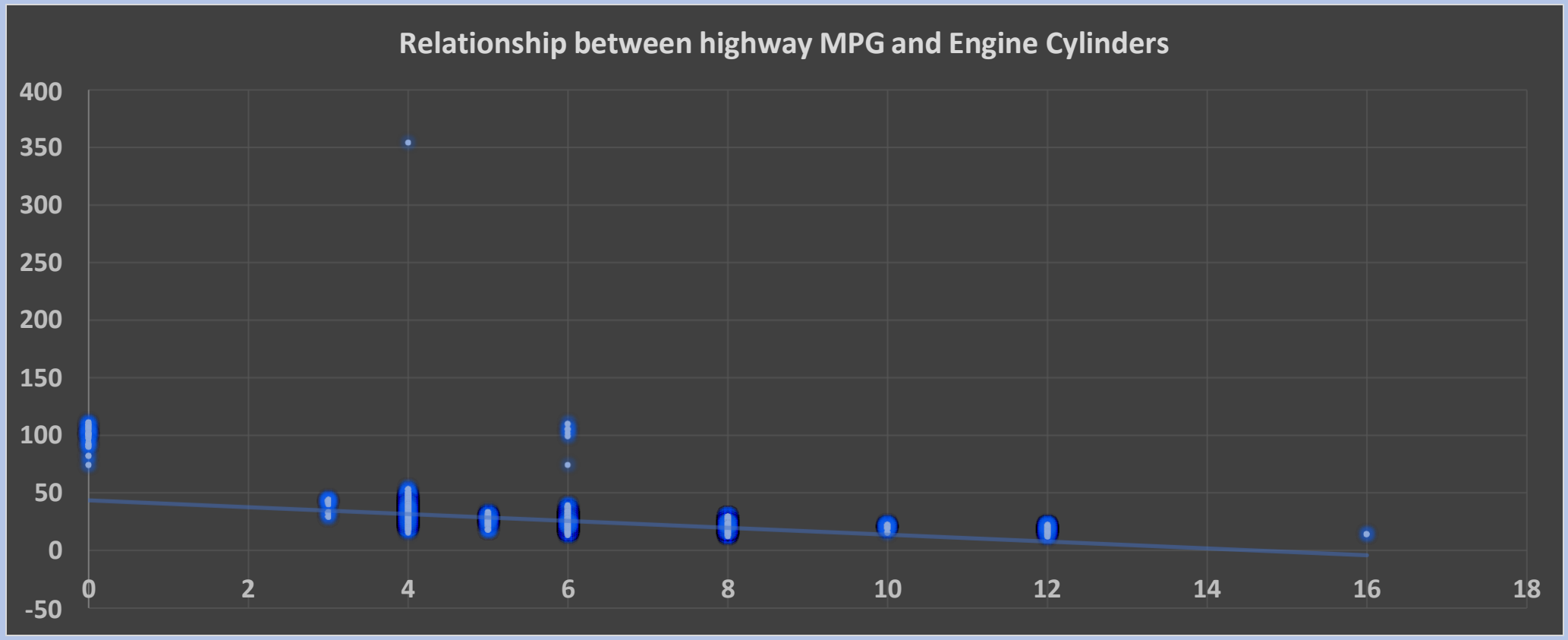
Findings - IV



Insights: - Bugatti has the highest Average price and Plymouth has the lowest average price.

