Data Analysis Portfolio

(Sourabh Kumar)

Professional Background

I am highly skilled data analytics professional with a Bachelor's degree in Computer Applications (BCA) from Birla Institute of Technology and currently pursuing a Master's degree in MSc in Data Science from Manipal Academy of Higher Education. My academic journey has provided me with a solid foundation in data analysis, machine learning, and data visualization.

To complement my academic achievements, I hold a Google Data Analytics Professional Certificate and have specialized in Data Visualization with Tableau. Additionally, I have received training in machine learning with Python, equipping me with the necessary knowledge and skills to deliver actionable insights that drive business success.

During my virtual internships with KPMG and TATA through the Forage platform, I gained valuable hands-on experience in data analytics consulting and data visualization. Leveraging my proficiency in Excel and Tableau, I successfully analyzed data, created visualizations, and delivered impactful insights to clients.

I have also undertaken various projects that showcase my ability to handle complex data sets and deliver insights that generate business value. For instance, in my Bellabeat case study, I analyzed smart device usage data to extract valuable insights into user behaviour. Furthermore, I conducted an analysis of the Superstore data set, identifying key indicators and relationships that enabled the company to optimize its operations and increase profits.

With a combination of technical skills, an analytical mindset, and a strong passion for utilizing data to drive business success, I am confident in my ability to deliver exceptional results as a data analytics professional. Whether working independently or as part of a team, my commitment lies in providing actionable insights that empower decision-makers and enable companies to thrive in today's data-driven world.

Table of Contents :

Professional Background ---------- 1

Table of Contents ---------- 2-3

Data Analytics Process

- Description ---------- 4
- Findings ---------- 4
- Conclusions ---------- 4

Instagram User Analytics

- Description ---------- 5
- Problem ---------- 5
- Design ---------- 5
- Findings ---------- 6-10
- Analysis ---------- 10-11
- Conclusions ---------- 11

Operation Analytics and Investigating Metric Spike

- Description ---------- 12
- Problem ---------- 12-13
- Findings ---------- 13-18
- Analysis ---------- 19-20
- Conclusions ---------- 20

Hiring Process Analytics

- Description ---------- 21
- Problem ---------- 21
- Findings ---------- 21-25
- Analysis ---------- 25-26
- Conclusions ---------- 26

IMDB Movies Analysis

- Description ---------- 27
- Problem ---------- 27-28
- Findings ---------- 28-33
- Analysis ---------- 33-34
- Conclusions ---------- 34

Bank Loan Case Study

Description ---------- 35

The Problem ---------- 35-36

Findings ---------- 36-51

Analysis ---------- 52

Conclusions ---------- 53

Impact of Car feature

Description ---------- 54

The Problem ---------- 54-55

Findings ---------- 55-61

Conclusions ---------- 61-62

ABC Call Volume Trend

Description ---------- 63

The Problem ---------- 63-64

Findings ---------- 64-68

Analysis ---------- 68-69

Conclusions ---------- 69-70

Data Analytics Process

Description:

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

Findings:

In this case study, I will perform real-world tasks of a data analyst. I will perform data analysis for a company who sells smart watches. I divide the data analysis process in different phase.

Ask Phase: I am focus on a smart watch company and analyze their device usage data in order to gain insight into how people are using their smart devices. Then, using this information, I would like to provide highlevel recommendations for marketing strategy.

Prepare Phase: The data set contains personal fitness tracker from users. Users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

Process Phase: In this phase, I need to check how much I want data from the data set. And the tools used for manipulation and processing of data sets. Apart from that, I have to ensured data integrity & data is clean to use for analysis.

Analyze Phase: In this phase, Analyzing the data to get insights how the users using their smart watches. Create multiple sheets for finding a trends by checking correlation between multiple variables and present it on dashboard.

Share Phase: In this phase, I share the dashboard to the manager and also provides report of how I analyze their data to find the outcomes.

Act Phase: Finally, company uses that outcomes to increase their sales as well as provides latest technology which a user's want in his smart watches

Conclusions:

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

- Ask
- Prepare
- Process
- Analyze
- Share
- Act

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

Instagram User Analytics

Description:

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

You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

The Problem:

A) Marketing: The marketing team wants to launch some campaigns, and they need your help with the following:

- Rewarding Most Loyal Users: People who have been using the platform for the longest time. Your Task: Find the 5 oldest users of the Instagram from the database provided Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo. Your Task: Find the users who have never posted a single photo on Instagram.
- Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner. Your Task: Identify the winner of the contest and provide their details to the team.
- Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform. Your Task: Identify and suggest the top 5 most commonly used hashtags on the platform
- Launch AD Campaign: The team wants to know, which day would be the best day to launch ADs. Your Task: What day of the week do most users register on? Provide insights on when to schedule an ad campaign.

B) Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds.

- User Engagement: Are users still as active and post on Instagram or they are making fewer posts Your Task: Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users
- Bots & Fake Accounts: The investors want to know if the platform is crowded with fake and dummy accounts Your Task: Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).

Findings:

Steps taken to load the data into the data base

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

Findings - I

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

1. We will use the data from the users table by selecting the username and created_at columns.

2. Then using the order by function, we will order the desired output by sorting with the created_at column in ascending order.

3. Then using the limit function, the output will be displayed for top 5 oldest Instagram users.

Output:

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

Findings - II

To Find the most inactive users i.e., the users who have never posted a single photo on Instagram:

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

Output:

username	user_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

Findings - III

To find the most the username, photo_id, image_url and total_number_of_likes of that image:

1. First we will select the users.username, photos.id, photos.image_url and count(*) as total.
2. Then, we will inner join the three tables wiz : photos, likes and users, on likes.photo_id = photos.id and photos.user_id = users.id
3. Then, by using group by function we will group the output on the basis of photos.id
4. Then, using order by function we will sorting the data on the basis of the total in descending order
5. Then, to find the most liked photo we will using limit function to view only the top liked photo's information

Output:

user_id	username	photo_id	image_url	total
52	Zack_Kemmer93	145	https://jarret.name	48

Findings - IV

To find the top 5 most commonly used hashtags on Instagram:

1. We need to select the tag_name column from the tag table and the count(*) as total function so as to count the number of tags used individually.

2. Then, we need to join tags table and photo_tags table, on tags.id = photo_tags.tag_id cause they contain the same content in them i.e. tag_id

3. Then using the group by function we need to group the desired output on the basis of tags.tag_name

4. Then using the order by function we need to sort the output on the basis of total (total number of tags per tag_name) in descending order

5. Then, to find the top 5 most used tag names we will use the limit 5 function.

Output:

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

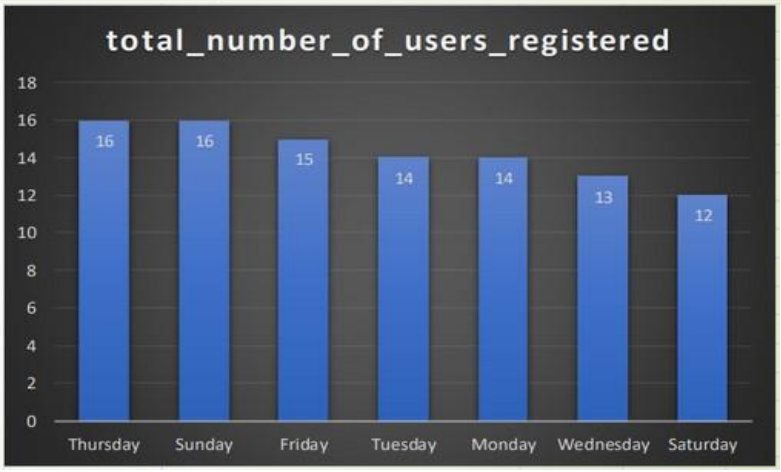
Findings - V

To find the day of week on which most users register on Instagram:

- First we define the columns of the desired output table using select dayname (created_at) as day_of_week and count(*) as total_number_of_users_registered from the users table
- Then using the group by function we group the output table on the basis of day_of_week
- Then using the order by function we order/sort the output table on the basis of total_number_of_users_registered in descending order

Output:

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



Findings - VI

To find the how many times does average posts on Instagram:

1. First, we need to find first the count number of photos(posts) that are present in the photos.id column of the photos table i.e. count(*) from photos
2. Similarly, we need to find the number of users that are present in the users.id column of the users table i.e. count(*) from users
3. Next, we need to divide both the values i.e. count(*) from photos/count(*) from users and hence we would get the total number of photos / total number of users
4. To find how many times the users posts on Instagram we need to find the total occurrences of each user_id in photos table

Output:

total_photos_divide_total_photos
2.57

user_id	user_post_count
1	5
2	4
3	4
4	3
6	5
8	4
9	4
10	3
11	5
12	4
13	5
15	4
16	4
17	3
18	1
19	2
20	1
22	1
23	12
26	5
27	1
28	4
29	8
30	2
31	1
32	4
33	5
35	2
37	1
38	2
39	1
40	1
42	3
43	5
44	4
46	4
47	5
48	1
50	3
51	5
52	5
55	1
56	1
58	8
59	10
60	2
61	1
62	2
63	4
64	5
65	5
67	3
69	1
70	1
72	5
73	1
77	6
78	5
79	1
82	2
84	2
85	2
86	9
87	4
88	11
92	3
93	2
94	1
95	2
96	3
97	2
98	1
99	3
100	2

Findings - VII

To find the bots and fake accounts :

1.First, we select the user_id column from the photos table

2.Then we select the username column from the users table

3.Then, we select the count(*) function to count total number of likes from the likes table

4.Then we inner join users and likes table on the basis of users.id and likes.user_id, using the on function/clause

5.Then by using the group by function we group the desired output table on the basis of likes.user_id

6.Then, we search for the values from the cout(*) from photos having equal values with the total_likes_per_user

Output:

user_id	username	total_likes_per_user
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

Analysis:

After performing the analysis I have the following points:

- The most loyal users i.e. the 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

- Out of the 100 total users there are 26 users who are inactive and 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 uers
- So, the user named Zack_Kemmer93 with user_id 52 is the winner of the contest cause his photo with photo_id 145 has the highest number of likes i.e. 48
- 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 and hence it would prove beneficial to start AD Campaign on these two days
- So, there are in total 257 rows i.e. 257 photos in the photos table and 100 rows i.e. 100 ids in the users table which makes the desired output to be 257/100 = 2.57 (avg. users posts on Instagram)
- Out of the total user id's there are 13 such user id's who have liked each and every post on Instagram (which is not practically possible) and so such user id's are considered as BOTS and Fake Accounts

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 What's keeping them away from using the Instagram.

- Why did the Marketing team wanted to know the top 5 #hashtags used?

--> May be the tech 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 most new 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

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 an 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 & Metric Analytics

Description :

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they collect.

Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that its very important to investigate metric spike.

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

Problem :

Case Study 1 (Job Data)

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

Your task: Calculate the number of jobs reviewed per hour per day for November 2020?

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

Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?

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

Your task: Calculate the percentage share of each language in the last 30 days?

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

Your task: Let's say you see some duplicate rows in the data.

Case Study 2 (Investigating metric spike)

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

Your task: Calculate the weekly user engagement?

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

Your task: Calculate the user growth for product?

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

Your task: Calculate the weekly retention of users-sign up cohort?

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

Your task: Calculate the weekly engagement per device?

- Email Engagement: Users engaging with the email service.

Your task: Calculate the email engagement metrics?

Findings:

Steps taken to load the data into the data base

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

Job Data Findings - I

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

1. We will use the data from job_id columns of the job_data table.

2. Then we will divide the total count of job_id (distinct and nondistinct) by (30 days * 24 hours)for finding the number of jobs reviewed per day

Output:

number_of_jobs_reviewed_per_day_non_distinct
0.0111

number_of_jobs_reviewed_per_day_distinct
0.0083

Job Data Findings - II

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

1. We will be first taking the count of job_id(distinct and non-distinct) andordering them w.r.t ds (date of interview)

2. Then by using the ROW function we will be considering the rowsbetween 6 preceding rows and the current row

3. Then we will be taking the average of the jobs_reviewed

Output:

date_of_review	jobs_reviewed	throughput_7_rolling_average
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

date_of_review	jobs_reviewed	throughput_7_rolling_average_non_distinct_job_id
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

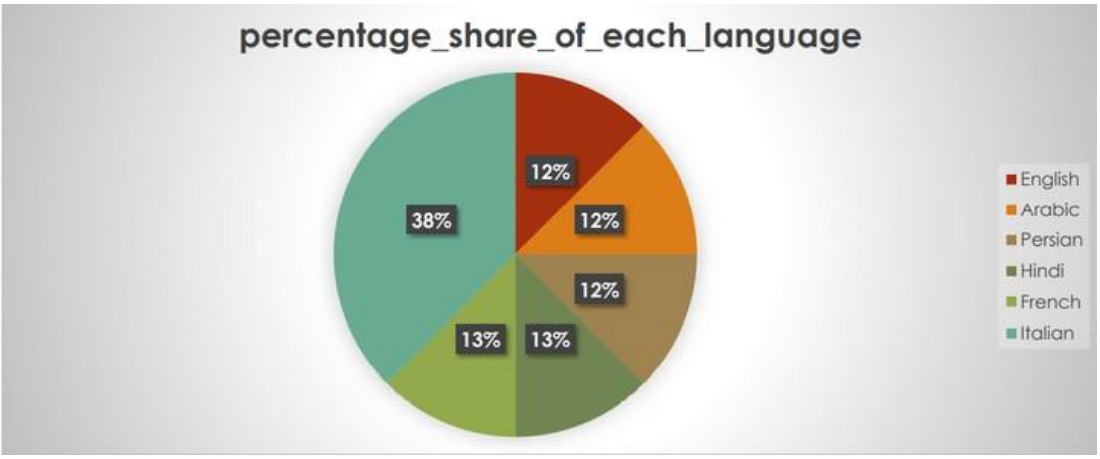
Job Data Findings - III

To Calculate the percentage share of each language (distinct and non-distinct):-

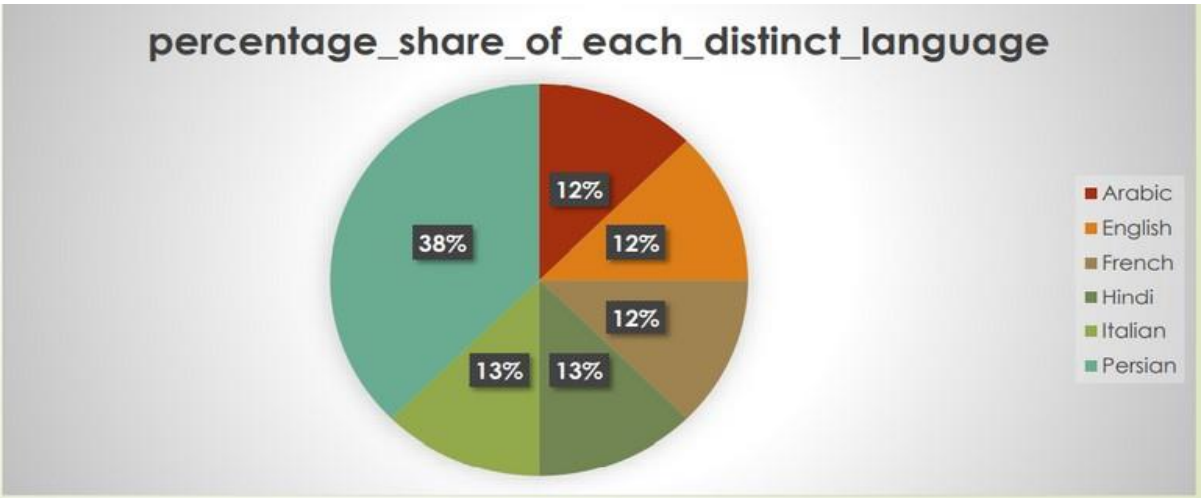
1. We will first divide the total number of languages (distinct/non-distinct) by the total number of rows presents in the table
2. Then we will do the grouping based on the languages.

Output:

job_id	language	total_of_each_language	percentage_share_of_each_language
21	English	1	12.5
22	Arabic	1	12.5
23	Persian	3	37.5
25	Hindi	1	12.5
11	French	1	12.5
20	Italian	1	12.5



job_id	language	total_of_each_language	percentage_share_of_each_distinct_language
22	Arabic	1	12.5
21	English	1	12.5
11	French	1	12.5
25	Hindi	1	12.5
20	Italian	1	12.5
23	Persian	1	37.5



Job Data Findings - IV

To view the duplicate rows having the same value we will:

1. First decide in which do we need to find the duplicate row values

2. After deciding the column(parameter) we will use the ROW_NUMBER function to find the row numbers having the same value

3. Then we will portioning the ROW_NUMBER function over the column (parameter) that we decided i.e. job_id

4. Then using the WHERE function we will find the row_num having value greater than 1 i.e. row_num > 1 based on the occurrence of the job_id in the table

Output:

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

Investigating Metric Spike Findings - I

To find the weekly user engagement:-

1. We will extract the week from the occurred_at column of the events table using the EXTRACT function and WEEK function

2. Then we will be counting the number of distinct user_id from the events table

3.Then we will use the GROUP BY function to group the output w.r.t week from occurred_at

Output:

week_number	number_of_users
18	791
19	1244
20	1270
21	1341
22	1293
23	1366
24	1434
25	1462
26	1443
27	1477
28	1556
29	1556
30	1593
31	1685
32	1483
33	1438
34	1412
35	1442

Investigating Metric Spike Findings - II

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

1. First we will the extract the year and week for the occurred_at column of the users table using the extract, year and week functions

2. Then we will group the extracted week and year on the basis of year and week number

3. Then we ordered the result on the basis of year and week number

4.Then we will find the cumm_active_users using the SUM, OVER and ROW function between unbounded preceding and current row

Output:

year_num	week_num	num_active_users	cum_active_users	year_num	week_num	num_active_users	cum_active_users
2013	1	67	67	2013	45	97	2564
2013	2	29	96	2013	46	94	2658
2013	3	47	143	2013	47	82	2740
2013	4	36	179	2013	48	103	2843
2013	5	30	209	2013	49	96	2939
2013	6	48	257	2013	50	117	3056
2013	7	41	298	2013	51	123	3179
2013	8	39	337	2013	52	104	3283
2013	9	33	370	2014	1	91	3374
2013	10	43	413	2014	2	122	3496
2013	11	33	446	2014	3	112	3608
2013	12	32	478	2014	4	113	3721
2013	13	33	511	2014	5	130	3851
2013	14	40	551	2014	6	132	3983
2013	15	35	586	2014	7	135	4118
2013	16	42	628	2014	8	127	4245
2013	17	48	676	2014	9	127	4372
2013	18	48	724	2014	10	135	4507
2013	19	45	769	2014	11	152	4659
2013	20	55	824	2014	12	132	4791
2013	21	41	865	2014	13	151	4942
2013	22	49	914	2014	14	161	5103
2013	23	51	965	2014	15	166	5269
2013	24	51	1016	2014	16	165	5434
2013	25	46	1062	2014	17	176	5610
2013	26	57	1119	2014	18	172	5782
2013	27	57	1176	2014	19	160	5942
2013	28	52	1228	2014	20	186	6128
2013	29	71	1299	2014	21	177	6305
2013	30	66	1365	2014	22	186	6491
2013	31	69	1434	2014	23	197	6688
2013	32	66	1500	2014	24	198	6886
2013	33	73	1573	2014	25	222	7108
2013	34	70	1643	2014	26	210	7318
2013	35	80	1723	2014	27	199	7517
2013	36	65	1788	2014	28	223	7740
2013	37	71	1859	2014	29	215	7955
2013	38	84	1943	2014	30	228	8183
2013	39	92	2035	2014	31	234	8417
2013	40	81	2116	2014	32	189	8606
2013	41	88	2204	2014	33	250	8856
2013	42	74	2278	2014	34	259	9115
2013	43	97	2375	2014	35	266	9381
2013	44	92	2467				

Investigating Metric Spike Findings - III

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

1.Firstly we will extract the week from occurred_at column using the extract,week functions

2. Then, we will select out those rows in which event_type = 'signup_flow' and event_name = 'complete_signup'

3. If finding for a specific week we will specify the week number using the extract function

4. Then using the left join we will join the two tables on the basis of user_id where event_type = 'engagement'

5. Then we will use the Group By function to group the output table on the basis of user_id

6. Then we will use the Order By function to order the result table on the basis of user_id

Output :

Google Drive Link for saved result: Click Here

Investigating Metric Spike Findings - IV

To find the weekly user engagement per device:-

1. Firstly we will extract the year_num and week_num from the occurred_atcolumn of the events table using the extract, year and week function

2. Then we will select those rows where event_type = 'engagement' usingthe WHERE clause

3. Then by using the Group By and Order By function we will group andorder the result on the basis of year_num, week_num and device

Output :

Google Drive link for saved result : Click Here

Investigating Metric Spike Findings - V

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

1. We will first categorize the action on the basis of email_sent,email_opened and email_clicked using the CASE, WHEN, THEN functions

2. Then we select the sum of category of email_opened divide by the sumof the category of email_sent and multiply the result by 100.0 and name is as email_opening_rate

3. Then we select the sum of category of email_clicked divide by the sum of the category of email_sent and multiply the result by 100.0 and name is as email_clicking_rate

4. email_sent = ('sent_weekly_digest','sent_reengagement_email')

5. email_opened = 'email_open'

6. email_clicked = 'email_clickthrough'

Output :

email_opening_rate	email_clicking_rate
33.58338805	14.78988838

Analysis :

From the tables and Bar plot I have infer the following:

- number of distinct job reviewed per day is 0.0083 number of non-distinct jobs
- reviewed per day is 0.0111
- 7 day rolling average throughput for 25, 26, 27, 28, 29 and 30 Nov 2020 are 1, 1, 1, 1.25, 1.2 and 1.3333 respectively(for both distinct and non-distinct)
- Percentage Share of each language i.e. Arabic, English, French, Hindi, Italian and Persian are 12.5, 12.5, 12.5, 12.5, 12.5 and 37.5 respectively (for both distinct and non-distinct)
- There are 2 duplicates values/rows having job_id = 23 and language = Persian in both the rows

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

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

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

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

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

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

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

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

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

From the tables and Bar plot I have infer the following: -

- The weekly user engagement is the highest for week 31 i.e. 1685
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014
- The email_opening_rate is 33.5833 and email_clicking_rate is 14.78988

I have used the Why's approach to gain few more insights:-

- ❖ 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 it's 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 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.

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 the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!

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

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

Problem:

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

Your task: How many males and females are Hired?

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

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

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

Your task: Draw the class intervals for salary in the company?

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

Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department?

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

Your task: Represent different post tiers using chart/graph?

