Data Analysis Portfolio

Prepared By:

Nagendra Pratap Singh

Professional Background

My name is Nagendra Pratap Singh. I completed my b.tech. In 2018. After that, I am working for a joint venture company, Mahindra and Mahindra Tractor and Spare Parts Company, which is based in Kanpur. It was selling tractors and their spare parts to people and to different agencies or vendors, where I was taking care of market research and operations to increase sales. I was negotiating prices with these people, and I analyzed the target market and my competitors. The company sells tractors and their spare parts in a particular area in Kanpur, where I analysis market segments to select correct pricing, I also analysis competitors in that market, like what are the current competitors, what is the exact pricing, and I analysis sales data in that company and do the revenue analysis. I also worked there for three years. I was data-driven over there, even though I am from a mechanical background. I was working on data, and I liked it. Then I look for some courses. I am looking for a profile that is data-driven. When I do so, I realize there is more opportunity in this field. After this, I took a data science course from Internshala to make myself a better candidate in this field. Right now, I have taken a course and I am currently looking for an opportunity.

I have completed several certificates in Advanced Excel, Python, SQL, Tableau, and Data Science, which have equipped me with the necessary skills to excel in this field.During my Data Science coursework, I had the opportunity to work on a machine learning project which further developed my analytical and problem-solving skills. My background in engineering, combined with my experience in sales and operations, has provided me with a unique perspective and valuable experience in data science and interpretation.

Table Of Contents

IMDB Movies Analysis

Bank Loan Case Study

Analyzing the Impact of Car Features on Price and Profitability

ABC Call Volume Trend

Data Analytics Process

Description

Without even realising it, we utilise data analytics in daily life. Your objective is to provide an example (or examples) of a real-world scenario in which we employ data analytics and how that circumstance relates to the data analytics process.

Design

Scenario: The day after tomorrow, I have a job interview. Since the business is a sort of MNC, business-casual attire is appropriate. I want to buy from the market since I don't have a good dress code. I have to research each data analysis method used during the purchasing process. so that I may do my assignment correctly.

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

1.Plan

a) **Dress code :** The company you are interviewing with likely has a clothing code, so be aware of it. Choose a modest outfit or suit if the setting is formal or business casual. Here I am looking for an formal dress code .
b) **Fit :** Make sure the dress is flattering to your body type and that it fits you nicely. Particularly in a high-pressure setting like a job interview, it's critical to feel confident and at ease with what you're wearing.
c) **Color :** Don't stray from traditional hues like black, navy, grey, or beige. Avoid using loud patterns or colors that could be distracting.
d) **Shoes:** Select closed-toe footwear that fits your clothing appropriately and is comfy. Although you should make sure you can walk easily in them, heels are a fantastic option.

2. Prepare

a) **Examine your finances :** Consider your existing financial condition and estimate the amount you have to spend on a dress. Think about your present income, spending, and any additional debts you may have.
b) **Establish a budget:** Decide how much you're willing to spend on a dress based on your judgement.
c) **Examine your financial choices:** If you don't have enough money to pay for the dress outright, think about financing options like a personal loan or credit card. Use these choices only if you can afford the payments and can pay off the loan right away.

d) **Look for bargains:** Look for formal dresses that meet your budget via sales, discounts, or other promotions. Without spending a fortune, you might be able to discover a dress that meets your requirements.

3. Process

a) **Think about the situation:** Consider the shoes' intended use. Are you purchasing shoes for a particular event or occasion? If so, the formality or dress code of the occasion may have an impact on the shoes you wear . Here I am looking a dress code for interview.
b) **Consider your comfort:** When choosing footwear, comfort is essential. The quantity of walking or standing you will be performing should be taken into account while selecting footwear that will be comfortable and supportive enough for the activity. Mainly quantity of sitting is more as compare to walking for an interview .
c) **Think about your individual taste:** Consider your personal style when making your footwear selection. For a more casual style, you can select trainers or athletic shoes, but dress shoes or heels might be more appropriate for a formal occasion. Here I want dress shoes for interview because it is a formal occasion .
d) **Type of footwear :** After giving the aforementioned aspects some thought, decide what kind of footwear you require. Do you need sneakers for working out, slippers for around the house or dress shoes for a formal occasion ? Here I want dress shoes.

4. Analysis

It's crucial to first take into account your individual preferences and taste. While keeping up with fashion trends can be thrilling and entertaining, it's ultimately crucial to wear clothes that make you feel good about yourself. You shouldn't force yourself to wear a trend or a particular style if you don't like it. it might be useful to think about how several pieces of clothing will combine to form a coherent style. Consider whether a new t-shirt, for instance, will look good with your existing pair of jeans or trousers before making the purchase. By doing this, you can build a diverse wardrobe that allows you to combine and match different items to create a variety of looks. it's important to remember that some classic fashion essentials never go out of style. A timeless pair of jeans, a well-fitting blazer, or a plain white t-shirt are examples of flexible items that you can use for several seasons. While experimenting with new looks and trends is entertaining, it's also a good idea to get some timeless, classic pieces that will never go out of style.

5. Share

I would first describe the type of clothing item I am searching for, such as a formal dress for an interview, when speaking with the shopkeeper to discover the greatest fit for me. I would also include any particular demands I have, such the necessity for a particular size or colour. I would ask the shopkeeper for advice on what styles and fits will suit my

body type and sense of style the best. Any further details about my tastes or requirements that could assist reduce the alternatives are welcome. I would ask the shopkeeper for help in choosing the proper size and making any necessary adjustments to achieve the best fit once they had given me a few options to try on. I would also seek their opinion on any extras or coordinating items that would round off the look. Together with the shop owner, I can pick a formal outfit that suits my demands for the interview and looks beautiful while also being comfortable.

6. Act

I would go ahead and make the purchase as soon as I had selected the ideal formal outfit for my interview. I would look at the cost to make sure it is within my pricing range. If it's too pricey, I might inquire about any sales or discounts that might be going on. I would then go ahead and pay for it if I am satisfied with the pricing and have verified that the dress fits nicely and meets all of my needs. I might decide to pay with cash, a credit card, or another type of payment that the shop accepts. I would bring my brand-new evening gown home and make sure it was stored safely until the day of my interview. I can feel confident and prepared for my forthcoming interview by taking the time to locate the ideal dress and making sure it fits properly.

Conclusions

Hence, we have seen how we can use the 6 steps of Data Analytics while making any decision in real life scenarios (finding the best Suitable dress code for any company interview)

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

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

Instagram User Analytics

Description

In order to get business insights for the marketing, product, and development teams, we track how consumers connect with and interact with our digital product (software or mobile application). Teams from throughout the company utilise these information to develop new marketing campaigns, choose which features to include in apps, gauge the performance of the apps by looking at user interaction, and generally improve the user experience while assisting in business expansion.

You are a member of Instagram's product team, and the product manager has requested you to provide your thoughts on the queries posed 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).

Design

Procedures for loading data into the database

- Create a database with MySQL's "create db" function.
- Add tables and column names after that.
- Then, using MySQL's "insert into" function, add the values to them.
- The "select" command allows us to query the desired output.

Software used for querying the results

I am using MySQL 8.0.33. Businesses may improve their operations, make better choices, and get a competitive edge in their industries by adopting MySQL 8.0.33 for analysis.

Findings – I

To determine the top 5 first Instagram users who are the most devoted:

- By choosing the columns for username and created at, we will use the information from the users table.
- The intended output will then be sorted using the created at column in ascending order using the order by function.
- The output will then be provided for the top 5 oldest Instagram users after utilising the limit function.

Output/Result:

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

Findings – II

To find Instagram users that are the most inactive, that is, those who have never shared a single photo:

- From the users table, we will first choose the username column.
- Then, since users.id and photographs.user_id both contain the same information, we will do a left join between the users table and the photos table on the condition that users.id = photos.user_id.
- Then we'll locate the rows in the users database that contain the photographs Id is empty.

Output/Result

username	id
Aniya_Hackett	5
Kasandra_Homenick	7
Jaclyn81	14
Rocio33	21
Maxwell.Halvorson	24
Tierra.Trantow	25
Pearl7	34
Ollie_Ledner37	36
Mckenna17	41
David.Osinski47	45
Morgan.Kassulke	49
Linnea59	53
Duane60	54
Julien_Schmidt	57
Mike.Auer39	66
Franco_Keebler64	68
Nia_Haag	71
Hulda.Macejkovic	74
Leslie67	75
Janelle.Nikolaus81	76
Darby_Herzog	80
Esther.Zulauf61	81
Bartholome.Bernhard	83
Jessyca_West	89
Esmeralda.Mraz57	90
Bethany20	91

Findings – III

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

1. We will choose the users first. user, photo id, photo 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. The output will then be grouped based on photographs using the group by function photos.id.
4. The data will then be sorted based on the sum in descending order using the order by function.

5. Then, in order to identify the most popular image, we will use the limit function to display just the data for the most popular image.

Output /Result

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

Findings – IV

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

1. To count the number of tags used individually, we must use the tag name column from the tag database and the count(*) as total function.
2. Consequently, we must link the tags database with the photo tags table so that tags.id = photo tags.tag id because they both contain the same information, namely tag id.
3. The intended output must then be grouped based on tags using the group by function tag name.
4. The result must then be sorted in decreasing order according to total (total number of tags per tag name) using the order by function.
5. The limit 5 function will then be used to determine the top 5 most often used tag names.

Output/Result

	tag_name	total
▶	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:

1. First, we use select dayname(created at) as the day of the week and count(*) as the total number of users registered from the users table to establish the columns of the required output table.
2. Next, based on day of week, we organise the result table using the group by function.
3. Next, using the order by function, the output table is sorted and ordered in descending order based on the total number of people who have registered.

Output/Result

day_of_week	total
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14

Findings – VI

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

1. First, we must determine how many photographs (posts) are present in the photos.id column of the photos table, which may be done by counting (*) from photos.
2. Similarly, we must count (*) from users to get the number of users that are shown in the users.id column.
3. After that, we must divide both numbers, count(*) from photographs/count(*) from users, to obtain the total number of photos / total number of users.
4. We must count the total instances of each user id in the photographs database in order to determine how frequently a user publishes on Instagram.

Output/Result

photos_per_user_ratio
2.5700

user_id	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

user_id	post_count
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

user_id	post_count
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

user_id	post_count
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 id column from the users 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/Result

id	username	total_likes_by_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:-

- Out of the 100 people altogether, 26 individuals are inactive and have never uploaded anything on Instagram, including any sort of text, video, or photo. Therefore, the Instagram marketing team must remind such inactive users.
- Therefore, user Zack Kemmer93 with user id 52 wins the competition since his photo with photo id 145 gets the most likes (48).
- Along with the overall number of uses, the top 5 #hashtags are grin (59), beach (42), party (39), fun (38), and concert (24)
- Since the majority of people (16) registered on Thursday and Sunday, it would be wise to launch the AD campaign on these two days.
- Consequently, there are 100 rows (100 ids) in the users database and 257 rows (257) altogether, making the required output equal to 257/100 = 2.57. (avg. users posts on Instagram)
- 13 user ids—out of the total—have liked each and every post on Instagram, which is nearly impossible. As a result, these user ids are regarded as BOTS and Fake Accounts.

I'm identifying the underlying reason of the following using the 5 Whys method:

Q. Why was it important to the marketing team to identify the most inactive users?

Ans. They can thus write those people and inquire as to what prevents them from utilising Instagram.

Q. Why were the top 5 hashtags used important to the marketing team?

Ans .The technical team could have wished to include some filter options for images and videos shared with the top 5 stated #hashtags.

Q. Why did the marketing team want to know which day of the week saw the most brand-new customers register on the platform?

Ans . so that they may broadcast more advertisements for different products on days like this and benefit from it.

Q. Why did the investors want to know how many posts on Instagram an average user makes?

Ans .The user engagement on social media platforms, which affects every brand and business, is a known truth. Investors also needed to know if the platform had the appropriate and reliable user base. Additionally, it assists the tech team in figuring out how to manage such traffic on the platform using the most recent technology without interfering with the platform's efficient and effective operation.

Q. Why were investors interested in learning how many BOTS and fake accounts there were, if any?

Ans .To provide investors peace of mind that they are funding an asset rather than a potential liability

Conclusion

I'd want to draw the following conclusion: Not just Instagram, but many other social media and commercial companies employ this type of analysis to get insights from their customer data, which in turn enables the companies to identify clients who will be an asset to them rather than a liability.

According to the demands of the business enterprises, this analysis and sorting of the client base is carried out weekly, monthly, quarterly, or yearly to optimise future revenues at the lowest possible cost to the organisation.

Operation Analytics and Investigating Metric Spike

Description

Operation analytics is the analysis performed for a company's whole end-to-end operations. This helps the business identify the areas where it needs to make improvements. You collaborate closely with the operations team, the support team, the marketing team, etc. and assist them in drawing conclusions from the data they gather.

Being one of the most crucial components of a business, this form of analysis is also utilised to forecast the general upward or downward trend in a company's fortune. Better automation, improved communication among cross-functional teams, and more efficient workflows are the results.

Investigating metric spikes is a crucial component of operational analytics since a data analyst has to be able to answer queries such, "Why is there a decline in daily engagement?" or at least help other teams answer these questions. Why have sales decreased? Etc. Daily answers to questions like these are required, thus it is crucial to look at metric increase.

You have the title of "Data Analyst Lead" and are employed by a corporation like Microsoft. You are given access to various data sets and tables from which you must draw particular conclusions and respond to inquiries from various departments.

The Problem

Case Study 1 (Job Data)

1. 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?
2. 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?
3. 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?
4. Duplicate rows: Rows that have the same value present in them. Your task: Let’s say you see some duplicate rows in the data. How will you display duplicates from the table?

Case Study 2 (Investigating metric spike)

1. 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?
2. User Growth: Amount of users growing over time for a product. Your task: Calculate the user growth for product?
3. Weekly Retention: Users getting retained weekly after signing-up for a product. Your task: Calculate the weekly retention of users-sign up cohort?
4. 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?
5. Email Engagement: Users engaging with the email service. Your task: Calculate the email engagement metrics?

Design

Steps taken to load the data into the data base

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

Software used for querying the results

MySQL Workbench 8.0 CE

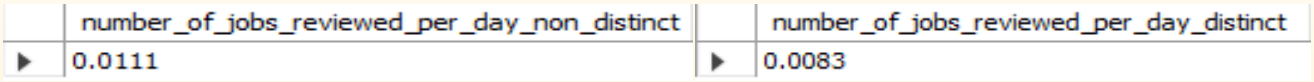
Job Data

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 /Result



number_of_jobs_reviewed_per_day_non_distinct	number_of_jobs_reviewed_per_day_distinct
0.0111	0.0083

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

Output/Result

Review_date	Reviewed_jobs	rolling_average
2020-11-25	1	1.0000
2020-11-26	1	1.0000
2020-11-27	1	1.0000
2020-11-28	2	1.2500
2020-11-29	1	1.2000
2020-11-30	2	1.3333

Findings – III

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

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 /Result

job_id	language	total_of_each_language	percentage_share_of_each_language
11	French	1	12.5000
20	Italian	1	12.5000
21	English	1	12.5000
22	Arabic	1	12.5000
23	Persian	1	37.5000
25	Hindi	1	12.5000

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 /Result

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

Investigating Metric Spike

Findings – 1

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 /Result

week	num_of_weekly_engaged_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

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 /Result

year_num	week_num	user_id	cum_active_users
2013	1	67	67
2013	2	29	96
2013	3	47	143
2013	4	36	179
2013	5	30	209
2013	6	48	257
2013	7	41	298
2013	8	39	337
2013	9	33	370
2013	10	43	413
2013	11	33	446
2013	12	32	478
2013	13	33	511
2013	14	40	551
2013	15	35	586
2013	16	42	628
2013	17	48	676

2013	18	48	724
2013	19	45	769
2013	20	55	824
2013	21	41	865
2013	22	49	914
2013	23	51	965
2013	24	51	1016
2013	25	46	1062
2013	26	57	1119
2013	27	57	1176
2013	28	52	1228
2013	29	71	1299
2013	30	66	1365
2013	31	69	1434
2013	32	66	1500
2013	33	73	1573
2013	34	70	1643
2013	35	80	1723
2013	36	65	1788
2013	37	71	1859
2013	38	84	1943
2013	39	92	2035
2013	40	81	2116
2013	41	88	2204
2013	42	74	2278
2013	43	97	2375
2013	44	92	2467
2013	45	97	2564
2013	46	94	2658
2013	47	82	2740
2013	48	103	2843
2013	49	96	2939
2013	50	117	3056
2013	51	123	3179
2013	52	104	3283
2014	1	91	3374
2014	2	122	3496
2014	3	112	3608
2014	4	113	3721
2014	5	130	3851
2014	6	132	3983
2014	7	135	4118
2014	8	127	4245
2014	9	127	4372
2014	10	135	4507
2014	11	152	4659

2014	12	132	4791
2014	13	151	4942
2014	14	161	5103
2014	15	166	5269
2014	16	165	5434
2014	17	176	5610
2014	18	172	5782
2014	19	160	5942
2014	20	186	6128
2014	21	177	6305
2014	22	186	6491
2014	23	197	6688
2014	24	198	6886
2014	25	222	7108
2014	26	210	7318
2014	27	199	7517
2014	28	223	7740
2014	29	215	7955
2014	30	228	8183
2014	31	234	8417
2014	32	189	8606
2014	33	250	8856
2014	34	259	9115
2014	35	266	9381

Output /Result

total_active_user
9381

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. First, we will use the extract and week functions to retrieve the week from the occurred at column.
2. Then, we will select 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 Then using the left join we will join the two tables on the basis of user_id where event_type = 'engagement'

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

Output /Result

Google Drive Link for saved result

- Without week number query output link kindly press to see result
- With week number kindly press to see result

Findings – IV

To find the weekly user engagement per device:-

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

Output /Result

Google Drive link for saved result

- User engagement per device kindly press to see result

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 sum of the category of email_sent and multiply the result by 100.0 and name is as email_opening_rate
3. Then we select the sum of category of email_clicked divide by the sum of the category of email_sent and multiply the result by 100.0 and name is as email_clicking_rate
4. email_sent = ('sent_weekly_digest','sent_reengagement_email')
5. email_opened = 'email_open'
6. email_clicked = 'email_clickthrough'

Output /Result

email_opening_rate	email_clicking_rate
33.58339	14.78989

Analysis

From the tables I have infer the following:-

1. number of distinct job reviewed per day is 0.0083
2. number of non-distinct jobs reviewed per day is 0.0111
3. For November 25, 26, 27, 28, 29, and 30 of 2020, the 7-day rolling average throughput is 1, 1, 1, 1.25, 1.2, and 1.3333, respectively (for both distinct and non-distinct)
4. Percentage Arabic, English, French, Hindi, Italian, and Persian each have a share of 12.5, 12.5, 12.5, 12.5, and 12.5 correspondingly (for both distinct and nondistinct)
5. 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

Q: Why are the numbers for the number of distinct jobs evaluated each day and the number of non-distinct jobs reviewed each day different?

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

Q: Why should throughput be calculated using a 7-day rolling average rather than a daily metric average?

- We will use the 7-day rolling method to calculate throughput since it provides us with average data for all days, from day 1 to day 7, whereas daily metrics just provide us with data for the current day.

Q: Why is Persian's percentage share of all languages 37.5 percent yet all other languages' is 12.5 percent?

- There are two possibilities in these situations: either there were duplicate rows with the language listed as "Persian," or there were actually two or more distinct individuals conversing in that language.

Q: Why do we need to search a dataset for duplicate rows?

- To prevent duplicates from having a negative impact on the analysis and maybe resulting in poor business decisions that result in losses for the firm or any other entity, one must search for duplicates and eliminate them as needed.

From the tables I have infer the following:-

1. The weekly user engagement is the highest for week 31 i.e. 1685
2. There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014
3. 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:-

Q: Why did weekly user involvement grow after starting off so low?

- It is a recognised truth that when a new product or service is first introduced to the market, fewer people are aware of it. Only a small number of individuals utilise the product, and depending on how they felt about it, the product or service's popularity may rise or fall. In this instance, the fact that user involvement grew after the product or service had been available for two to three weeks indicates that customers were satisfied with it.

Q: Why is weekly retention so important?

- If visitors who just complete the sign-up process or abandon it in the middle are appropriately directed and persuaded, they may turn into future customers. Weekly retention helps businesses persuade and assist these people.

Q: Why is weekly engagement per device plays an important role?

- Based on user feedback, weekly engagement per device informs businesses as to which devices to enhance and where to spend their attention in order to receive positive feedback from consumers.

Q: Why is Email Engagement plays an important role?

- The decision-making process for discounts and offers on certain items is aided by email engagement. In this instance, the email opening rate is 33.58, meaning that only 34 of the 100 emails sent were viewed, and the email clicking rate is 14.789, meaning that only 15 of the 100 emails opened were clicked to view further information about the deal or product. This indicates that the present company has to have a catchier subject line for emails as well as rigorous preparation and content selection before sending the emails.

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

The hiring process is the foundational and crucial part of a business. The MNCs learn about the key underlying trends relating to the hiring process here. Before employing freshmen or anybody else, a corporation should consider trends such as the number of rejections, interviews, sorts of positions, openings, etc. Hence, there is a chance for a Data Analyst employment here as well!

As a data analyst, it is your responsibility to examine these trends and derive insights that the hiring department may use.

As a lead data analyst for a multinational corporation (MNC) like Google, your employer has given you access to the recruiting history data and asked you to make sense of it in order to respond to a series of questions.

The Problem

- Hiring: Process of intaking of people into an organization for different kinds of positions.
- Your task: How many males and females are Hired ?
- Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.
- Your task: What is the average salary offered in this company ?
- Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.
- Your task: Draw the class intervals for salary in the company ?
- Charts and Plots: This is one of the most important part of analysis to visualize the data.
- Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department?
- Charts: Use different charts and graphs to perform the task representing the data.
- Your task: Represent different post tiers using chart/graph?

Design

Before starting the actual analysis I have:-

1. To ensure that any modifications I made would not damage the original data, I first created a duplicate of the raw data on which I could do the analysis.
2. Second, I checked to see whether there were any empty spaces or NULL values.

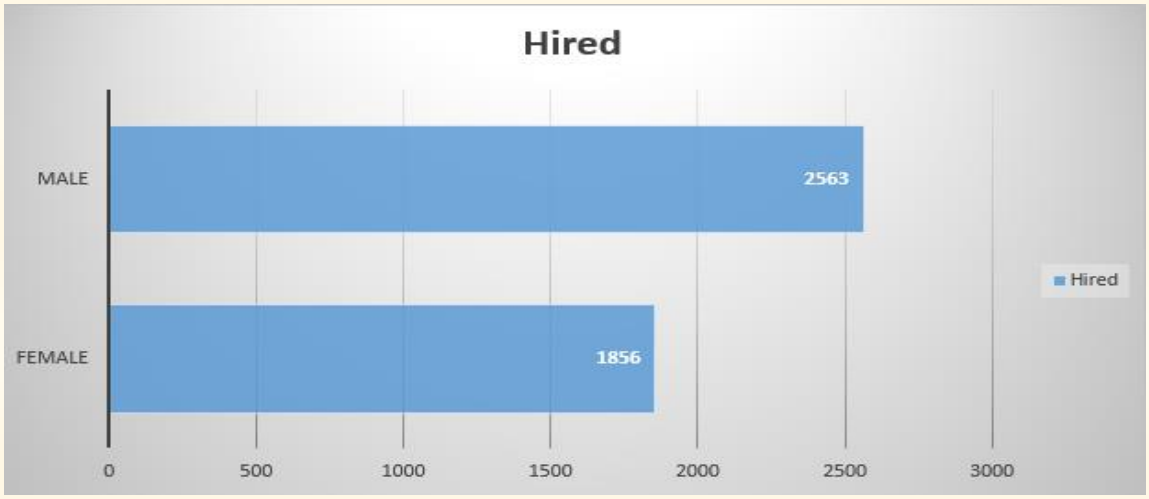
3. Then, assuming there were no outliers for that specific column, I had either imputed the mean of the column or the median to the numerical blank and NULL cells (if outliers existed for that column).
4. After that, I checked to see if there were any outliers and replaced them with the median of the specific column where the outlier had been found.
5. Then I had the variable with the highest count substituted in all category variables that had blank cells.
6. Then, if any duplicate rows were found, I eliminated them.
7. In order to do the analysis, I then deleted the unnecessary columns (data) from the dataset.