Analyzing the Impact of Car Features on Price and Profitability

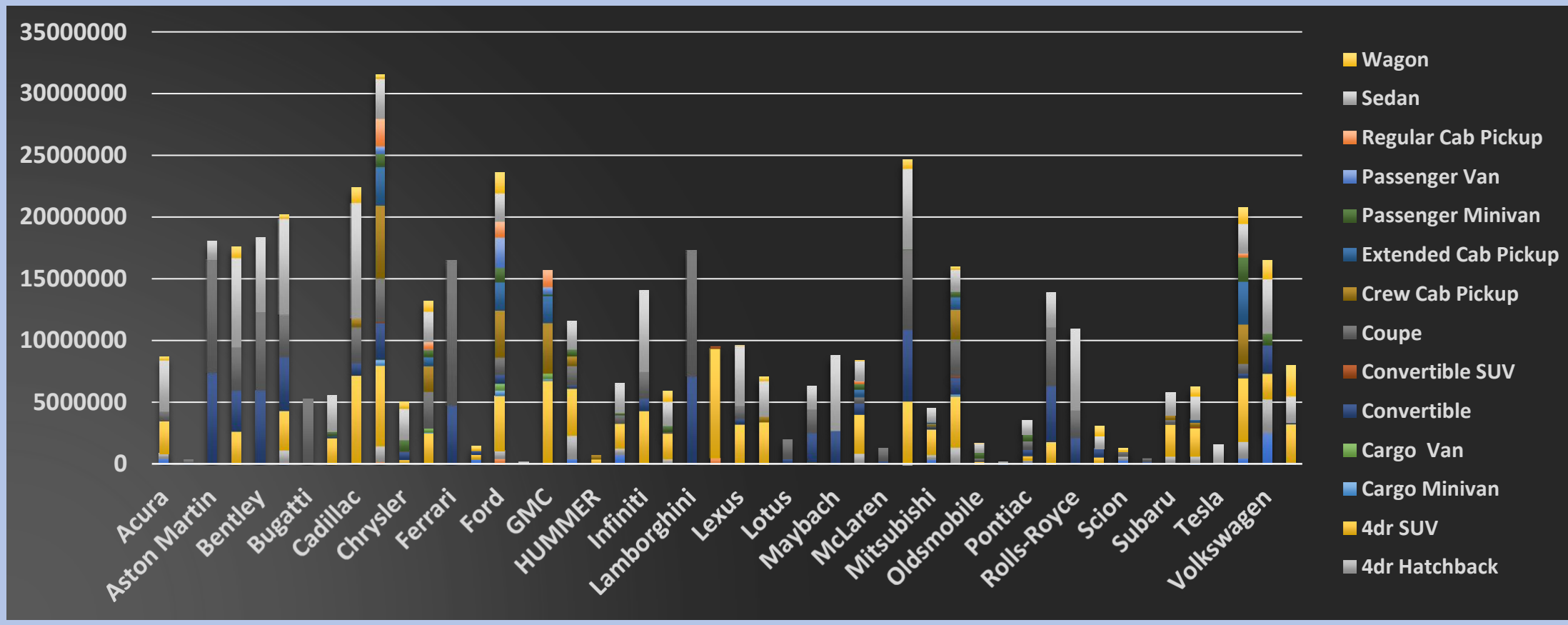
Findings – V



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

Analyzing the Impact of Car Features on Price and Profitability

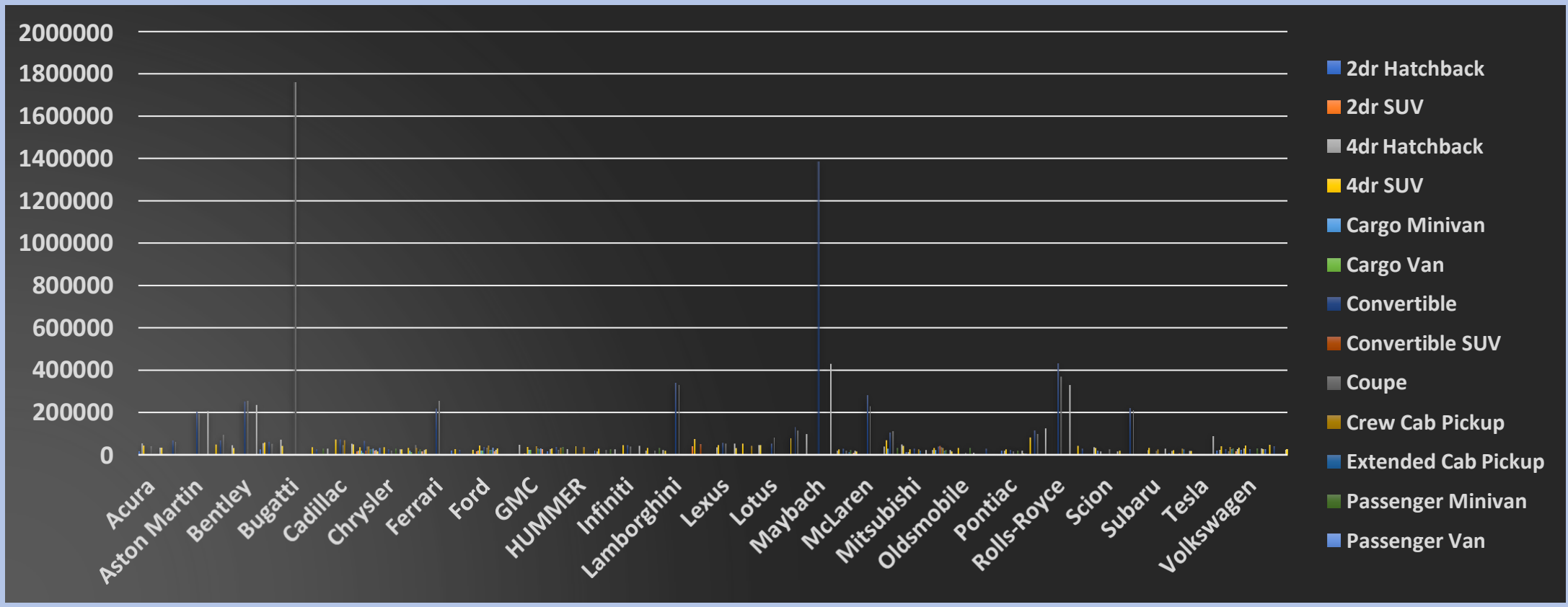
Findings - VI



Insights: - Chevrolet has the highest price distribution by body style.