Findings:

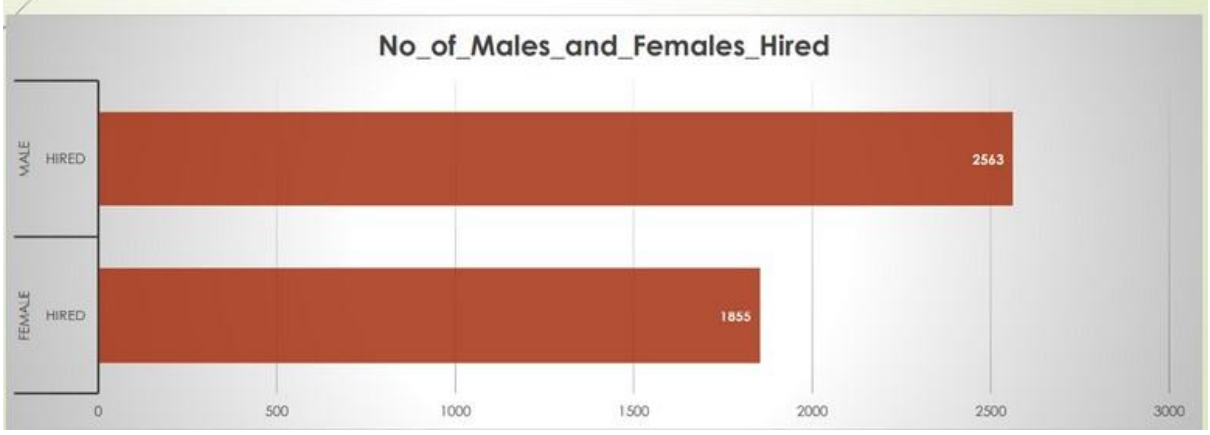
Before starting the actual analysis I have:-

- Firstly, I made a copy of the raw data where I can perform the Analysis so that whatever changes I made it will not affect the original data
- Secondly, I looked for blank spaces and NULL values if any.

- Then I had imputed the numerical blank and NULL cells with mean of the column(if no outliers existed for that particular column) or with median (if outliers existed for that column)
- Then I looked for if any outliers exists and replaced them with the median of the particular column where the outlier existed
- Then for blank cells of categorical variables I had replaced with the variable with the highest count
- Then I looked for duplicate rows and removed them if any
- Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis

Findings - I

event_name	Status	No_of_males_and_females_hired
Female	Hired	1855
Male	Hired	2563



From the above table and bar plot I have inferred that:-

There are 2563 Males hired for different roles in the company While there are only 1855 Females hired for different roles in the company.

Findings - II

To find the average salary offered in this company: -

1. First, we need to remove the outliers i.e., to remove the salaries below 1000 and above 100000

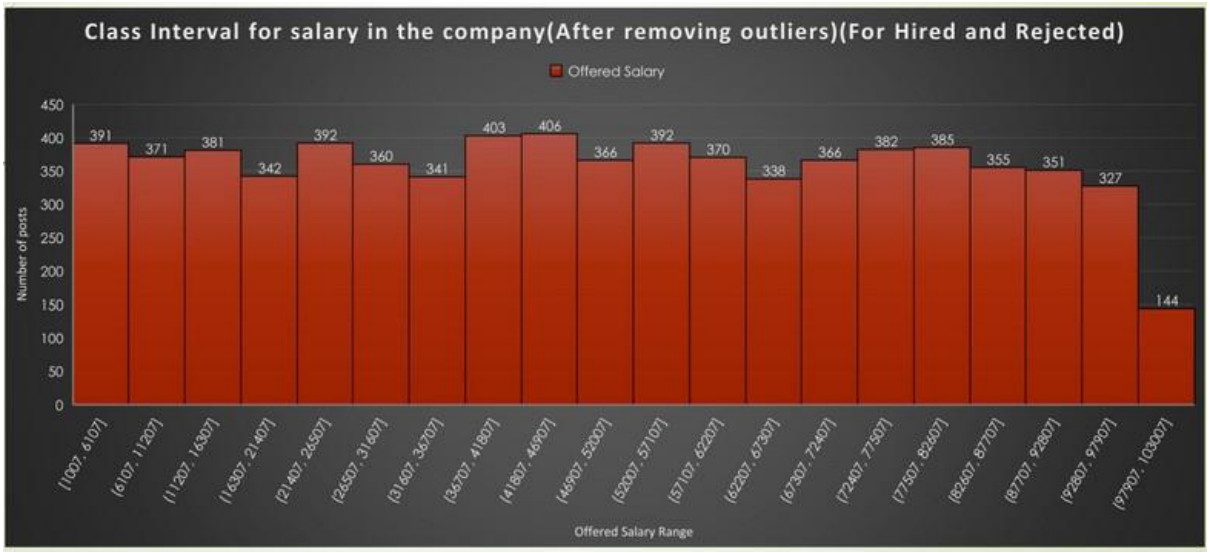
2.Then using the formula

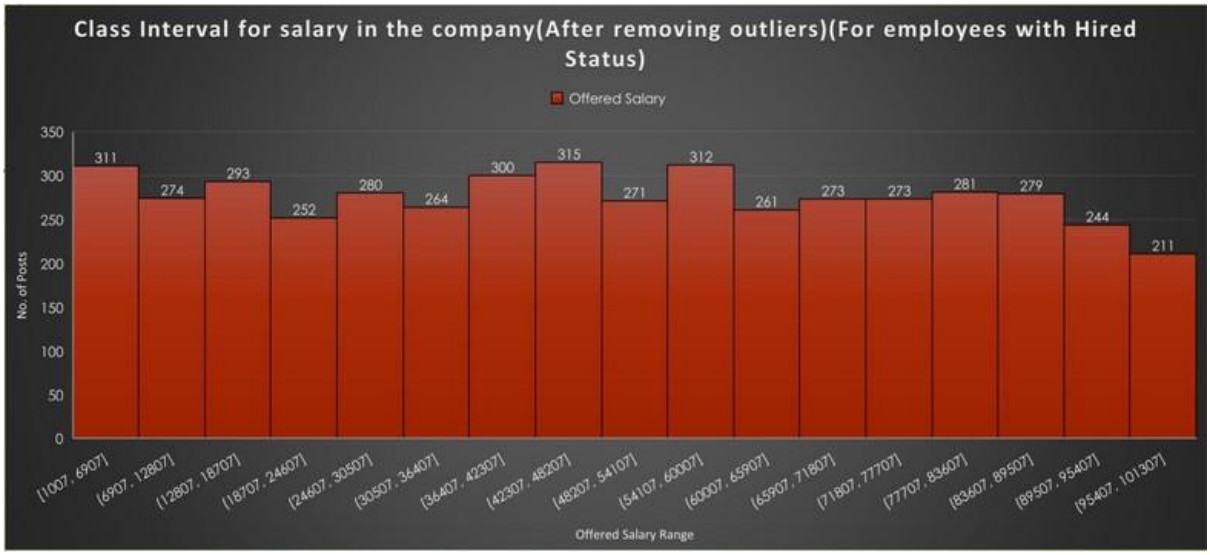
=AVERAGE(entire_column_of_salary_after_removing_outliers)

Output/Result: 49983.03223

Findings - III

From the Bar plot I have inferred that the highest number of posts (both hired and rejected) is 406 for the salary range 41807 to 46907

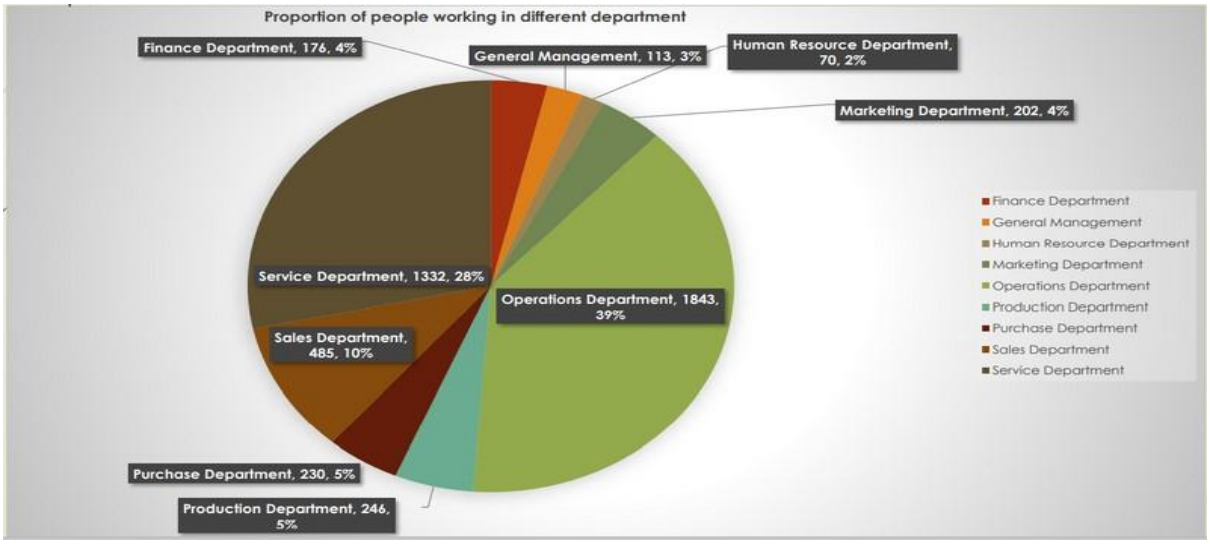


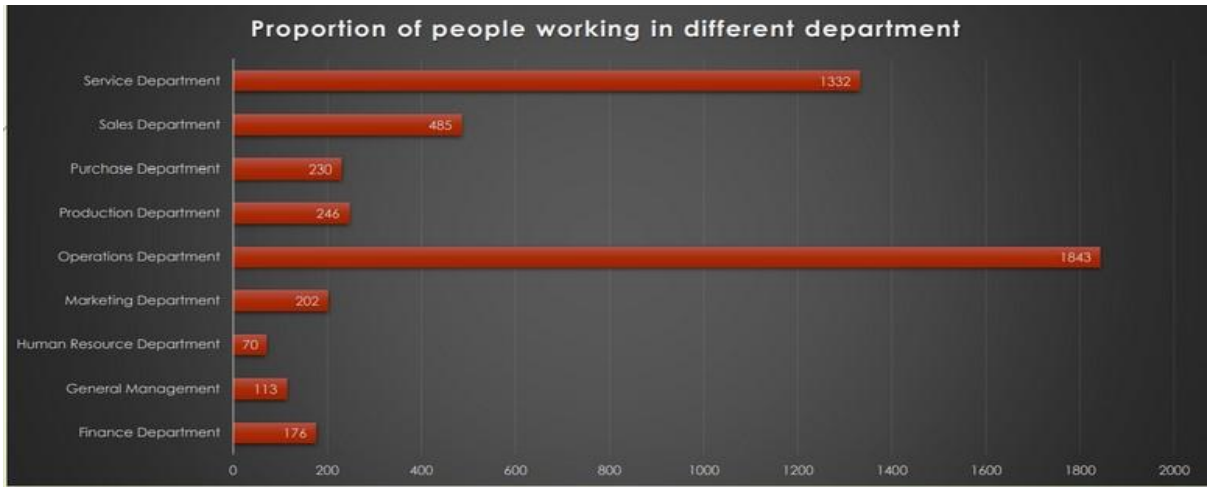


From the above Bar plot I have inferred that the highest number of posts (hired) is 315 for the salary range 42307 to 54107.

Findings - IV

Department	Status	Count of Department
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

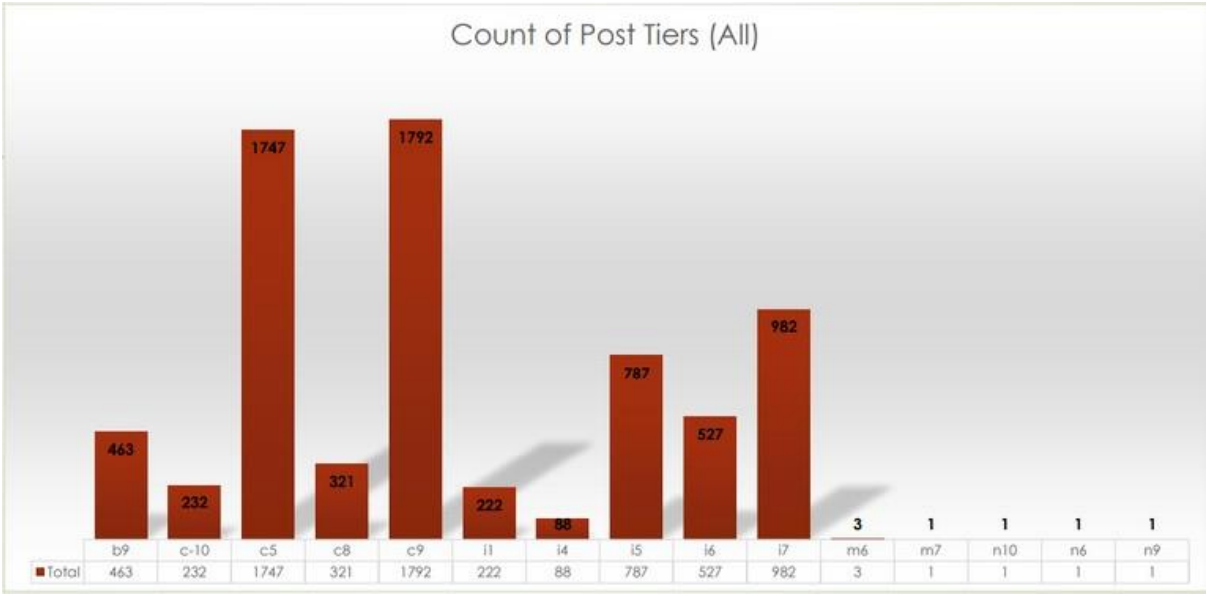


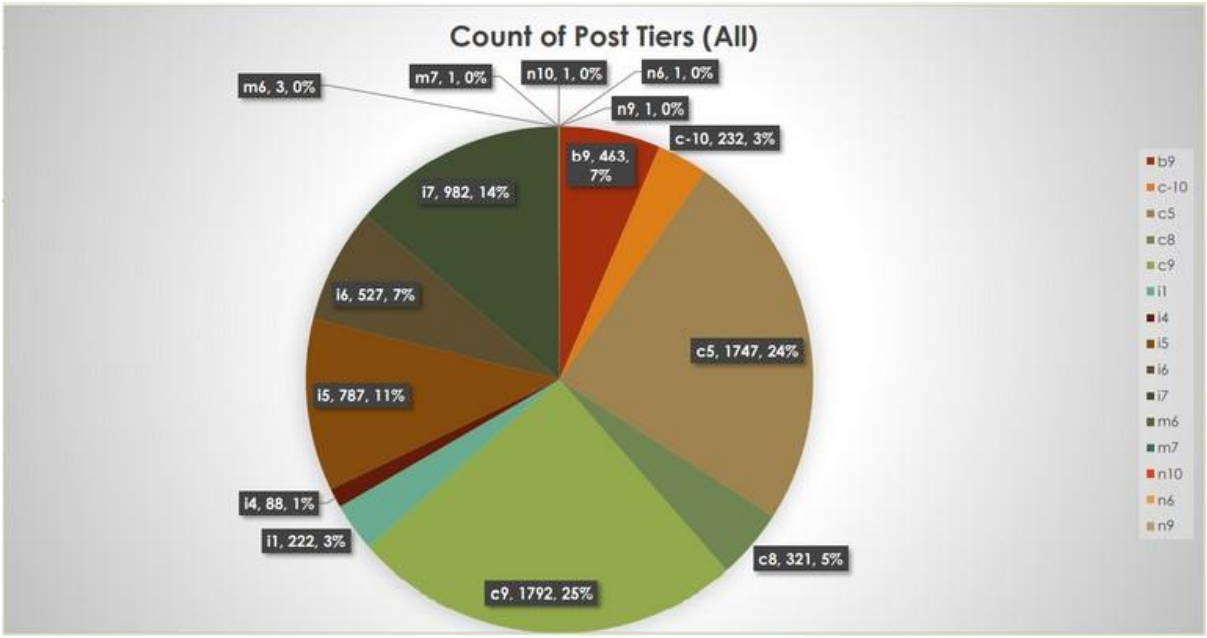


From the above table, pie chart and Bar Plot I have inferred that the Highest number of people were working in the Operations Department i.e. 1843 which accounts for almost 39% of the total workforce of the company.

Findings - V

Post Name	Status	Count of Post Tiers (All)
b9		463
c-10		232
c5		1747
c8		321
c9		1792
i1		222
i4		88
i5		787
i6		527
i7		982
m6		3
m7		1
n10		1
n6		1
n9		1





From the above table, Bar plot and Pie chart I have inferred that the c9 post has the highest number of openings i.e. 1792 which accounts for 25% of the total job openings of the company/firm.

Analysis:

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

- 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 men-women 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 is it that there are less number of people who have salaries more than 85000 and there are more number of people who have salaries 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

Conclusion:

In the conclusion part, I would like to conclude that Hiring Process Analytics plays an important part 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 it's current employees.

IMDB Movie Analysis

Description:

You are provided with dataset having various columns of different IMDB Movies. You are required to Frame the problem.

For this task, you will need to define a problem you want to shed some light on.

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

Problem:

- Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

Your task: Find the movies with the highest profit?

- Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

- Best Directors: Group the column using the director_name column. Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

Your task: Find the best directors

- Popular Genres: Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres

- Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined. Group the combined column using the actor_1_name column.

- Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

- Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the critic-favorite and audience-favorite actors

Findings:

1. Firstly I made a copy of the raw data where I can perform the Analysis so that whatever changes I made it will not affect the original data

2. Then dropping the columns which have no use for the analysisthat we will be doing

3. 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_posts', 'plot_keywords', 'movie_imdb_link', 'content_rating', 'actor_2_facebook_likes', 'aspect_ratio', 'movie_facebook_likes' are the columns containing irrelevant data for the analysis tasks provided. So, these columns need to be dropped.

4. After dropping the irrelevant columns now we need to remove the rows from the dataset having anyone of its column value as blank/NULL

5. Then we need to get rid off the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab

Findings - I

To find the movies with the highest profit: -

1. First we need to subtract the budget value from the gross value to get the profit.

2. Then, by using the scatter plot option we will plot values of profit(y_axis) and budget (x_axis)

3. Then with the help of graph we will be finding the outliers

director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score	Profit
James Cameron	723	760505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ	886204	3054	English	USA	PG-13	237000000	2009	7.9	523505847
Colin Trevorrow	644	652177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic WorldÂ	418214	1290	English	USA	PG-13	150000000	2015		502177271
James Cameron	315	658672302	Drama|Romance	Leonardo DiCaprio	TitanicÂ	793059	2528	English	USA	PG-13	200000000	1997	7.7	458672302
George Lucas	282	460935665	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New HopeÂ	911097	1470	English	USA	PG	11000000	1977	8.7	449935665
Steven Spielberg	215	434949459	Family|Sci-Fi	Henry Thomas	E.T. the Extra-TerrestrialÂ	281842	515	English	USA	PG	10500000	1982	7.9	424449459

After removing the outliers, from the above table I have inferred that 'Avatar' was the highest profit making movie ever with a profit of 523505847.

Findings - II

To find the IMDB Top 250 we will:-

1. First we will filter out those rows whose num_voted_users > 25000 using thesort and filter option

2. Then we will arrange the dataset on the basis of imdb_score in descendingorder

3. Then we will select only the top 250 rows for the further analysis

4. Then we will create a new column rank using the RANK() function and using the formula =RANK(N2,$N$2:$N$251,0)+COUNTIFS($N$2:N2,N2)-1

5. Then we will filter out (unselect ‘English’) from the language column and wewill get the desired output

Top - 5 IMDB Movies all languages

director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score	Rank
Frank Darabont	199	28341469	Crime|Drama	Morgan Freeman	The Shawshank RedemptionÂ	1689764	4144	English	USA	R	25000000	1994	9.3	1
Francis Ford Coppola	208	134821952	Crime|Drama	Al Pacino	The GodfatherÂ	1155770	2238	English	USA	R	6000000	1972	9.2	2
Christopher Nolan	645	533316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark KnightÂ	1676169	4667	English	USA	PG-13	185000000	2008	9	3
Francis Ford Coppola	149	57300000	Crime|Drama	Robert De Niro	The Godfather: Part IIÂ	790926	650	English	USA	R	13000000	1974	9	4
Peter Jackson	328	377019252	Action|Adventure|Drama|Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the KingÂ	1215718	3189	English	USA	PG-13	94000000	2003	8.9	5

From the above table I have inferred that 'The Shawshank Redemption' had the highest IMDB ratings.

Top - 5 IMDB Movies all languages (except English)

director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score	Rank
Sergio Leone	181	6100000	Western	Clint Eastwood	The Good, the Bad and the UglyÂ	503509	780	Italian	Italy	Approved	1200000	1966	8.9	8
Fernando Meirelles	214	7563397	Crime|Drama	Alice Braga	City of GodÂ	533200	749	Portuguese	Brazil	R	3300000	2002	8.7	19
Akira Kurosawa	153	269061	Action|Adventure|Drama	Takashi Shimura	Seven SamuraiÂ	229012	596	Japanese	Japan	Unrated	2000000	1954	8.7	20
Hayao Miyazaki	246	10049886	Adventure|Animation|Family|Fantasy	Bunta Sugawara	Spirited AwayÂ	417971	902	Japanese	Japan	PG	19000000	2001	8.6	25
Florian Henckel von Donnersmarck	215	11284657	Drama|Thriller	Sebastian Koch	The Lives of OthersÂ	259379	407	German	Germany	R	2000000	2006	8.5	45

From the above table I have inferred that the movie 'The Good, the Bad and the Ugly' had the highest IMDB ratings w.r.t movies with all other languages (except English); it's country of origin in Italy.

Findings - III

To find the best top 10 directors on the basis of mean of imdb_score we will:-

1. First select the imdb_score column of the cleaned dataset

2. Then we will click on pivot table

3. We will add director_name into the series section of the pivot table

4. Then we will add average imdb_score into the values section of thepivot table

5. Then we will first sort the data on the basis of average ofimdb_score in descending order and then on the basis of director name alphabetically.

Top 10 director having the highest IMDB mean	
Director Name	mean of Imdb Score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

From the above table I have inferred that Charles Chaplin and Tony Kaye had the highest mean of IMDB Score i.e. 8.6.

Findings - IV

To find the Popular Genres we will:-

1. First select the genres column of the cleaned dataset

2. Then we will go for the pivot table option

3. Then we will Select the genres name as row labels

4. Then we will the values as the count of the number of genres and then sort it in descending order on the basis of count of the number of genres

Top 10 Popular genres are:-	
Genre name	count
Drama	153
Comedy | Drama | Romance	151
Comedy | Drama	147
Comedy	145
Comedy | Romance	135
Drama | Romance	119
Crime | Drama | Thriller	82
Action | Crime | Thriller	55
Action | Crime | Drama | Thriller	50
Action | Adventure | Sci-Fi	46

From the above table I have inferred that genre named 'Drama' was the most popular with a count of 153.