Software used for doing the overall Analysis:-

- Microsoft Excel 2013

Findings – I

Count of event_name	Column Labels	
Row Labels	Hired	Grand Total
Female	1856	1856
Male	2563	2563
Grand Total	4419	4419



Observation :

- There are 2563 Males hired for different roles in the company
- While there are only 1856 Females hired for different roles in the company

Findings – II

To find the average salary offered in this company:-

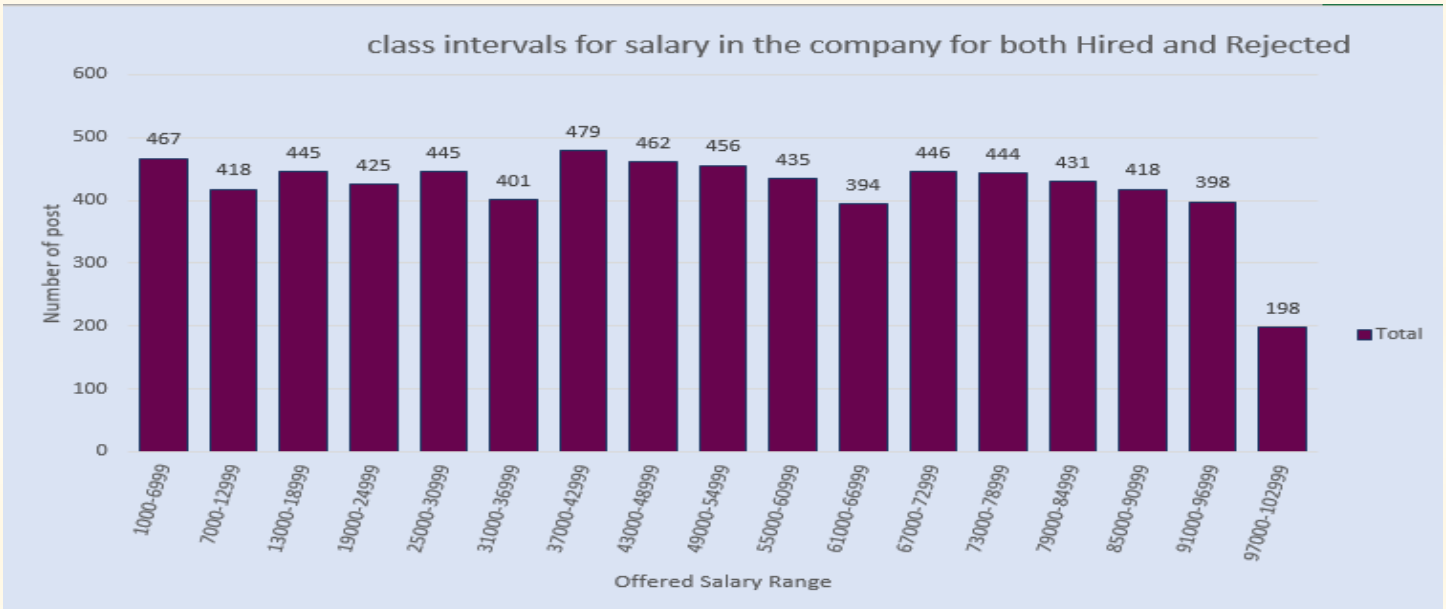
- First, we need to remove the outliers i.e. to remove the salaries below 1000 and above 100000

- Then using the formula
  =AVERAGE(Whole_salary_column_after_removing_outliers)

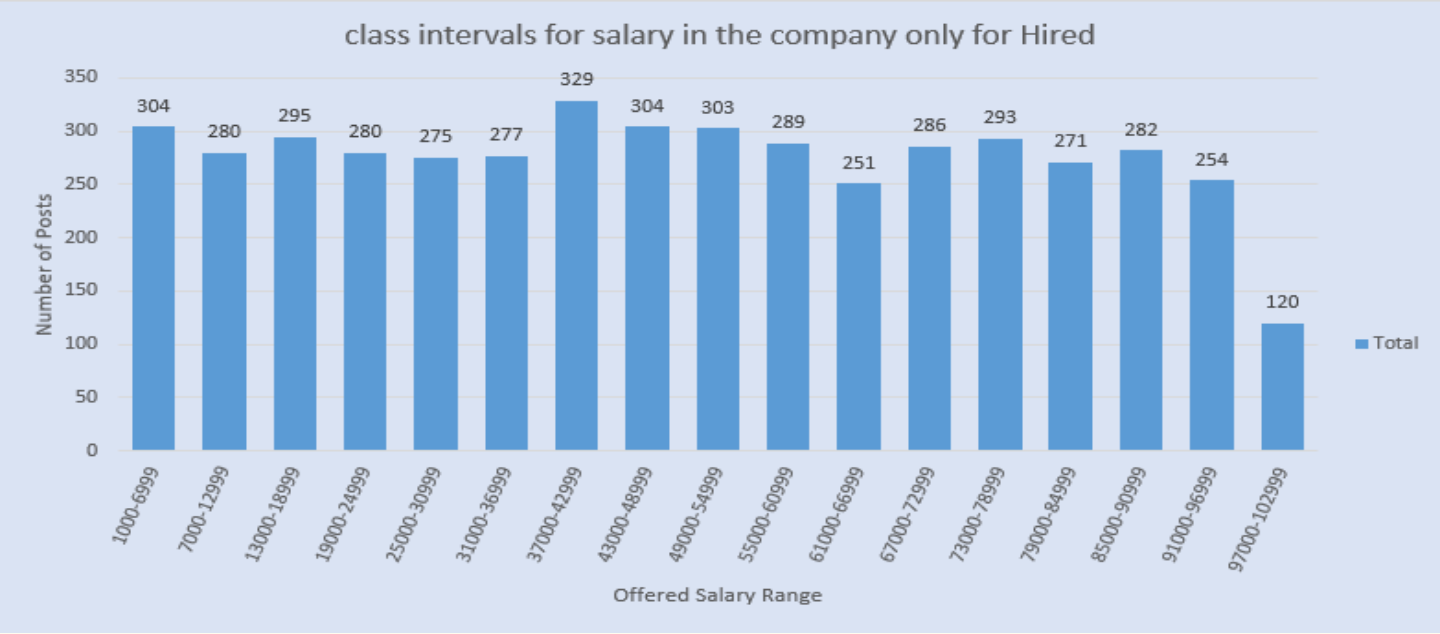
Output :

49885.28

Findings – III



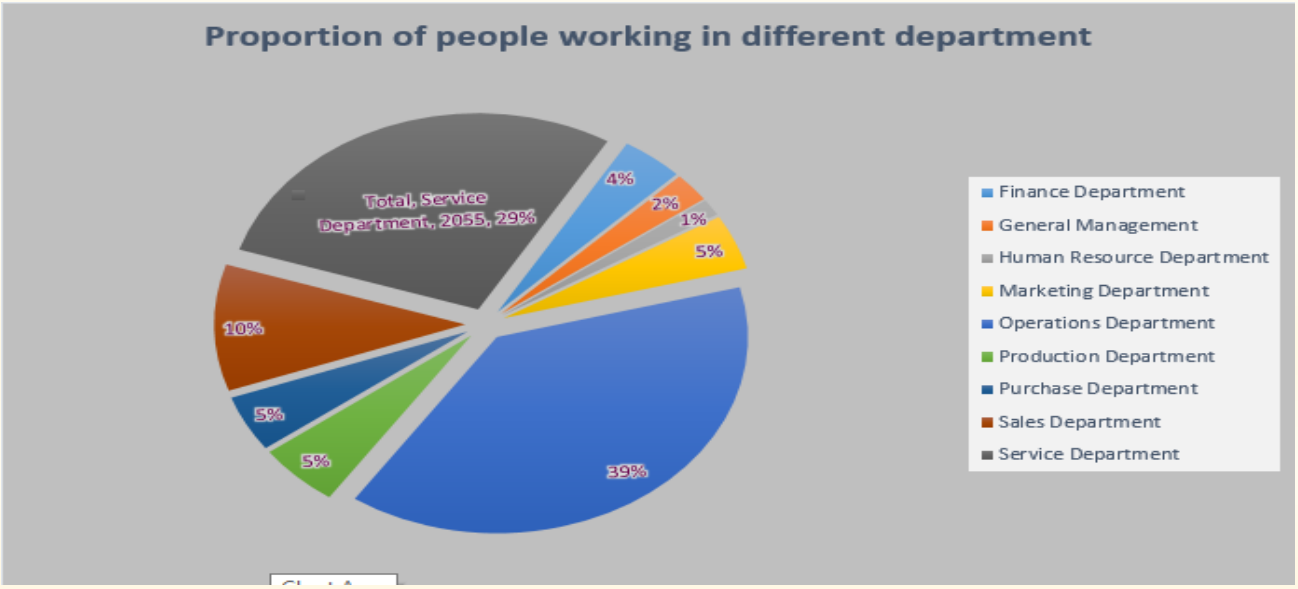
From the above column plot I have inferred that the highest number of posts (both hired and rejected) is 479 for the salary range 37000 to 42999.



From the above column plot I have inferred that the highest number of posts (hired) is 329 for the salary range 37000 to 42999.

Findings – IV

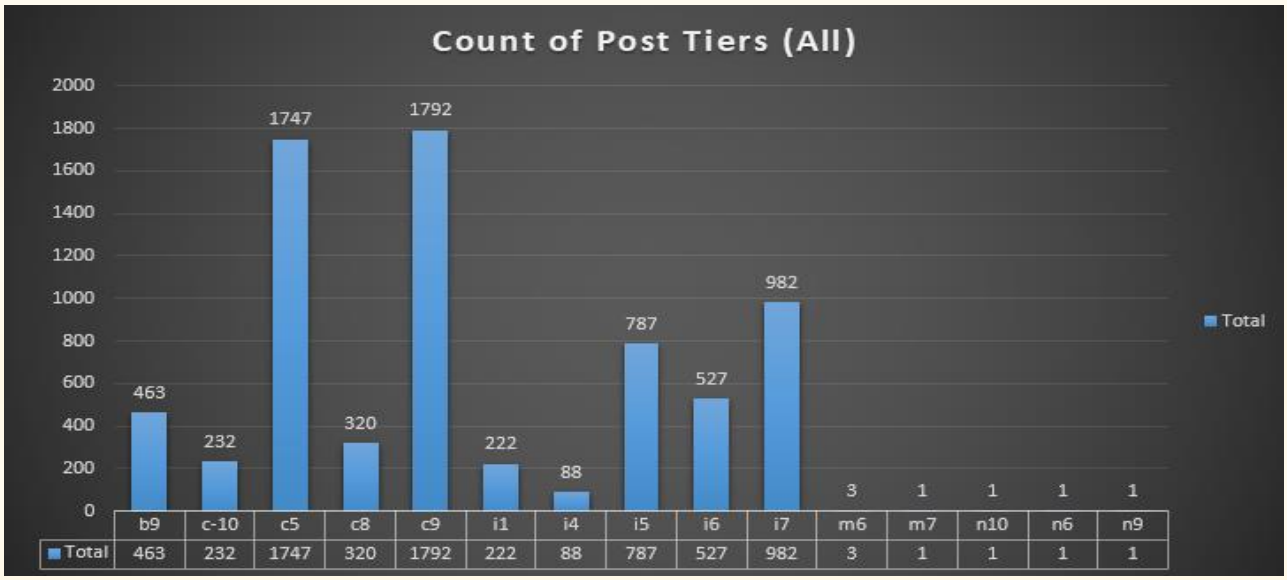
Row Labels	Count of Department
Finance Department	288
General Management	172
Human Resource Department	97
Marketing Department	325
Operations Department	2771
Production Department	380
Purchase Department	333
Sales Department	747
Service Department	2055
Grand Total	**7168**

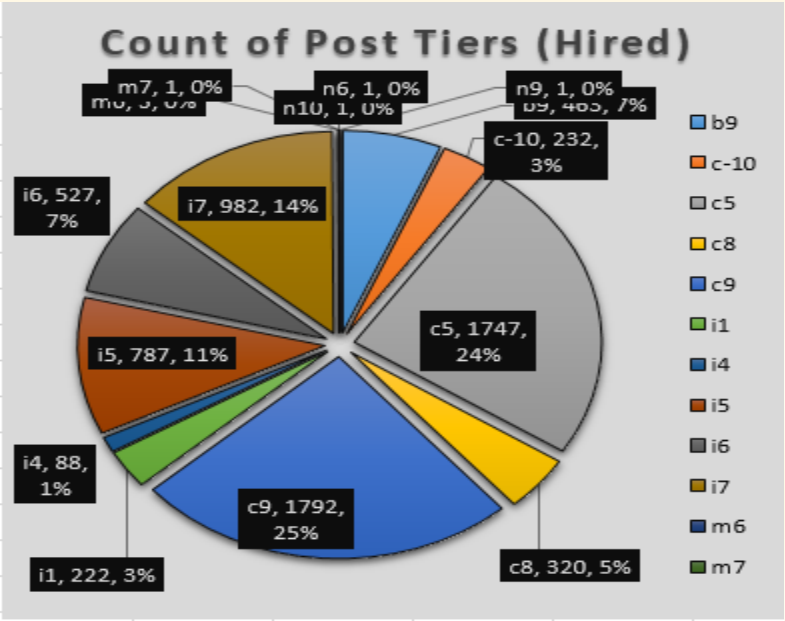


From the above table and pie chart I have inferred that the Highest number of people were working in the Operations Department i.e. 2771 which accounts for almost 39% of the total workforce of the company.

Findings – V

Row Labels	Count of Department
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1
Grand Total	**7167**





From the above table, Column 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:

Q : 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.

Q . 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.

Q.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 employess

IMDB Movie Analysis

Description

We are giving you a dataset with multiple columns of different IMDB movies for your final project. It is necessary for you to frame the issue. You must identify the issue you wish to illuminate for this activity.

It is necessary for you to frame the issue. You must identify the issue you wish to illuminate for this activity. Once an issue has been identified, you should clean the data as appropriate before using data analysis techniques to examine the data set and draw conclusions.

The 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

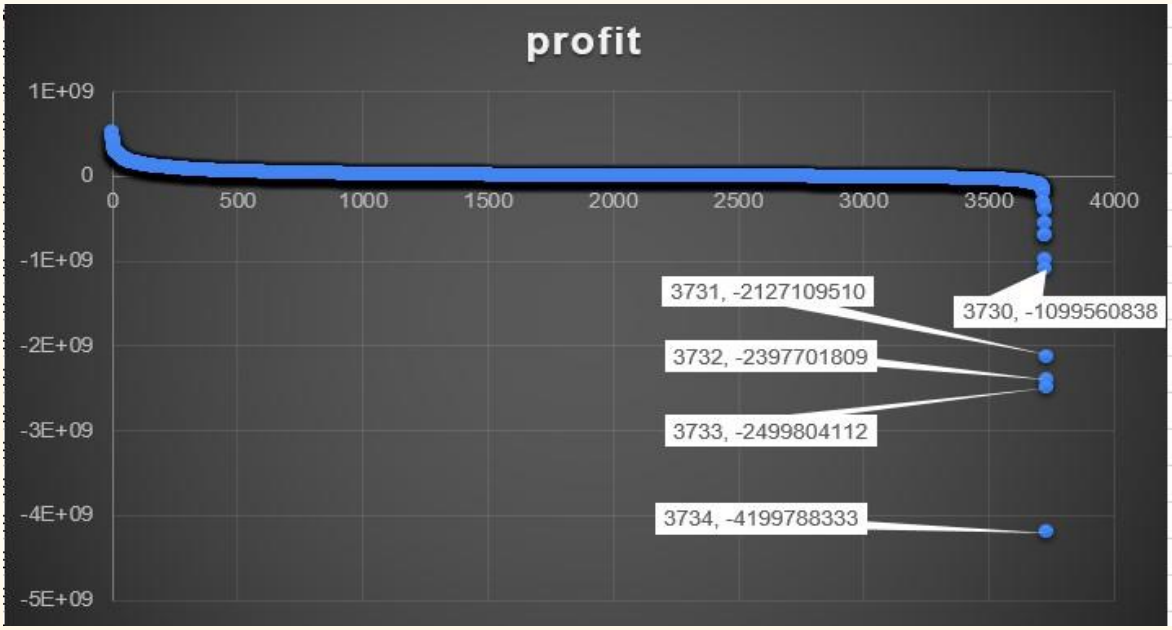
Design

- To ensure that any modifications I made would not damage the original data, I first created a duplicate of the raw data on which I could do the analysis.
- After that, delete any columns that won't be used in the analysis we'll be undertaking.
- 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 needs to be dropped.
- We now need to eliminate the rows from the dataset that have any of their column values as blank or NULL after removing the unnecessary columns.
- Then, we must remove the duplicate values from the dataset using the 'Remove Duplicate Values/Cells' option found in the 'Data' tab.

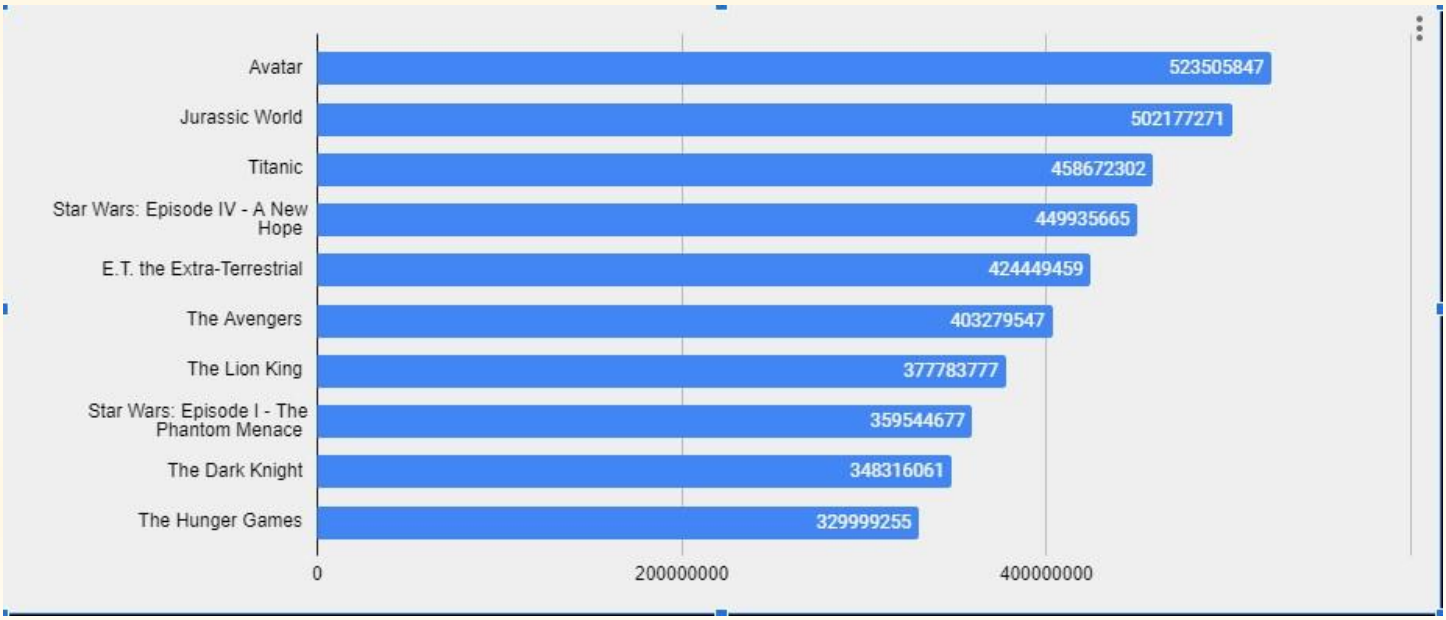
Findings – I

To find the movies with the highest profit: -

- First we need to subtract the budget value from the gross value to get the profit.
- Then, by using the scatter plot option we will plot values of profit(y_axis) and budget(x_axis)
- Then with the help of graph we will be finding the outliers



Outliers are
-4199788333
-2499804112
-2397701809
-2127109510
-989962610
-1099560838



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. Using the sort and filter option, we will first remove any rows with num_voted_users more than 30000.
2. Next, the dataset will be arranged in decreasing order using the imdb_score.
3. Only the top 250 rows will be chosen for further research.
4. Next, we'll use the RANK() function and the formula to generate a new column for rank. =RANK(O2,$O$2:$O$251,0)+COUNTIFS($O$2:O2,O2)-1
5. The desired output will then be obtained once we filter out (unselect "English" from the language column).

Top 10 IMDB Movies(all language) are:-

director_name	num_critic_for_review	duration	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_review	language	budget	title_year	imdb_score	profit	Rank
Frank Darabont	199	142	28341469	Crime|Drama	Morgan Freeman	The Shawshank Redemption	1689764	4144	English	25000000	1994	9.3	3341469	1
Francis Ford Coppola	208	175	1.35E+08	Crime|Drama	Al Pacino	The Godfather	1155770	2238	English	6000000	1972	9.2	128821952	2
Christopher Nolan	645	152	5.33E+08	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	4667	English	185000000	2008	9	348316061	3
Francis Ford Coppola	149	220	57300000	Crime|Drama	Robert De Niro	The Godfather: Part II	790926	650	English	13000000	1974	9	44300000	4
Peter Jackson	328	192	3.77E+08	Action|Adventure|Drama|Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the King	1215718	3189	English	94000000	2003	8.9	283019252	5
Quentin Tarantino	215	178	1.08E+08	Crime|Drama	Bruce Willis	Pulp Fiction	1324680	2195	English	8000000	1994	8.9	99930000	6
Steven Spielberg	174	185	96067179	Biography|Drama|History	Liam Neeson	Schindler's List	865020	1273	English	22000000	1993	8.9	74067179	7
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	780	Italian	1200000	1966	8.9	4900000	8
Robert Zemeckis	149	142	3.3E+08	Comedy|Drama	Tom Hanks	Forrest Gump	1251222	1398	English	55000000	1994	8.8	274691196	9
Irvin Kershner	223	127	2.9E+08	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode V - The Empire Strikes Back	837759	900	English	18000000	1980	8.8	272158751	10

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

Top - 10 IMDB Movies all languages (except English)

director_name	num_critic_for_review	duration	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_review	language	budget	title_year	imdb_score	profit	Rank
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugl	503509	780	Italian	1200000	1966	8.9	4900000	1
Fernando Meirelles	214	135	7563397	Crime|Drama	Alice Braga	City of God	533200	749	Portugues	3300000	2002	8.7	4263397	2
Akira Kurosawa	153	202	269061	Action|Adventure|Drama	Takashi Shimura	Seven Samurai	229012	596	Japanese	2000000	1954	8.7	-1730939	3
Hayao Miyazaki	246	125	10049886	Adventure|Animation|Family|Fantasy	Bunta Sugawara	Spirited Away	417971	902	Japanese	19000000	2001	8.6	-8950114	4
Florian Henckel von Donnersm	215	137	11284657	Drama|Thriller	Sebastian Koch	The Lives of Others	259379	407	German	2000000	2006	8.5	9284657	5
Asghar Farhadi	354	123	7098492	Drama|Mystery	Shahab Hosseini	A Separation	151812	264	Persian	500000	2011	8.4	6598492	6
Chan-wook Park	305	120	2181290	Drama|Mystery|Thriller	Min-sik Choi	Oldboy	356181	809	Korean	3000000	2003	8.4	-818710	7
Wolfgang Petersen	96	293	11433134	Adventure|Drama|Thriller|War	Jürgen Prochnow	Das Boot	168203	426	German	14000000	1981	8.4	-2566866	8
Jean-Pierre Jeunet	242	122	33201661	Comedy|Romance	Mathieu Kassovitz	Amélie	534262	1314	French	77000000	2001	8.4	-43798339	9
Hayao Miyazaki	174	134	2298191	Adventure|Animation|Fantasy	Minnie Driver	Princess Mononoke	221552	570	Japanese	2.4E+09	1997	8.4	-2.398E+09	10

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 Itlay.