Analyzing the Impact of Car Features on Price and Profitability

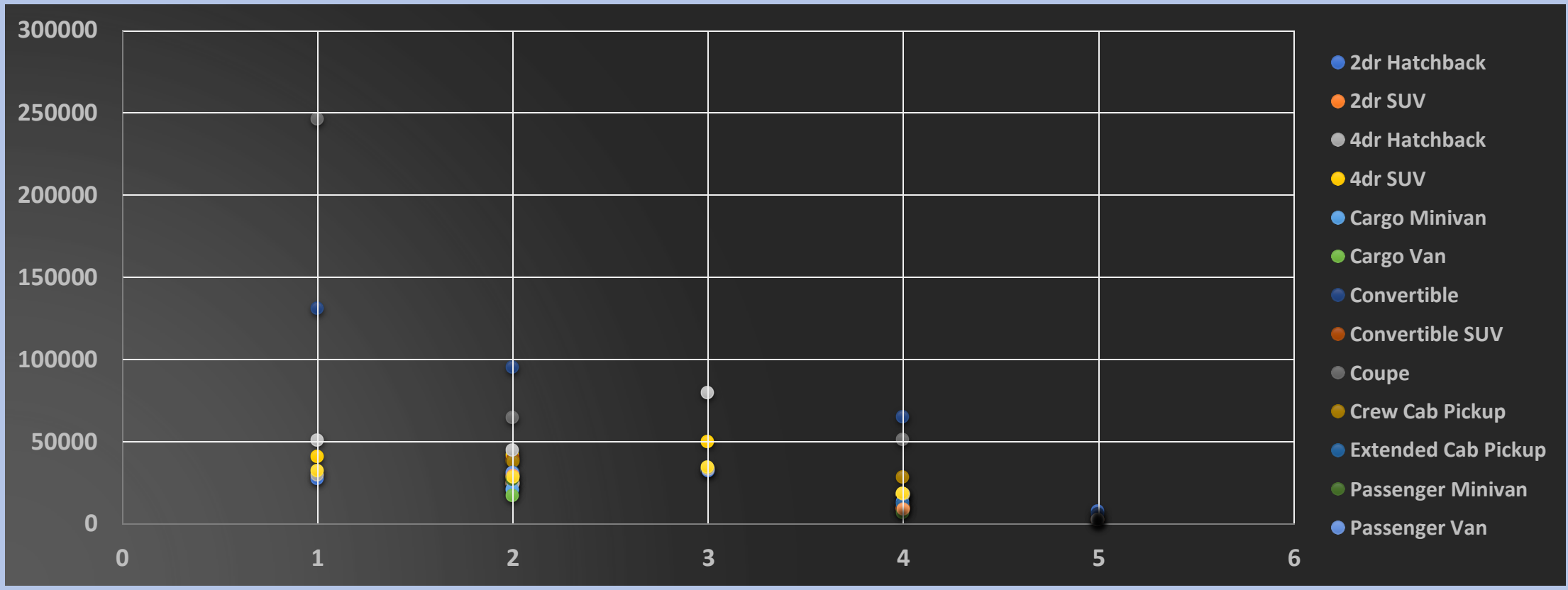
Findings - VII



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

Analyzing the Impact of Car Features on Price and Profitability

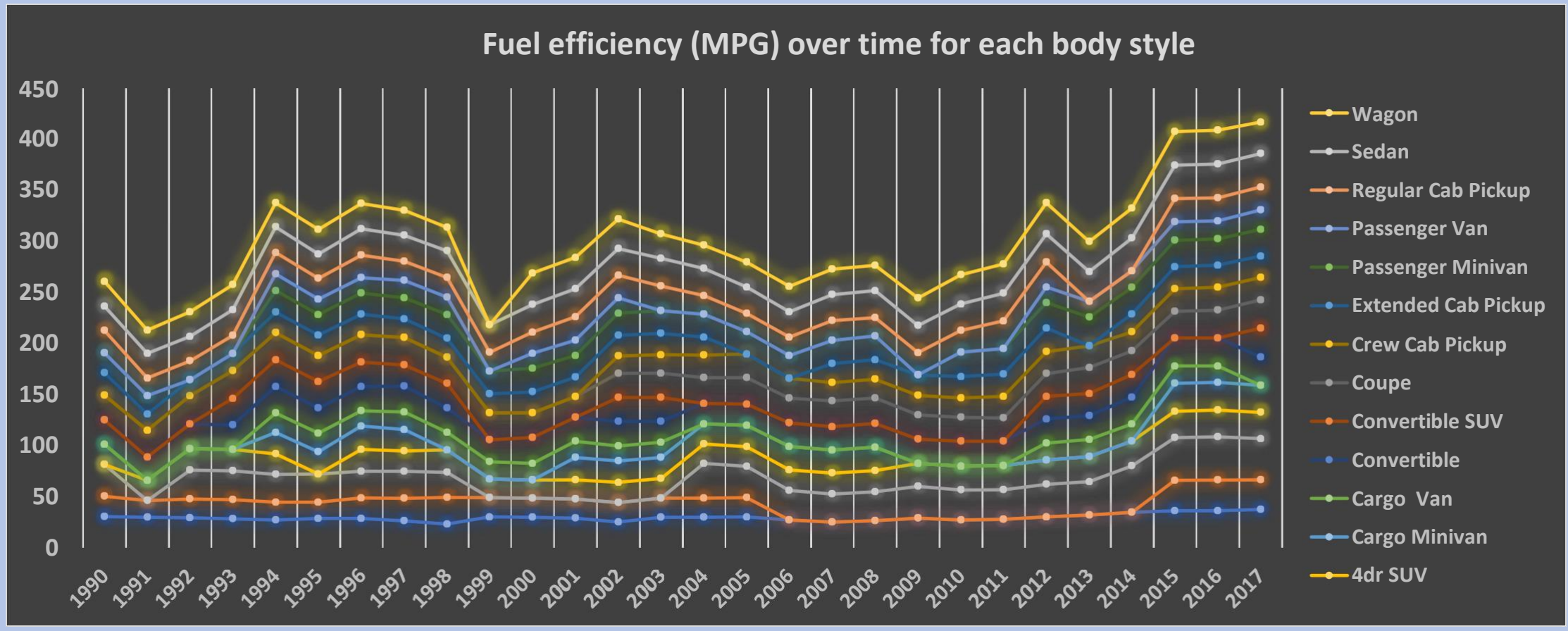
Findings – VIII



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

Analyzing the Impact of Car Features on Price and Profitability

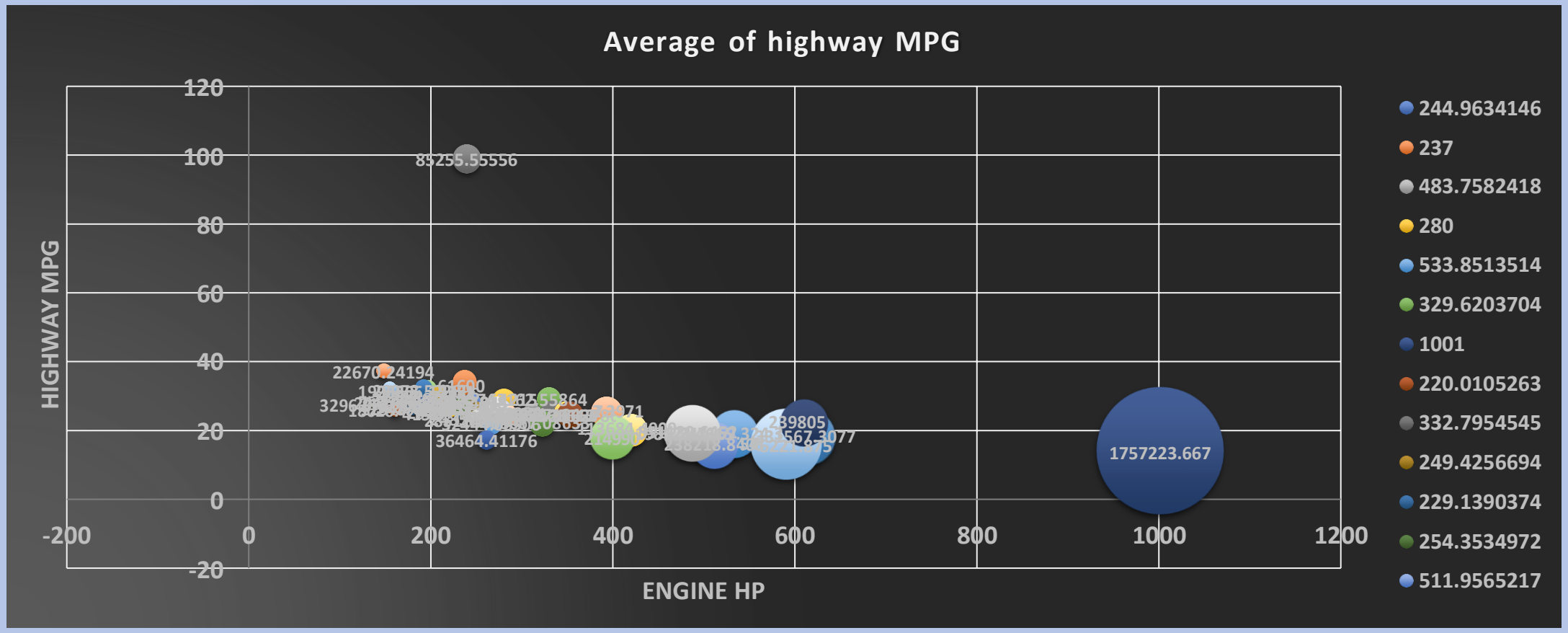
Findings - IX



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

Analyzing the Impact of Car Features on Price and Profitability

Findings - X



Insights: - If Engine HP goes up, Highway MPG goes down but price increases.

Analyzing the Impact of Car Features on Price and Profitability

Conclusion

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

- Flex Fuel, Diesel, Hatchback, Crossover, Performance is the most popular market category for car models.
- If Engine power increases Price will also increase.
- Engine Cylinders are the most important features in determining a car's price.
- Bugatti has the highest Average price and Plymouth has the lowest average price.
- Number of Cylinders will increase then highway MPG will decrease.
- Chevrolet has the highest price distribution by body style.
- AUTOMATED_MANUAL with Coupe body style is the most expensive transmission.
- Wagon body style has the highest fuel efficiency.
- If Engine HP goes up, Highway MPG goes down but price increases.

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).