Findings - V

To find the critic-favorite and audience-favorite actors we will:-

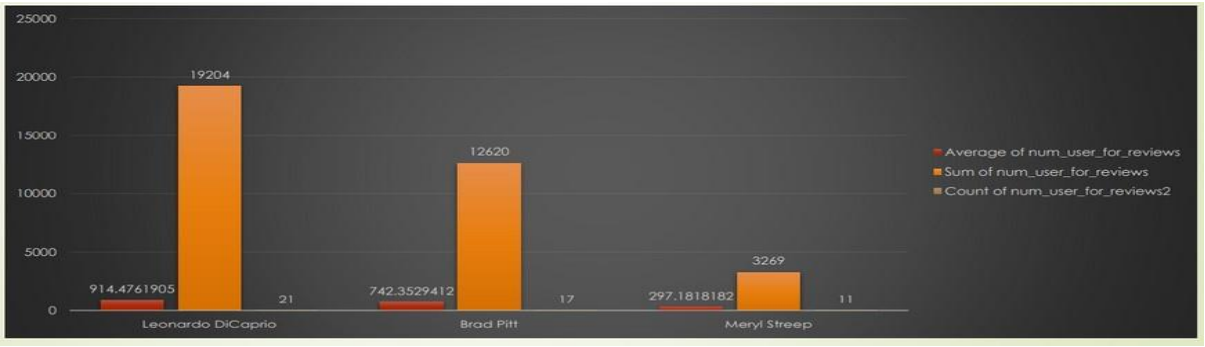
1. First three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors from the actor_1_name column

2. Then we will append the above 3 created columns into 1 column named actor_1_name_combine

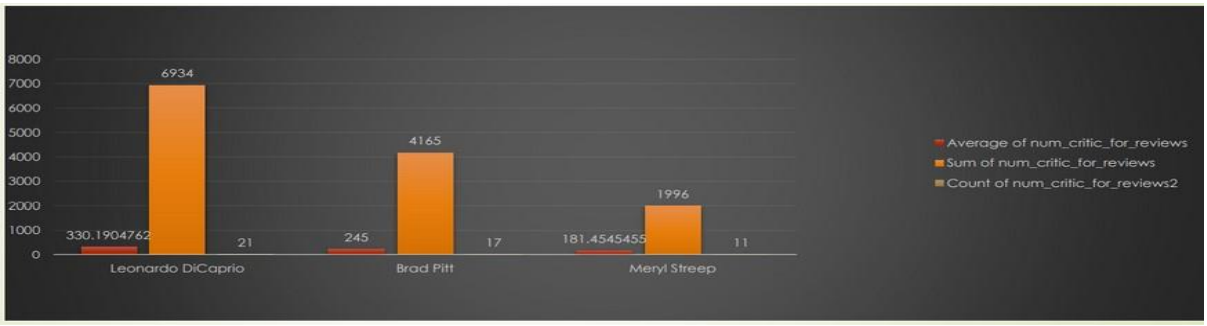
3. Then we will group the 3 columns of critic-favorite and audience-favorite actors

4. Then using the pivot table we will find the average, sum and count of critic favorite and audience-favorite actors

Row Labels	Average of num_user_for_reviews	Sum of num_user_for_reviews	Count of num_user_for_reviews2
Leonardo DiCaprio	914.4761905	19204	21
Brad Pitt	742.3529412	12620	17
Meryl Streep	297.1818182	3269	11
Grand Total	716.1836735	35093	49

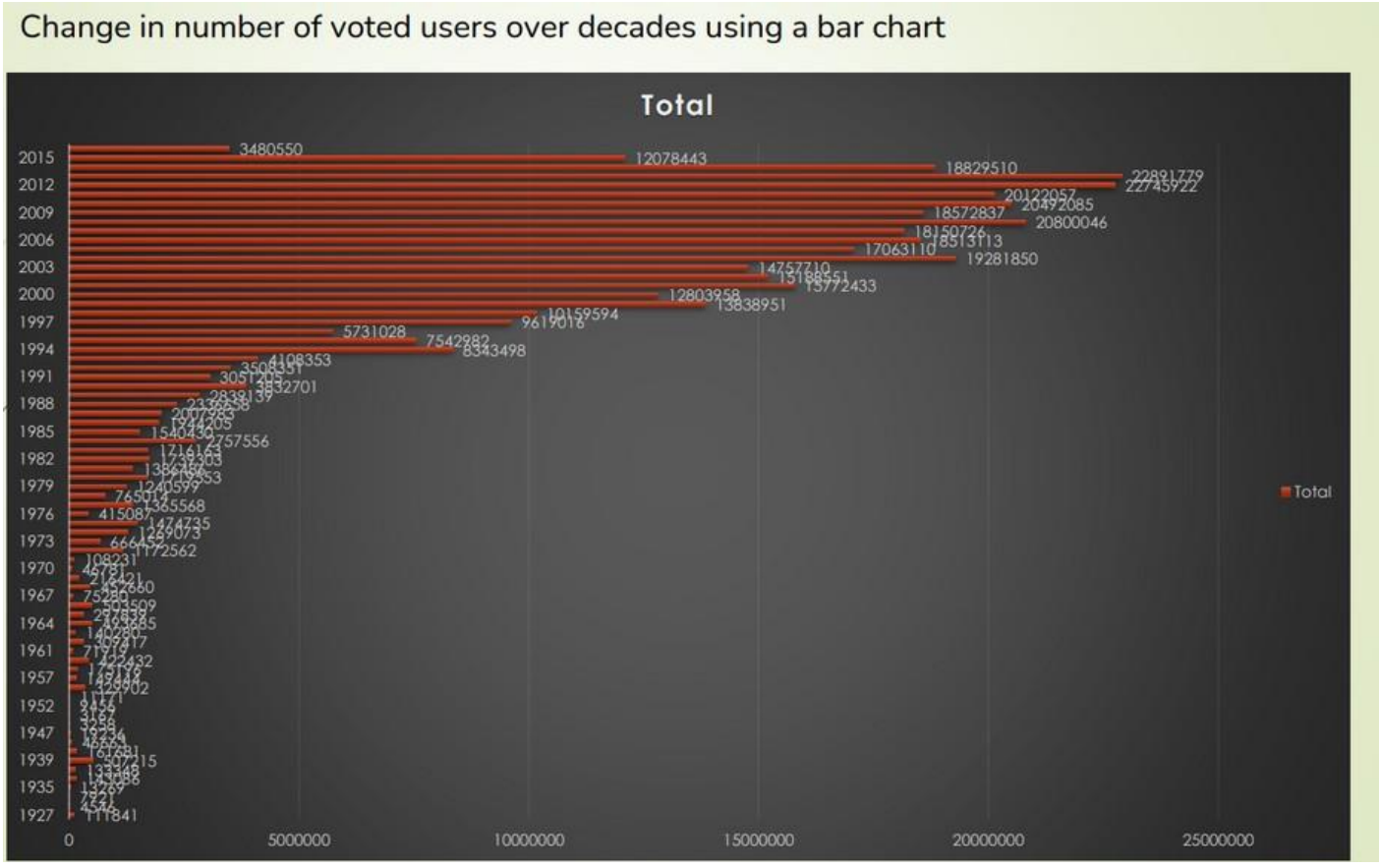


Row Labels	Average of num_critic_for_reviews	Sum of num_critic_for_reviews	Count of num_critic_for_reviews2
Leonardo DiCaprio	330.1904762	6934	21
Brad Pitt	245	4165	17
Meryl Streep	181.4545455	1996	11
Grand Total	267.244898	13095	49



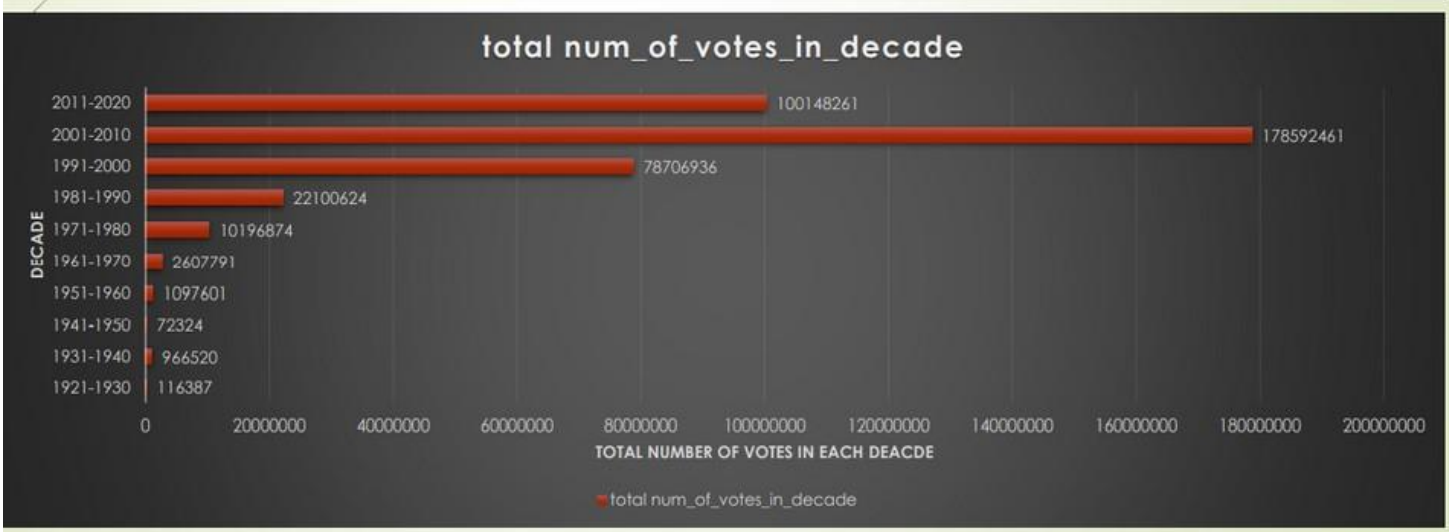
From the above two graphs I have inferred that 'Leonardo DiCaprio' was both critic-favorite and audience favorite.

Findings - VI



Change in number of voted users over decades using a bar chart

Decade	total num_of_votes_in_decade
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2607791
1971-1980	10196874
1981-1990	22100624
1991-2000	78706936
2001-2010	178592461
2011-2020	100148261
Total	394605779



From the above table and Bar plot I have inferred that most number of votes were in the decade 2001-2010 with a count of 178592461

Analysis:

Using the Why's approach I am trying to find some useful insights

- 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.

- Why there are more number of votes during the decade 2001-2010?

-----> The period 2001-2010 saw many scientific advancements and computer graphics advancement, also during this interval there was a splendid increase in the production of movies all over the world, so huge number of movies were produced and released during this decade. Also before 2000 there were no laws around the world that had a separate ministry/board/committee from the Government side that looked into the matters of film production and release

- Why is it that only movies having language as 'English' are the top 5 ranked movies on the basis of IMDB?

-----> 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 so as to gain some financial gains.

- 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 Comedy or Drama genre or both. But, most of them preferred Comedy genre films

- Why is it that there were more number of votes for the decade 2001-2010 than compared to 2011-2020, though there was advancement in graphics and animation during 2011-2020?

----> It is a fact that there was a great and immense growth of technology not only in the graphics and animation sector but in all aspects of life; Also it was during this interval VPN was introduced; VPN led to piracy (illegal distribution of film) due to which most of people avoided going to theatres.

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 consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1.Approved: The company has approved loan application

2.Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

3.Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

4.Unused Offer: Loan has been cancelled by the client but on different stages of the process.

Problem:

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

It aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics – understanding the types of variables and their significance should be enough).

Findings:

Firstly, create a copy of the raw data

Then the percentage of null values needs to be analyzed and those columns that have more than 50% of the null data have to be dropped And those columns with less than 50% of the null data have to be replaced with mean or median or the highest occurring categorical variables.

The following columns needs o be dropped as they have more than 50% of the NULL values

- OWN_CAR_AGE
- EXT_SOURCE_1
- APARTMENTS_AVG
- BASEMENTAREA_AVG
- YEARS_BUILD_AVG
- COMMON_AREA_AVG
- ELEVATORS_AVG ENTRANCES_AVG
- FLOORSMAX_AVG
- FLOORSMIN_AVG
- LANDAREA_AVG
- LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG
- NONLIVINGAPARTMENTS_AVG
- NONLIVINGAREA_AVG
- APARTMENTS_MODE
- BASEMENTAREA_MODE
- YEARS_BUILD_MODE
- COMMON_AREA_MODE
- ELEVATORS_MODE
- ENTRANCES_MODE
- FLOORSMAX_MODE
- FLOORSMIN_MODE
- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MODE
- NONLIVINGAPARTMENTS_MODE
- NONLIVINGAREA_MODE
- APARTMENTS_MEDIAN
- BASEMENTAREA_MEDIAN
- YEARS_BUILD_MEDIAN
- COMMON_AREA_MEDIAN
- ELEVATORS_MEDIAN
- ENTRANCES_MEDIAN FLOORSMAX_MEDIAN FLOORSMIN_MEDIAN
- LANDAREA_MEDIAN

- LIVINGAPARTMENTS_MEDIAN
- LIVINGAREA_MEDIAN
- NONLIVINGAPARTMENTS_MEDIAN
- NONLIVINGAREA_MEDIAN
- FONDKAPREMONT_MODE
- HOUSETYPE_MODE
- WALLSMATERIAL_MODE

Then drop those columns which are irrelevant for doing the Data Analysis. The following columns needs to be dropped:

- FLAG_MOBILE
- FLAG_EMPLOY_PHONE
- FLAG_WORK_PHONE
- FLAG_CONT_MOBILE
- FLAG_PHONE
- FLAG_EMAIL
- CNT_FAMILY_MEMBERS
- REGION_RATING_CLENT
- REGION_RATING_CLENT_W_CITY
- EXT_SOURCE_3
- YEAR_BEGINEXPLUATATION_AVG
- YEAR_BEGINEXPLUATATION_MODE
- YEAR_BEGINEXPLUATATION_MEDIAN
- TOTAL_AREA_MODE
- EMERGENCYSTATE_MODE
- DAYS_LAST_PHONE_CHANGE
- FLAG DOC 2
- FLAG DOC 3 FLAG DOC 4 FLAG DOC 5
- FLAG DOC 6
- FLAG DOC 7
- FLAG DOC 8
- FLAG DOC 9
- FLAG DOC 10
- FLAG DOC 11
- FLAG DOC 12
- FLAG DOC 13
- FLAG DOC 14
- FLAG DOC 15
- FLAG DOC 16
- FLAG DOC 17
- FLAG DOC 18
- FLAG DOC 19
- FLAG DOC 20
- FLAG DOC 21

Replacing Blanks in Occupation_Type column of the Application Dataset with the highest occurring categorical variable --> Highest occurring categorical variable is ‘Laborers’

Replacing Blanks in AMT_ANNUTIY column of the Application Dataset with the median of the AMT_ANNUITY as there exists outliers in the AMT_ANNUITY column --> Median of AMT_ANNUITY = 24903

Replacing Blanks in AMT_GOODS_PRICE column of the Application Dataset with the median of the AMT_GOODS_PRICE as there exists outliers in the AMT_GOODS_PRICE column --> Median of AMT_GOODS_PRICE = 450000

Replacing Blanks in Name_Type_Suite column of the Application Dataset with the highest occurring categorical variable --> Highest occurring categorical variable is ‘Unaccompanied’

Replacing Blanks in Organization_type column of the Application Dataset with the highest occurring categorical variable --> Highest occurring categorical variable is ‘Business Entity Type 3’

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

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

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

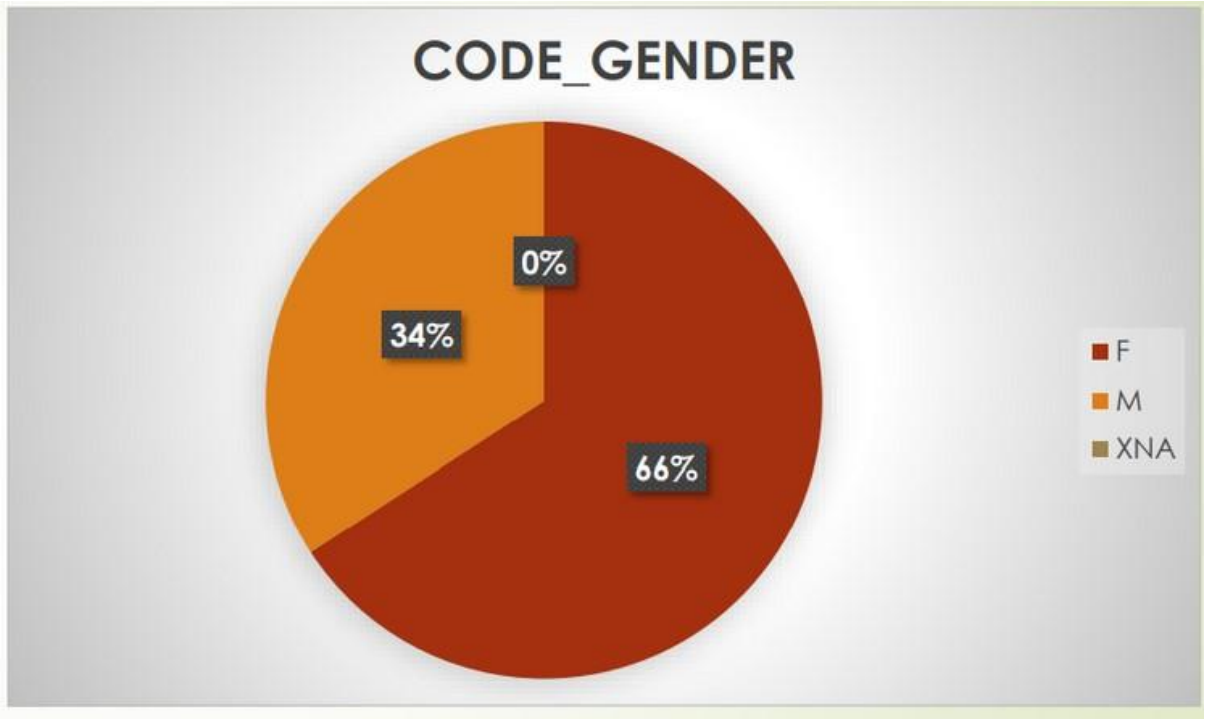
----> Replace Blanks with Unaccompained

AMT_ANNUITY :- Replace Blanks with 21340(median)

Findings - I

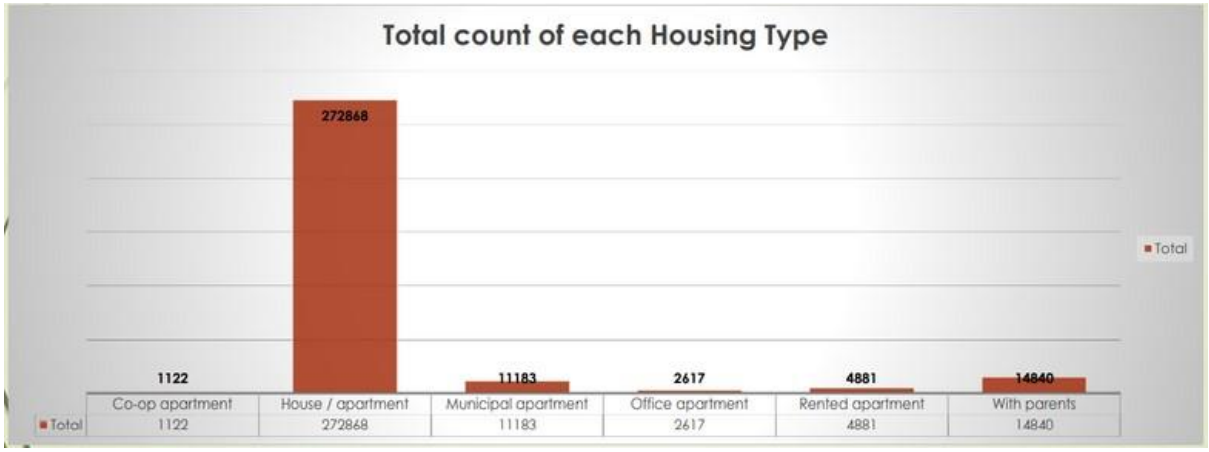
The Target Variable Pie chart shows that almost 92% of the total clients had no problem during payment while 8% of the clients had some or the other problem.

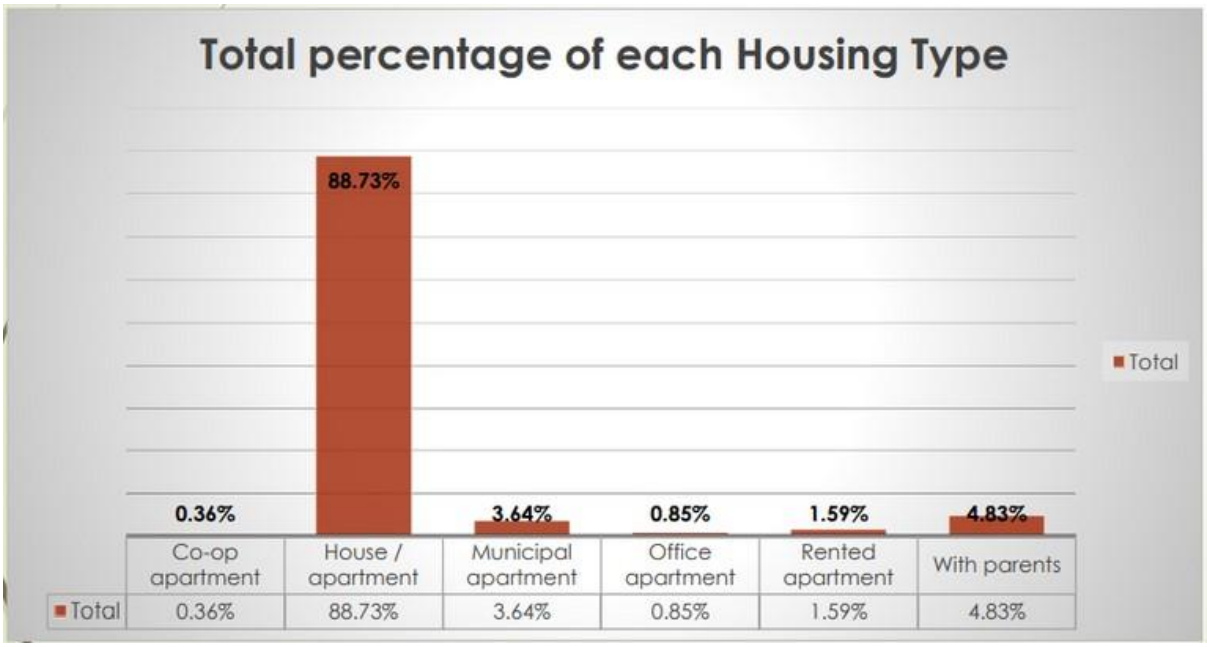
Findings - II



From the GENDER_VARIABLE pie chart we can infer that almost 66% of the clients are female and 34% of the clients are Male. The 0% of the applicants have gender as XNA which can be ignored.