Findings - III

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

1. First, choose the cleaned dataset's imdb_score column.
2. Next, we'll select the pivot table.
3. We will include director_name in the pivot table's series section.
4. Next, we'll add the average imdb_score to the pivot table's values section. The information will then be sorted alphabetically by director name after being sorted first based on average imdb_score in descending order.

Row Labels	Average of imdb_score
Akira Kurosawa	8.7
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

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

Row Labels	Count of genres
Comedy|Drama|Romance	148
Drama	147
Comedy	143
Comedy|Drama	142
Comedy|Romance	135
Drama|Romance	117
Crime|Drama|Thriller	79
Action|Crime|Thriller	54
Action|Crime|Drama|Thriller	48
Action|Adventure|Sci-Fi	45
Comedy|Crime	45
Grand Total	**1103**

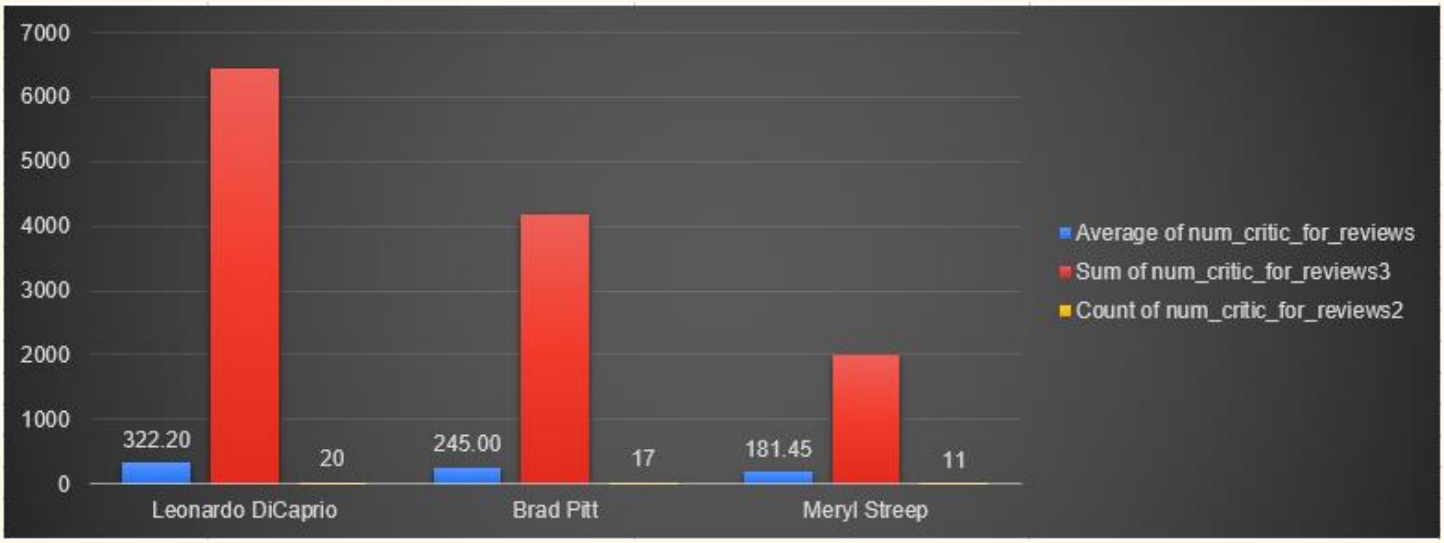
From the above table I have inferred that genre named Comedy|Drama|Romance was the most popular with a count of 148.

Findings – V

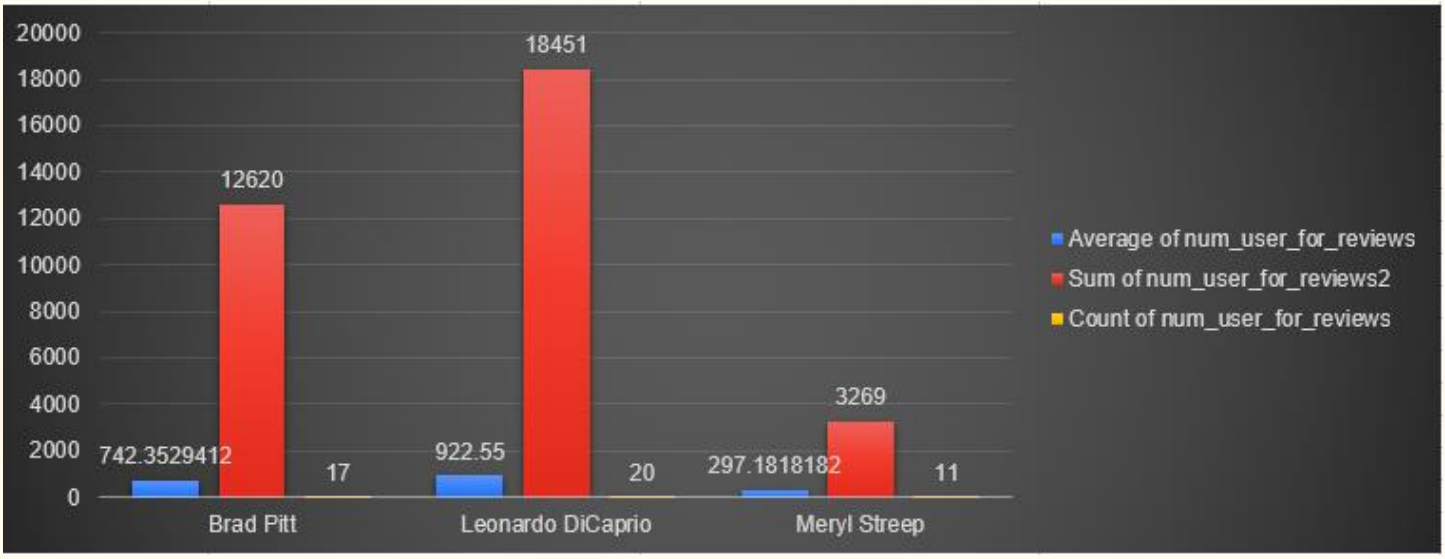
To find the critic-favorite and audience-favorite actors we will:-

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_critic_for_reviews	Sum of num_critic_for_reviews3	Count of num_critic_for_reviews2
Leonardo DiCaprio	322.2	6444	20
Brad Pitt	245	4165	17
Meryl Streep	181.4545455	1996	11
Grand Total	**262.6041667**	**12605**	**48**

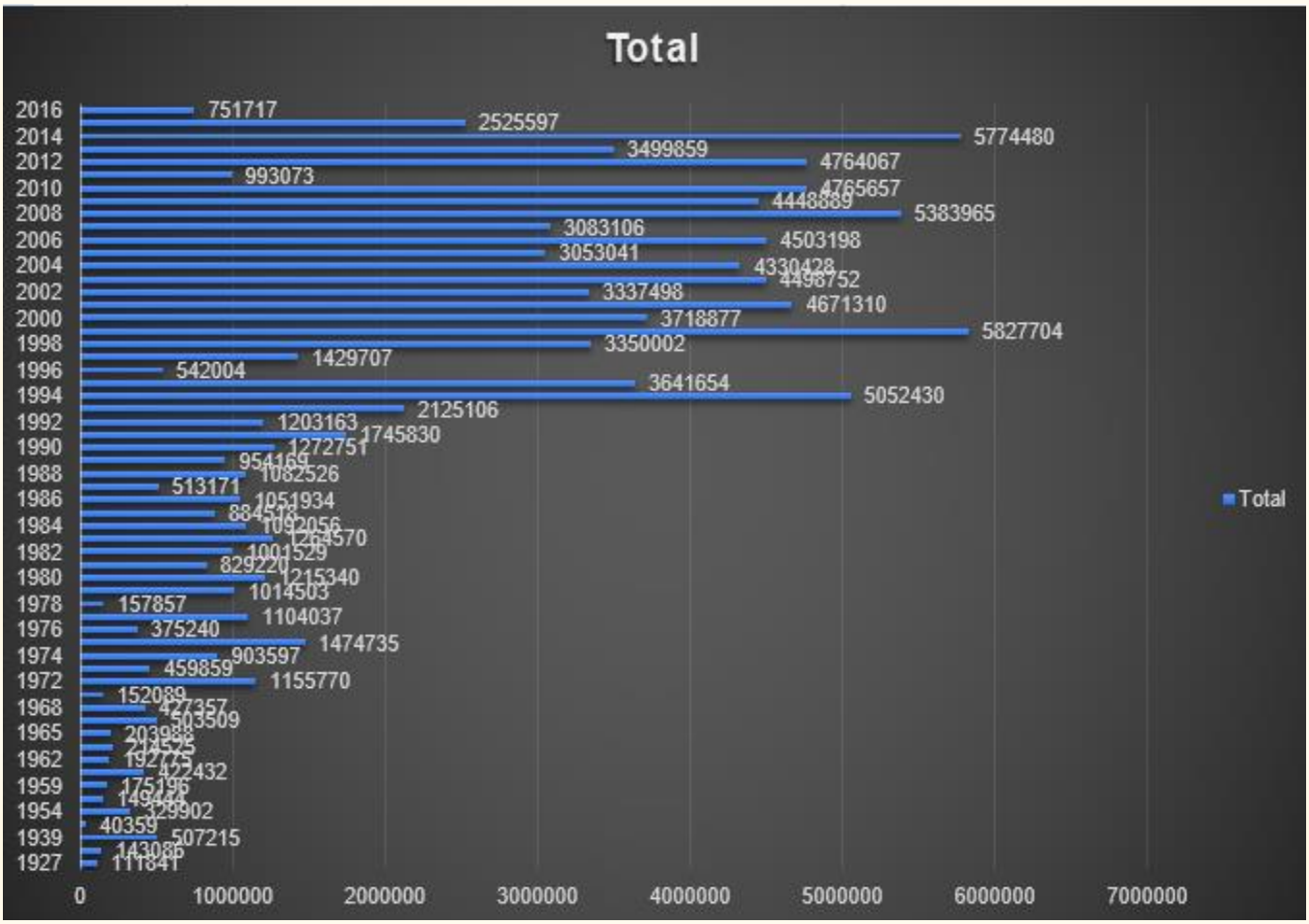


Row Labels	Average of num_user_for_reviews	Sum of num_user_for_reviews2	Count of num_user_for_reviews
Brad Pitt	742.3529412	12620	17
Leonardo DiCaprio	922.55	18451	20
Meryl Streep	297.1818182	3269	11
Grand Total	**715.4166667**	**34340**	**48**

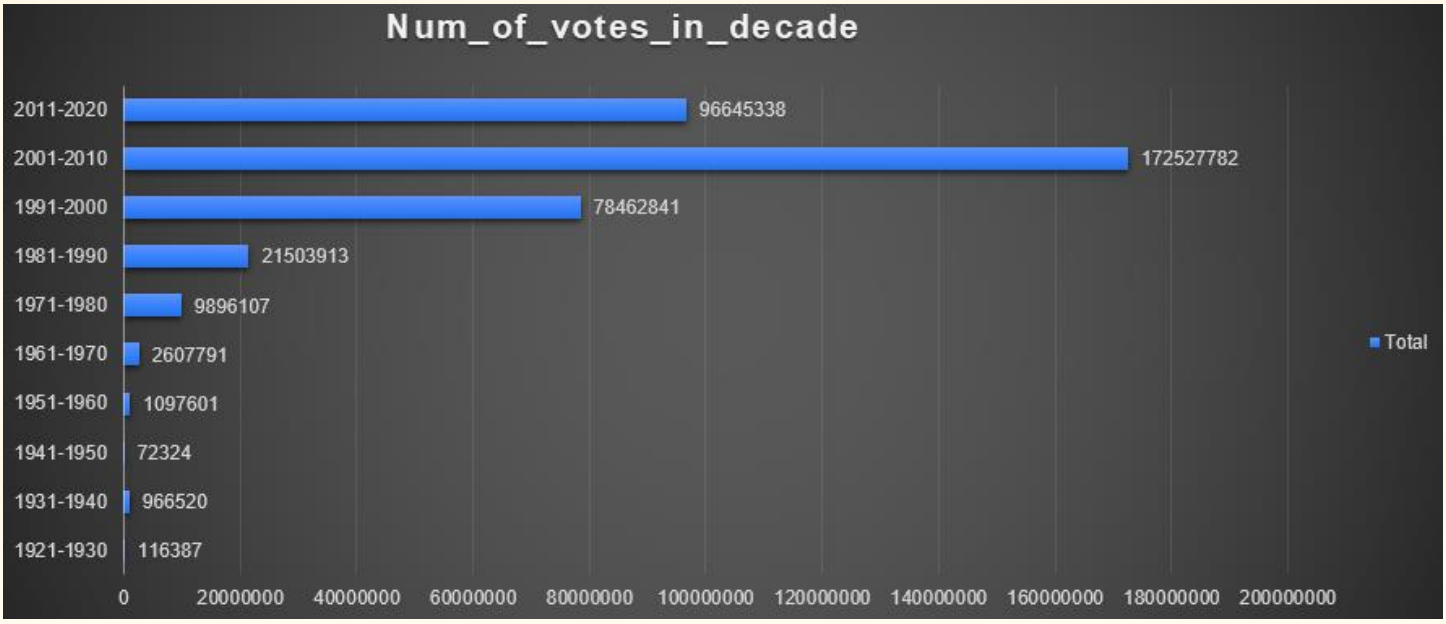


From the above two graphs I have inferred that 'Leonardo DiCaprio' was both critic-favorite and audiencefavorite.

Findings – VI



Row Labels	Sum of num_voted_users
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2607791
1971-1980	9896107
1981-1990	21503913
1991-2000	78462841
2001-2010	172527782
2011-2020	96645338
Grand Total	383896604



From the above table and Column plot I have inferred that most number of votes were in the decade 2001-2010 with a count of 172527782.

Analysis

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

- Why are the highest-rated movie on IMDB and the one that made the most money different?
- Perhaps as a result of the fact that during the IMDB rating only authorised users had access to the IMDB portal. On the other side, the amount of money made is determined by how many tickets were sold throughout the world in theatres.
- Why were there more votes cast between 2001 and 2010?
- Many scientific and computer-graphics improvements were made between 2001 and 2010, and there was a tremendous growth in global movie production during this time. As a result, a significant number of films were made and released throughout this decade. Additionally, before to 2000, there were no rules in place anywhere in the globe with a distinct ministry, board, or committee from the

government side that examined issues related to the creation and dissemination of films.

- Why are the top 5 movies on IMDB based solely on titles with the language set to "English"?
- The USA was the source of origin for English-language films at the time, and it is well-known that the USA's economy was booming at the time. In order to make money, social media investors searched for filmmakers who have produced films.
- Why did Comedy and Drama have the highest levels of popularity?
- The majority of people worldwide are under stress from their jobs, thus they want a calming diversion rather than an action or horror film. People therefore like viewing comedies, dramas, or both types of movies. However, majority of them liked comedies.
- Why did the decade 2001-2010 receive more votes than the decade 2011-2020, despite the fact that the latter had advances in animation and graphics?
- It is undeniable that technology advanced greatly and significantly during this time, not only in the graphics and animation industry but in all facets of daily life. Additionally, VPN was introduced during this time, and VPN caused film piracy (illegal distribution of films), which discouraged most people from visiting movie theatres.

Conclusion

As a conclusion, I would like to state that, prior to the creation of a film, not only movie producers but also a variety of financiers, stakeholders, and theatre outlet owners perform IMDB Movie Analysis or any other similar analysis.

Normal individuals wouldn't mind performing such analyses, but such analyses are vital to both the pre-production and post-production stages of the filmmaking process.

Additionally, it's not a given that the film with the highest IMDB rating will also make the most money. The quantity of tickets sold by theatres throughout the world is the real basis for profit calculation.

The majority of individuals favour comedies, dramas, or both because they find their daily lives to be exhausting, and they avoid action and horror films.

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

Due to their weak or nonexistent credit histories, loan providers find it challenging to grant loans to individuals. Because of this, some customers take advantage of it by defaulting. Let's say you work for a consumer finance business that specialised in providing urban consumers with different kinds of loans. To examine the patterns found in the data, you must utilise EDA. By doing this, it will be ensured that only those applicants who can repay the loan would be accepted.

When a loan application is received, the business must evaluate whether to approve the loan based on the applicant's profile. The bank's choice is subject to two different kinds of risks:

- If the borrower is likely to pay back the loan, refusing to approve the loan will cost the firm money.
- If the borrower is not likely to pay back the loan, or is likely to default, then accepting the loan may result in a loss of revenue for the business.

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.

The 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).

Design

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

Filling in blanks in the application dataset's Occupation Type column with the categorical variable with the highest frequency

- Highest occurring categorical variable is ‘Laborers’

Replacing blanks in the Application Dataset's AMT ANNUTIY column with the median value of AMT ANNUITY since the column contains outliers

- Median of AMT_ANNUITY = 24903

Blanks in the Application Dataset's AMT GOODS PRICE column should be replaced with the median of AMT GOODS PRICE since the column contains outliers.

- Median of AMT_GOODS_PRICE = 450000

Filling in blanks in the Name Type Suite column of the Application Dataset with the categorical variable with the highest frequency

- Highest occurring categorical variable is ‘Unaccompanied’

Filling in blanks in the Application Dataset's Organization type column with the most frequent categorical variable

- Highest occurring categorical variable is ‘Business Entity Type 3’

The following columns from the preceding datasets for applications must be removed since they are unnecessary for doing data analysis.

- HOUR_APPR_PROCESS_START
- WEEKDAY_APPR_PROCESS_START_PREV
- FLAG_LAST_APPL_PER_CONTRACT
- NFLAG_LAST_APPL_IN_DAY

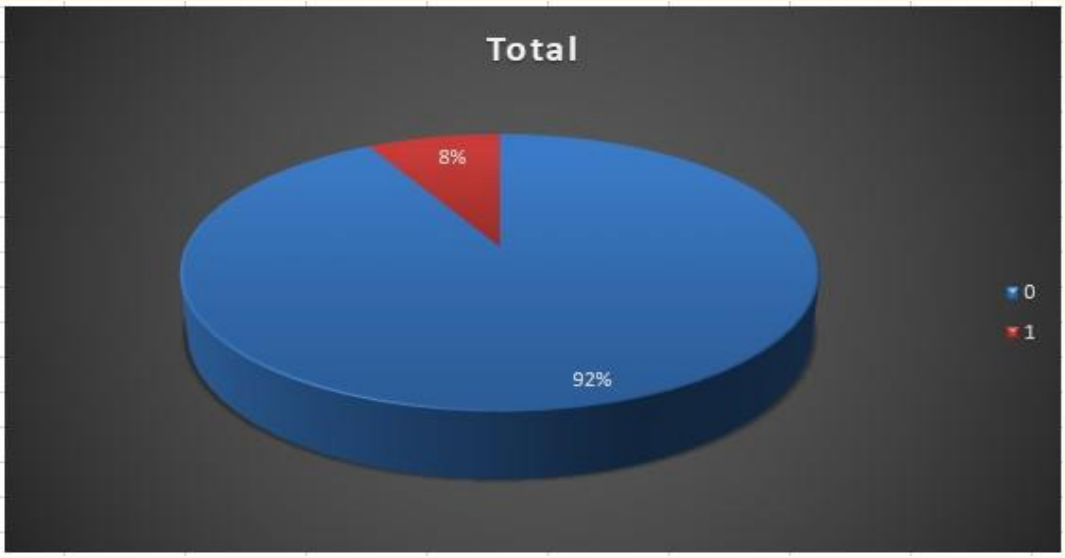
- SK_ID_CURR
- WEEKDAY_APPR_PROCESS_START

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

- Replace Blanks with Unaccompained

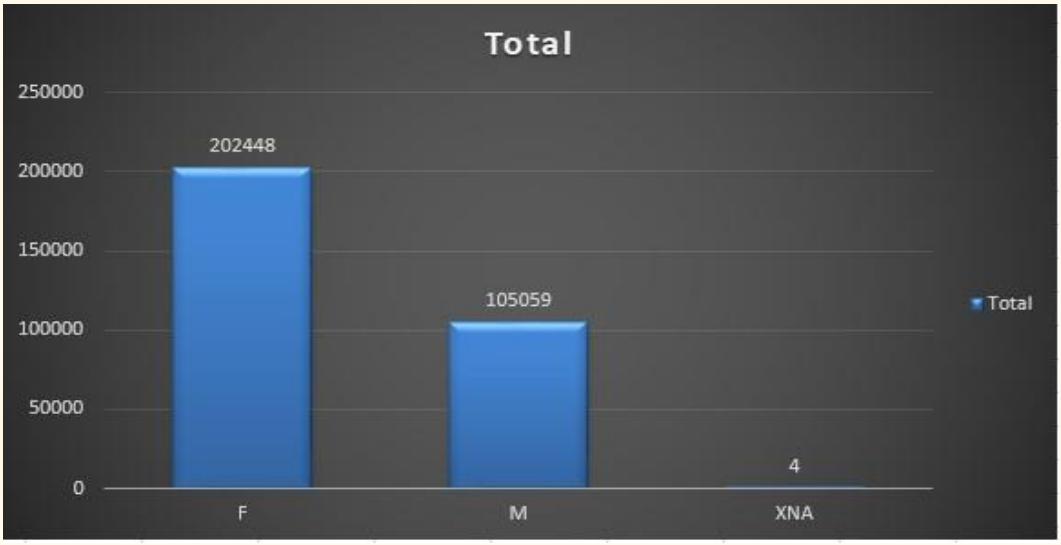
AMT_ANNUITY :- Replace Blanks with 24903(median)

Findings – I



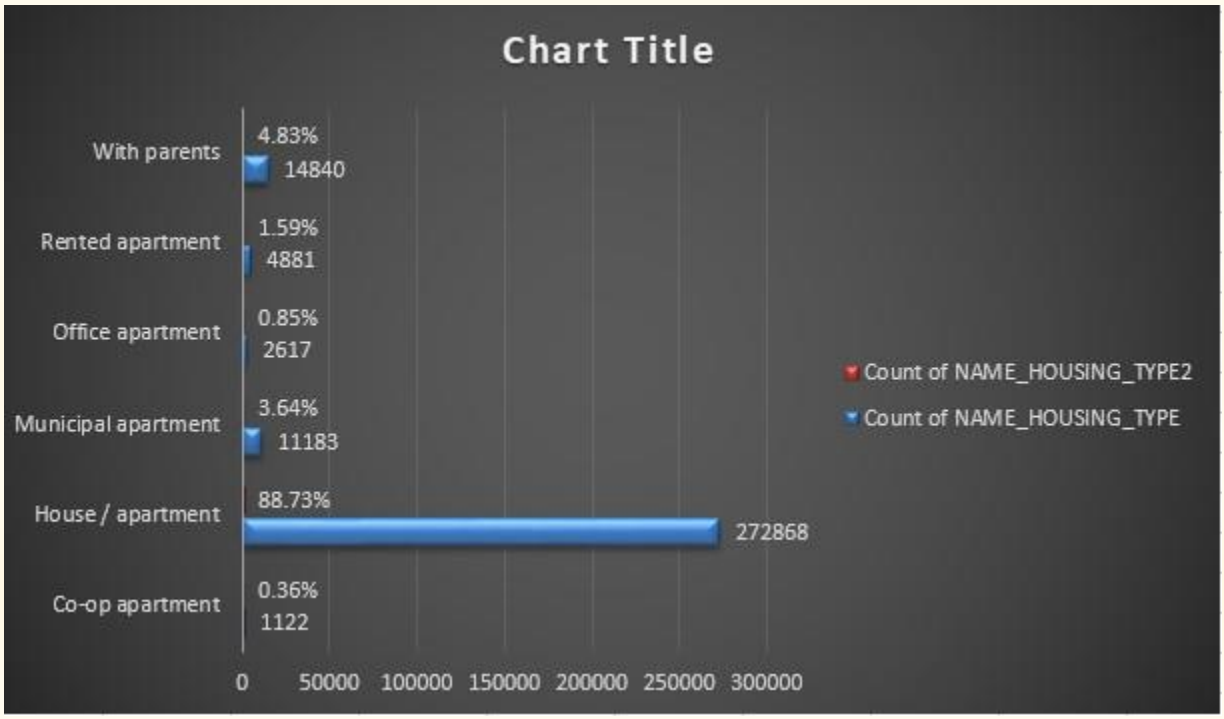
Nearly 92% of all clients had the target variable, according to the Target Variable Pie Chart.Whereas 8% of clients had some sort of issue upon payment, there were no problems.