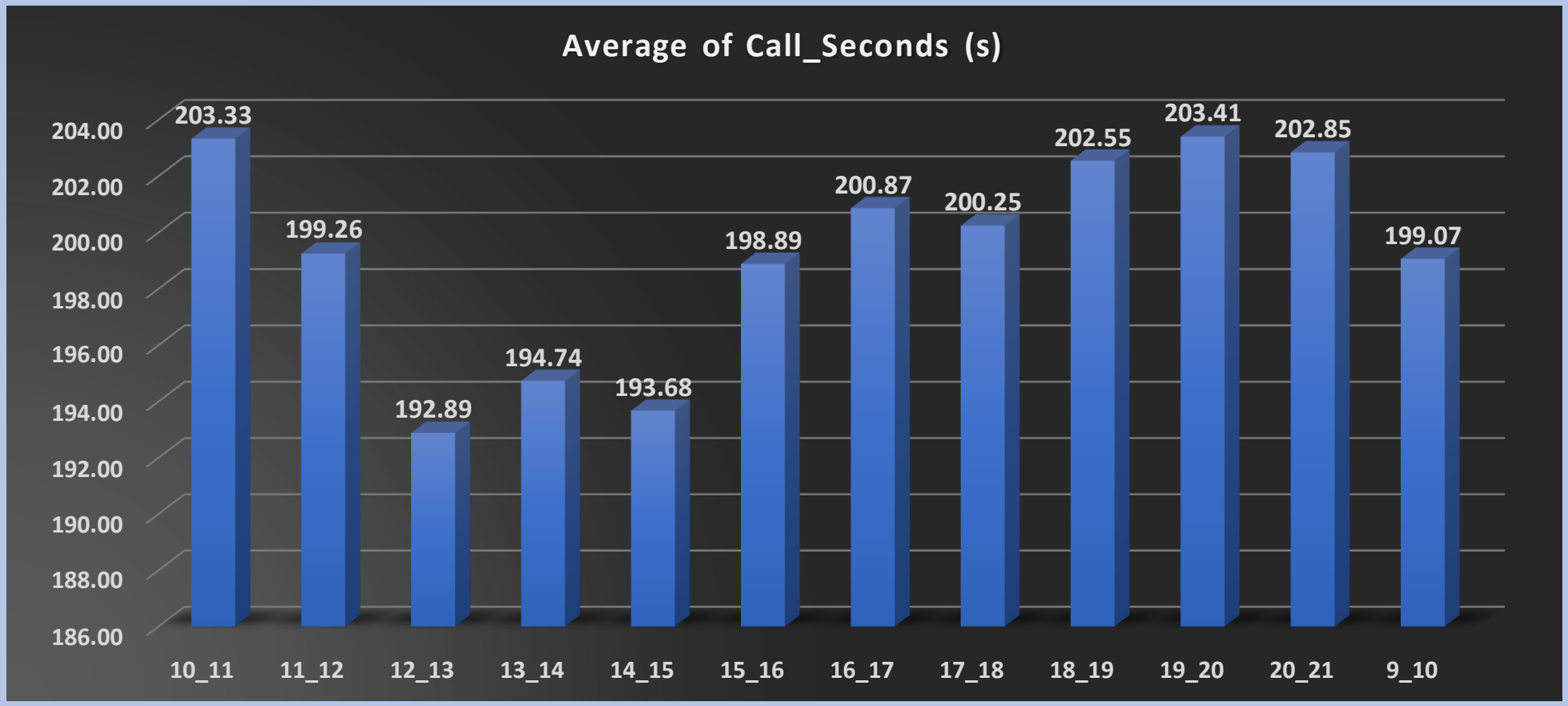
ABC Call Volume Trend Analysis

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%.