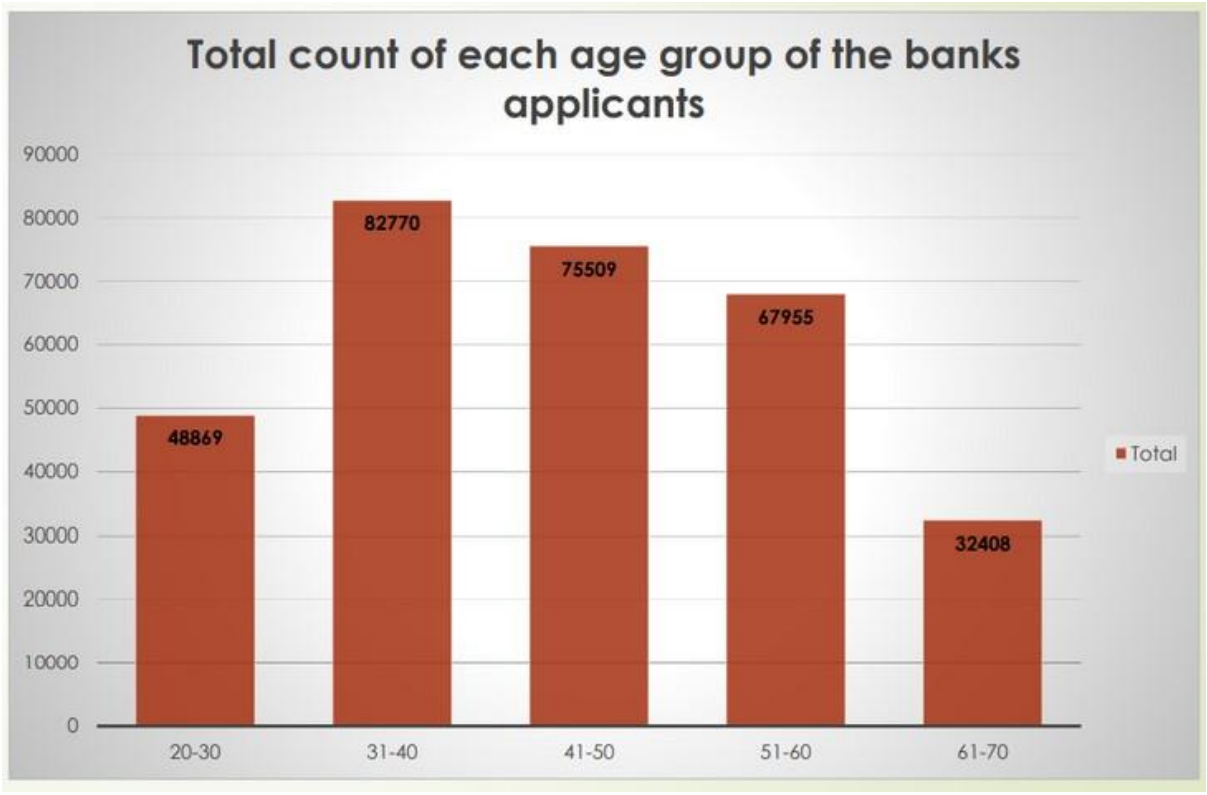
Findings - III





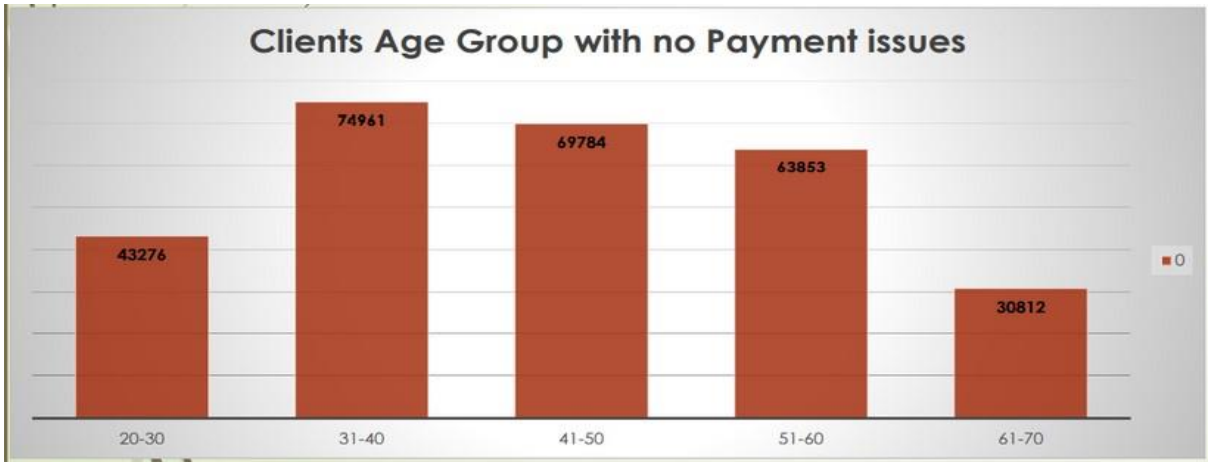
From the bar graphs of count and percentage The bank can target those groups who do not have their own apartment i.e. the bank may consider the people living in Co-op apartment, Municipal Apartment, Rented Apartment and people living with their parents.

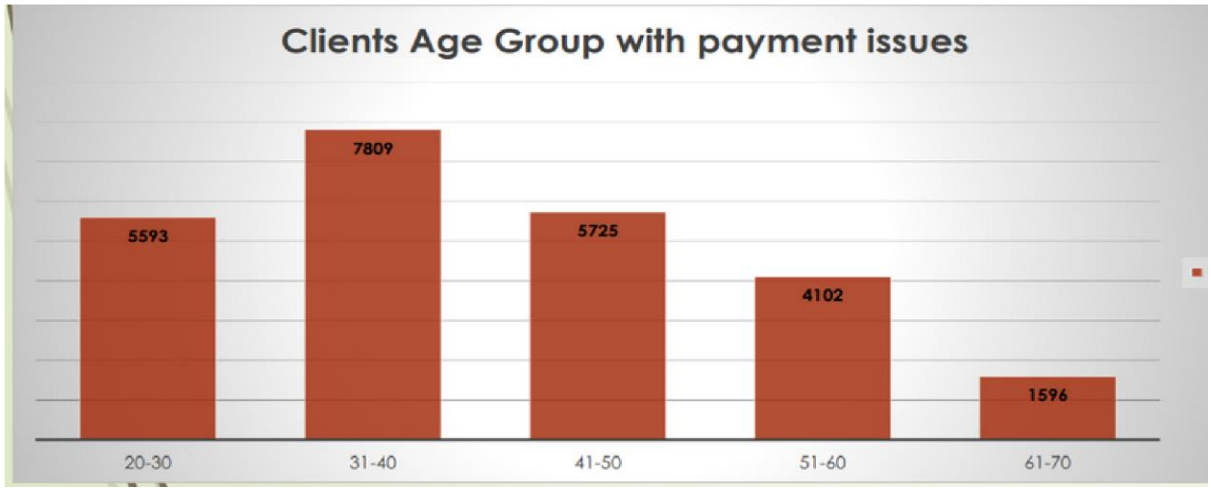
Findings - IV



From the above bar plot we can infer that most of the applicants belong to the Age Group '31-40'.

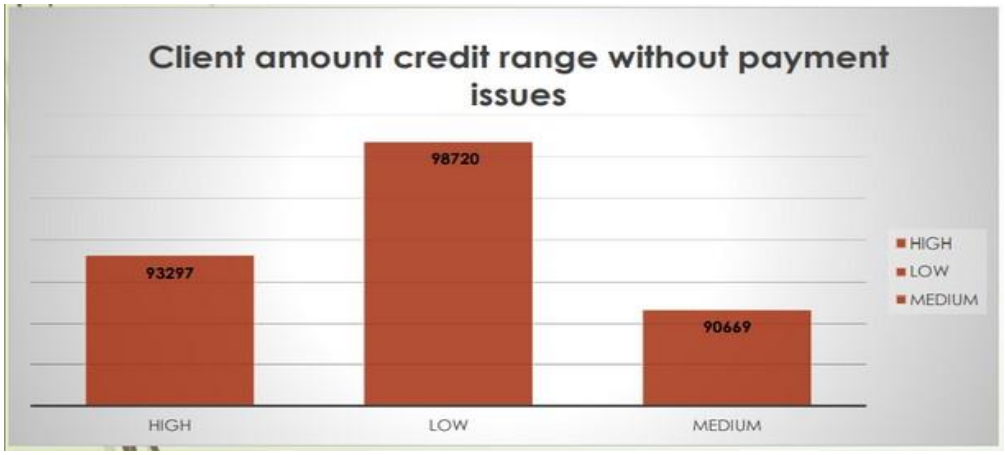
Findings - V



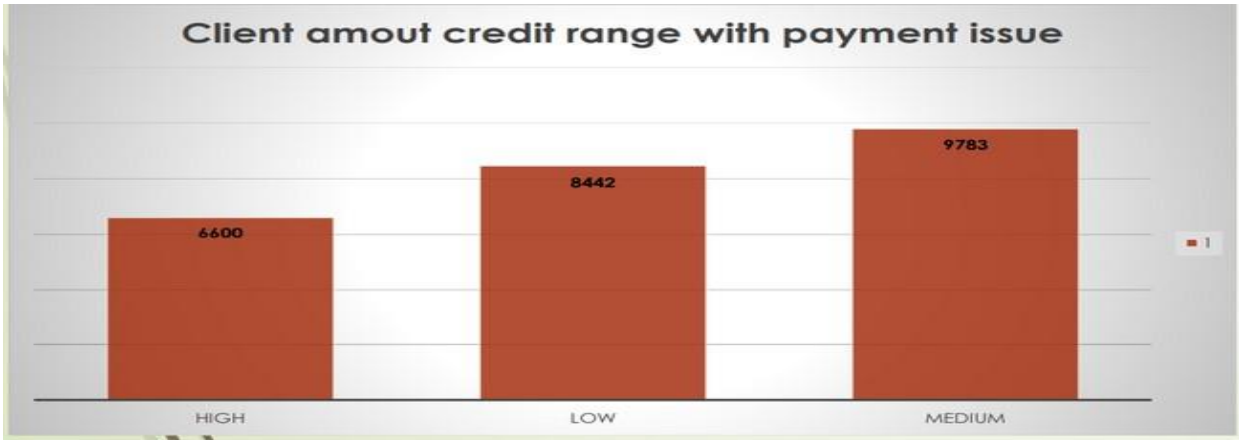


From the above Bar plots we can infer that clients/applicants in the Age Group '31-40' are having the highest number when it comes to doing/returning Payment to Banks also they have they highest count of payment issues when returning money to the bank.

Findings - VI

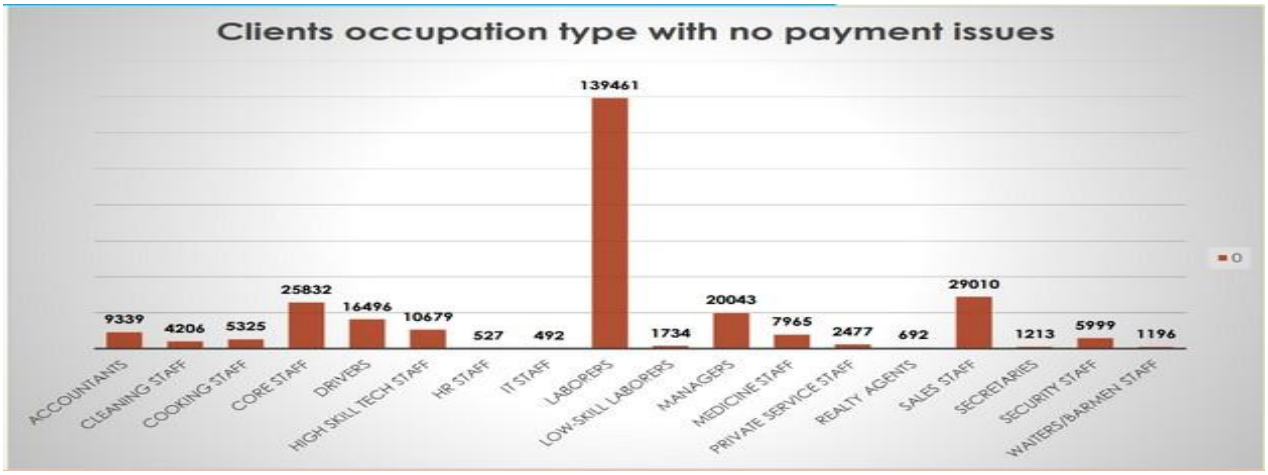


From the above Bar plot we can infer that clients belonging to 'Low' income range have the highest count when it comes to clients with no payment issues.

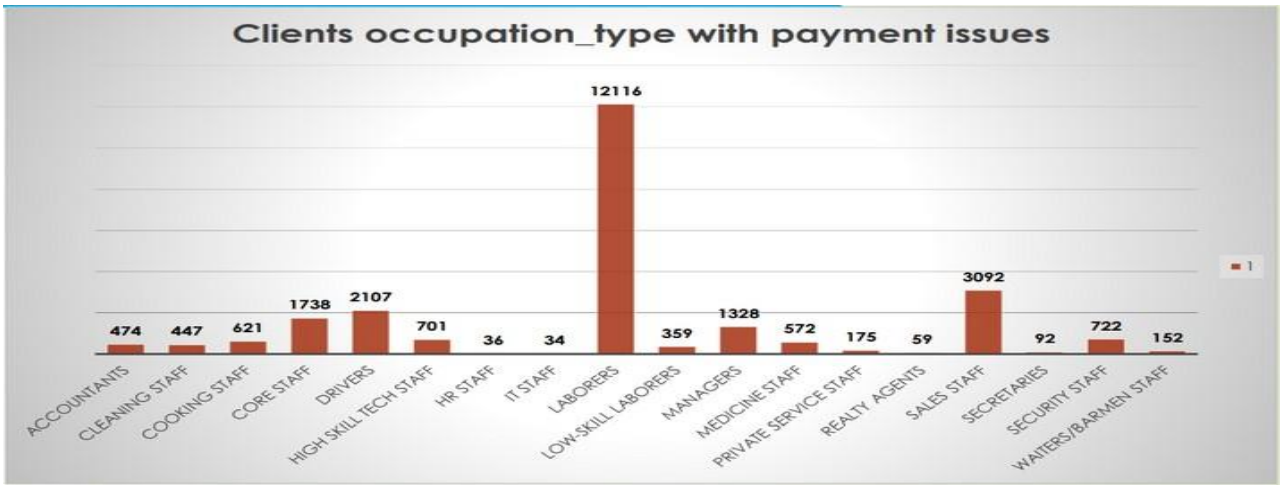


From the above Bar plot, we can infer that clients belonging to 'Medium' income range have the highest count when it comes to clients with payment issues.

Findings - VII

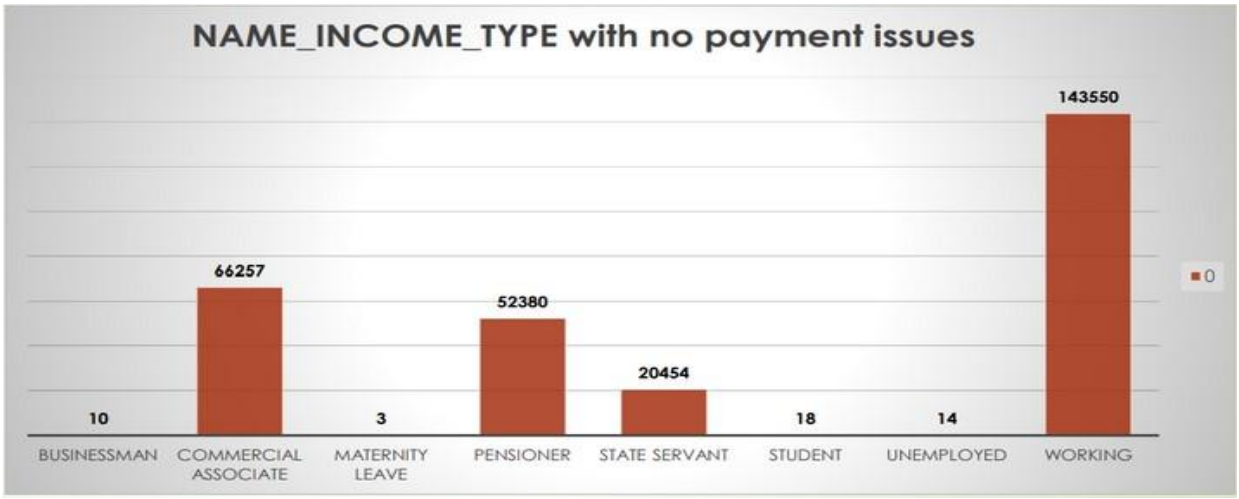


From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with no payment issues

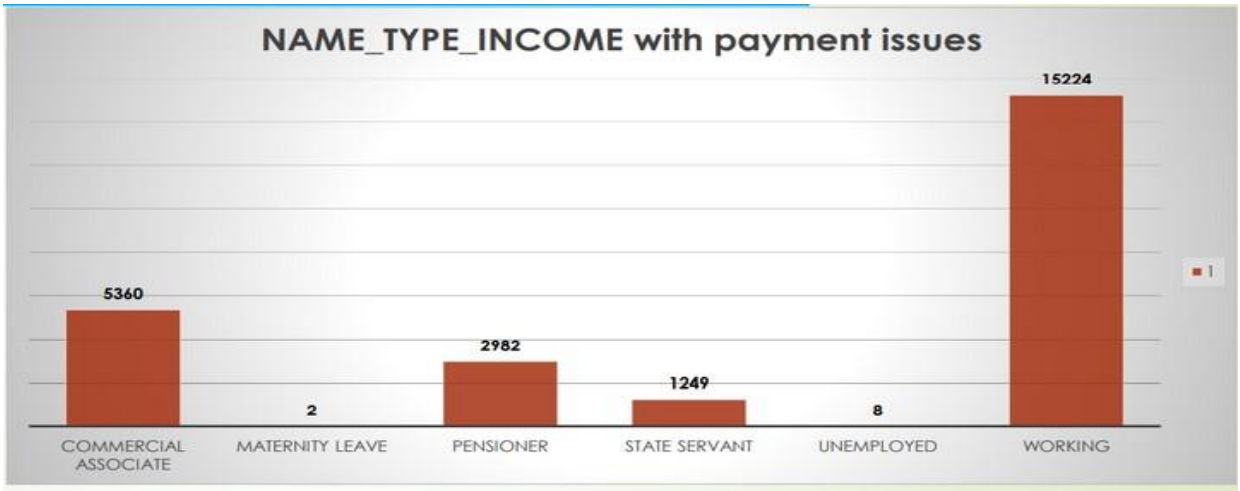


From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with payment issues.

Findings - VIII

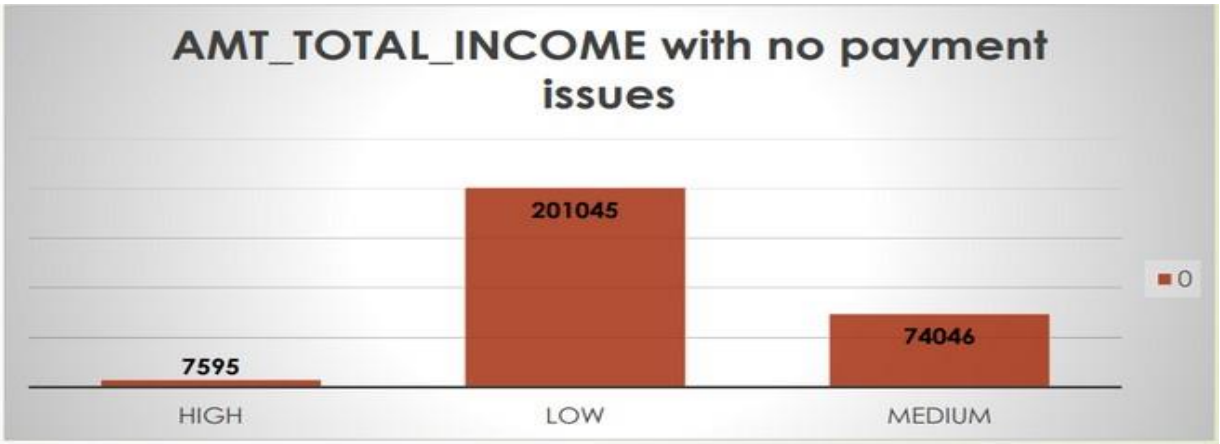


From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with no payment issues.

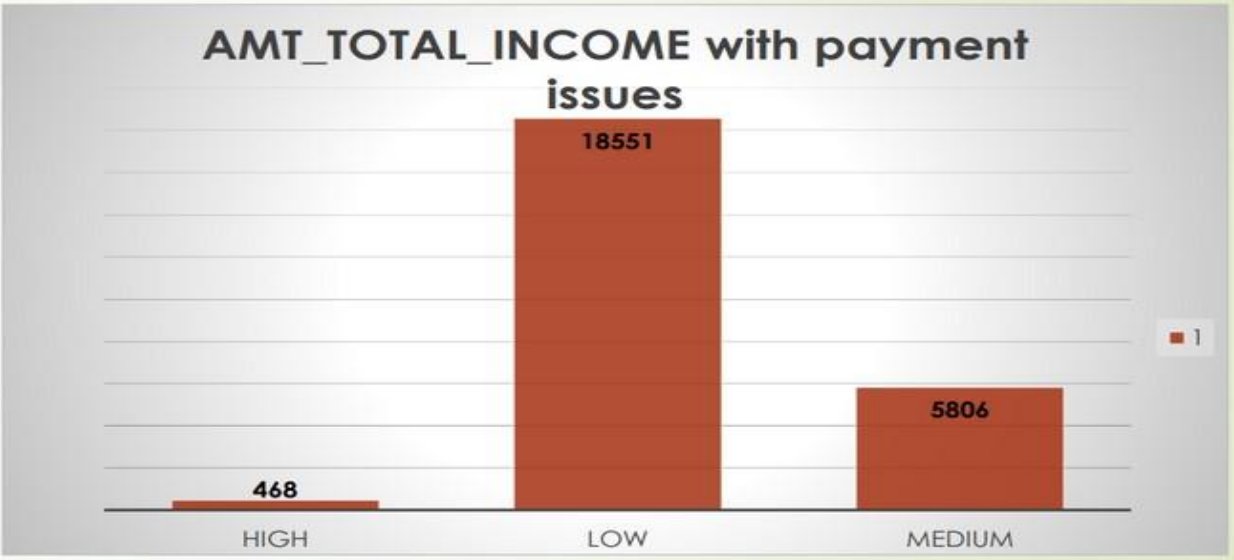


From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with payment issues.