Findings – II



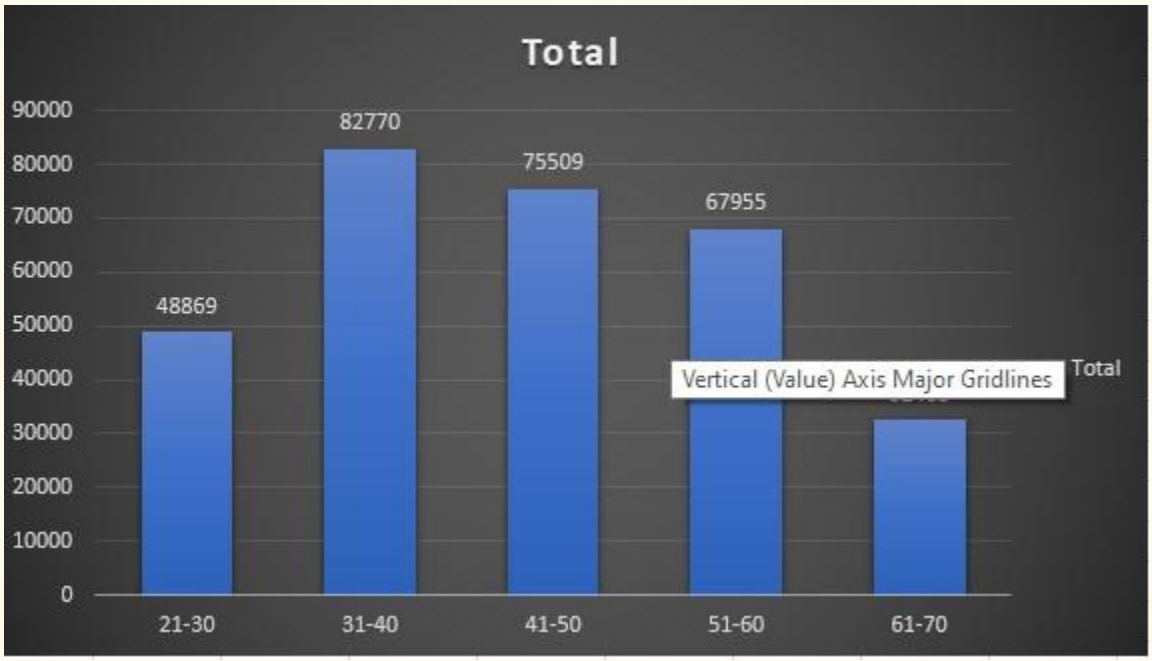
We may deduce that around 66% of clients are female and 34% are male based on the GENDER_VARIABLE column chart.The four applicants' XNA gender designations can be disregarded.

Findings – III



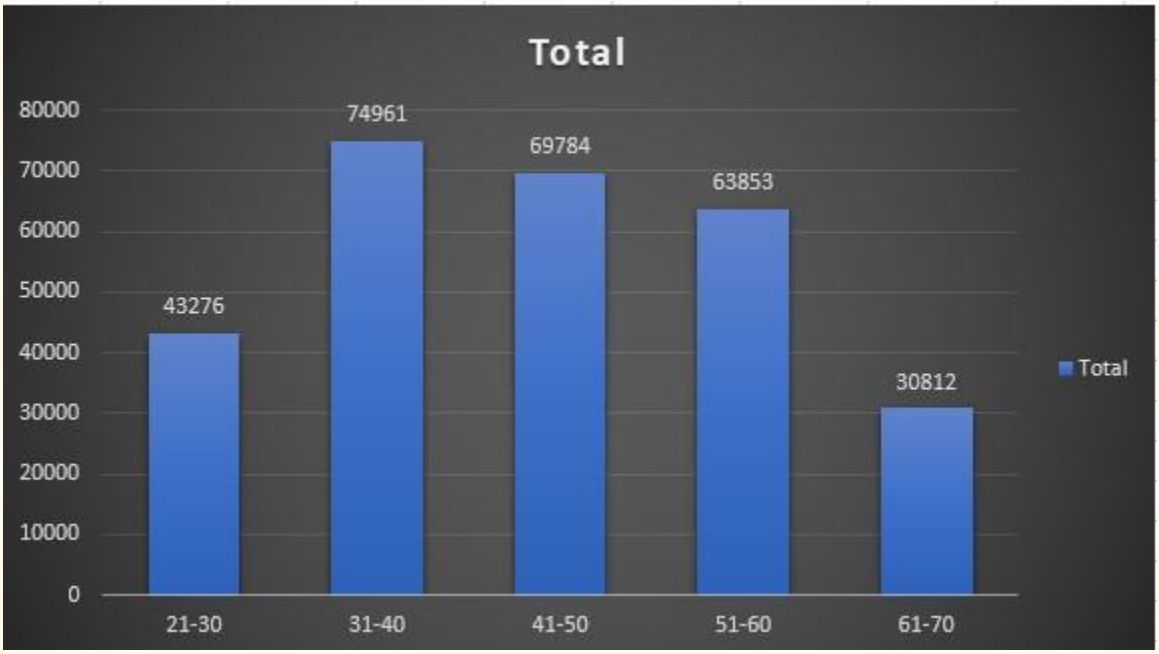
Based on the count and percentage bar graphs.The bank can focus on those populations that do not have their own apartments, such as those who live in co-ops, municipal apartments, rented apartments, and those who live with their parents.

Findings – IV

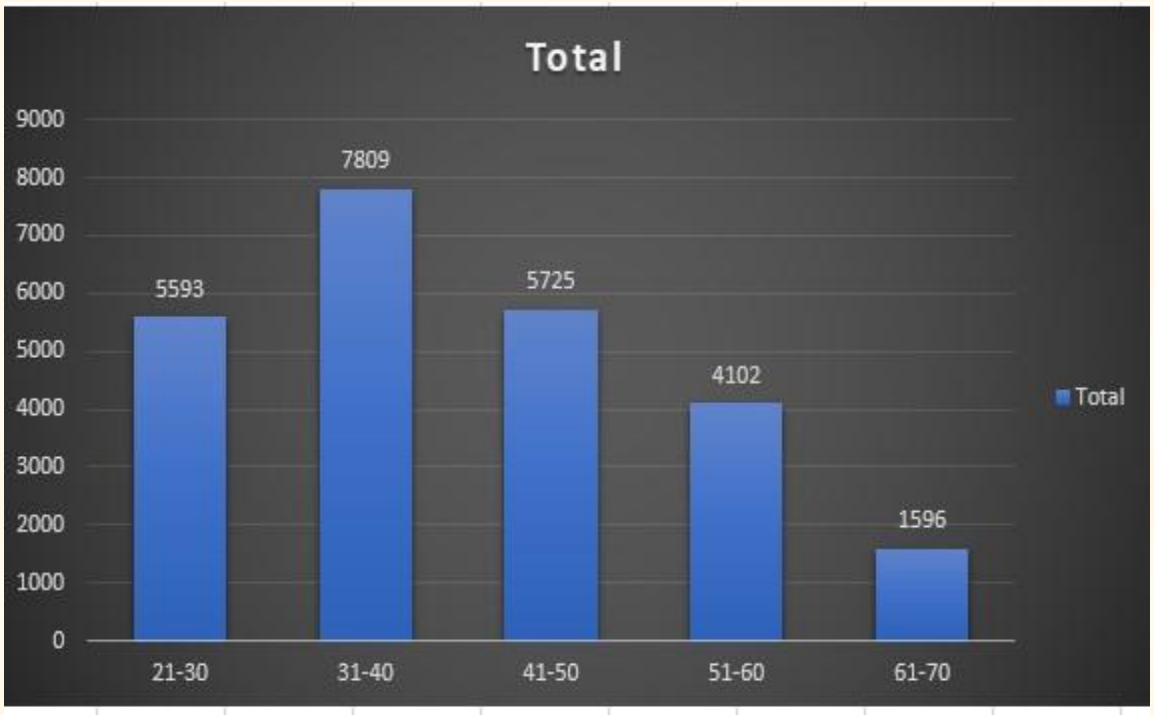


We may deduce that the majority of applicants fall into the Age Group "31-40" from the nearby pub layout.

Findings – V

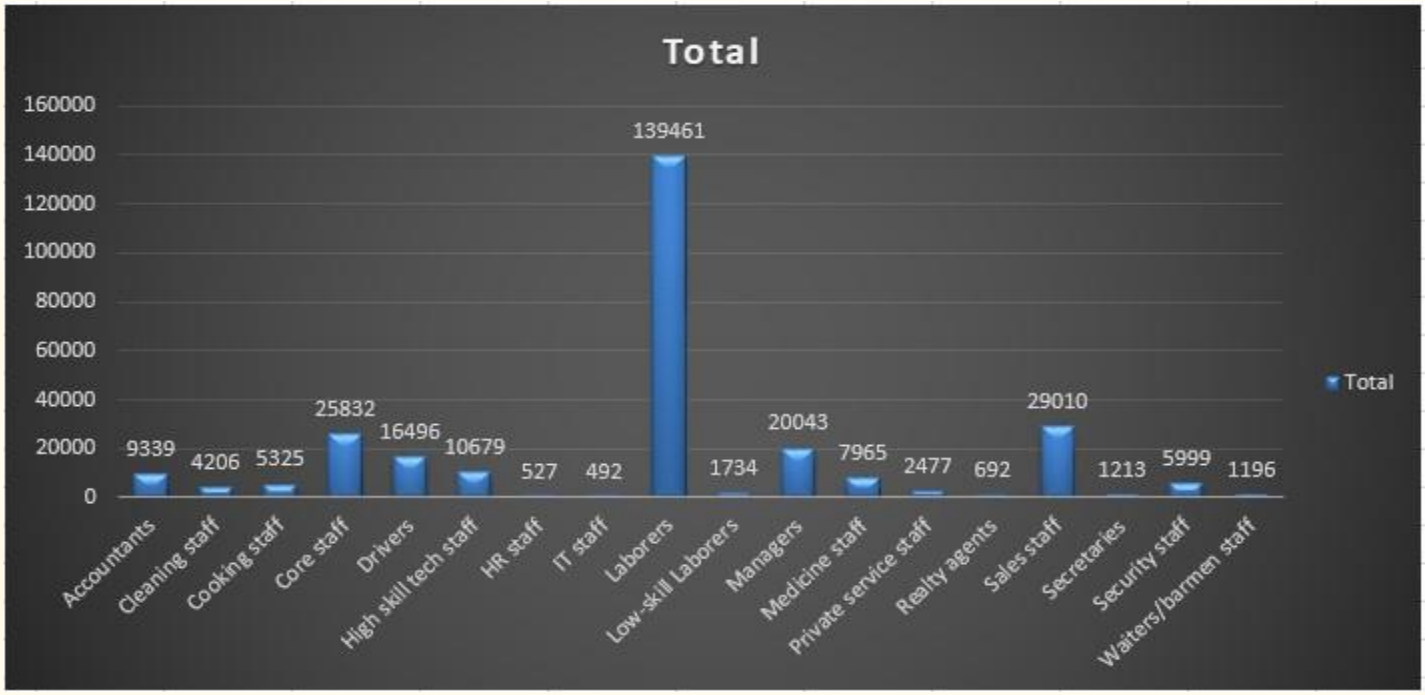


We may deduce from the adjacent bar plot that customers/applicants in the age group "31-40" have the biggest number when it comes to making or returning payments to banks.

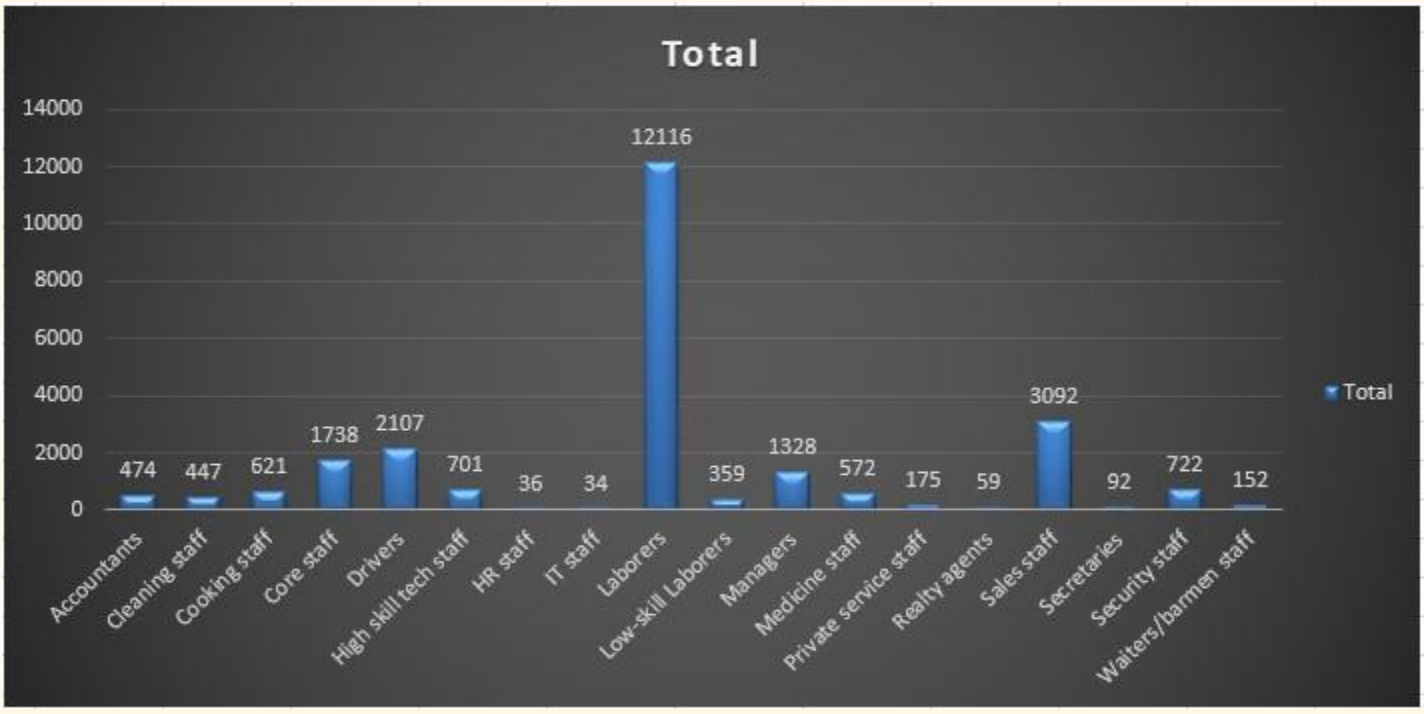


According to the adjacent bar plot, customers/applicants in the age group "31-40" experience the most payment problems while making or returning payments to banks.

Findings – VI

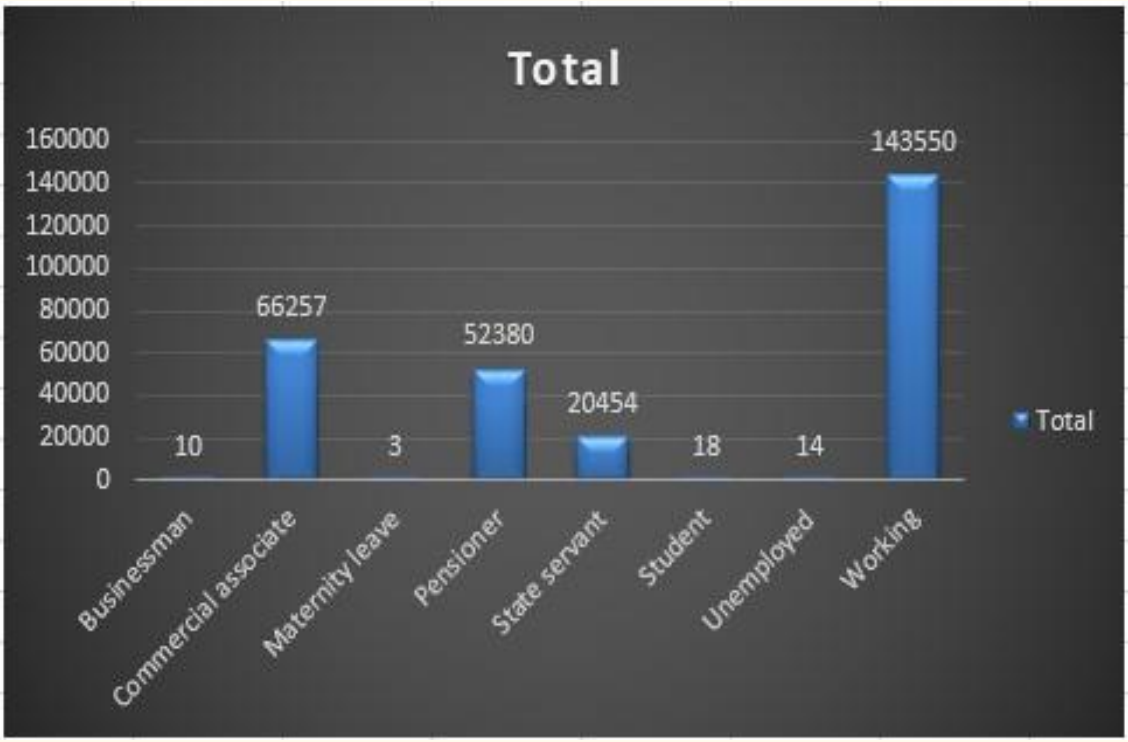


'Labourers' occupation_type clients have the largest count when it comes to clients with no payment concerns, according to the above bar plot, it can be deduced.

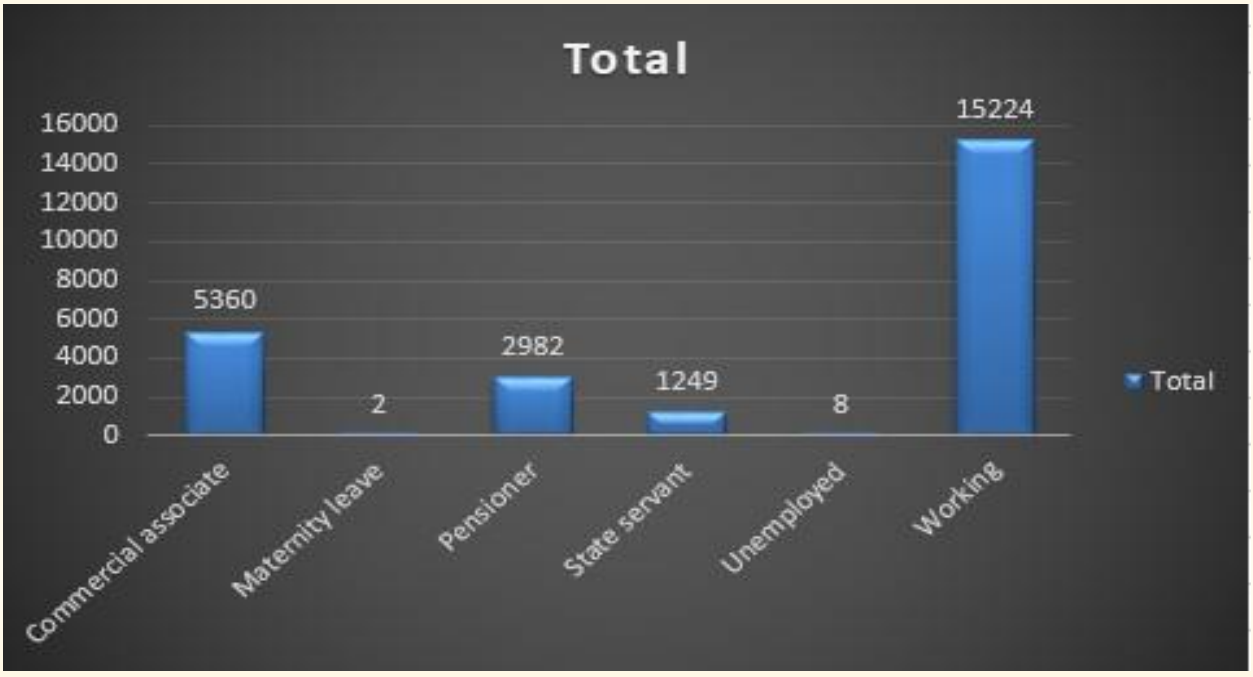


From the above bar plot we can infer that clients with occupation_type ‘Laborers’ have the highest number of count when it comes to clients with payment issues.