ABC Call Volume Trend Analysis

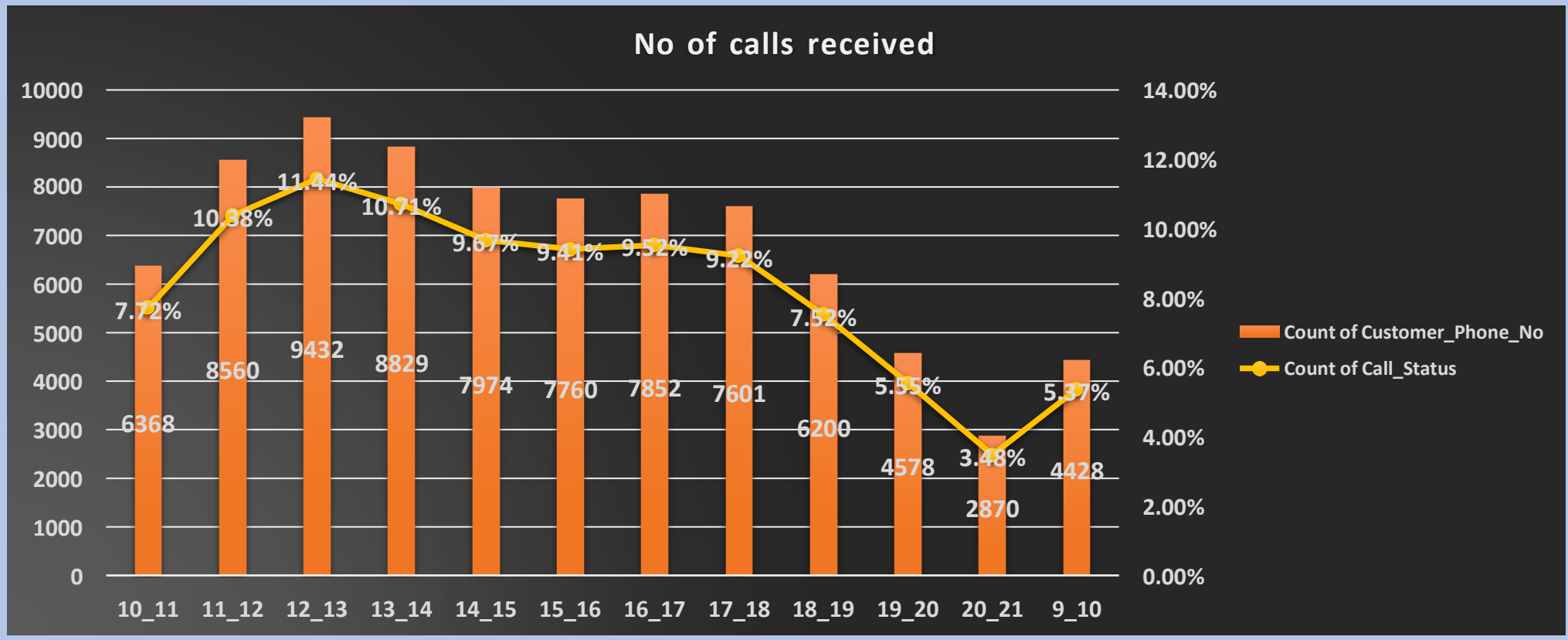
Findings - I



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

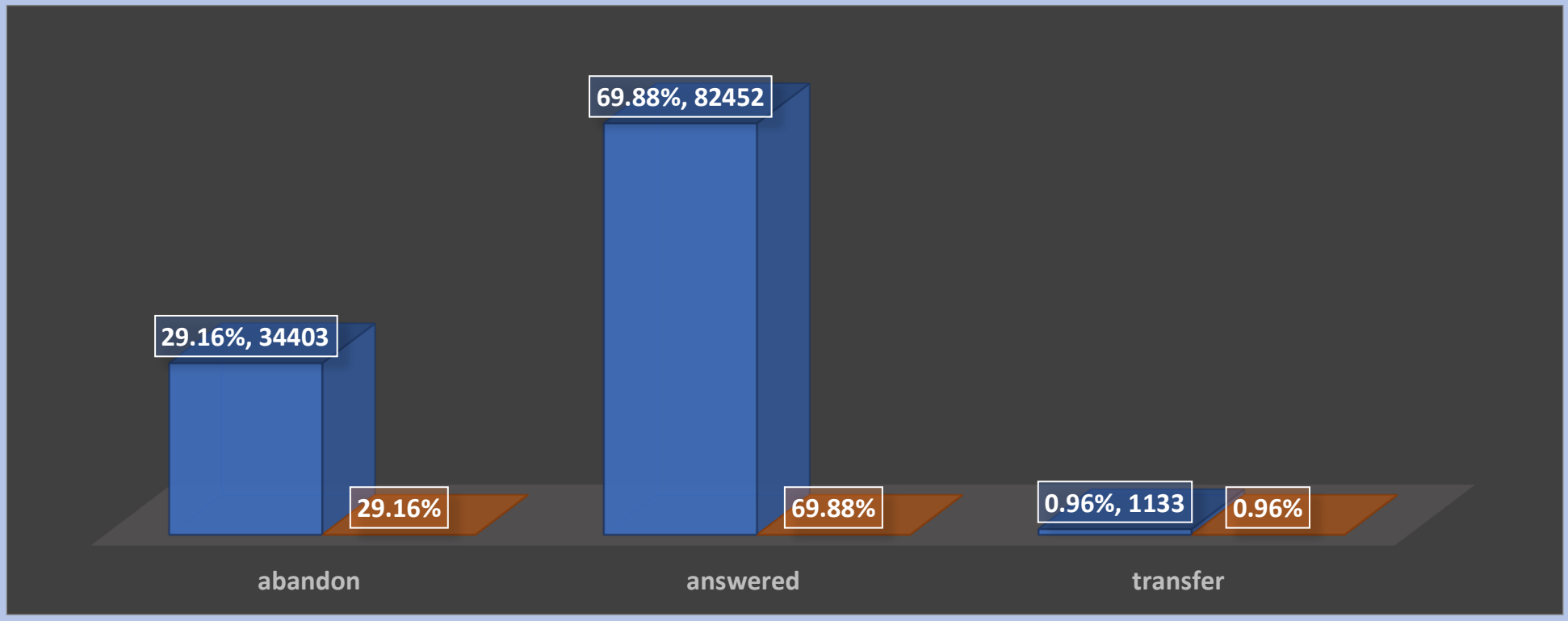
Findings - II



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

ABC Call Volume Trend Analysis

Findings - III



We can see in graph almost 30% calls are getting abandoned.

ABC Call Volume Trend Analysis

Findings - III

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.