Findings - IX



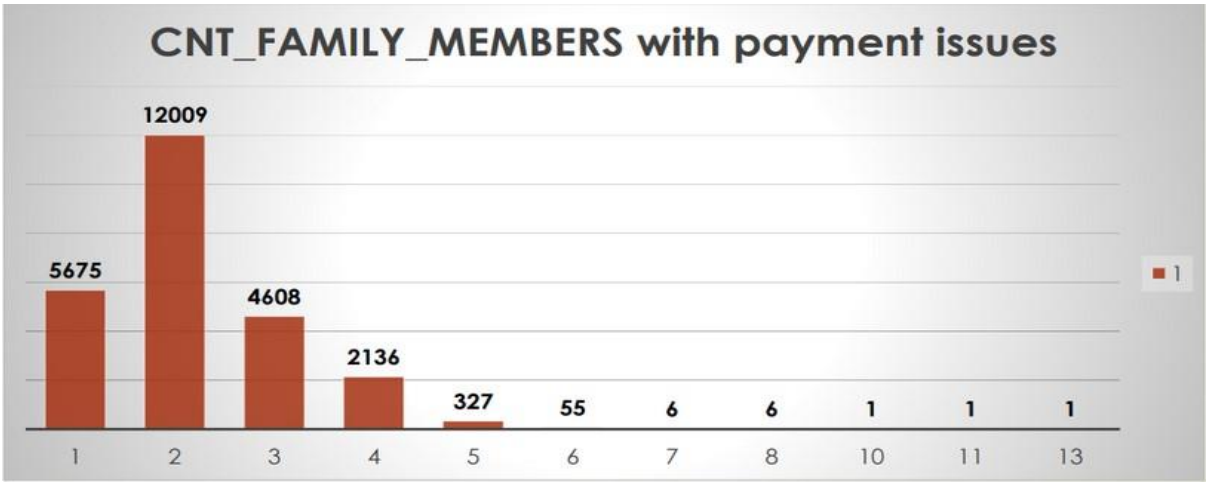
From the above Bar plot we can infer that client having the total income range as ‘LOW’ have the highest count when it comes to clients having no payment issues.



From the above Bar plot we can infer that client having the total income range as ‘LOW’ have the highest count when it comes to clients having payment issues.

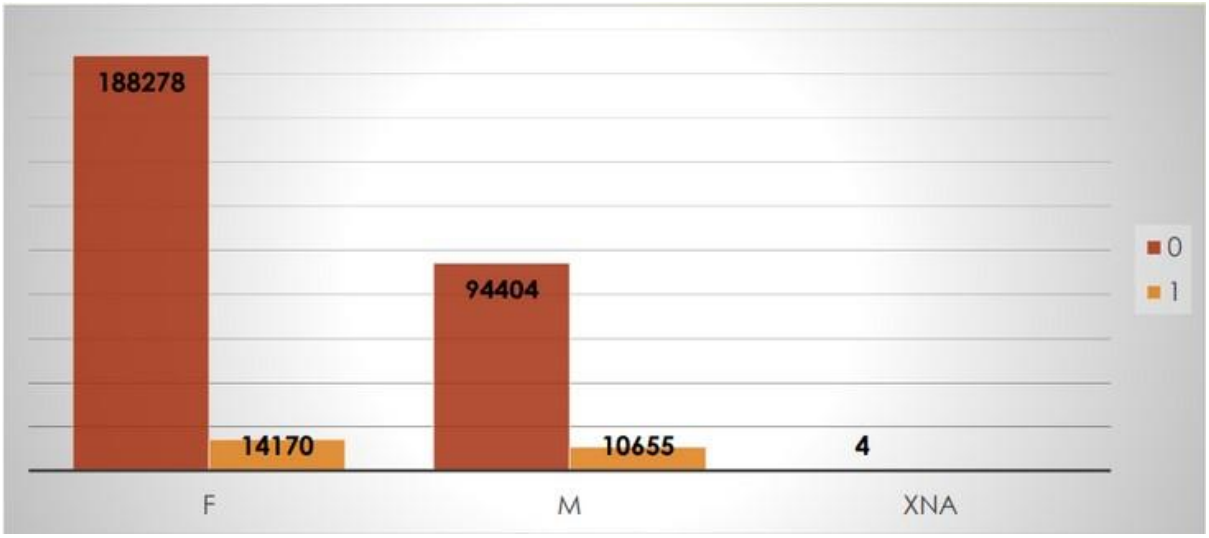
Findings - X

From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having no payment issues



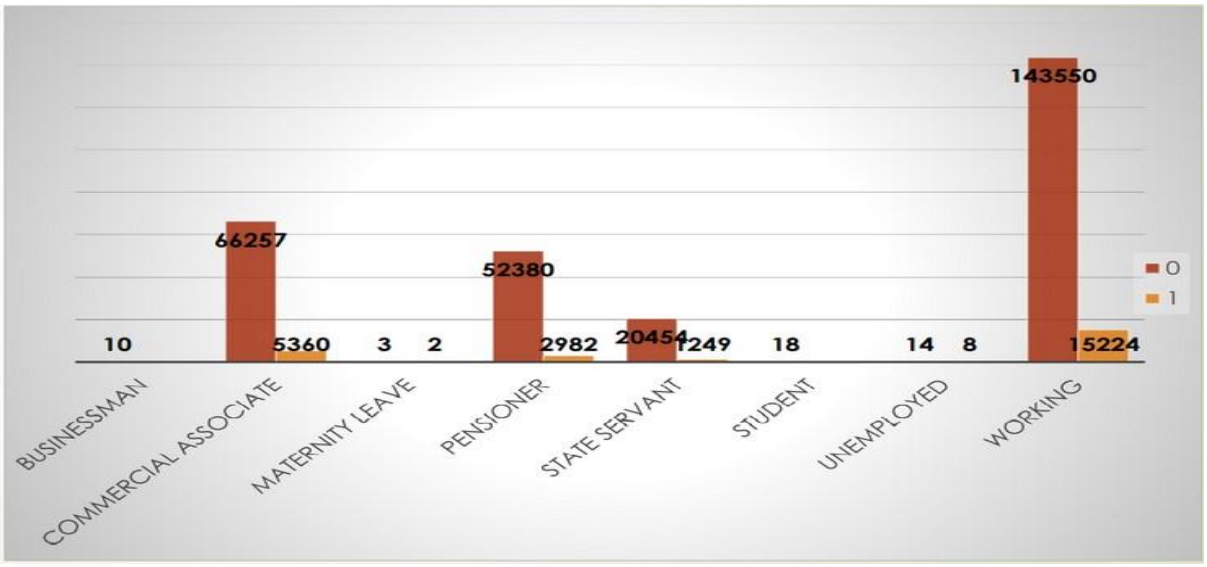
From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having payment issues.

Findings - XI



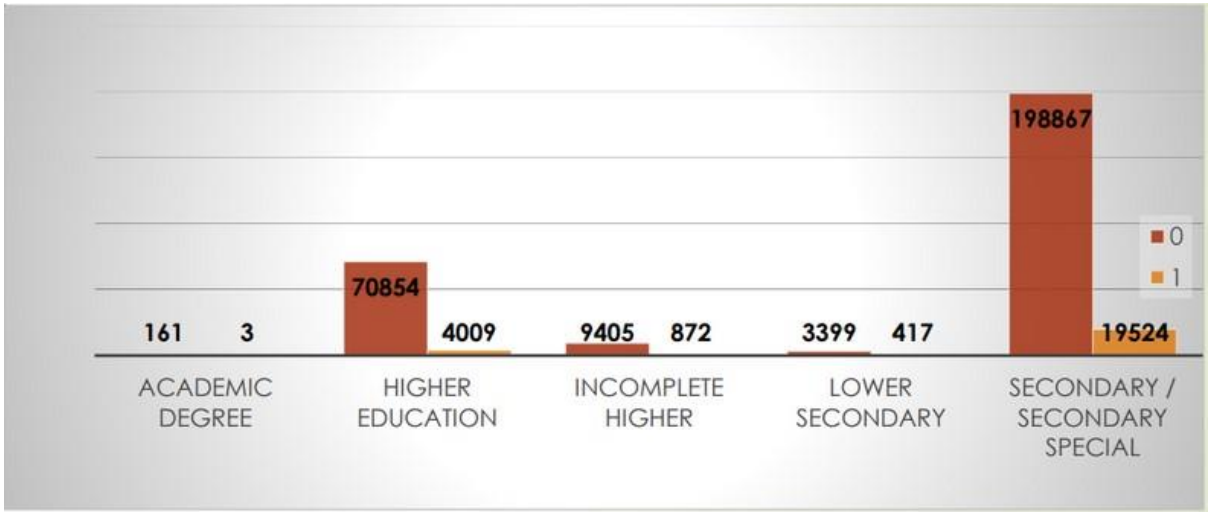
From the above Bar Plot we can infer that Clients with CODE_GENDER = 'F' have the highest number of non-defaulters i.e. 188278-14170 = 174108

Findings - XII



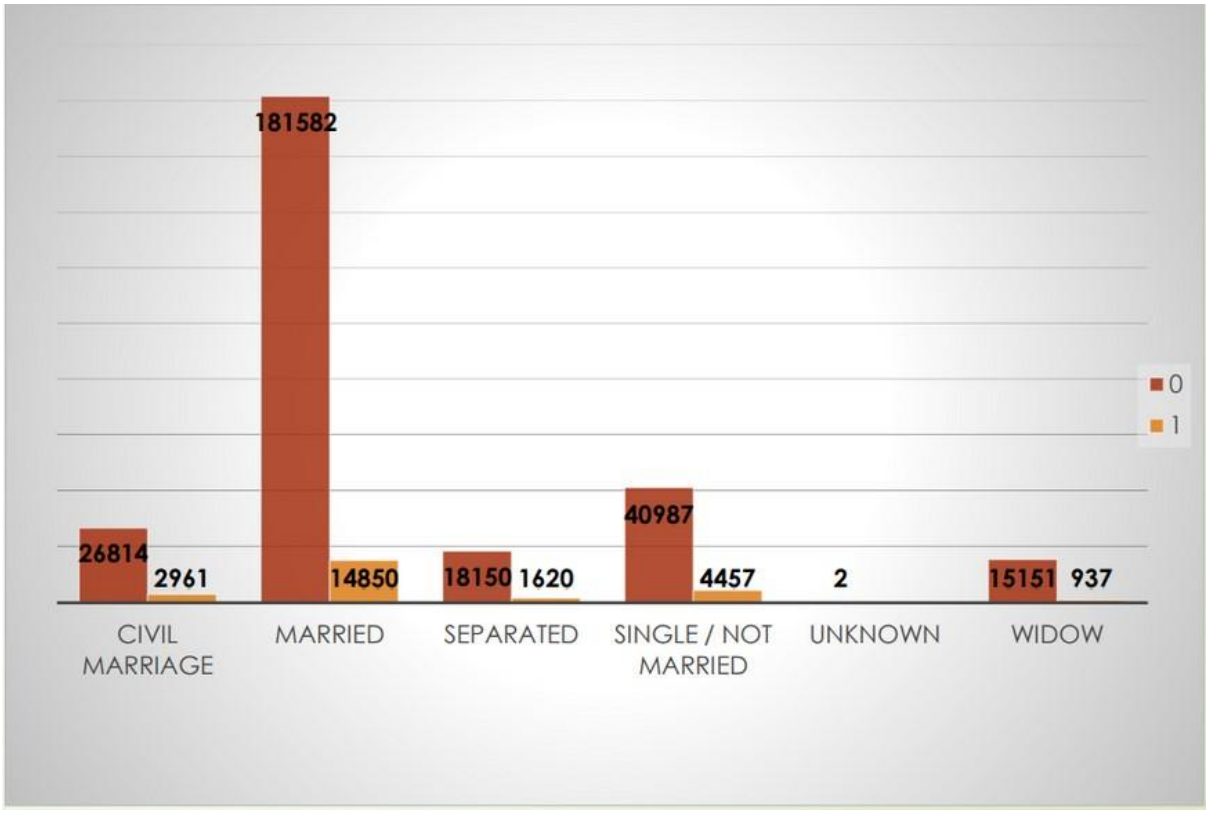
From the above Bar Plot we can infer that clients having NAME_INCOME_TYPE = 'WORKING' having the highest count of Non-defaulters i.e. 143550-15224 = 128326

Findings - XIII



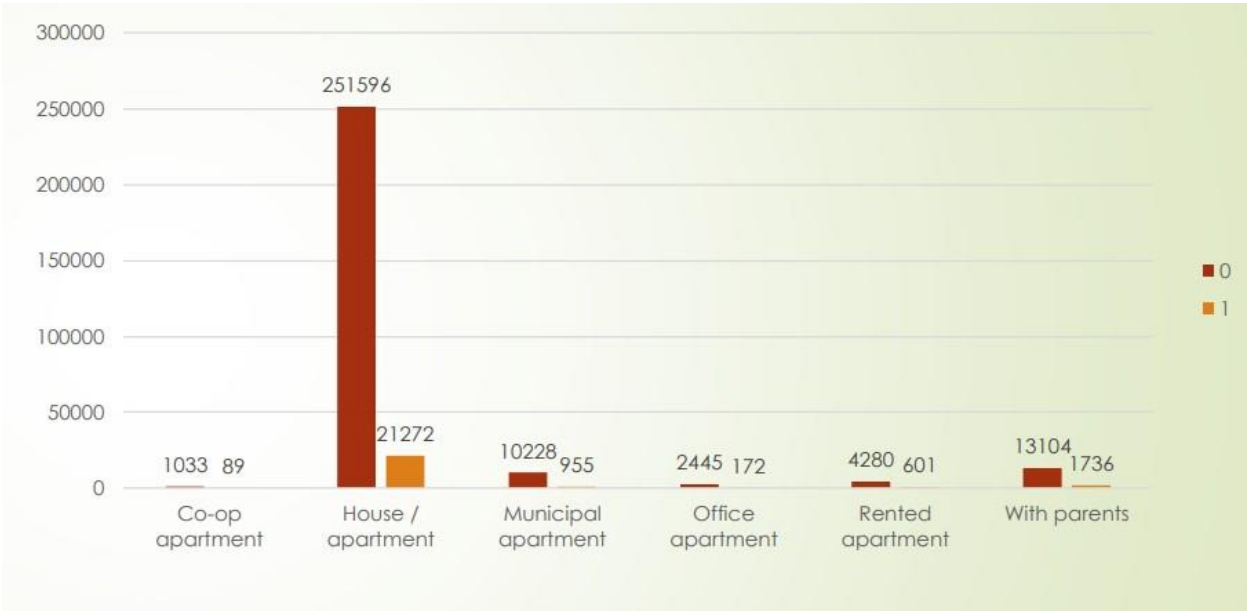
From the above Bar Plot we can infer that clients having NAME_EDUCATION_TYPE = 'SECONDARY/SECONDARY SPECIAL' have the highest count for Non-defaulters i.e.198867-19524 = 179343

Findings - XIII



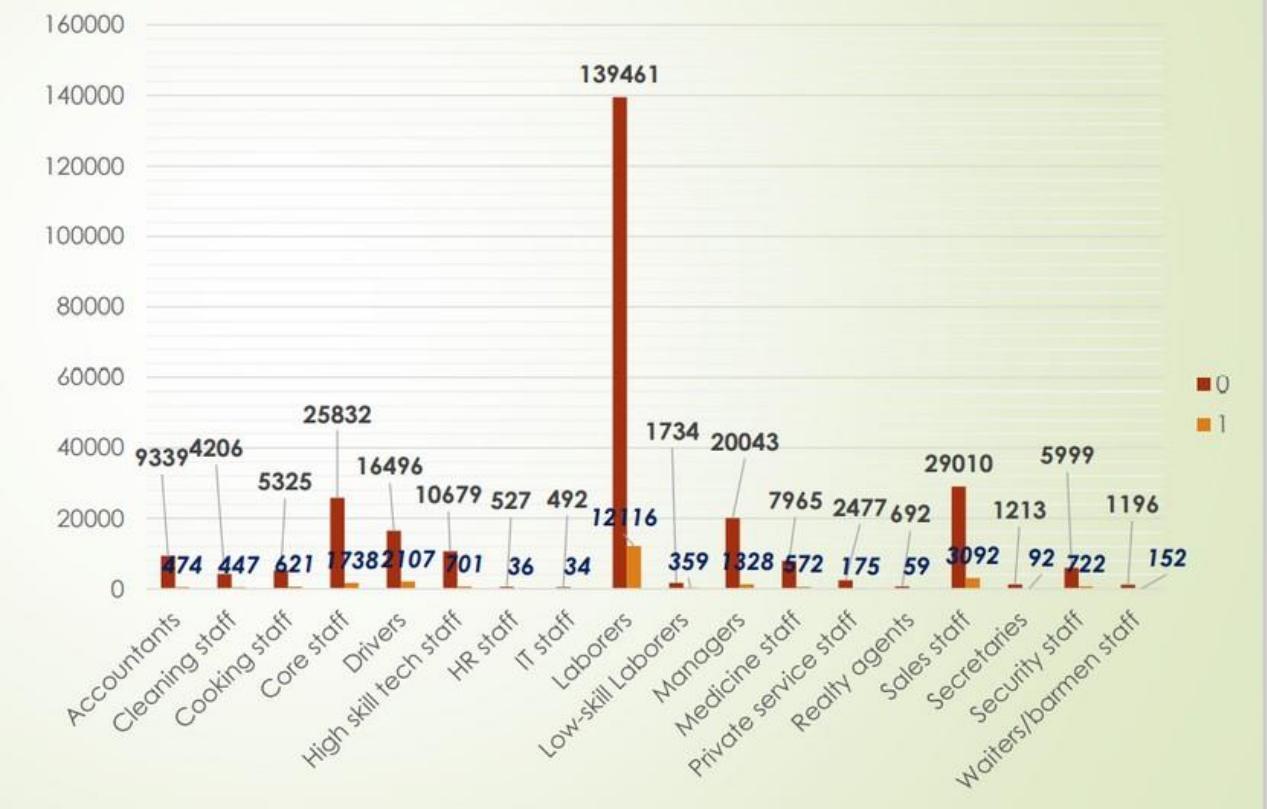
From the adjacent Bar Plot we can infer that clients having NAME_FAMILY_STATUS = 'MARRIED' have the highest count of Non_defaulters i.e.181582 – 14850 = 166732

Findings - XIV



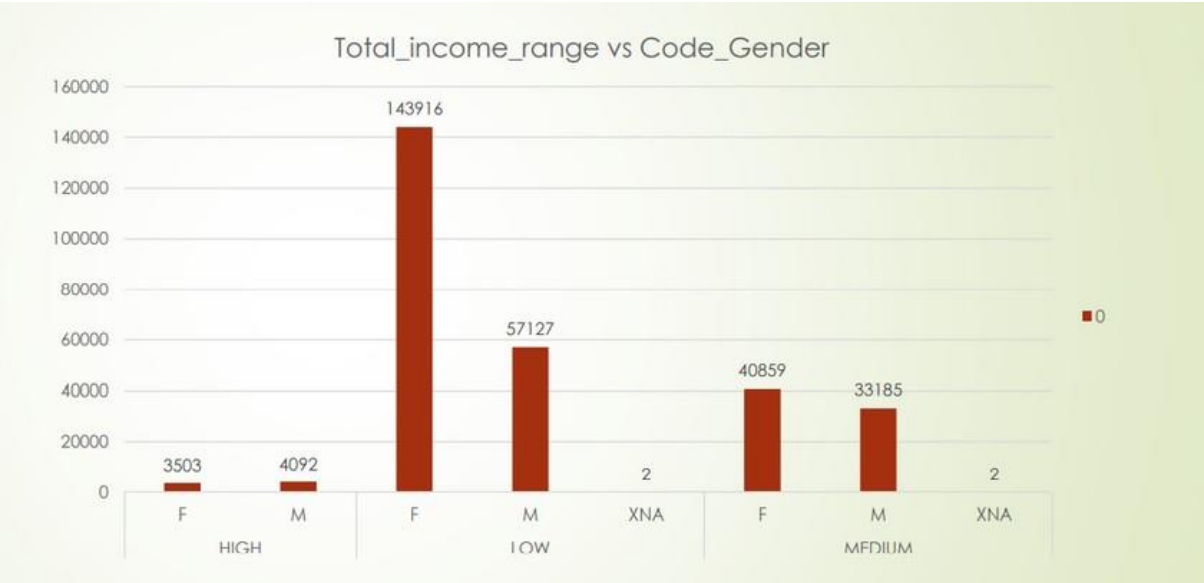
From the above Bar Plot we can infer that clients having NAME_HOUSING_TYPE = 'House/Apartment' have the highest count of Non-defaulters i.e. 251596-21272 = 230324

Findings - XV



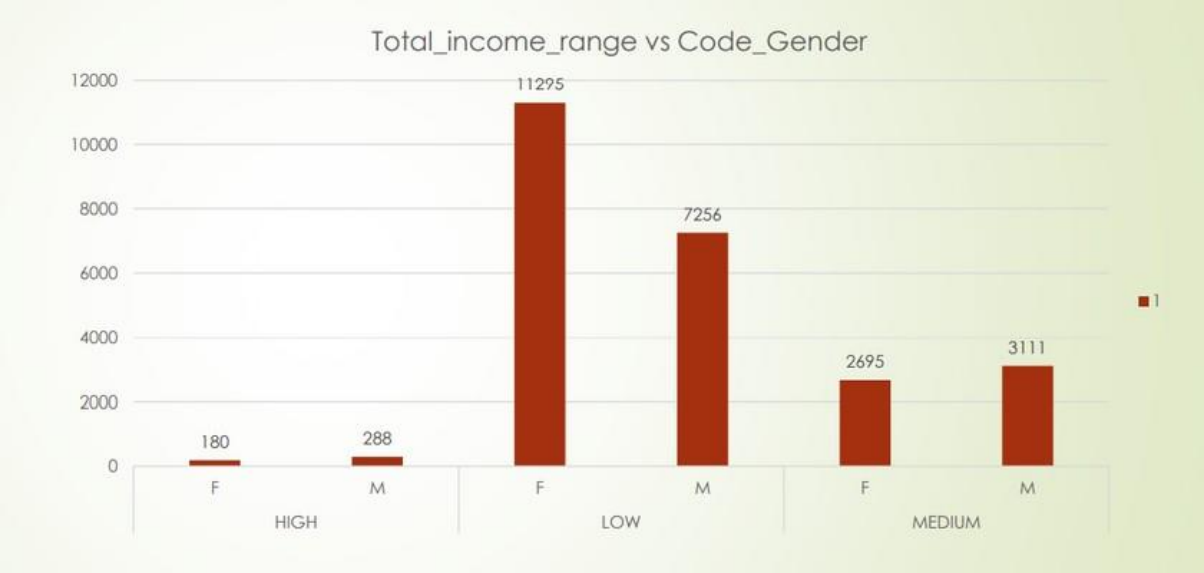
From the adjacent Bar plot we can infer that clients having occupation_type = 'Laborers' have the highest count for Non-defaulters i.e.139461-12116 = 127345

Findings - XVI



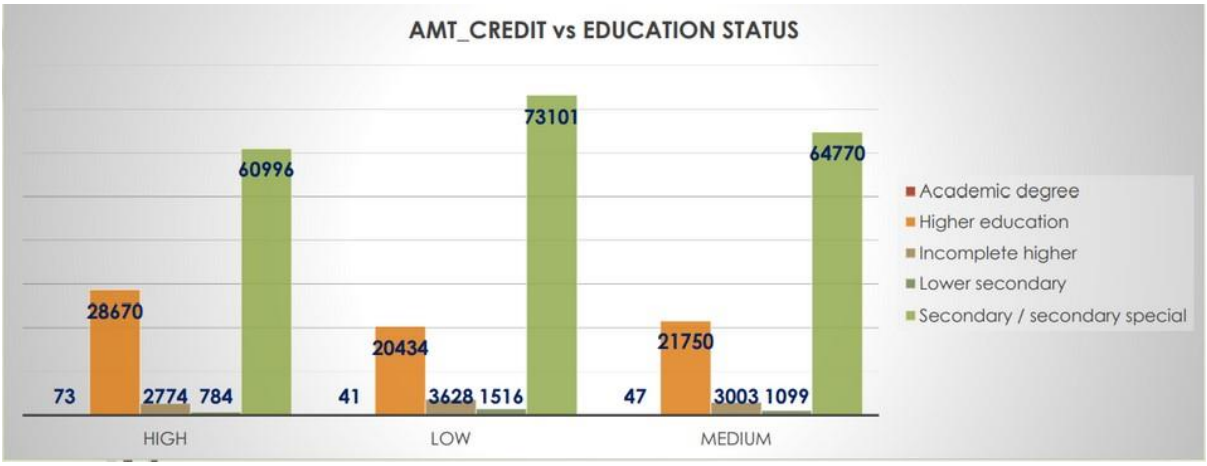
From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with no payment issues.