Findings – VII

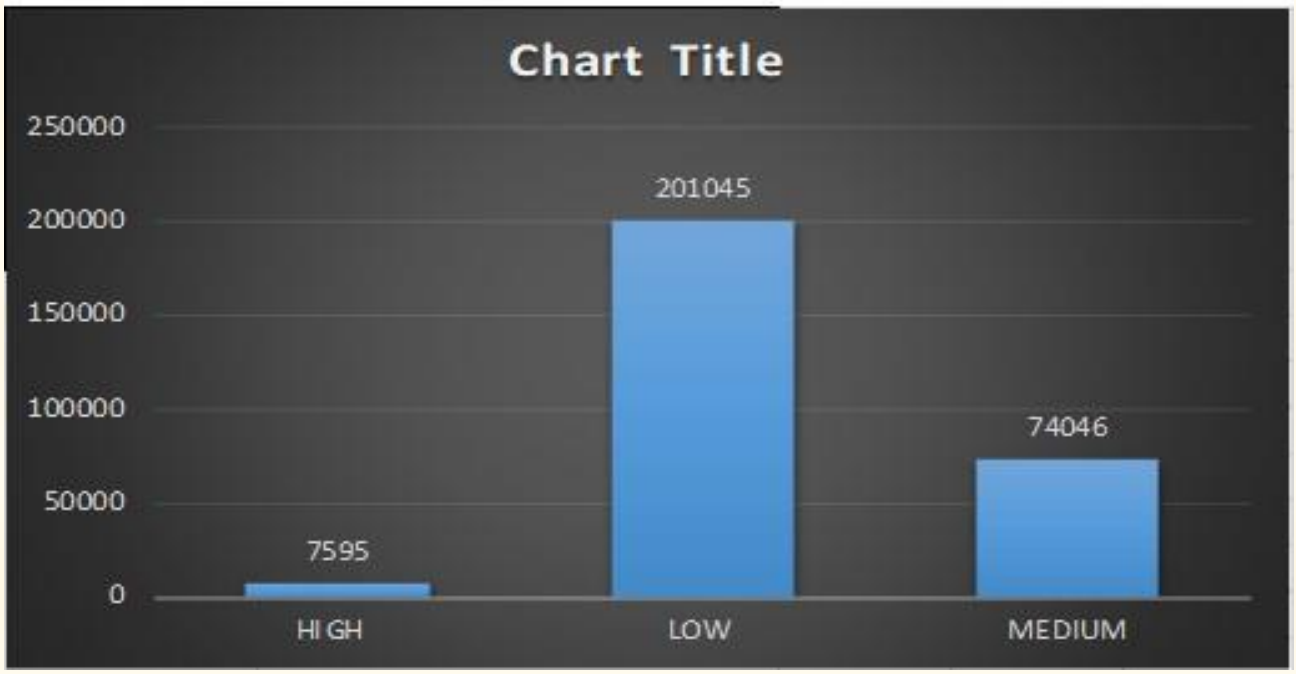


The 'WORKING' income_type customers have the largest count of those who have no payment concerns, according to the aforementioned Bar plot.

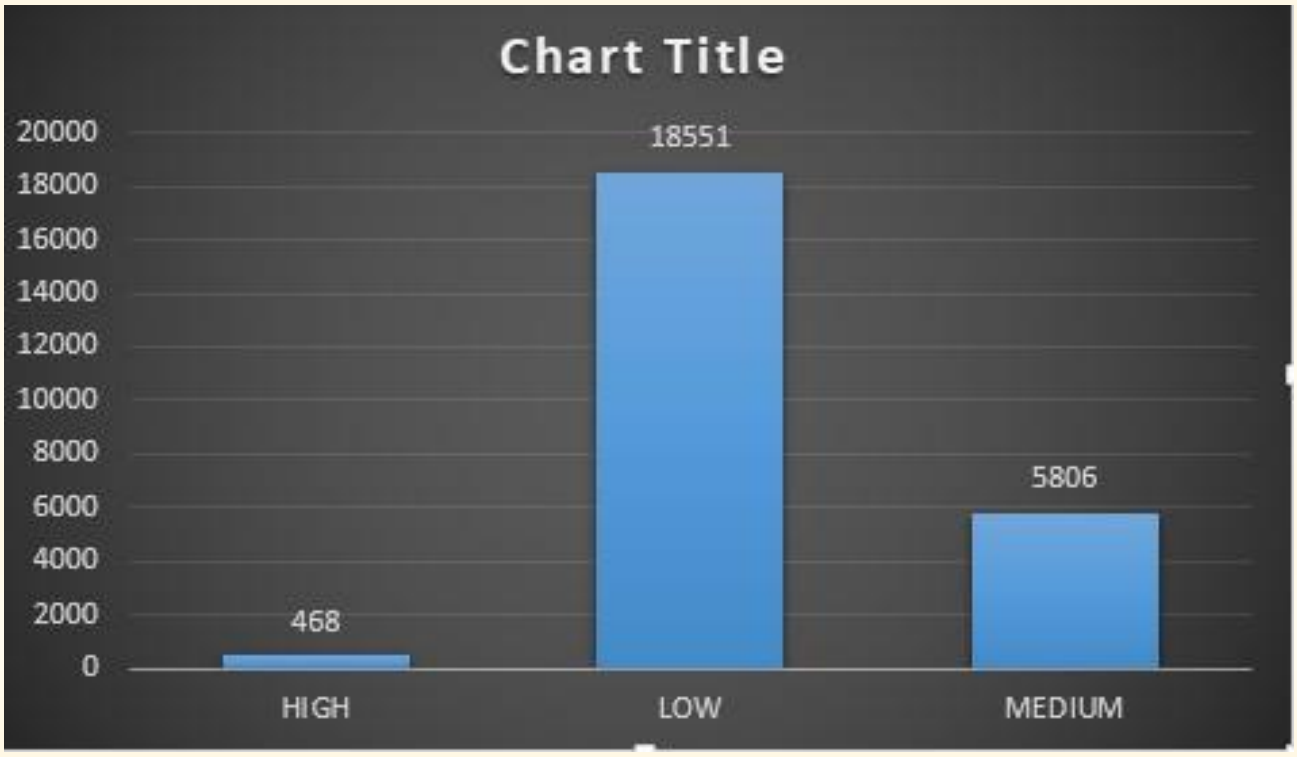


The 'WORKING' income_type customers have the largest count of clients experiencing payment concerns, according to the aforementioned Bar plot.

Findings – VIII

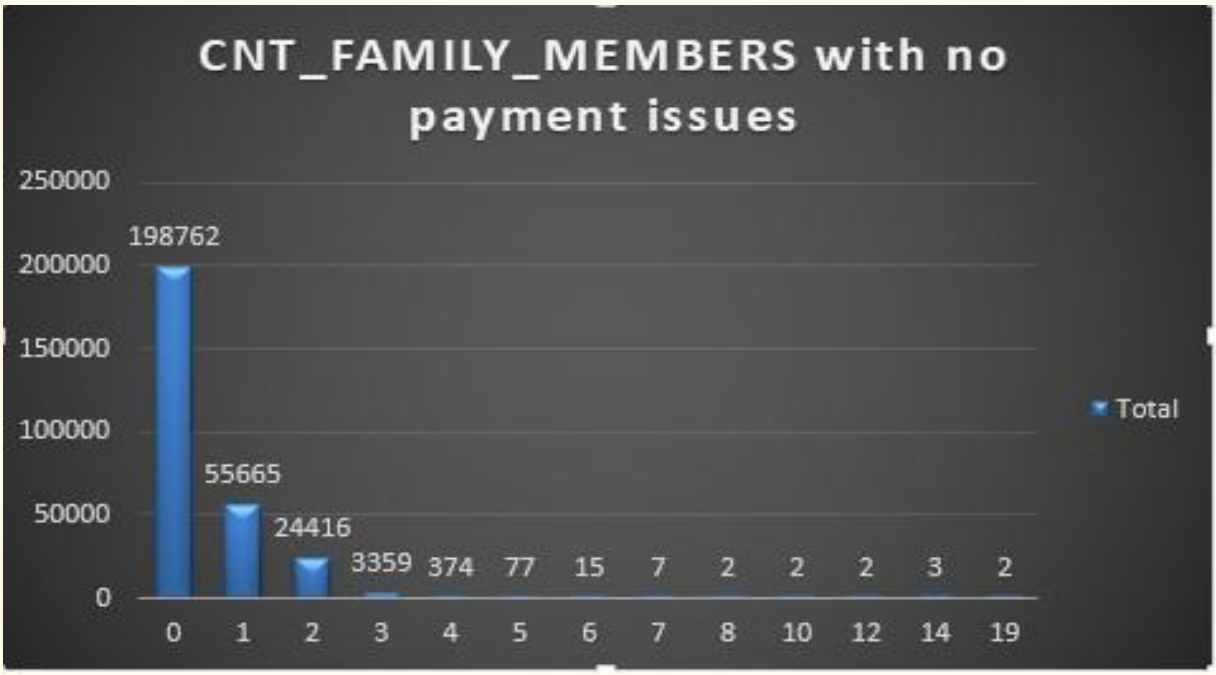


The customer having the whole income range as 'LOW' has the largest count when it comes to clients having no payment concerns, according to the aforementioned bar plot.

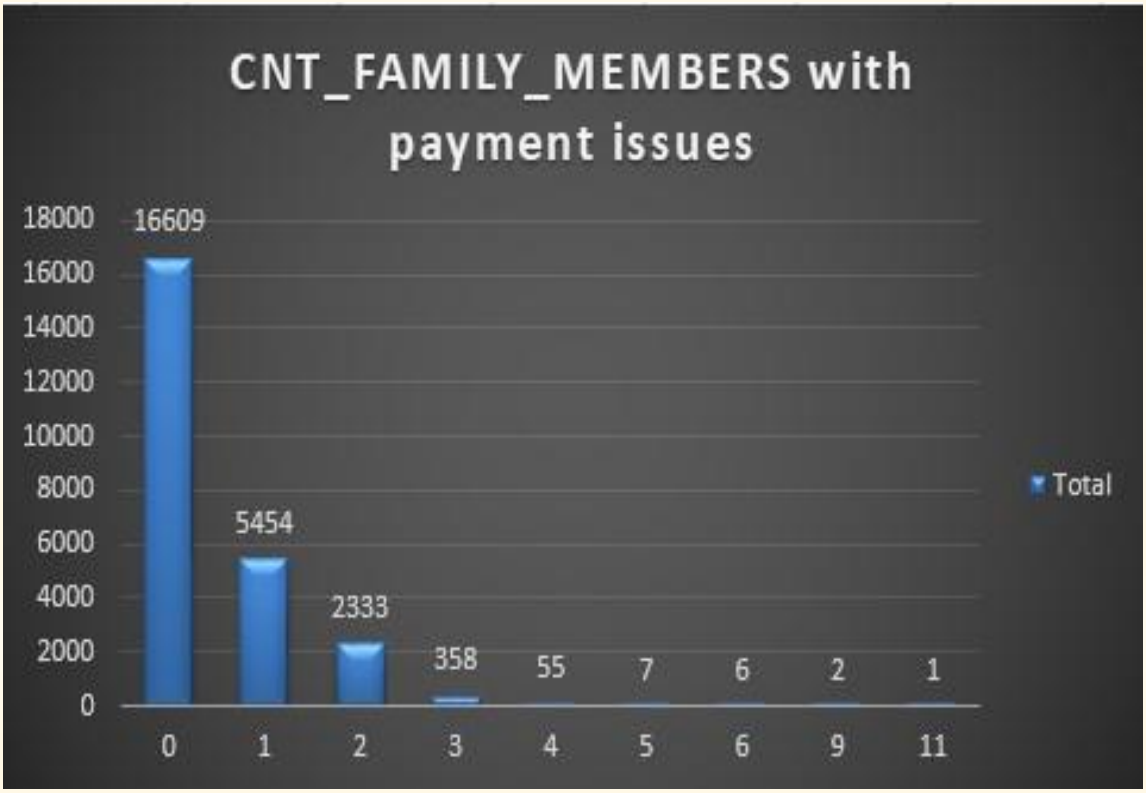


The accompanying bar plot indicates that clients with total income ranges that are "LOW" have the largest percentage of clients with payment difficulties.

Findings – IX

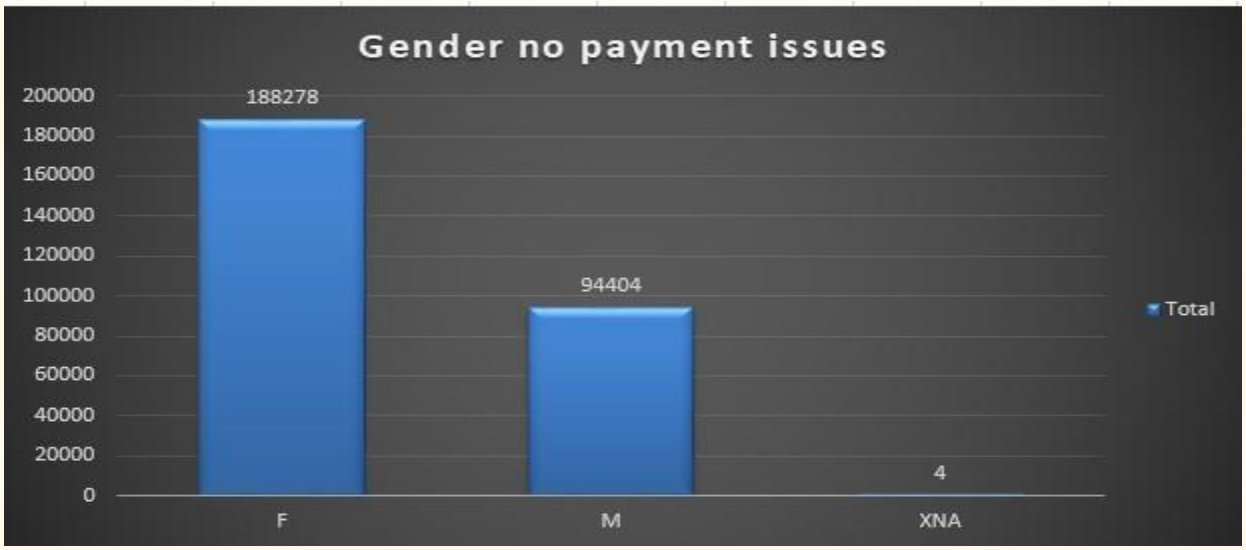


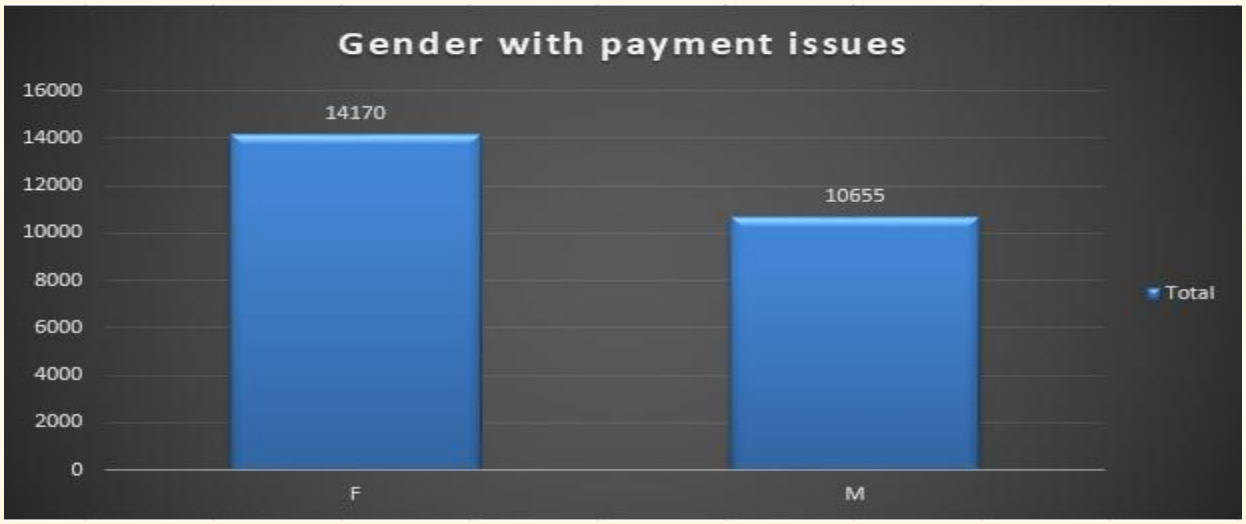
According to the above Bar Plot, clients with no family members have the highest percentage of clients with no payment concerns.



According to the aforementioned bar plot, customers with no family members are the ones who have the most number of payment problems.

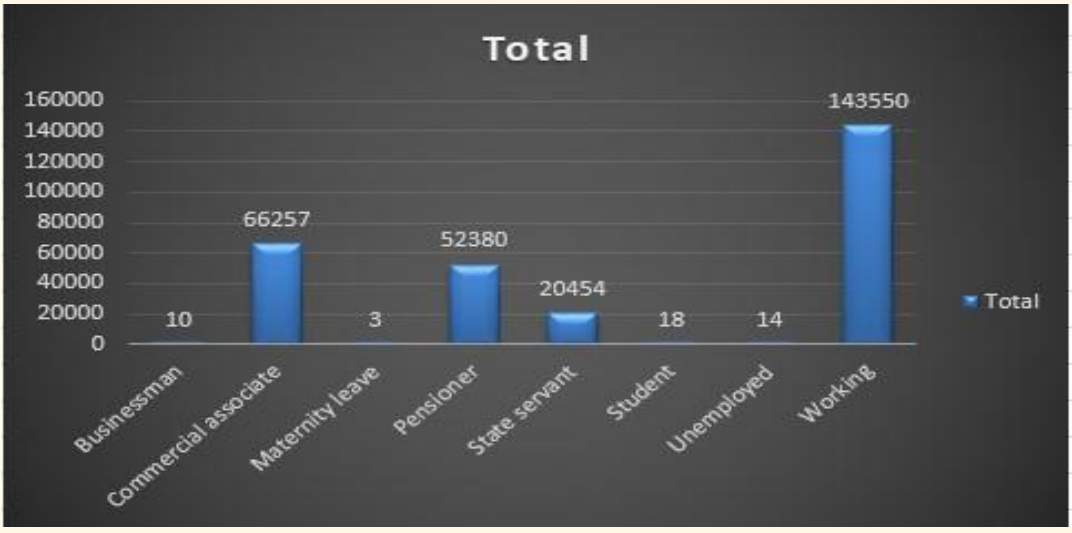
Findings – X

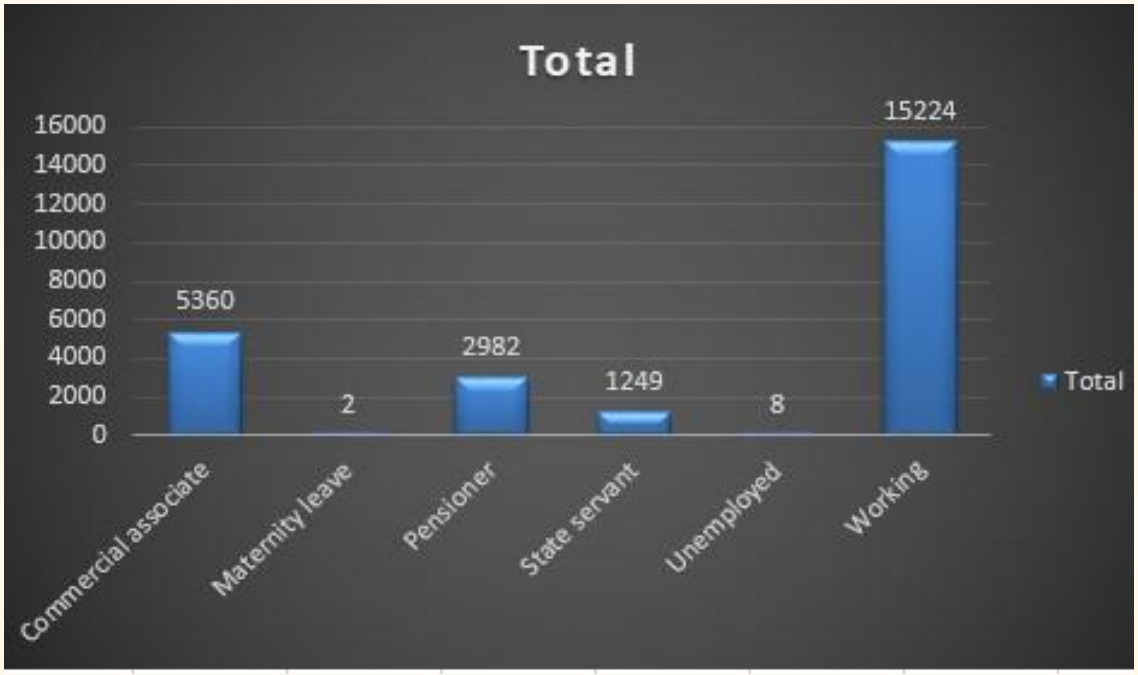




According to the above bar plot, clients with CODE_GENDER = 'F' had the most non-defaulters (188278-14170 = 174108).

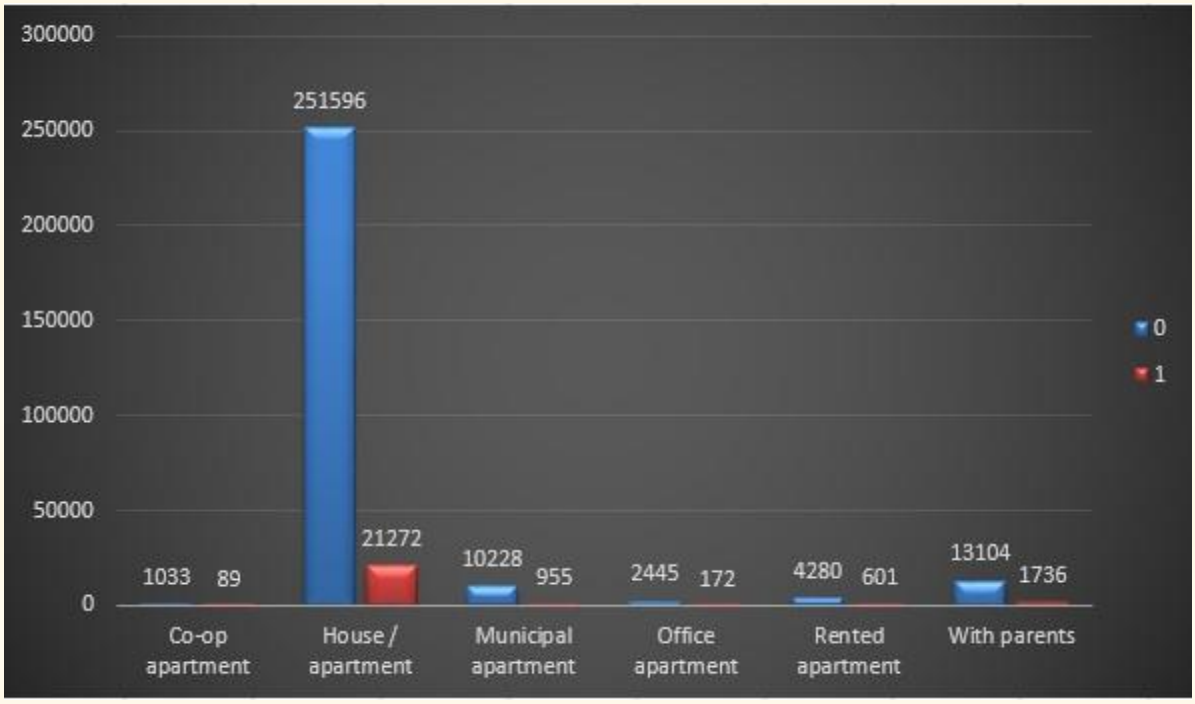
Findings – XI





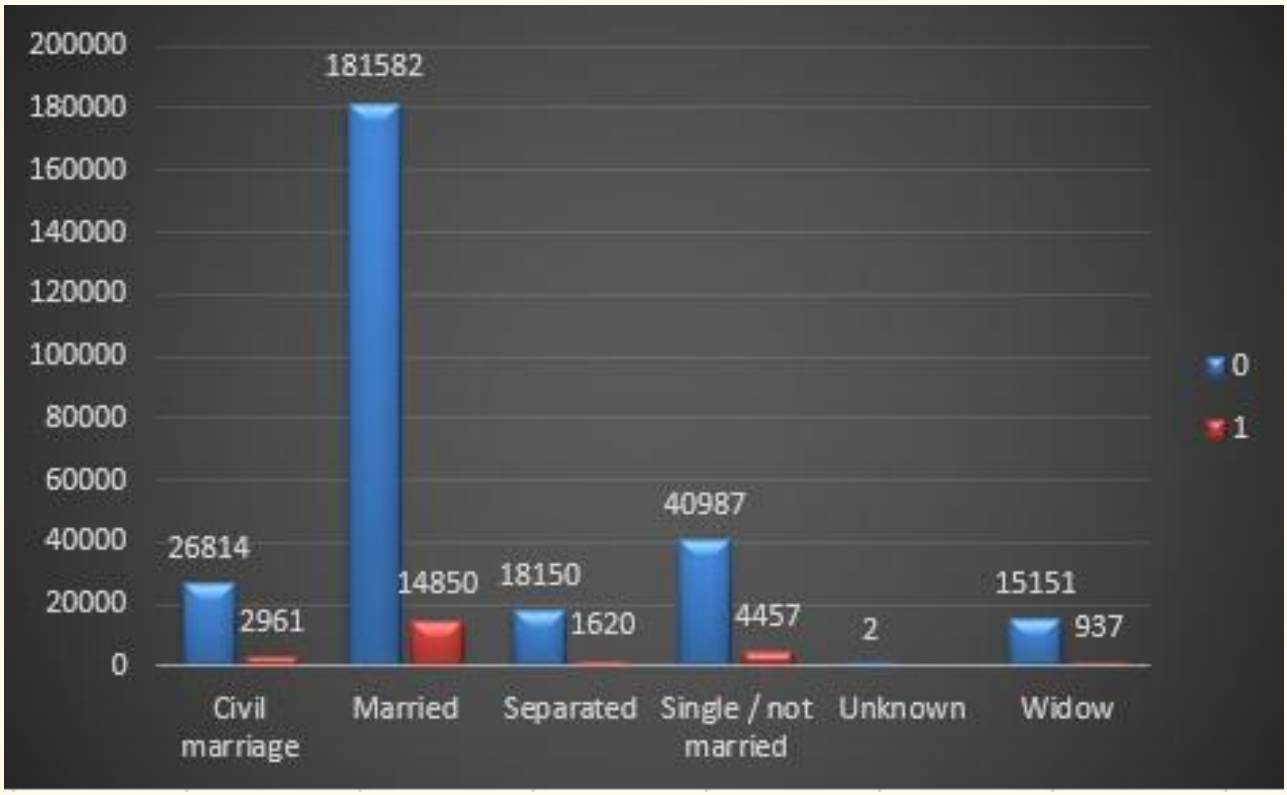
According to the adjacent bar plot, customers with NAME_INCOME_TYPE = "WORKING" have the largest number of non-defaulters, or 143550-15224 = 128326.