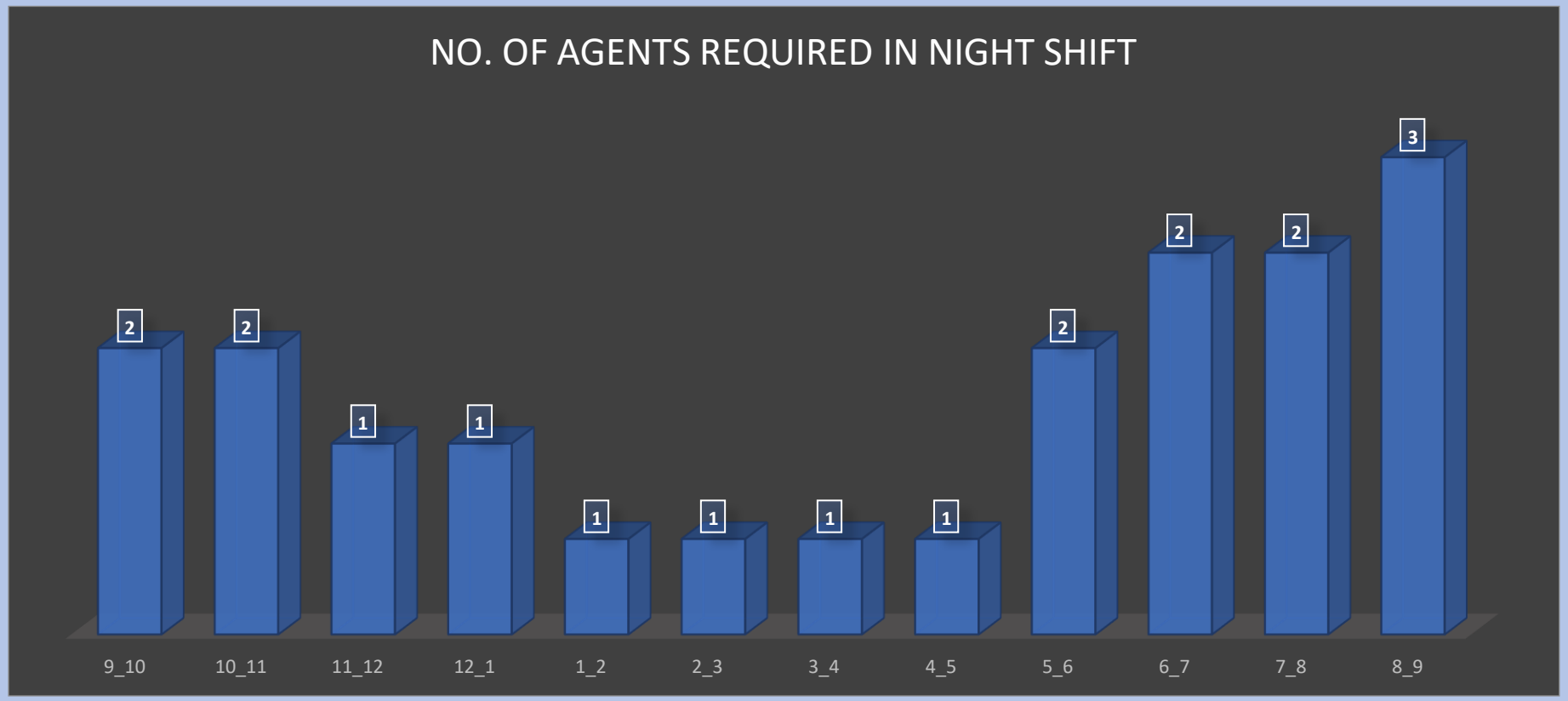
ABC Call Volume Trend Analysis

Findings - IV

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**

ABC Call Volume Trend Analysis

Findings - IV



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

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

- Company can divide workforce into three shifts to ensure 24/7 availability for addressing customer's queries and concerns.
- Total average call duration answered by agents is 198.62 seconds.
- Further analysis reveals that maximum average duration of calls for incoming calls is at 10_11 AM and 7_8 PM.
- Based on analysis minimum average call duration for incoming calls received by agents is at 12_1 PM.
- Based on analysis highest number of calls received is between 12 PM and 1 PM.
- Further analysis also revealed that least number of calls answered is between 8 PM and 9 PM.
- Based on analysis incoming calls in evening are less. So, Company can optimize workforce by reducing the number of agents in evening for call handling.
- Total no. of agents required to reduce the abandon rate to 10% is 54.
- Company can hire 17 agents who will be available during night hours from 9 PM to 9 AM to handle the calls or shift some of the day workers to the night shift.
- These insights provide the company with actionable strategies for optimizing workforce allocation, enhancing customer service efficiency, and ensuring continuous availability to address customer needs.

Appendix

Data Analytics Process: -

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

Data Analytics process

Instagram User Analytics: -

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

Instagram User Analytics

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

IMDB Movie Analysis-

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

Data Analytics Trainee Task - 5.pdf - Google Drive

Appendix

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