Findings - XVII



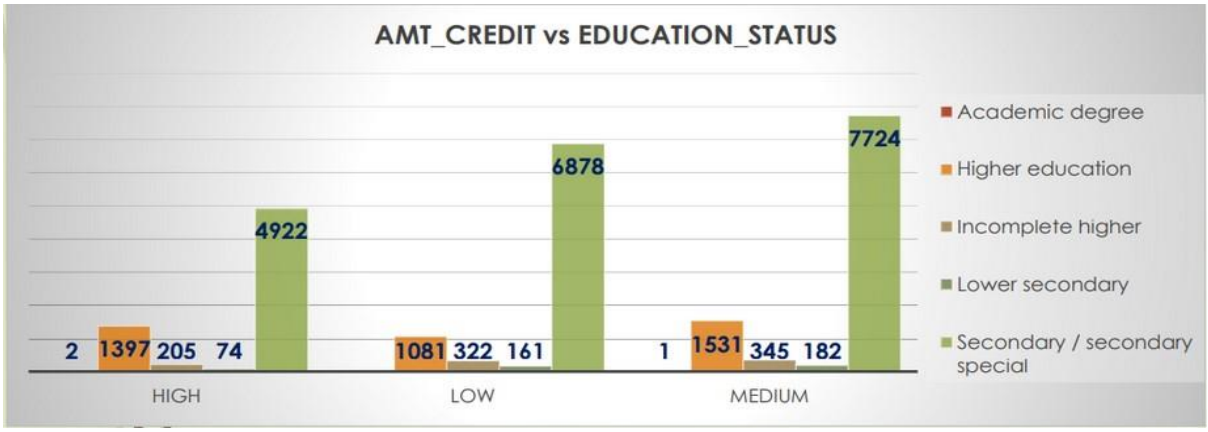
From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with payment issues

Findings - XVIII



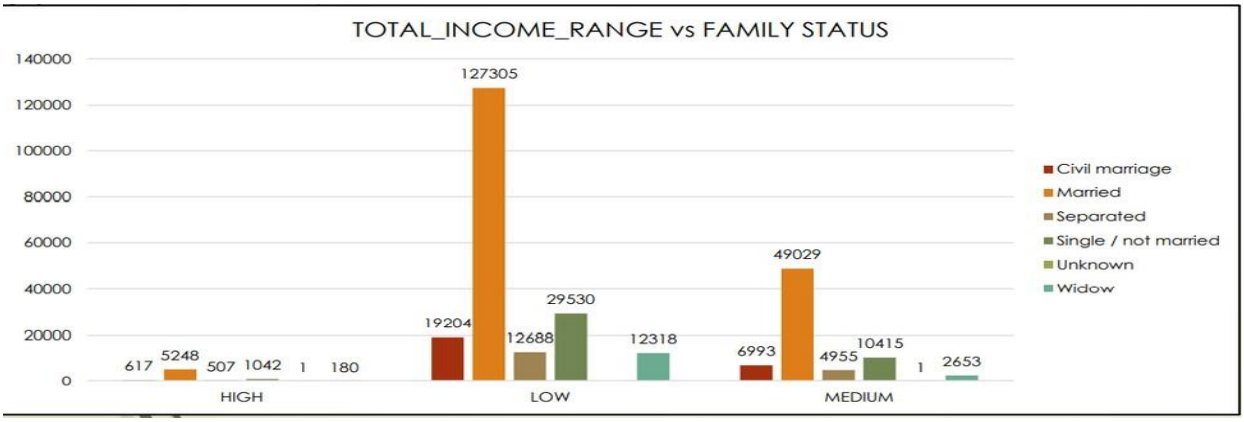
From the above Bar Plot we can infer that clients having credit amt range as ‘Low’ and education status as ‘Secondary/ Secondary Special’ have the highest count for clients with no payment issues

Findings - XIX



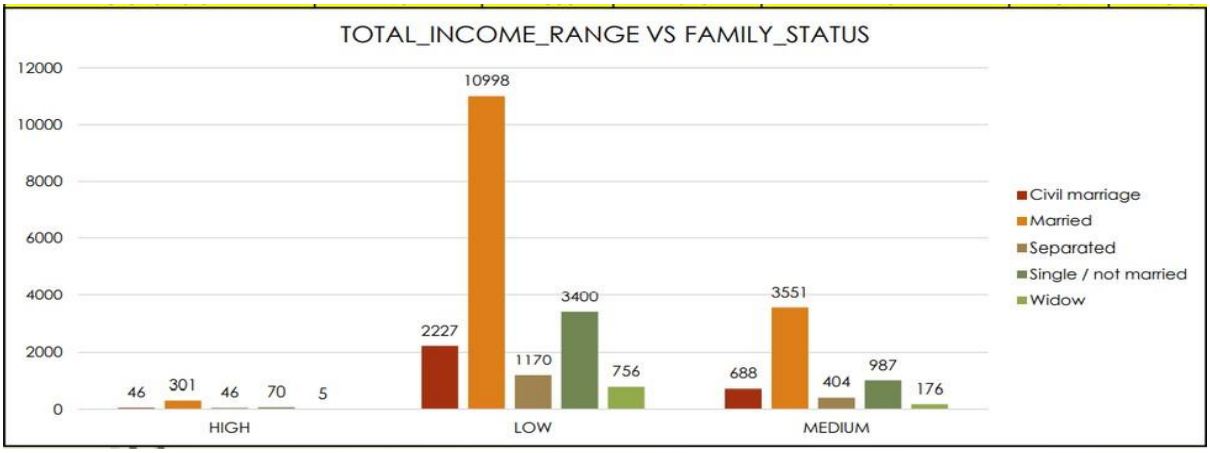
From the above Bar Plot we can infer that clients having credit amt range as 'Medium' and education status as 'Secondary/ Secondary Special' have the highest count for clients with payment issues

Findings - XX



From the above Bar plot we can infer that clients with total_income_range as 'Low' and family_status as 'Married' have the highest count for clients having no payment issues.

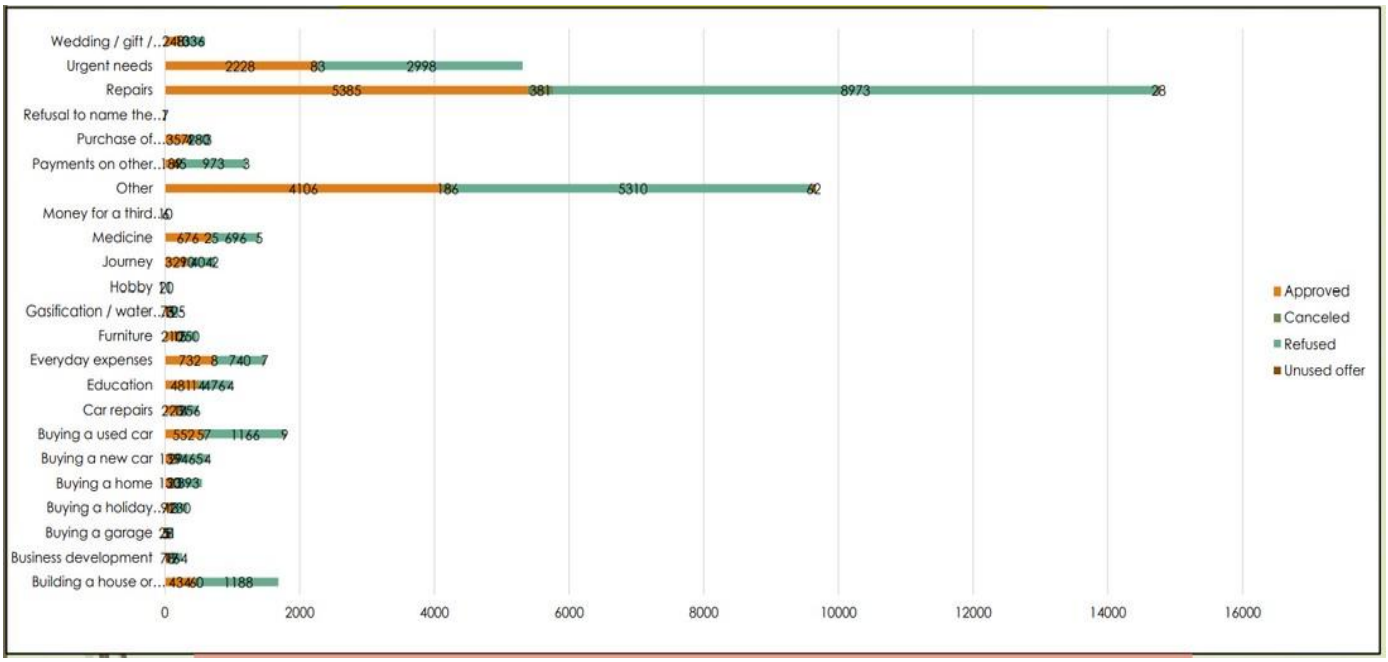
Findings - XXI



From the adjacent Bar plot we can infer that clients with total_income_range as ‘Low’ and family_status as ‘Married’ have the highest count for clients having payment issues

Findings - XXII

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



From the above Table and Bar Plot we can infer that Name of Contract status i.e. Repairs work has the highest count of Approved Loans.

Analysis:

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

- ❖ Why is it that the target_variable is of so much importance?

---> In this dataset target_variable represents whether the client had some payment issues(1) or the client didn't had some payment issues(0); It is important because the target_variable decides whether the bank should increase/decrease it's 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 and it has very less or no Non-preforming Accounts.

- ❖ 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 rates; 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

- ❖ 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, Coop Apartment, Rented Apartment, with Parents are in the search of their own house of their own name plate; Also now a days 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

- ❖ 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

- ❖ 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/ parlor services and they usually enjoy benefit from government schemes for such purpose.

Conclusion:

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

- The proportion/percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92% The Bank generally lends more loan to Female clients as compared to Males clients as the count of Female clients in the defaulter's list is less than that of Males. Still Bank can look for more Male clients if their credit amount is satisfied
- Also, the clients who belong to Working class tend to pay their loans on time followed by the clients who fall under Commercial Associate
- Clients having Education status like Secondary/ Higher Secondary or more tend to pay loan on time so bank can prefer lending loans to clients having such Education Status
- Clients who fall in the Age Group 31-40 have the highest count for paying off their loans on time followed by the clients who fall in the Age Groups 41-60
- Clients having LOW credit amount range tend to pay off their loans on time than compared to HIGH and MEDIUM credit range Clients living with their Parents tend to pay off their loans quickly as compared to other housing type. So Bank can lend loan to clients having housing type → Living with Parents
- Clients taking loan for purchasing New Home i.e. clients taking Home Loans or purchasing New Car i.e. Car Loans and clients who have a income type as State Servant tend to pay their loans on time and hence Bank should prefer clients having such background
- The Bank should be more cautious when lending money to clients with Repairs purpose because they have high count of Defaulters along with High count of Defaulters.

Impact of Car Feature

Description:

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

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

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

This problem could be approached by analyzing 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.

Problem:

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.

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.

Now for the Next portion of the Project, you need to create the Interactive Dashboard.

Use filters and slicers to make the chart interactive. The client has requested these questions given below:

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?

Findings:

Before diving into the analysis of the given dataset, it is important to perform thorough data cleaning to ensure accurate and reliable results.

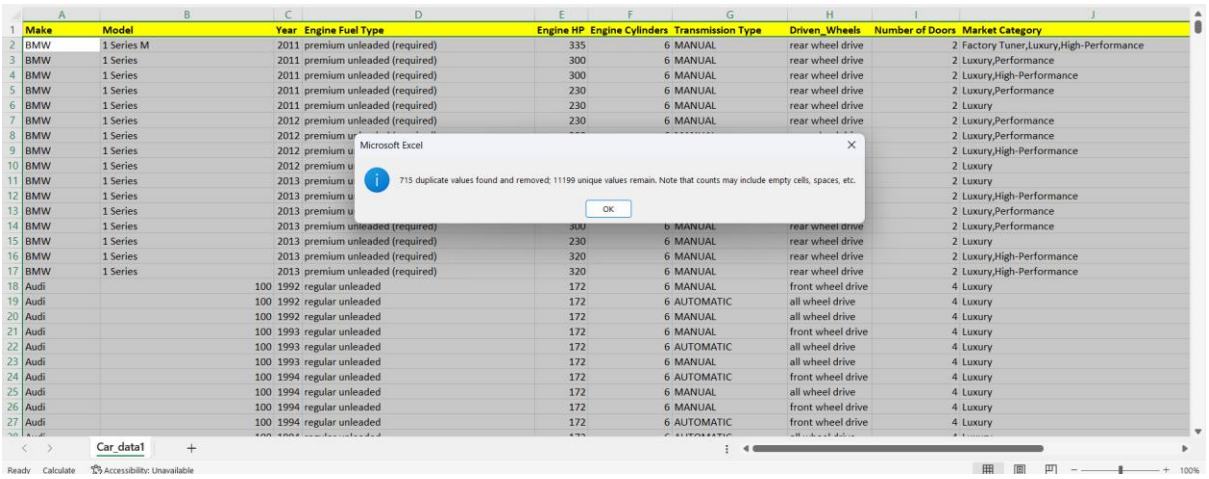
1)Cleaning the Dataset:

At first, removed all the rows which were empty. Found out the number of blank cells in the particular column. To find the blank values we used COUNTBLANK function in Excel.

After using the formula, we found the data to be mostly in good shape as there were hardly any null values in the column.

Checking null values	
Make	0
Model	0
Year	0
Engine Fuel Type	3
Engine HP	69
Engine cylinders	30
Transmission Type	0
Driven_wheels	0
Numbers of Doors	6
Market Category	0
Vechicle size	0
Vechicle Style	0
city mpg	0
Popularity	0
MSRP	0

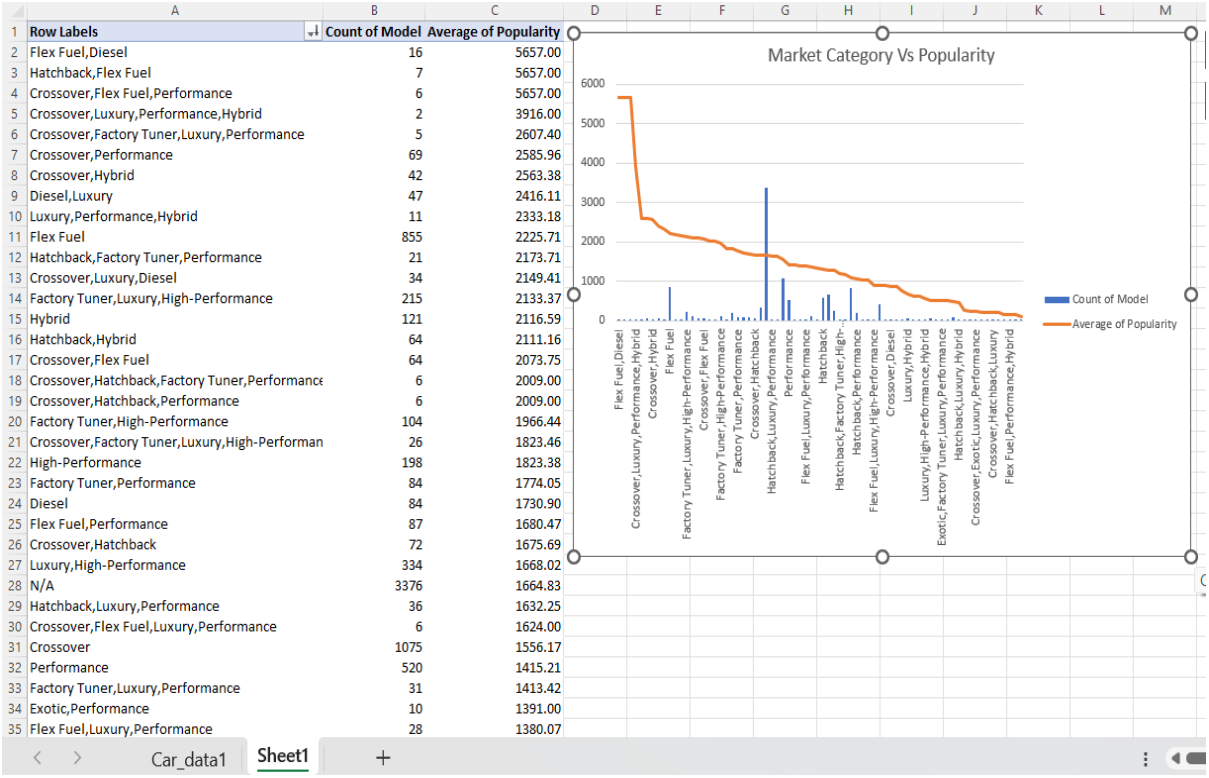
Next, we removed all the duplicate values in the dataset. Here, is the excel result for that.



Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category
BMW	1 Series M	2011	premium unleaded (required)	335	6	MANUAL	rear wheel drive	2	Factory Tuner,Luxury,High-Performance
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury
BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance
BMW	1 Series	2012	premium u					2	Luxury,Performance
BMW	1 Series	2012	premium u					2	Luxury,High-Performance
BMW	1 Series	2012	premium u					2	Luxury
BMW	1 Series	2013	premium u					2	Luxury
BMW	1 Series	2013	premium u					2	Luxury,High-Performance
BMW	1 Series	2013	premium u					2	Luxury,Performance
BMW	1 Series	2013	premium unleaded (required)	300		MANUAL	rear wheel drive	2	Luxury,Performance
BMW	1 Series	2013	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury
BMW	1 Series	2013	premium unleaded (required)	320	6	MANUAL	rear wheel drive	2	Luxury,High-Performance
BMW	1 Series	2013	premium unleaded (required)	320	6	MANUAL	rear wheel drive	2	Luxury,High-Performance
Audi		100	1992 regular unleaded	172	6	MANUAL	front wheel drive	4	Luxury
Audi		100	1992 regular unleaded	172	6	AUTOMATIC	all wheel drive	4	Luxury
Audi		100	1992 regular unleaded	172	6	MANUAL	all wheel drive	4	Luxury
Audi		100	1993 regular unleaded	172	6	MANUAL	front wheel drive	4	Luxury
Audi		100	1993 regular unleaded	172	6	AUTOMATIC	all wheel drive	4	Luxury
Audi		100	1993 regular unleaded	172	6	MANUAL	all wheel drive	4	Luxury
Audi		100	1994 regular unleaded	172	6	AUTOMATIC	front wheel drive	4	Luxury
Audi		100	1994 regular unleaded	172	6	MANUAL	all wheel drive	4	Luxury
Audi		100	1994 regular unleaded	172	6	MANUAL	front wheel drive	4	Luxury
Audi		100	1994 regular unleaded	172	6	AUTOMATIC	front wheel drive	4	Luxury

Project Insights

Insight 1-

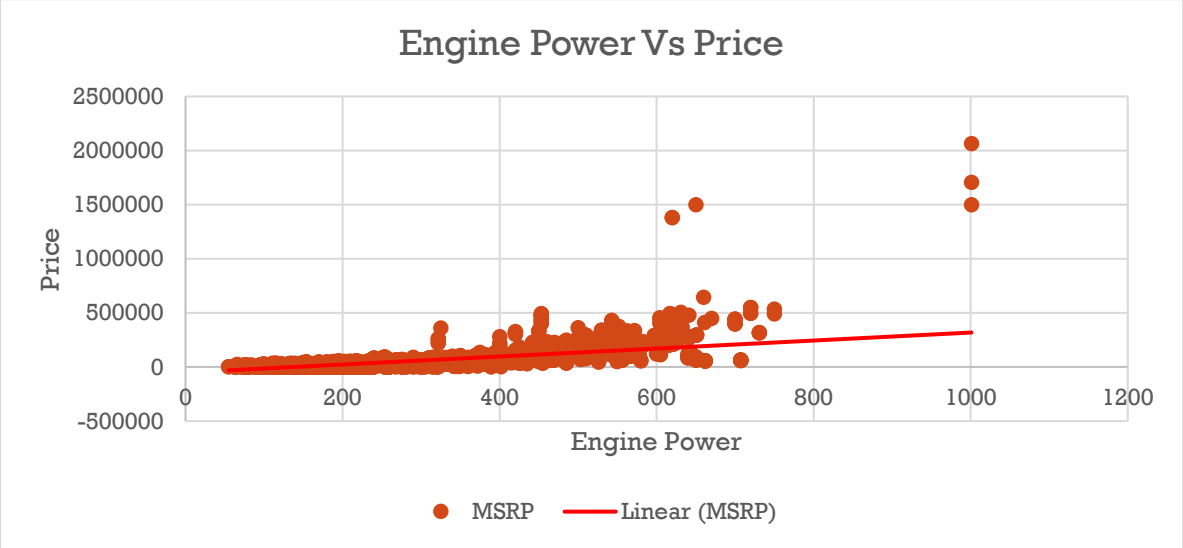


Row Labels	Count of Model	Average of Popularity
Flex Fuel,Diesel	16	5657.00
Hatchback,Flex Fuel	7	5657.00
Crossover,Flex Fuel,Performance	6	5657.00
Crossover,Luxury,Performance,Hybrid	2	3916.00
Crossover,Factory Tuner,Luxury,Performance	5	2607.40
Crossover,Performance	69	2585.96
Crossover,Hybrid	42	2563.38
Diesel,Luxury	47	2416.11
Luxury,Performance,Hybrid	11	2333.18
Flex Fuel	855	2225.71
Hatchback,Factory Tuner,Performance	21	2173.71
Crossover,Luxury,Diesel	34	2149.41
Factory Tuner,Luxury,High-Performance	215	2133.37
Hybrid	121	2116.59
Hatchback,Hybrid	64	2111.16
Crossover,Flex Fuel	64	2073.75
Crossover,Hatchback,Factory Tuner,Performance	6	2009.00
Crossover,Hatchback,Performance	6	2009.00
Factory Tuner,High-Performance	104	1966.44
Crossover,Factory Tuner,Luxury,High-Performan	26	1823.46
High-Performance	198	1823.38
Factory Tuner,Performance	84	1774.05
Diesel	84	1730.90
Flex Fuel,Performance	87	1680.47
Crossover,Hatchback	72	1675.69
Luxury,High-Performance	334	1668.02
N/A	3376	1664.83
Hatchback,Luxury,Performance	36	1632.25
Crossover,Flex Fuel,Luxury,Performance	6	1624.00
Crossover	1075	1556.17
Performance	520	1415.21
Factory Tuner,Luxury,Performance	31	1413.42
Exotic,Performance	10	1391.00
Flex Fuel,Luxury,Performance	28	1380.07