Findings – XII



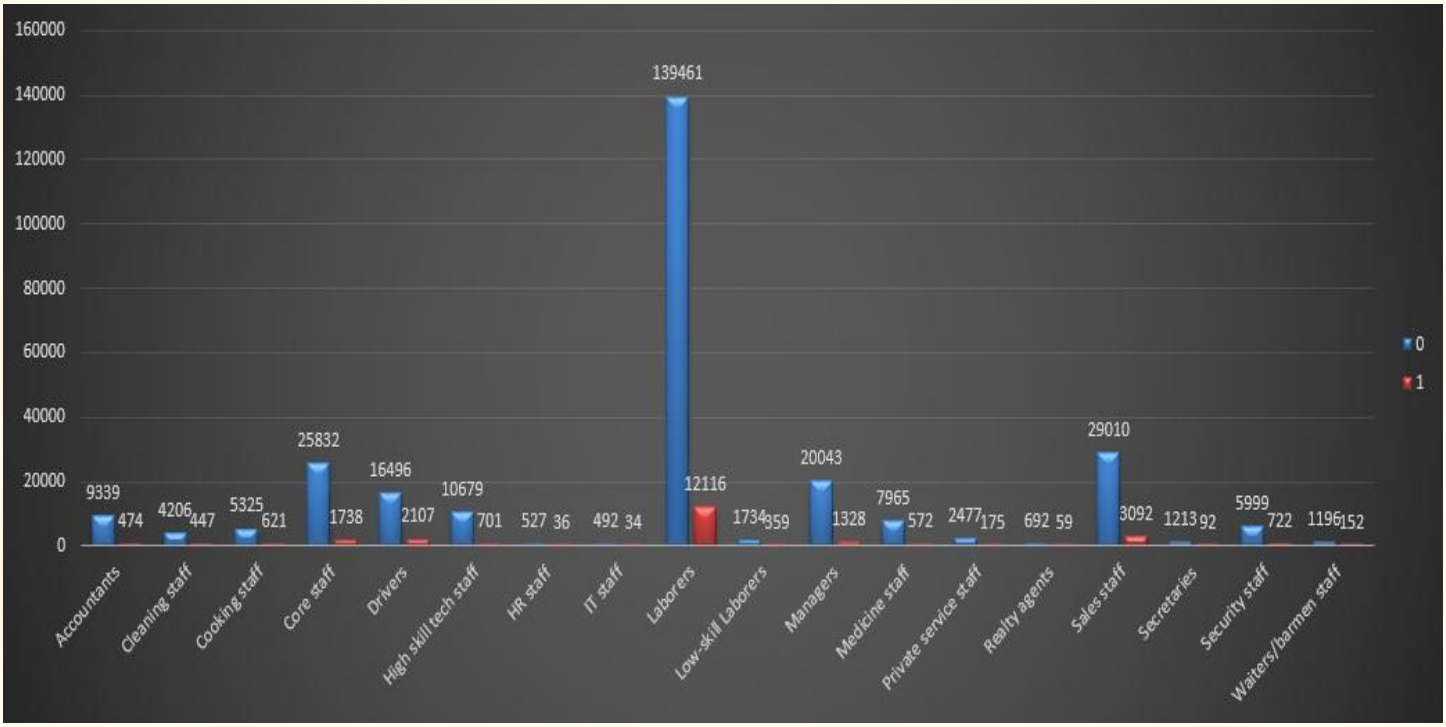
According to the above bar plot, clients with NAME_HOUSING_TYPE = "House/Apartment" had the largest number of non-defaulters, with a total of 251596-21272 = 230324.

Findings – XIII



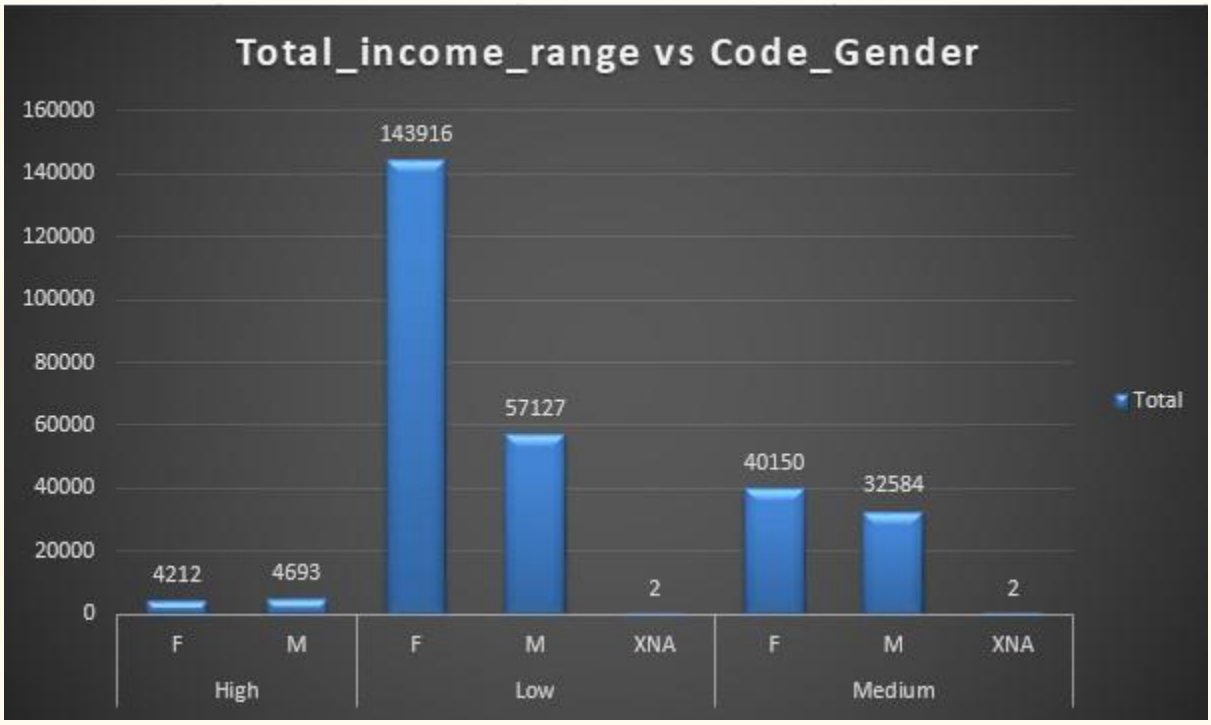
customers with NAME_FAMILY_STATUS = 'MARRIED' are, according to the adjacent Bar Plot, customers had the most nondefaulters, with a total of 166732 (181582 - 14850).

Findings – XIV



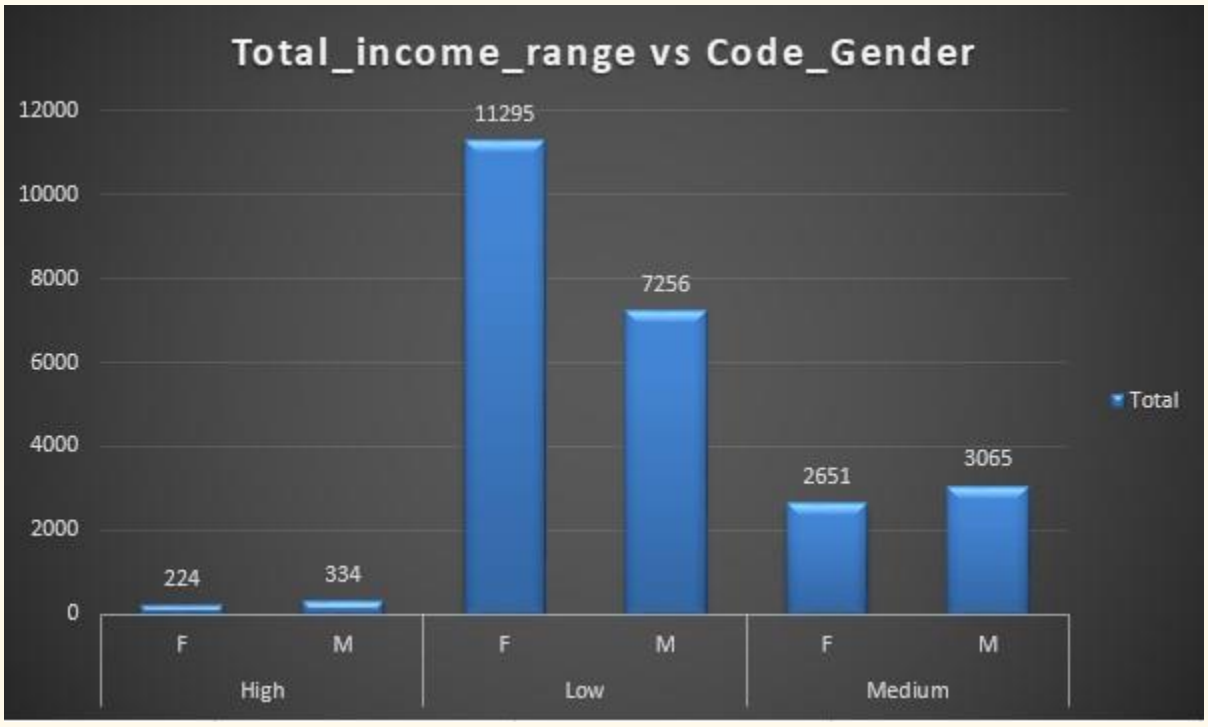
The customers with occupation_type = "Labourers" have the greatest count for non-defaulters, which is 139461-12116 = 127345, according to the adjacent Bar plot.

Findings – XV



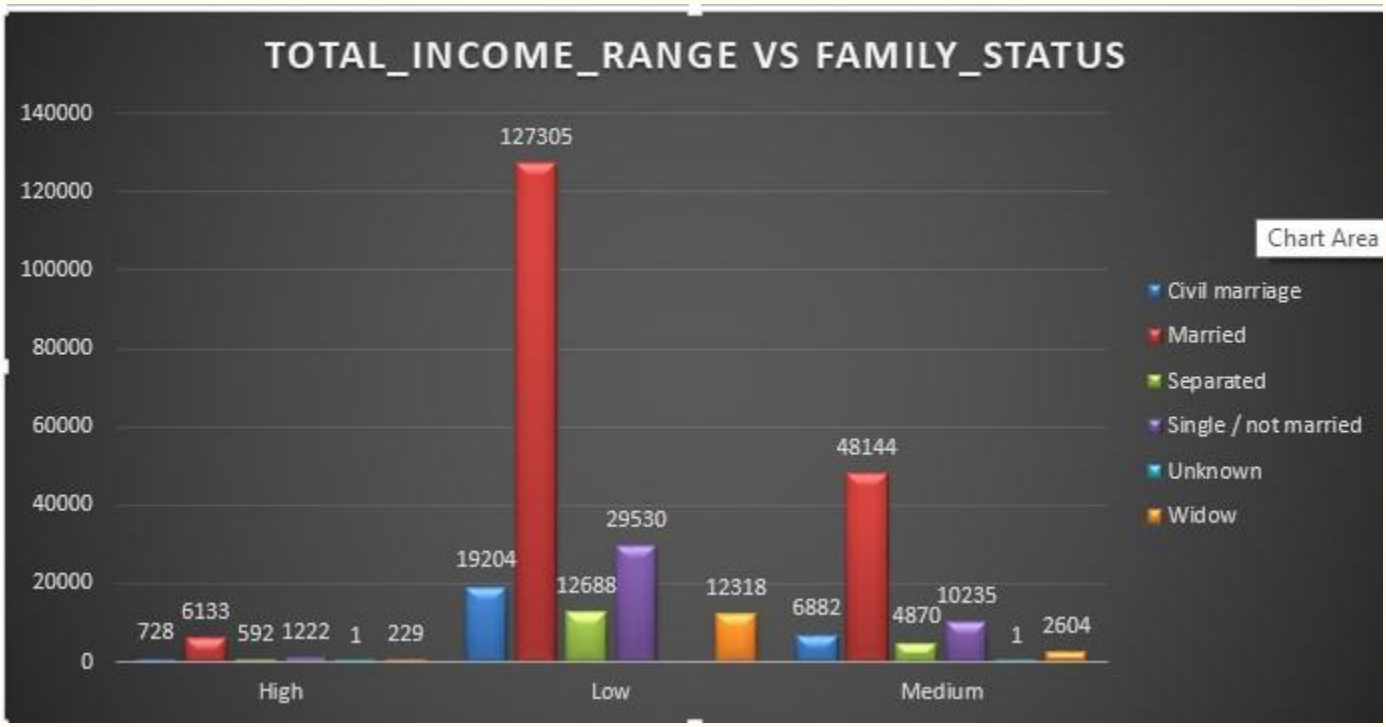
The majority of consumers without payment concerns are women who are in the low income bracket, according to the aforementioned bar map.

Findings – XVI



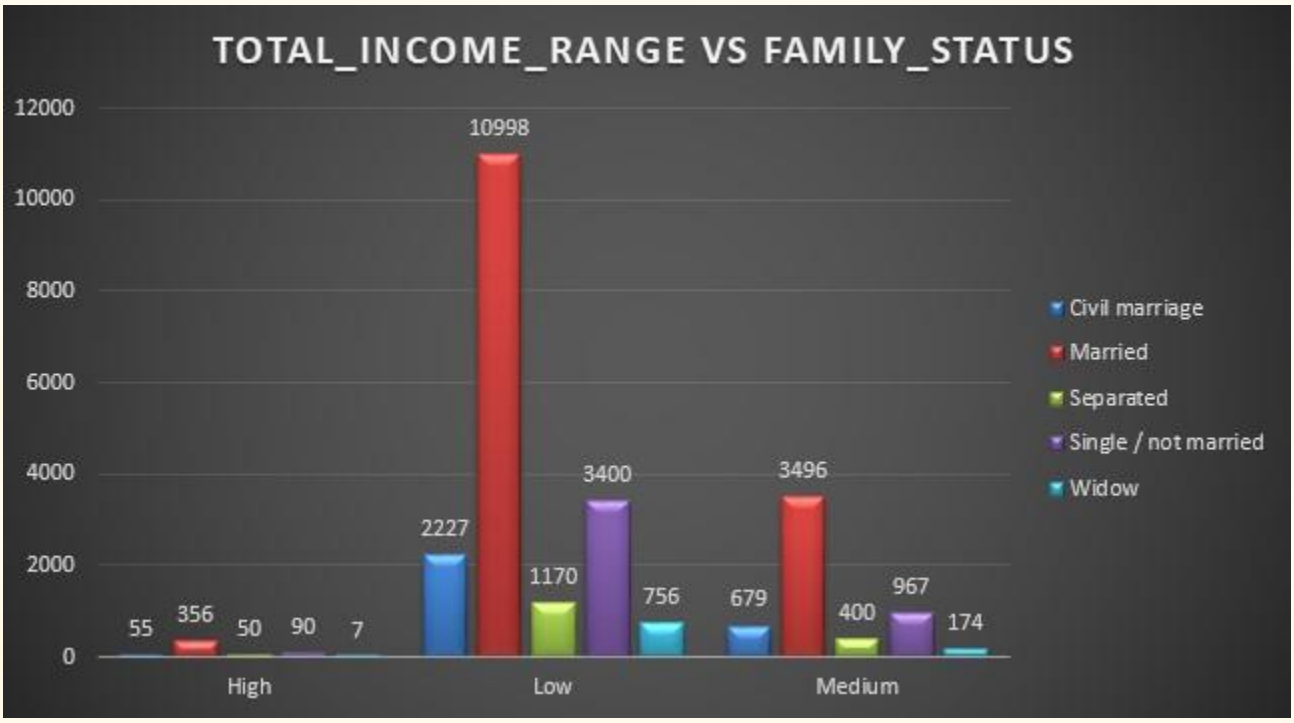
The accompanying bar plot indicates that the majority of clients that have payment concerns are women who fall into the low income group.

Findings – XVII



Clients with a family status of "Married" and a total income range of "Low" are those most likely to have no payment troubles, according to the adjacent Bar plot.

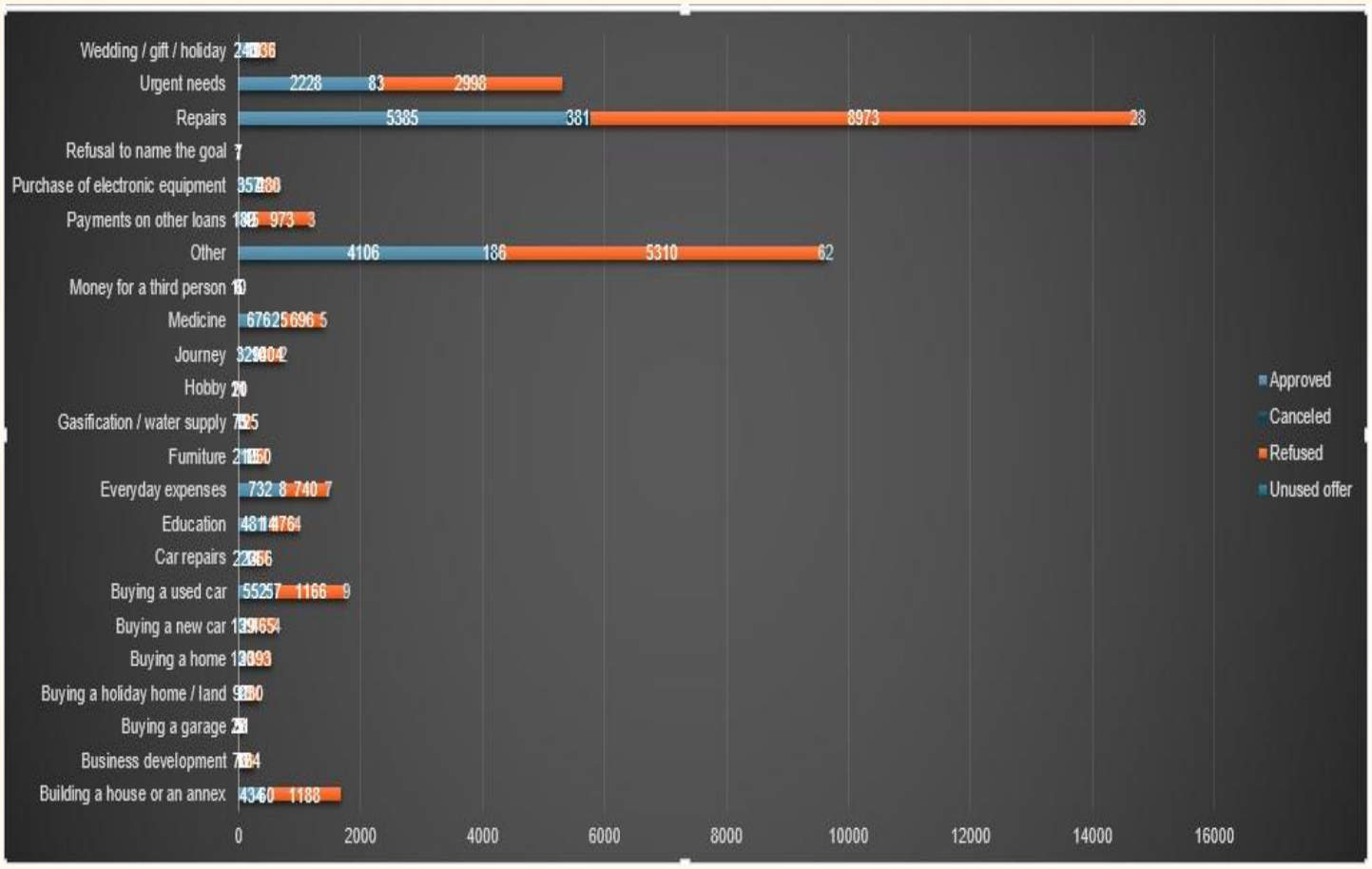
Findings – XVIII



Clients with a total income range of "Low" and a family status of "Married" are those most likely to experience payment troubles, according to the adjacent Bar plot.

Findings – XIX

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



The Name of Contract status, i.e., Repairs work, has the maximum number of Loans Approved according to the above Table and Bar Plot.

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?

- These laws provide loans to women clients at a relatively low interest rate; in some cases, people purposefully use their retired/household mother or household wife so they can receive some sort of concession, such as low interest rates when applying for home loans. This is especially common in countries like India.

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

- Because people in other groups are looking for their own home with their own nameplate, such as those who live with their parents in municipal apartments, cooperative apartments, or rented apartments. Additionally, joint families are becoming less common in India today, and the next generation prefers to live in their own 1/2 BHK instead of a large family apartment.

Why should a bank favour working class customers over those in the state government, even when state employees are paid regularly and receive several benefits?

- While it is true that those who work for the state government receive many perks, they also receive housing allowances that are higher than those of the working class, and in certain circumstances, they even receive apartments where they may live with their family; The working class, on the other hand, does not receive such housing allowances or receives very little of them, nor do they receive an apartment to live in for the duration of their professional lives (i.e., until retirement). As a result, the working class chooses to buy their own home by taking out a mortgage.

Why shouldn't the bank approve loans to customers who identify as "Laborers" even if this group has the greatest percentage of nondefaulters?