Most Popular market Category – Flex Fuel, Diesel |Hatchback, Flex Fuel | Crossover, Flex Fuel, Performance

Least Popular Market Category – Exotic, Luxury | Flex Fuel, Hybrid

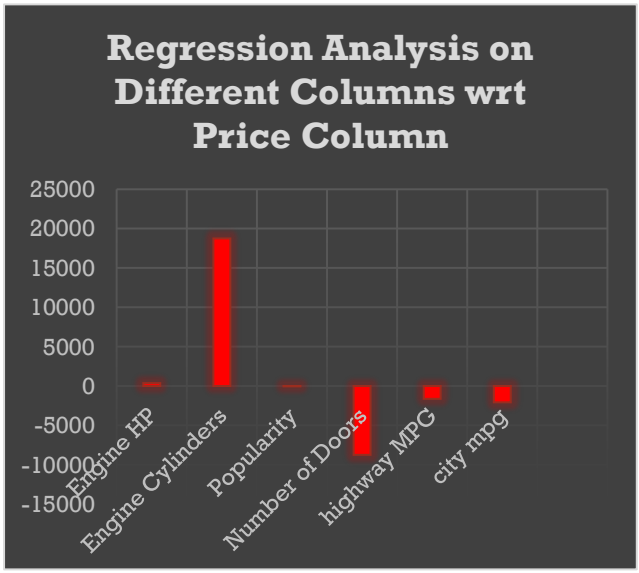
Insight 2-

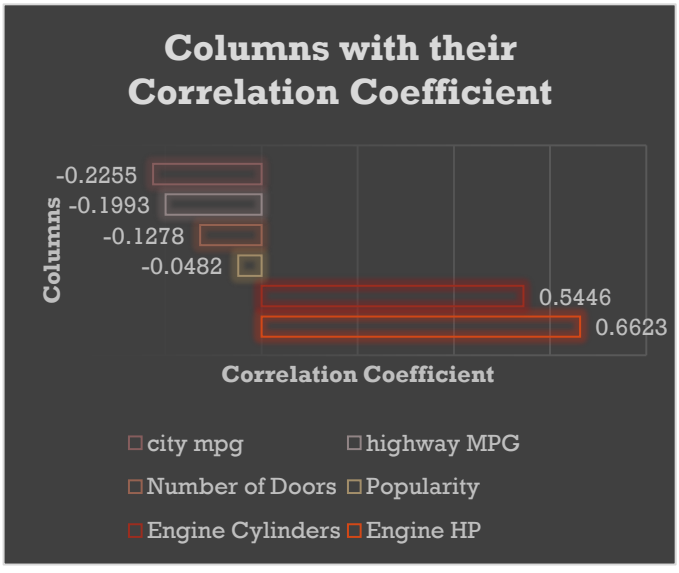


The Price will increase with the increasing number of Engine Power.

Insight 3-

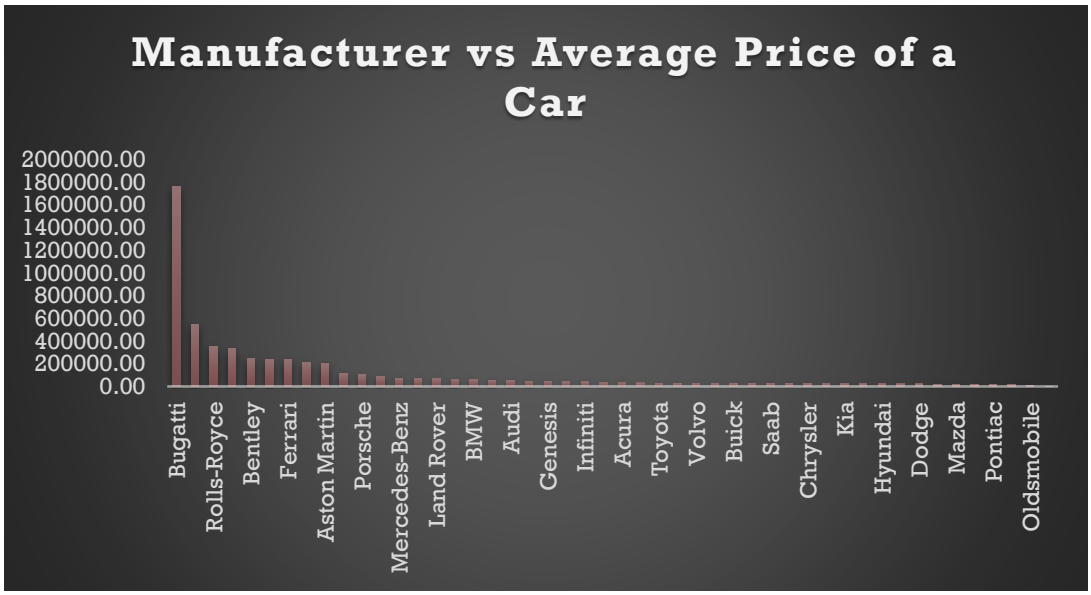
Columns	Regression Analysis (On Price) - b	a	Correlation Coefficient
Engine HP	365.6216009	-50651.97909	0.6623
Engine Cylinders	18745.09728	-65334.84992	0.5446
Popularity	-2.018886939	43712.2643	-0.0482
Number of Doors	-8733.708219	70554.78374	-0.1278
highway MPG	-1614.95867	83081.18734	-0.1993
city mpg	-2084.370422	80860.20555	-0.2255





Engine Horse Power and Engine Cylinders are having a positive relationship with Price whereas Highway MPG, City MPG, Number of Doors, and Popularity is having a negative relationship with Price.

Insight 4-

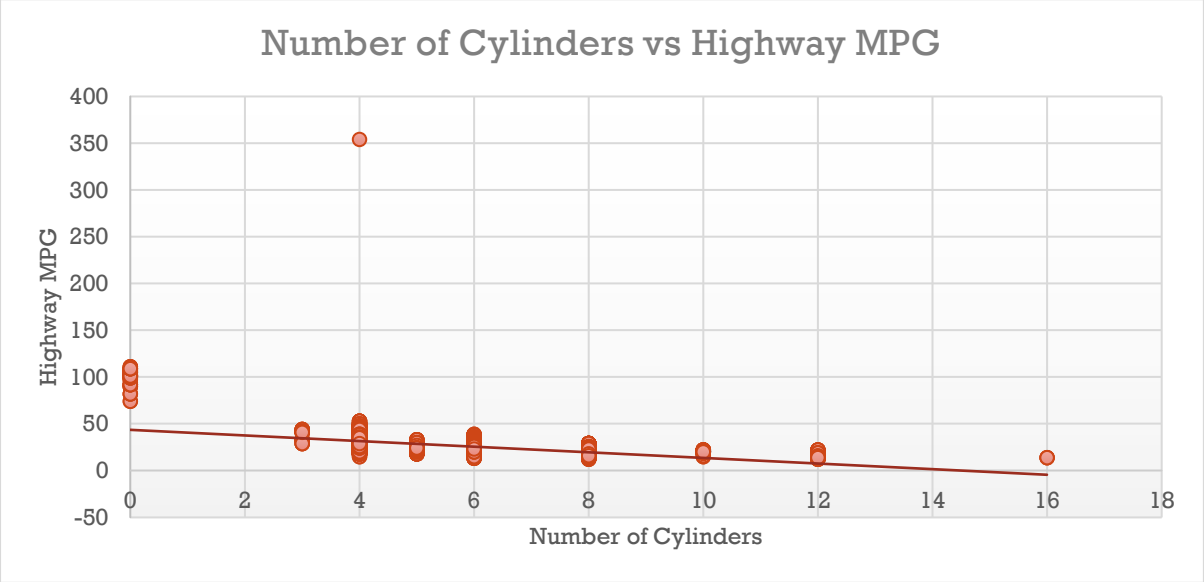


Manufacturer	Average Price of Car
Bugatti	1757223.67
Maybach	546221.88
Rolls-Royce	351130.65
Lamborghini	331567.31
Bentley	247169.32
McLaren	239805.00
Ferrari	238218.84
Spyker	213323.33
Aston Martin	197910.38
Maserati	114207.71
Porsche	101622.40
Tesla	85255.56
Mercedes-Benz	71476.23
Lotus	69188.28
Land Rover	67823.22
Alfa Romeo	61600.00
BMW	61546.76
Cadillac	56231.32
Audi	53452.11
Lexus	47549.07
Genesis	46616.67
Lincoln	42839.83
Infiniti	42394.21
HUMMER	36464.41
Acura	34887.59
GMC	30493.30
Toyota	29030.02
Nissan	28583.43
Volvo	28541.16
Chevrolet	28350.39
Buick	28206.61
Volkswagen	28102.38
Saab	27413.50
Ford	27399.27
Chrysler	26722.96
Honda	26674.34
Kia	25310.17
Subaru	24827.50
Hyundai	24597.04
FIAT	22670.24
Dodge	22390.06
Mitsubishi	21240.54
Mazda	20039.38
Scion	19932.50
Pontiac	19321.55
Suzuki	17907.21
Oldsmobile	11542.54
Plymouth	3122.90

Highest Average Price of Cars Manufacturer – Bugatti | Maybach | Rolls-Royce | Lamborghini

Lowest Average Price of Cars Manufacturer– Plymouth | Oldsmobile | Suzuki

Insight 5-

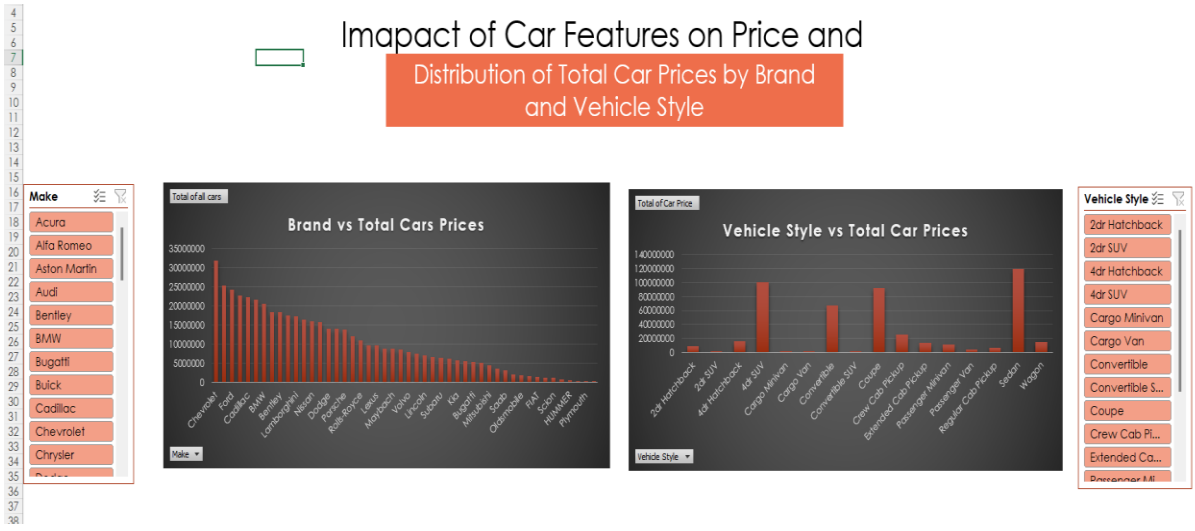


Insight-5 Task B					
Correlation coefficient between number of cylinders and highway mpg					
-0.620345938					

The Trendline for the relationship between the number of Cylinders and Highway MPG is negative which means a lesser number of cylinders will give more highway mpg.

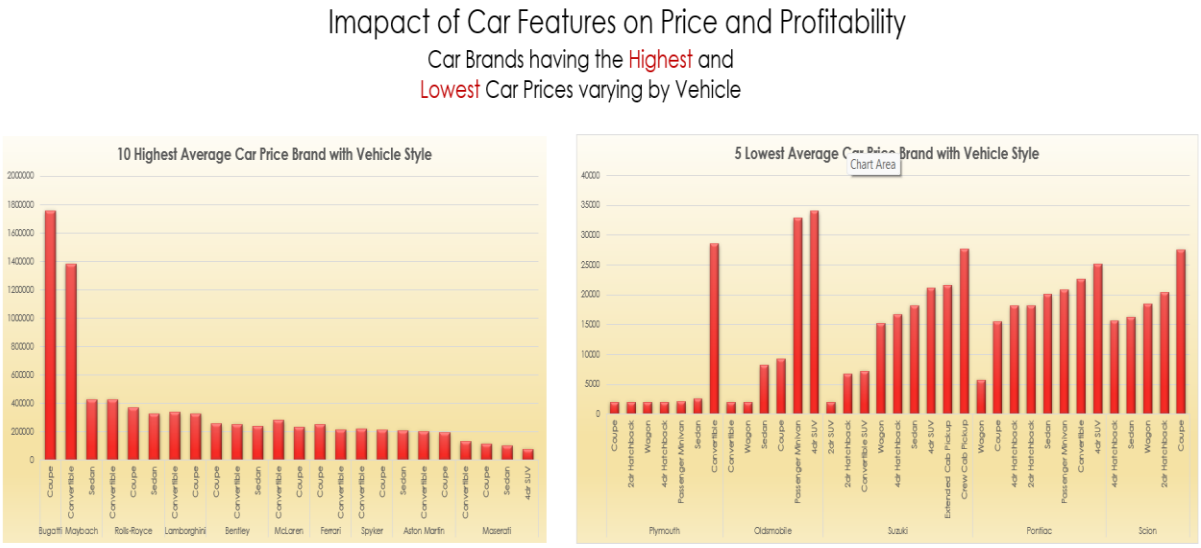
Building the Dashboards:

Dashboard 1-



Chevrolet Brand and Sedan Vehicle Types will likely have greater MSRP.

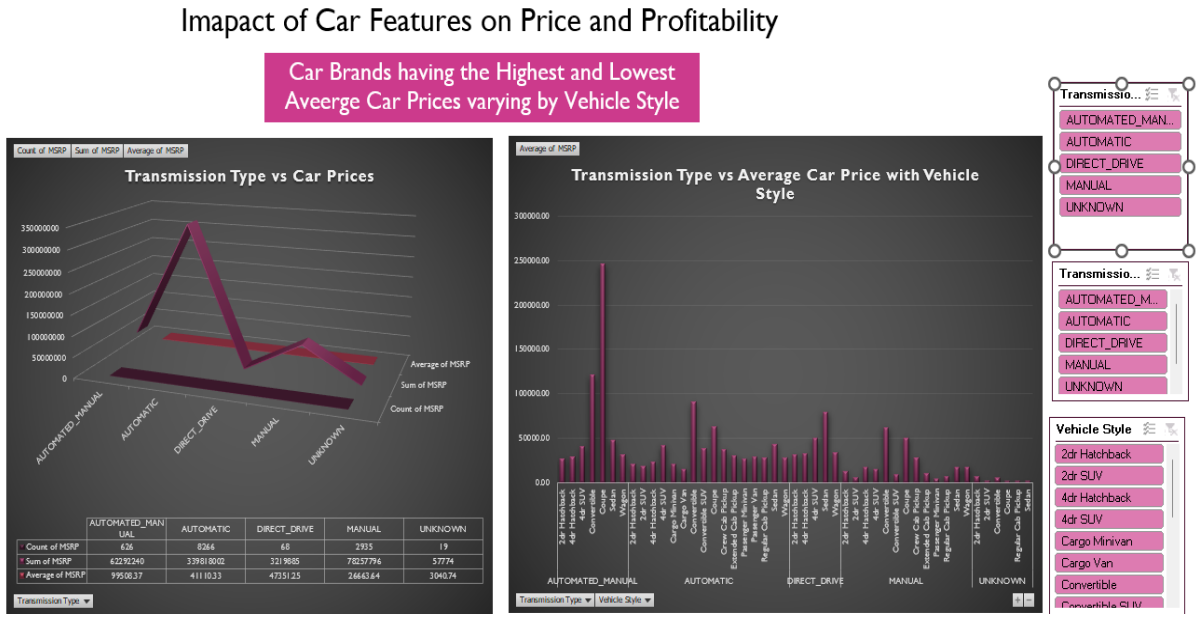
Dashboard 2-



The Highest Average Car Price Brand with Vehicle Style is Bugatti's Coupe and Maybach's Convertible.

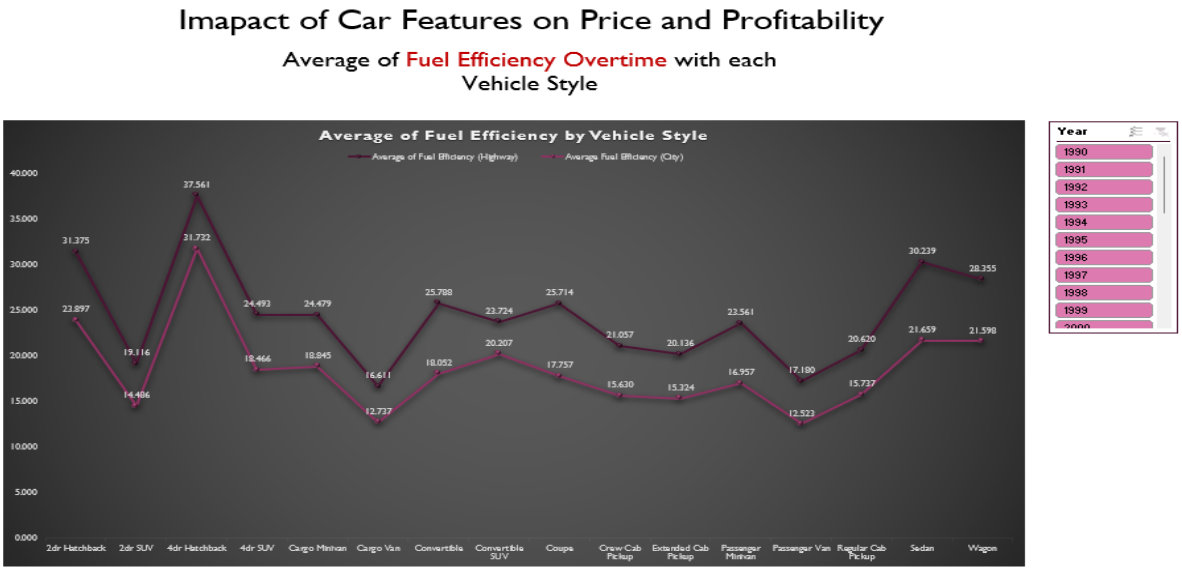
The Lowest Average Car Price Brand with Vehicle Style is Plymouth's Coupe & 2dr hatchback and Oldsmobile's Wagon & Convertible.

Dashboard 3-



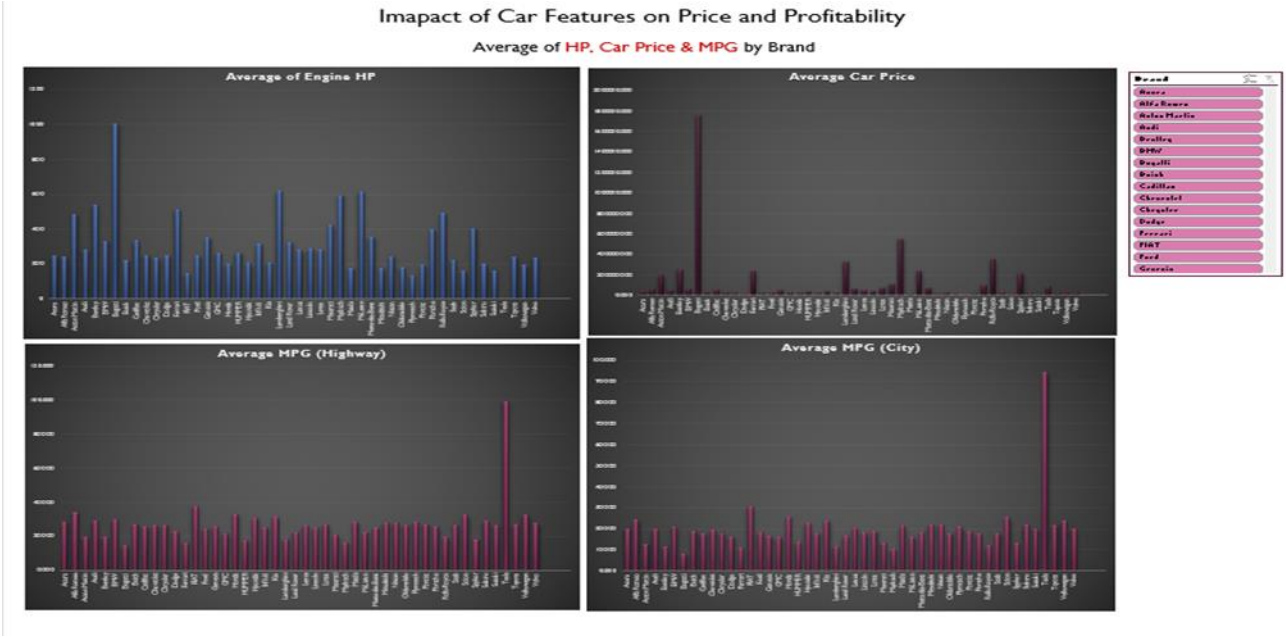
Upon analysis, we found out that Automated_Manual has the highest Average MSRP while Manual is having lowest Average MSRP.

Dashboard 4-



4dr Hatchback's fuel efficiency is the highest whereas Cargo Van and Passenger Van's fuel efficiency is the lowest.

Dashboard 5-



Bugatti is having the highest Engine HP and Car Price while Tesla is having the highest MPG (both on Highway and in City).

Conclusion:

- Most Popular market Category are Flex Fuel, Diesel |Hatchback, Flex Fuel | Crossover, Flex Fuel, Performance
- Least Popular Market Category are Exotic, Luxury | Flex Fuel, Hybrid

- The Price will increase with the increasing number of Engine Power.
- Engine Horse Power and Engine Cylinders are having a positive relationship with Price whereas Highway MPG, City MPG, Number of Doors, and Popularity is having a negative relationship with Price.
- Highest Average Price of Cars Manufacturer are Bugatti | Maybach | Rolls-Royce | Lamborghini
- Lowest Average Price of Cars Manufacturer are Plymouth | Oldsmobile |Suzuki
- Relationship between the number of Cylinders and Highway MPG is negative which means a lesser number of cylinders will give more highway mpg.
- Sedan Vehicle Types will likely have greater MSRP.
- Chevrolet Brand will likely have greater MSRP.
- The Highest Average Car Price Brand with Vehicle Style is Bugatti's Coupe and Maybach's Convertible.
- The Lowest Average Car Price Brand with Vehicle Style is Plymouth's Coupe & 2dr hatchback and Oldsmobile's Wagon & Convertible.
- Automated_Manual has the highest Average MSRP while Manual is having lowest Average MSRP.
- 4dr Hatchback's fuel efficiency is the highest whereas Cargo Van and Passenger Van's fuel efficiency is the lowest.
- Bugatti is having the highest Engine HP and Car Price while Tesla is having the highest MPG (both on Highway and in City).

ABC Call Volume Trend Analysis

Description:

A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. Typically, these teams fulfil various roles and responsibilities such as: Customer experience programs (CX programs), Digital customer experience, Design and processes, Internal communications, Voice of the customer (VoC), User experiences, Customer experience management, Journey mapping, Nurturing customer interactions, Customer success, Customer support, Handling customer data, Learning about the customer journey.

Let's look at some of the most impactful AI-empowered customer experience tools you can use today: Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, Intelligent Routing

In a Customer Experience team there is a huge employment opportunity for Customer service representatives A.k.a. call centre agents, customer service agents. Some of the roles for them include: Email support, Inbound support, Outbound support, social media support.

Inbound customer support is defined as the call centre which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of the existing customers or prospective customers for your business which are attended by customer care representatives.

Inbound customer service is the methodology of attracting, engaging, and delighting your customers to turn them into your business' loyal advocates. By solving your customers' problems and helping them achieve success using your product or service, you can delight your customers and turn them into a growth engine for your business

Problem:

Analysis done on the following points: -

A. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

B. Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]. You can select time in a bucket form (i.e., 1-2, 2-3, ….)

C. As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e., You have to calculate minimum number of agents required in each time bucket so that at least 90 calls should be answered out of 100.)

D. Let's say customers also call this ABC insurance company in night but didn't get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am]. Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

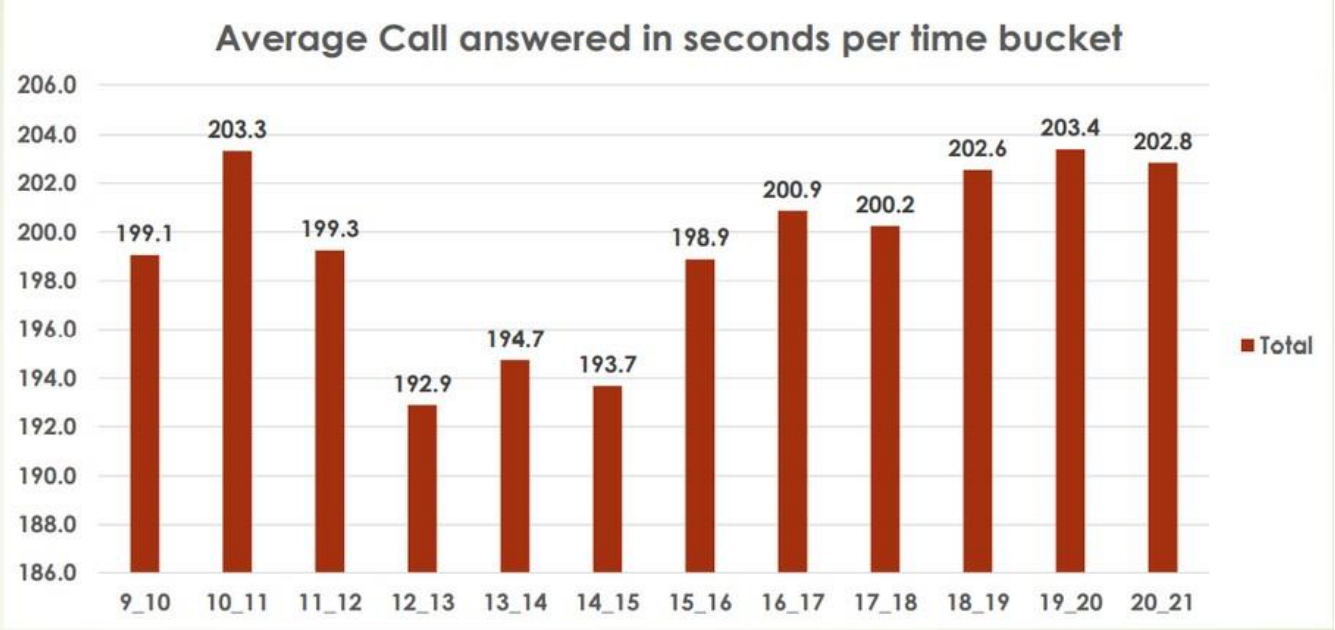
Assumption: An agent work for 6 days a week; On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into

lunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e 60% of 7.5 Hrs) on call with customers/ users. Total days in a month is 30 days.

Findings:

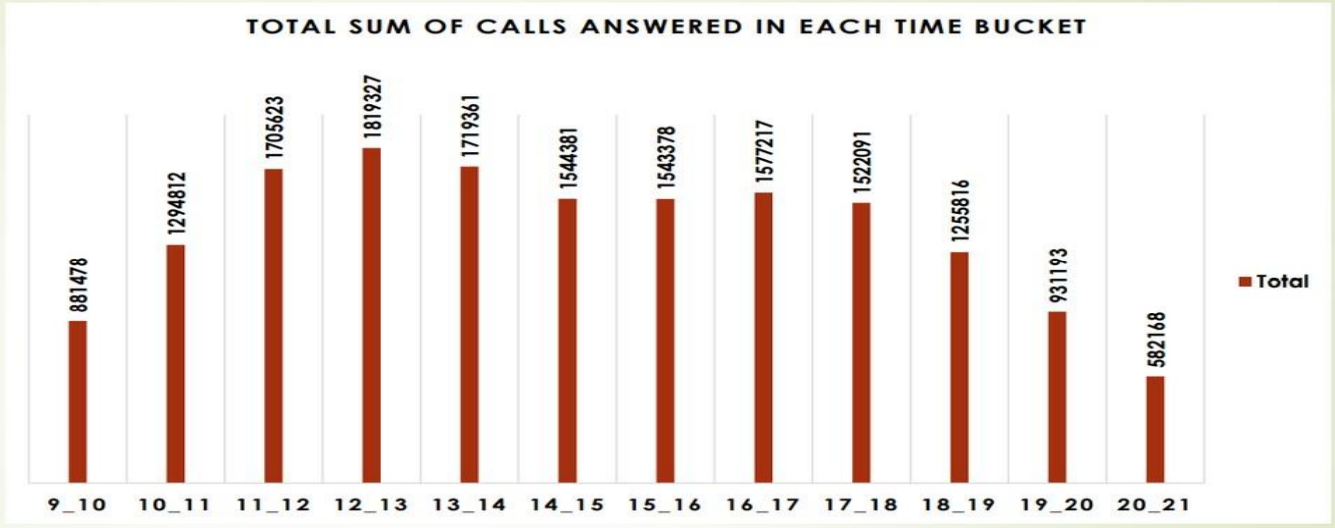
Findings – I

From the above bar plot we can infer that time_bucket 19_20 i.e. 7PM to 8PM had the



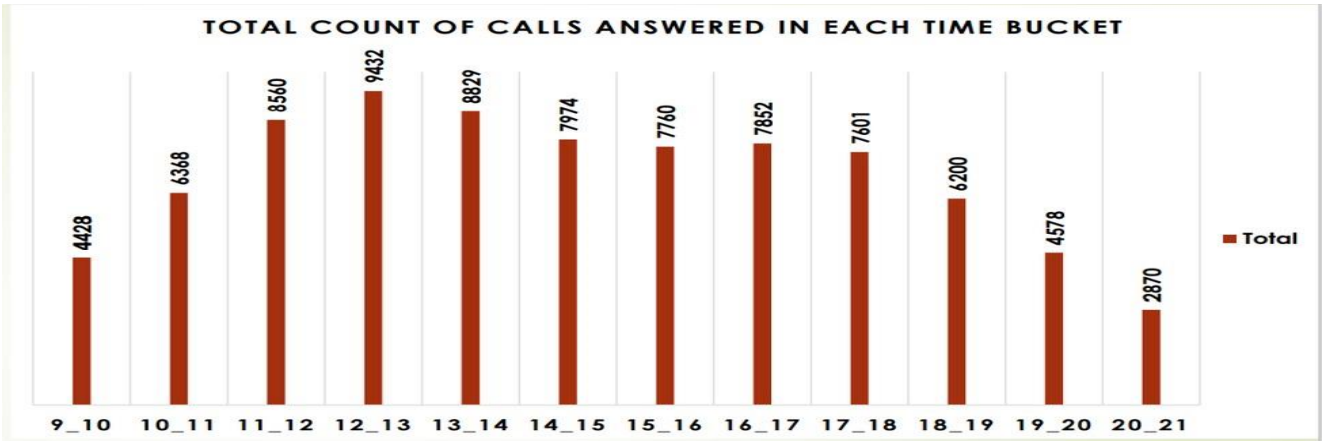
highest of average of calls answered in seconds i.e. 203.4

Findings – II



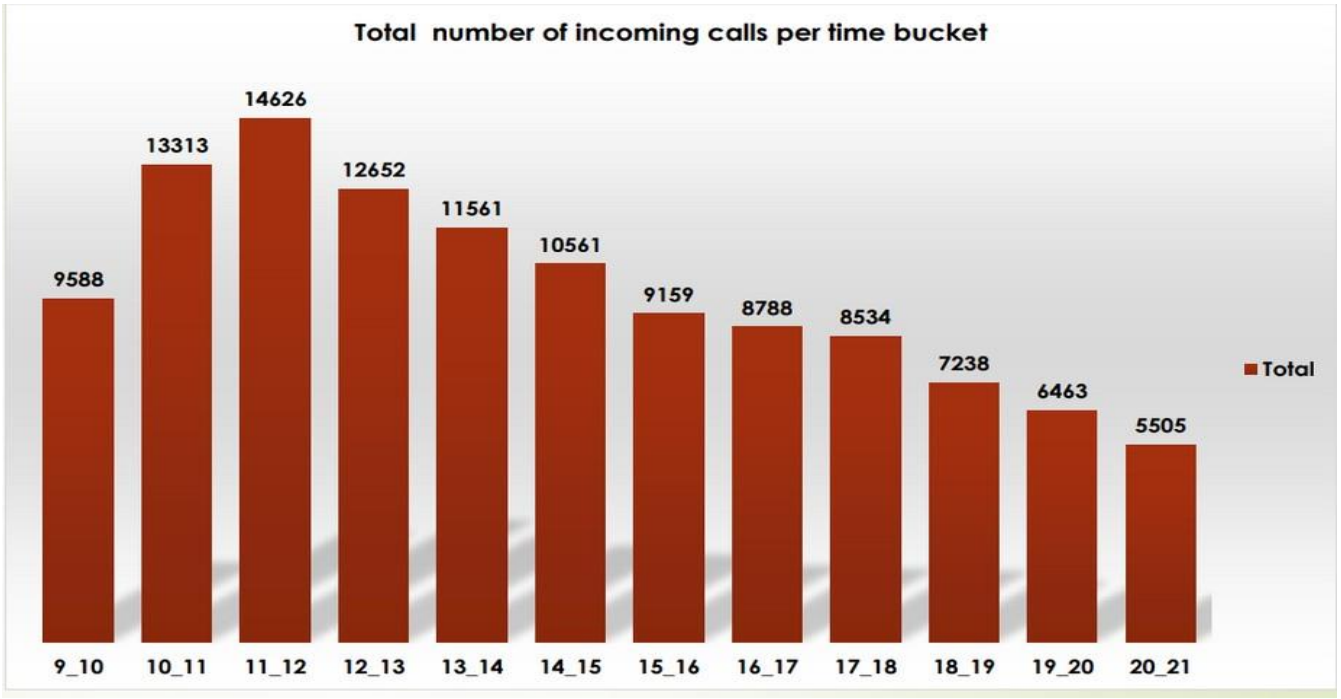
From the above Bar plot we can infer that the time_bucket 12_13 i.e. during the time period 12PM to 1PM had the highest total number of calls answered i.e. 1819327.

Findings – III



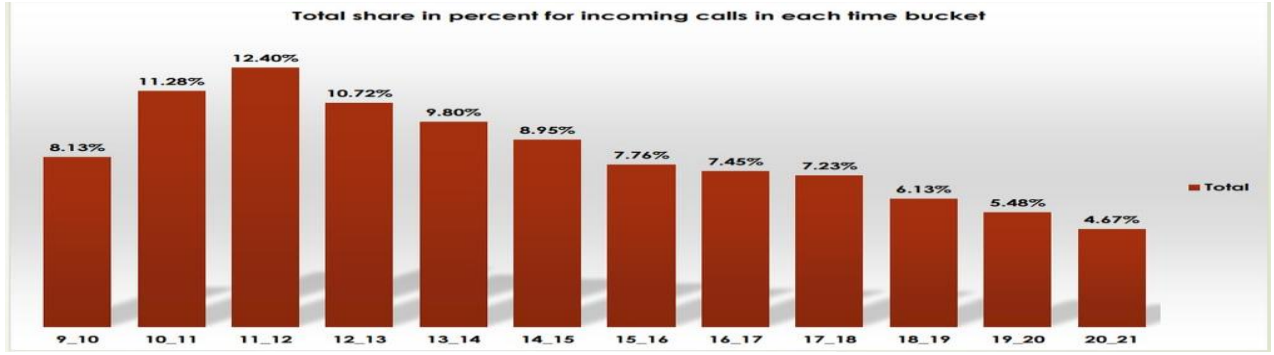
From the above bar plot we can infer that the time_bucket 12-13 i.e. 12PM to 1PM had the highest count of calls answered i.e. 9432

Findings – IV



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

Findings – V



From the above bar plot we can infer that the time bucket 11_12 i.e. 11 AM to 12 PM has the highest share for incoming calls i.e. 12.40%

Findings – VI

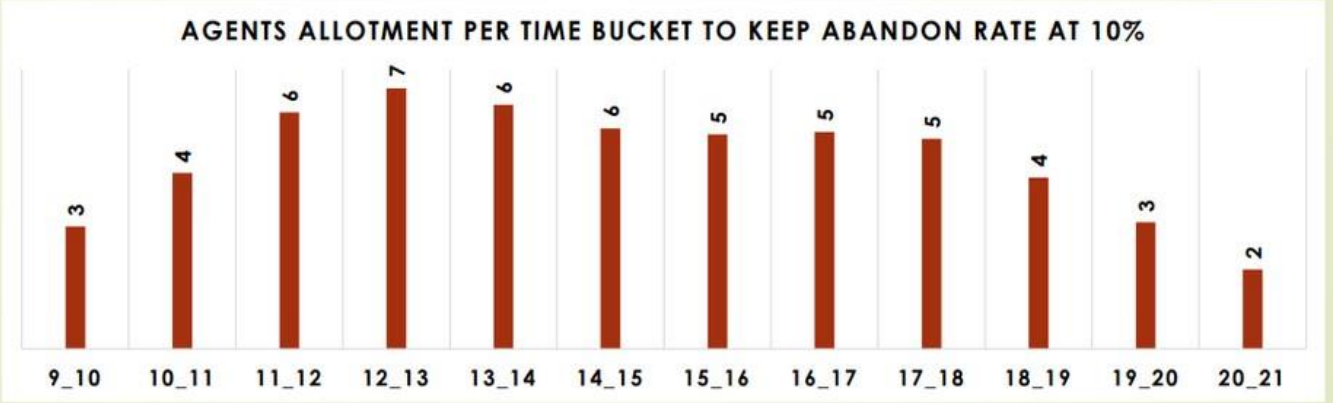
Count of Call_Status	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
01-Jan	684	3883	77	4644
02-Jan	356	2935	60	3351
03-Jan	599	4079	111	4789
04-Jan	595	4404	114	5113
05-Jan	536	4140	114	4790
06-Jan	991	3875	85	4951
07-Jan	1319	3587	42	4948
08-Jan	1103	3519	50	4672
09-Jan	962	2628	62	3652
10-Jan	1212	3699	72	4983
11-Jan	856	3695	86	4637
12-Jan	1299	3297	47	4643
13-Jan	738	3326	59	4123
14-Jan	291	2832	32	3155
15-Jan	304	2730	24	3058
16-Jan	1191	3910	41	5142
17-Jan	16636	5706	5	22347
18-Jan	1738	4024	12	5774
19-Jan	974	3717	12	4703
20-Jan	833	3485	4	4322
21-Jan	566	3104	5	3675
22-Jan	239	3045	7	3291
23-Jan	381	2832	12	3225
Grand Total	34403	82452	1133	117988
Avg calls on daily basis	1496	3585	49	5130
% of Avg calls on daily basis	29%	70%	1%	

From the table we can infer that the current abandon rate is around 30%

Findings – VII

The distribution of manpower plan per time bucket to keep abandon rate at 10% i.e. keeping call answered rate at 90% is as follows:-

Call_Status	answered	
Row Labels	Count of Customer_Phone_No	Agents allotment
9_10	4428.0	3
10_11	6368.0	4
11_12	8560.0	6
12_13	9432.0	7
13_14	8829.0	6
14_15	7974.0	6
15_16	7760.0	5
16_17	7852.0	5
17_18	7601.0	5
18_19	6200.0	4
19_20	4578.0	3
20_21	2870.0	2
Grand Total	82452.0	57



Findings – VIII

		Total hours needed		
Night time slot	Calls per slot	76.41135	Agents needed	Time distribution
21_22	3	7.641135	13	10%
22-23	3	7.641135	13	10%
23_24	2	5.09409	8	7%
00_01	2	5.09409	8	7%
01_02	1	2.547045	4	3%
02_03	1	2.547045	4	3%
03_04	1	2.547045	4	3%
04_05	1	2.547045	4	3%
05_06	3	7.641135	13	10%
06_07	4	10.18818	17	13%
07_08	4	10.18818	17	13%
08_09	5	12.735225	21	17%
Total	30	76.41135	126	100%

The table above shows the desired distribution of the night calls to keep the abandon rate at 10%

• Since we have only 17 agents during night we need to distribute in annon 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

• The 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%

Analysis:

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

- 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

- Why is it that the total number of incoming calls reached it's peak value during the time bucket 11_12 and got decreased from time bucket 12_13 onwards?

---> It is a general tendency of the customers(people) that they want their query/complaint get resolved on that particular day itself when they called the customer center; so most of the customers try to place their complaint/query before 12 Pm so that by the end of the day their complaint gets resolved depending upon the complexity of the problem faced by the customer

- Why is proportion if the monthly transfer rate is less than compared to monthly answered and abandon rate?

---> In most of the customer service centers they have the dedicated toll free number of the particular problem faced by the customer, also there are skilled people at the call center who are well versed with the problems they come across while handling and guiding thousands of customers on daily basis; And so most of the calls gets answered by providing an solution to the query, some of the calls get abandon due to unavailability or shortage of the skilled person, and very few calls gets transferred from the junior level to senior level if the problem is too complex for the junior level expertise

- 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 an 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.

Conclusion:

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

- From the previous analysis we can derive that Avg calls answered per agent is 198.6 in each time bucket
- We need to reduce the abandon rate by 30%(current) – 10%(desired) = 20% i.e. we need to increase call answered rate by 70% (current) + 20%(change) = 90% .So, we need to have 90% of the total calls to be answered so as to reduce the abandon rate to 10%
- Total avg calls incoming per day = 5130 • Avg calls answered per second = 198.6 • Answered rate = 90% i.e. 0.9 • Seconds per hour =3600 • So, time required to answer 90% of the incoming calls = 5130 * 198.6 * 0.9 / 3600 = • So, new total number of agents working per day is 255 divided by the number of hours an agent actually works(on a consumer call) i.e. 4.5 = 255/4.5 = 56.67 == 57. Agents working per day 254.7001826
- So, to have a 10% abandon rate we need 57 Agents working per day
- From the assumptions given the following points were noted:- In a day an agent work for 9 hours → Total Agent working hours = 9 HOURS
- Out of the total 9 hours , 1.5 hours goes for lunch and coffee/tea breaks; so remaining working hours = 9 – 1.5 = 7.5 HOURS Out of the remaining 7.5 hours per day an agent is occupied with consumers call for only 60% of the time i.e. 60% of 7.5 i.e. 0.6 * 7.5 = 4.5. So, an agent spends only 4.5 hours per day out of total 7.5 hours on consumer calls. An agent works 6 days a week. In a month of 30 days 6 days per week; In a month of 30 there are 4 weeks; 7 days per week means total 28 days out of which 4 days are unplanned leave
- Days of agent on floor = (20*7)/28 = 5 days. Now, total days left 28 – 4 = 24 days. Per week there is one Sunday which is an official holiday for all workplaces around the world; So in a month of 30 there are 4 Sundays. Now total days left for work = 24 -4 = 20 days. So, an agent is available to work for 20 days in a month of 30 days
- In a certain scenario there are calls from consumers not only during the day time but also during the night time and if there are no agents available during the night time to answer the call then it creates a bad impression on the consumer regarding the company Now we need to give the distribution of the total manpower available for each time bucket right from 9AM to 9 PM and then from 9 PM to 9 AM, keeping the abandon rate at 10% i.e. keeping the answered rate at 90%
- For each 100 day calls there are 30 night calls; then for 5130 day calls there will be : 5130*30/100 = 1539 night calls.
- So, there are 1539 night calls for a total of 5130 day calls
- So, the additional working hours keeping the answered rate at 90% will be 1539 * 198.6(avg calls answered per sec) * 0.9 / 3600(total seconds in each hour) = 76.41135

- So, additional agents needed by the company to answer nightcalls as well be 76.41135/4.5 = 16.98 == 17
- So, we need additional 17 agents to answer the night calls as well, making the total number of agents working per day keeping the answer rate to 90% will be 57(day call answer 90%) + 17(night call answer 90%) = 74 agents . So, we need 74 Agents per day to answer the consumer calls from day as well as the night time keeping the answered rate to 90% / Abandon rate to 10%.