- Workers only take out personal loans for things like marriage or home repairs, and because the amount they borrow is less and the interest rate is lower than it is for things like home loans, vehicle loans, etc., banks will make less money from them.

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

- Women who belong to these organisations frequently profit from government programmes for launching their own businesses, catering services, or parlours by taking out small loans for these purposes.

Conclusion

1. The Name of Contract status, i.e., Repairs work, has the largest number of Loans that have been approved, according to the above Bar Plot.
2. So, both the Applications Dataset and the Precious Applications Dataset are being used for the analysis.
3. The percentage of defaulters (target = 1) is around 8%, whereas the percentage of non-defaulters (target = 0) is approximately 92%.
4. The Bank often loans more money to female customers than to male customers since there are fewer female customers on the list of defaulters. If the credit amount is met, the bank may still hunt for additional male customers.
5. Additionally, customers from the Working Class are more likely than those from the Commercial Associate category to make their loan payments on time.
6. Consumers with education levels of secondary or higher secondary or above have a tendency to repay loans on schedule, allowing banks to prioritise lending to those consumers.
7. Clients with LOW credit amounts tend to pay off their loans on time as opposed to HIGH and MEDIUM credit amounts. Clients in the Age Groups 31–40 have the greatest rate of timely loan repayment, followed by clients in the Age Groups 41–60.
8. Compared to other housing types, customers who live with their parents often pay off their debts rapidly. Therefore, a bank may extend credit to customers who live with their parents.
9. consumers who are taking out loans to buy a new home, or who are taking out loans to buy a new car, or who have an income type like "State Servant," have a tendency to repay their debts on time, thus banks should favour consumers with this background.
10. The Bank should exercise greater caution when providing loans to customers for repairs since they have a high number of defaulters in addition to a high number of defaulters.

Analyzing the Impact of Car Features on Price and Profitability

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.

The Problem

A data analyst could use this dataset to gain insights into various aspects of the automotive industry, such as:

Analyzing trends in car features and pricing over time: By examining the variables in the dataset, a data analyst could identify how car features and prices have changed over time, which could help manufacturers make informed decisions about product development and pricing.

Comparing the fuel efficiency of different types of cars: By looking at the MPG variables in the dataset, a data analyst could compare the fuel efficiency of different types of cars

and identify which types are the most efficient. This could help consumers make informed decisions about which car to purchase.

Investigating the relationship between a car's features and its popularity: By examining the popularity variable in the dataset, a data analyst could identify which features are most popular among consumers and how they affect a car's popularity. This could help manufacturers make informed decisions about product development and marketing.

Predicting the price of a car based on its features and market category: By using the various features and market category variables in the dataset, a data analyst could develop a model to predict the price of a car. This could help manufacturers and consumers understand how different features affect the price of a car and make informed decisions about pricing and purchasing.

Overall, this dataset could be a valuable resource for data analysts interested in exploring various aspects of the automotive industry and could provide insights that could inform decisions related to product development, marketing, and pricing.

Design

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

Then dropping the columns which have no use for the analysis that we will be doing

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

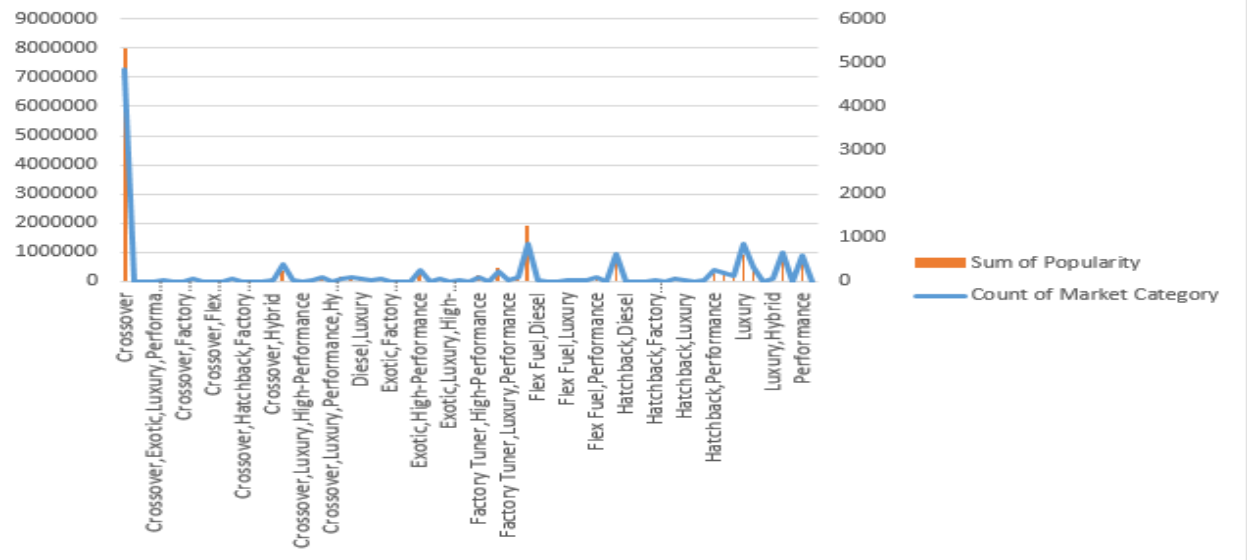
Then we need to get rid off the duplicate values in the dataset which can be achieved by using the ‘Remove Duplicate Values/Cells’ available in the ‘Data’ tab

Then, in order to uncover insightful information, I'll perform regression analysis along with analysis using a pivot table and graph.

Regression allows me to determine other factors' coefficents with regard to MSRP and detect correlation, which is helpful for figuring out how two variables are related to one another. Graphs can help you comprehend the distribution of variables better.

Findings – I

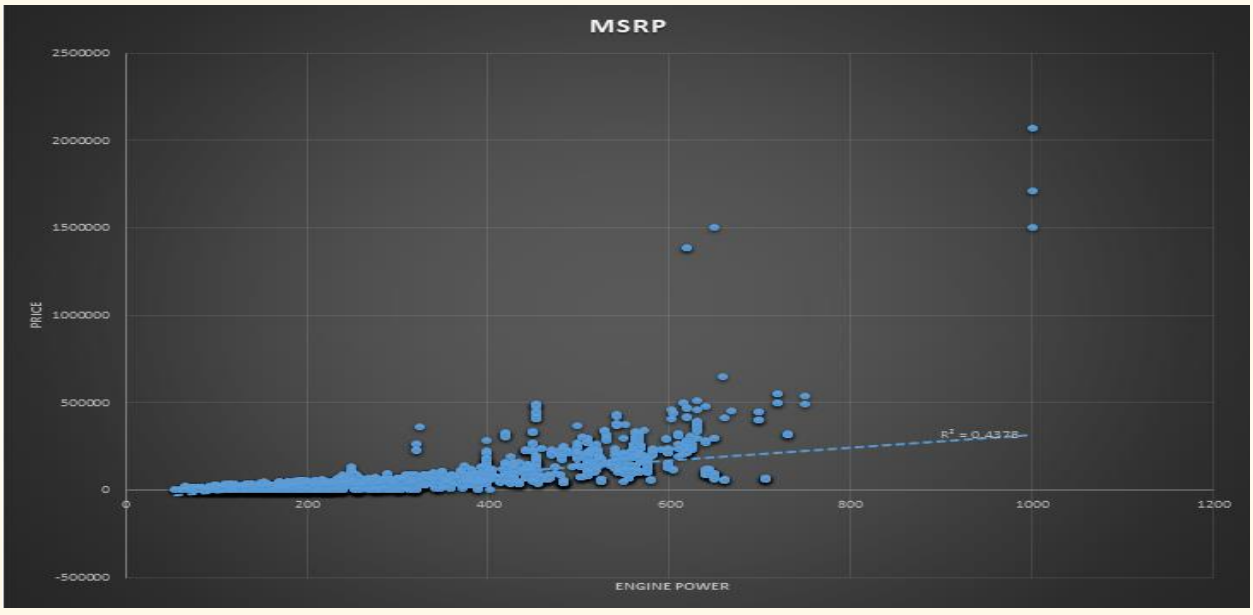
1.How does the popularity of a car model vary across different market categories?



In this graph we can see that count of market category is proportional to sum of popularity.so we can say that popularity depends on market category means crossover market is more popular as compare to others.

Findings – II

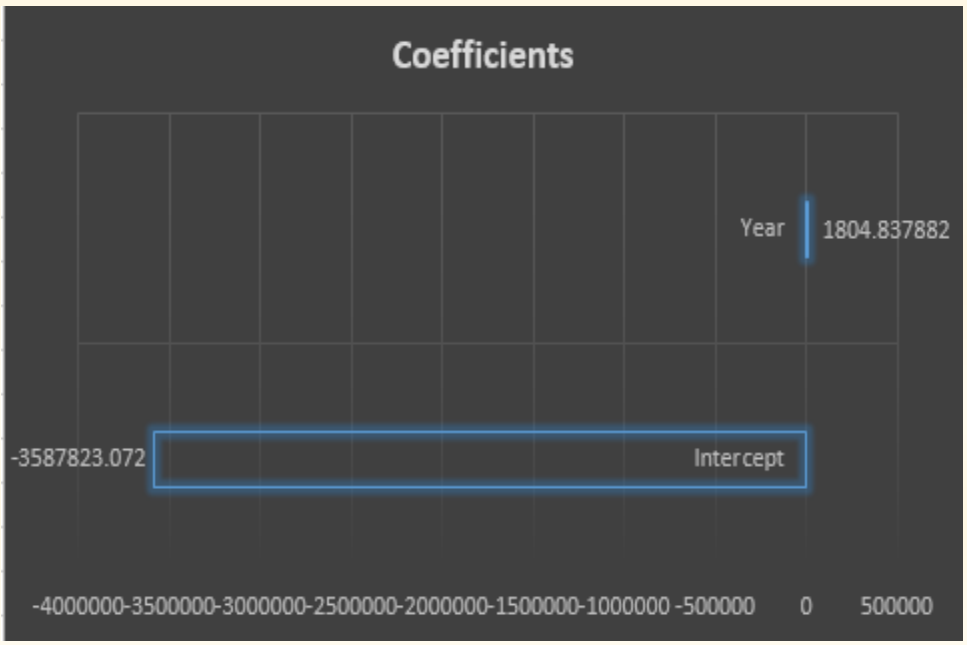
2. What is the relationship between a car's engine power and its price?

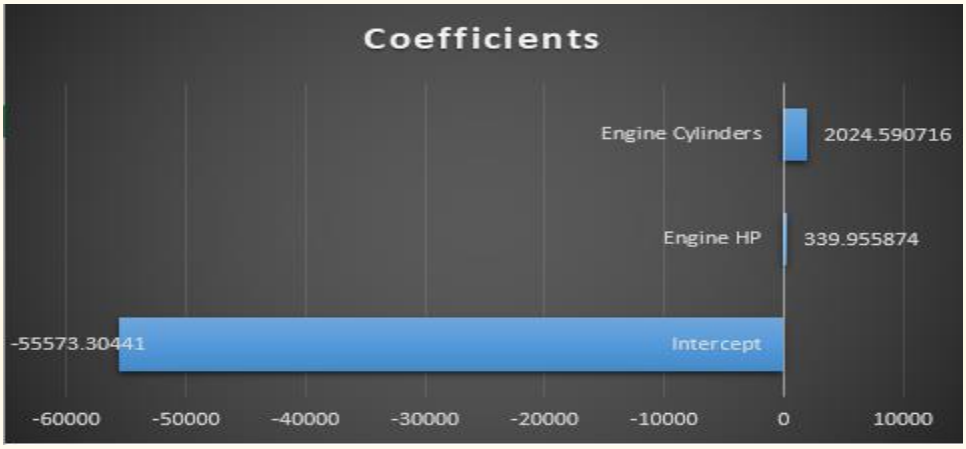


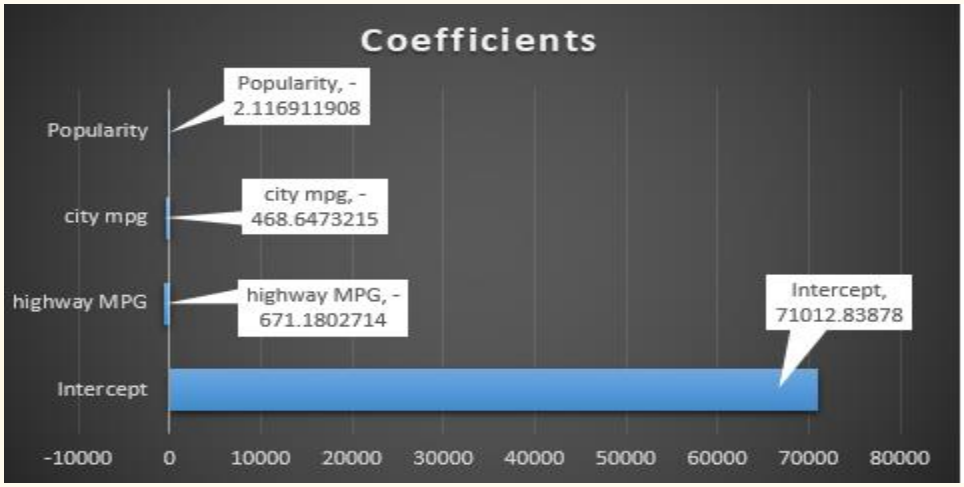
In this graph we can see that car engine power is linearly proportional to MSRP also we can see that some values in MSRP behave like a outliers but we know that some cars are expensive like Bugatti.so we can say that when MSRP increases when power of car increases.

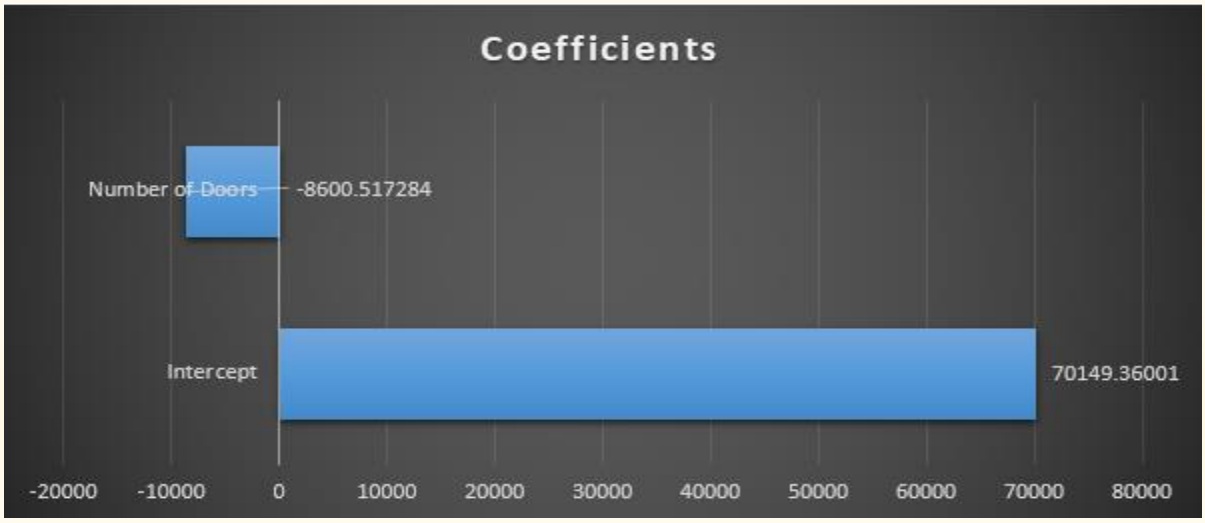
Findings – III

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





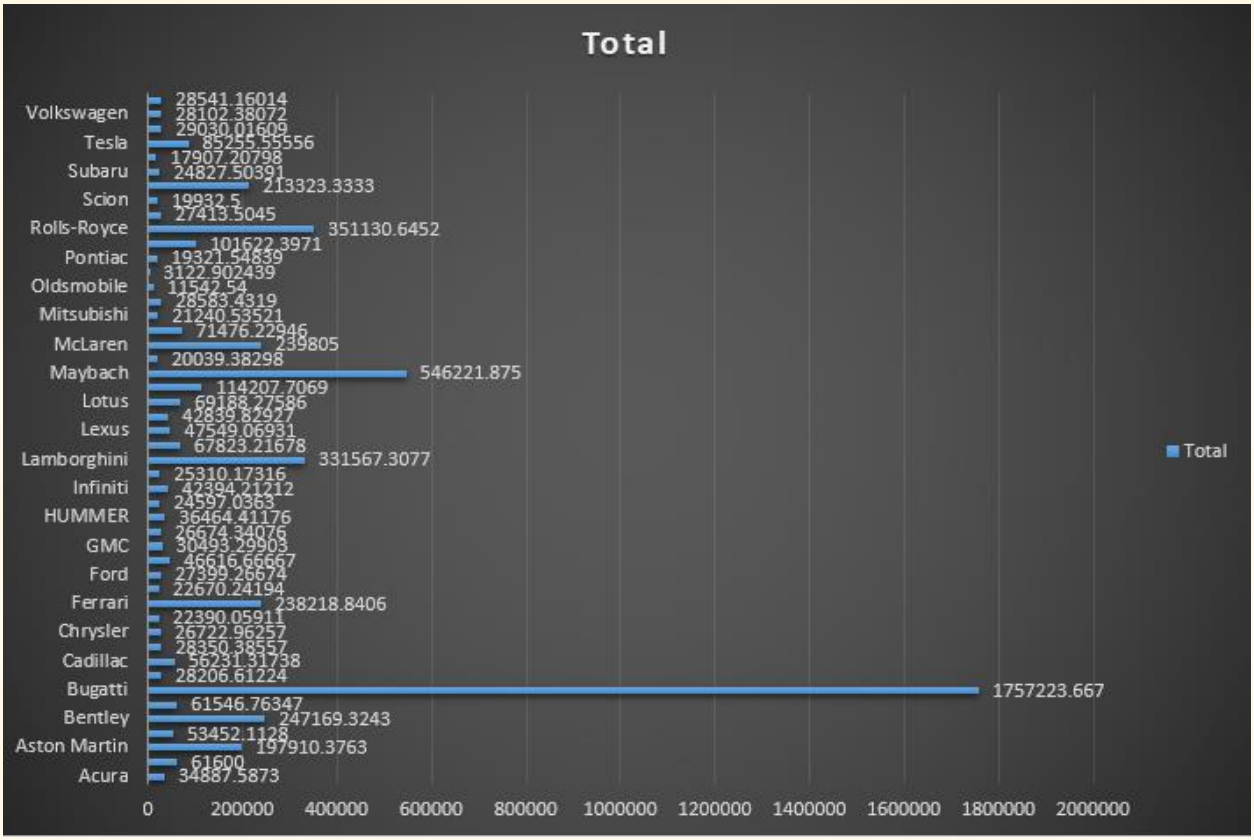




Here, we can see that the engine cylinders coefficient is larger than the year, city mpg, and high mpg for automobile features; thus, we can conclude that this is a crucial characteristic of any manufacturer as it is inversely related to the engine horsepower.

Findings – IV

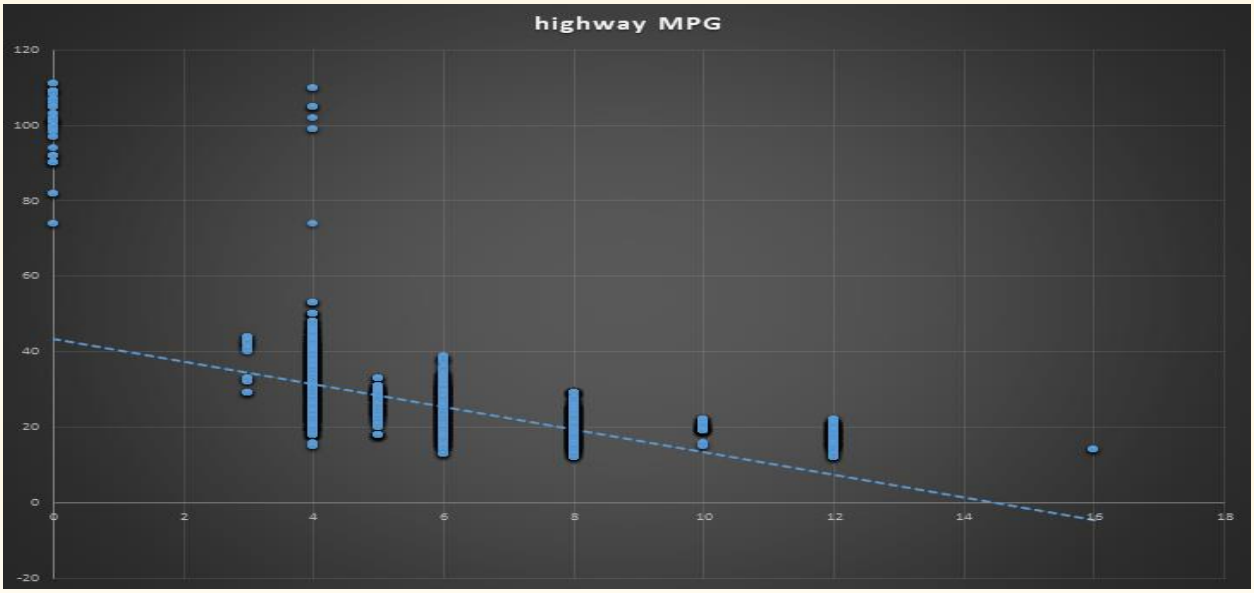
4.How does the average price of a car vary across different manufacturers?



Here we can see that Bugatti has a highest average price and second highest is Maybach.Also avg car price is depend on every brand some are expensive because they are dealing with luxury segment and high HP.

Findings – V

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



We are comparing the fuel economy in this scatter plot with regard to the number of cylinders, and since we can see that efficiency decreases as the number of cylinders increases, we may conclude that the relationship between the two is inverse or we can say that number of cylinders increases then power also increases and fuel efficiency decreases.

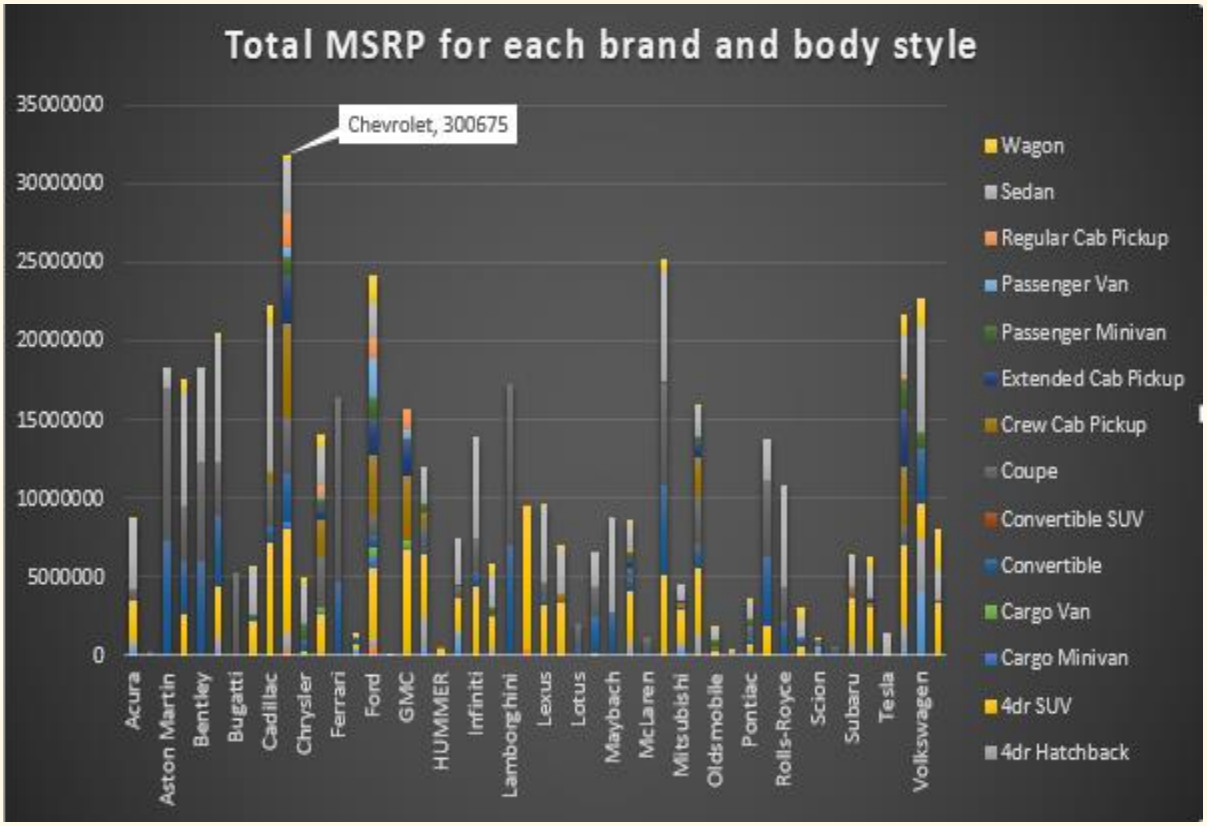
Correlation

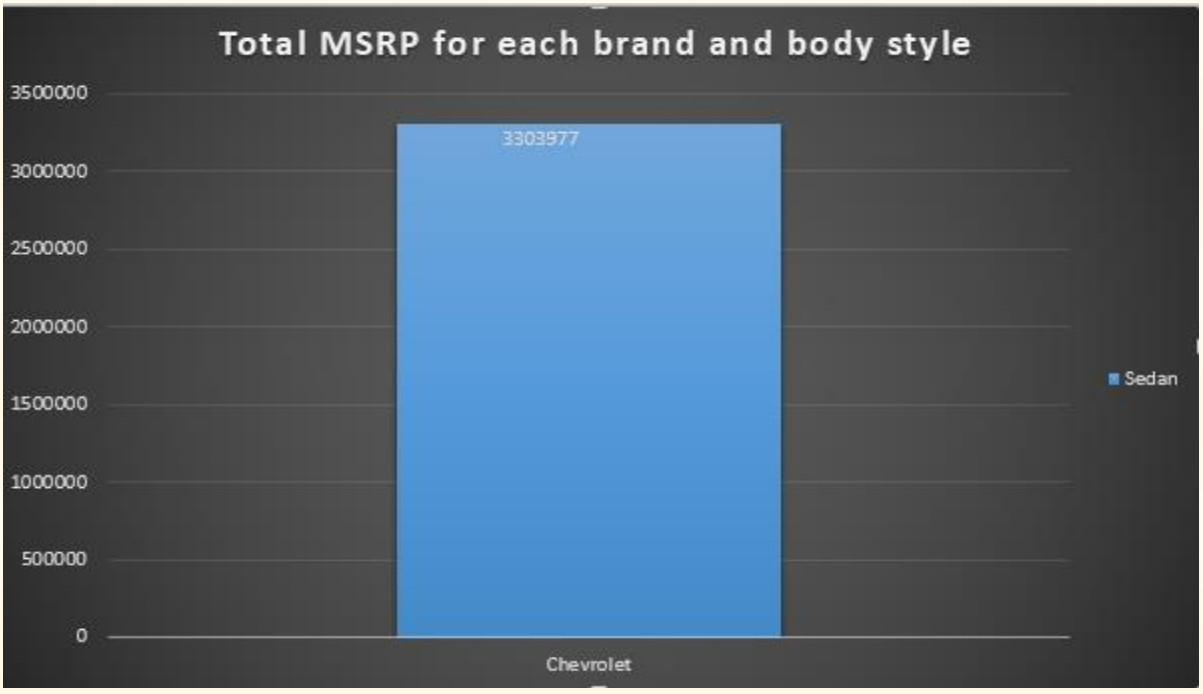
	Engine Cylinders	*highway MPG*
Engine Cylinders	1	
highway MPG	-0.6074225	1

Here we can see that engine cylinders and highway MPG are high negative correlated to each other so I can say that when engine cylinders are increases then MPG decreases.

Findings – VI

6. How does the distribution of car prices vary by brand and body style?

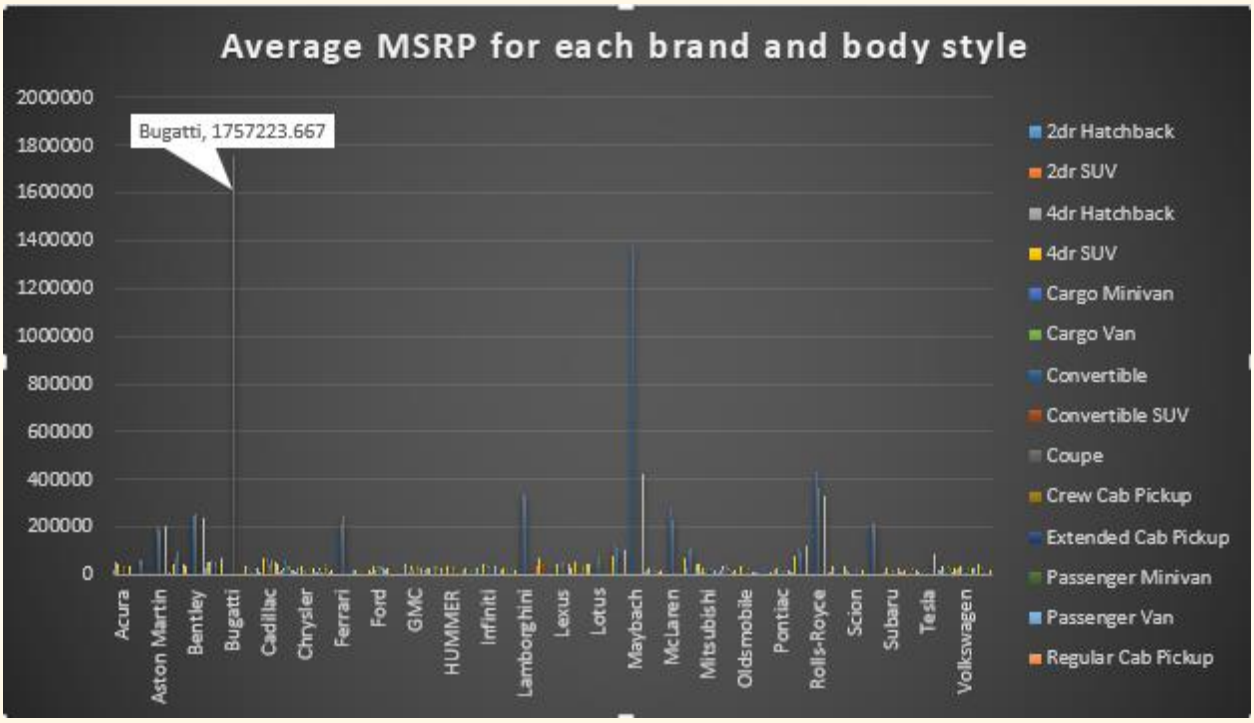




From both the graphs we can see that in Chevrolet Brand with sedan style have the highest MSRP.

Findings – VII

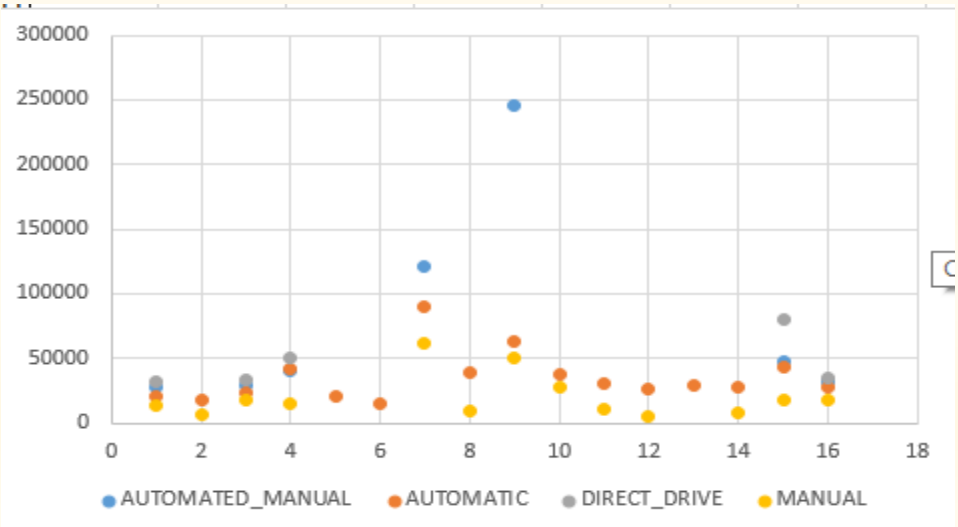
7. Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?



Highest Average MSRP is Bugatti of coupe body style and the lowest Average MSRP is Chrysler of coupe body style.

Findings – VIII

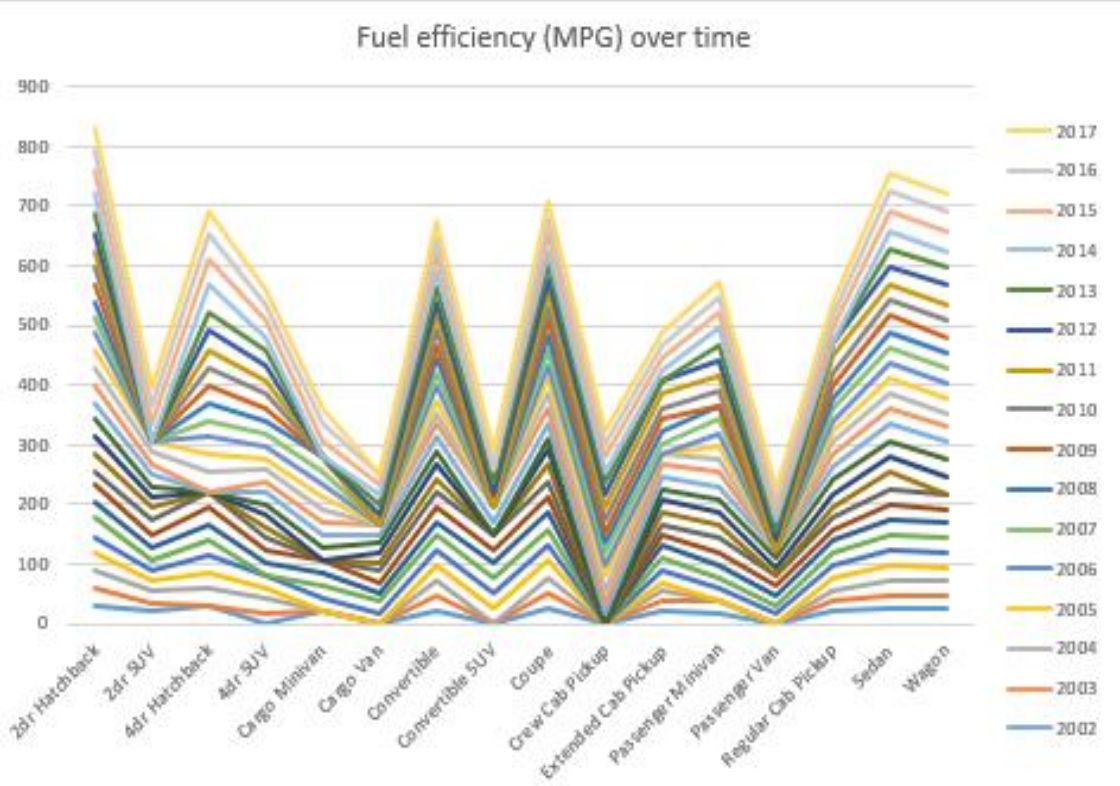
8. How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

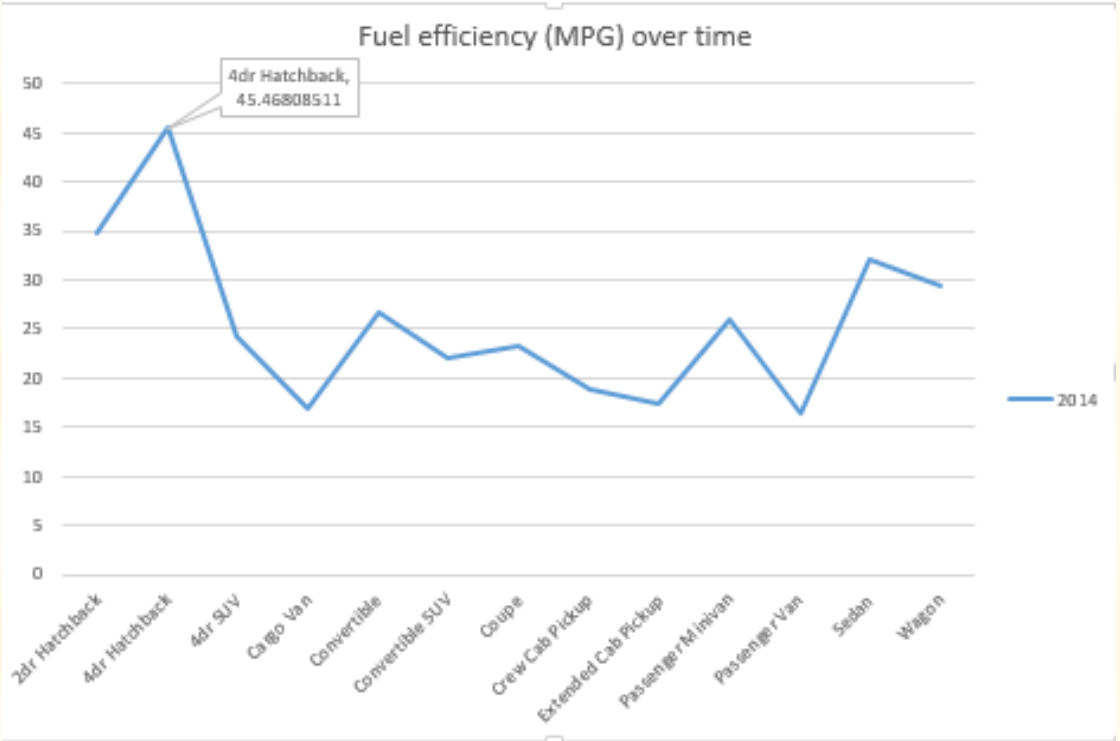


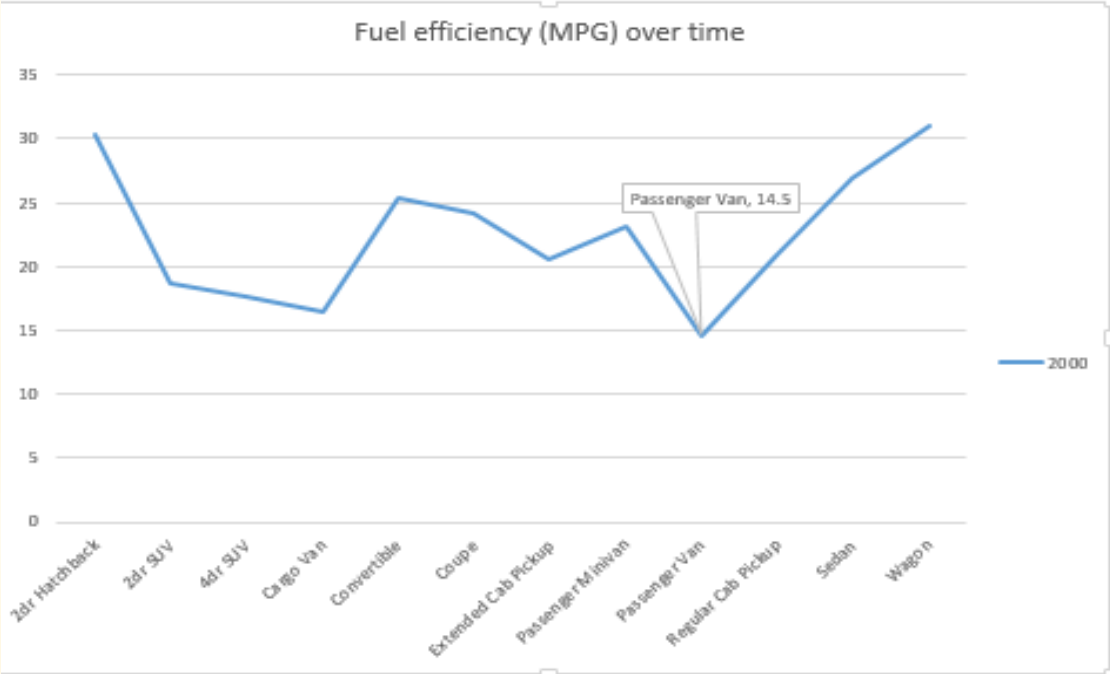
According to this graph we can see that MSRP is higher in Automatic as compare to other transmission type.

Findings – IX

9.How does the fuel efficiency of cars vary across different body styles and model years?







We can quickly comprehend the distribution of fuel economy over time from the three graphs above. We can observe that the passenger van in 2000 had the lowest efficiency, 14.5, while the 4dr hatchback in 2014 had the best efficiency, 45.46.

Findings – X

10. How does the car's horsepower, MPG, and price vary across different Brands?

Row Labels	Average of Engine HP	Average of highway MPG	Average of MSRP
Acura	244.797619	28.11111111	34887.5873
Alfa Romeo	237	34	61600
Aston Martin	484.3225806	18.89247312	197910.3763
Audi	277.7737003	27.82874618	53457.77676
Bentley	533.8513514	18.90540541	247169.3243
BMW	326.9071856	29.24550898	61546.76347
Bugatti	1001	14	1757223.667
Buick	219.244898	26.94897959	28206.61224
Cadillac	332.3098237	25.23677582	56231.31738
Chevrolet	246.985175	25.81567231	28350.38557
Chrysler	229.1390374	26.36898396	26722.96257
Dodge	244.4153355	22.34504792	22390.05911
Ferrari	511.9565217	15.72463768	238218.8406
FIAT	148.6802546	37.33870968	22670.24194
Ford	243.1908003	24.00681044	27399.26674
Genesis	347.3333333	25.33333333	46616.66667
GMC	259.8446602	21.4038835	30493.29903
Honda	195.9883828	32.57461024	26674.34076
HUMMER	261.2352941	17.29411765	36464.41176
Hyundai	201.9174917	30.39273927	24597.0363
Infiniti	310.0666667	24.77878788	42394.21212
Kia	207.748743	30.65367965	25310.17316
Lamborghini	614.0769231	18.01923077	331567.3077
Land Rover	322.0979021	22.12587413	67823.21678
Lexus	277.4158416	25.87623762	47549.06931
Lincoln	283.177655	24.48780488	42839.82927
Lotus	275.9655172	26.55172414	69188.27586
Maserati	420.7931034	20.29310345	114207.7069
Maybach	590.5	16	546221.875
Mazda	171.9929078	27.85106383	20039.38298
McLaren	610.4	22.2	239805

Mercedes-Benz	349.8962944	24.83002833	71476.22946
Mitsubishi	173.7858776	27.54460094	21240.53521
Nissan	240.0912532	27.79928315	28583.4319
Oldsmobile	177.4666667	26.23333333	11542.54
Plymouth	131.5609756	27.96341463	3122.902439
Pontiac	190.2956989	27.06989247	19321.54839
Porsche	392.7941176	25.36764706	101622.3971
Rolls-Royce	487.5483871	19.12903226	351130.6452
Saab	220.5225225	26.35135135	27413.5045
Scion	154.4333333	32.3	19932.5
Spyker	400	18	213323.3333
Subaru	197.3085938	28.68359375	24827.50391
Suzuki	160.2877493	26.03418803	17907.20798
Tesla	249.3919284	98.94444444	85255.55556
Toyota	236.1833564	26.45308311	29030.01609
Volkswagen	189.7577256	32.12855377	28102.38072
Volvo	230.9715302	27.20284698	28541.16014

Here we can see that all three variables are correlated to each other when HP engine is increases then MPG is decreases and MSRP is increases so we can say that fuel efficiency is dependent on HP when it increases fuel consumption increases.

Analysis

Why is the popularity of a car model vary across different market categories?

- The popularity of a car model can vary across different market categories due to factors such as consumer preferences, demographics, geographic location, economic conditions, and marketing strategies. Different types of consumers have varying preferences for factors like size, performance, features, and price, which influences their choice of car model. Geographic factors, such as climate and terrain, can also impact the suitability of certain car models in specific regions. Economic factors, including income levels and affordability, play a significant role in determining popularity. Additionally, marketing efforts and brand image contribute to the appeal of a car model within specific market segments. Understanding these factors helps manufacturers and marketers cater to the preferences of different market categories.

Why the car price is depends on its engine power?

- The price of a car often depends on its engine power due to several reasons. Higher-powered engines involve more advanced technology and higher-quality materials, leading to increased manufacturing costs that are reflected in the car's price. Cars with powerful engines offer better performance and driving experience, which appeals to enthusiasts and individuals seeking thrill and acceleration. Brand positioning also plays a role, as high-performance or luxury brands command a premium based on their reputation and perceived performance associated with powerful engines. The research and development costs involved in designing and

refining high-performance engines contribute to the higher pricing. Additionally, market demand for powerful engines and the need for upgraded components and technologies further justify the higher prices. While engine power is a significant factor, other factors like brand, features, technology, design, and overall quality also influence car pricing.

Why car feature is important for any company and how it decide the pricing of cars?

- Car features are important for any company because they significantly influence consumer purchasing decisions and provide a competitive edge in the market. Features such as safety systems, infotainment technology, advanced driver-assistance systems (ADAS), comfort features, and fuel efficiency contribute to the overall appeal and value of a car.

Why does fuel efficiency increase or decrease when we increase the number of cylinders?

- The impact of increasing the number of cylinders on fuel efficiency can vary depending on various factors. Modern engine technologies can optimize fuel delivery and combustion efficiency, leading to improved fuel efficiency despite having more cylinders. However, increasing the number of cylinders can also result in a larger engine size and displacement, potentially leading to increased fuel consumption. The weight of the engine and the overall vehicle, as well as driving conditions and driving style, also play a role in determining fuel efficiency. Therefore, the relationship between the number of cylinders and fuel efficiency is complex and depends on a combination of factors.

Conclusions

1. This graph demonstrates how the number of market categories is related to the total level of popularity.
2. In this graph, we can see that the MSRP and engine power of automobiles are linearly related. We can also observe that some MSRP values behave strangely, but we know that some cars are pricey, like the Bugatti.
3. Here, we can see that the engine cylinders coefficient is more than the year, city mpg, and high mpg for vehicle characteristics; thus, we can draw the conclusion that this is an essential quality of any manufacturer as it is inversely connected to the engine horsepower.
4. Here, it is clear that Bugatti has the highest average price, with Maybach coming in second.
5. The fuel economy and the number of cylinders are being compared in this scatter plot, and since we can see that efficiency falls as the number of cylinders rises, we may infer that the connection between the two is inverse.

6. As you can see, engine cylinders and highway MPG have a strong negative correlation. Therefore, as engine cylinders grow, MPG falls.
7. The two graphs show that the Chevrolet Brand's MSRP for sedans is the highest.
8. The Bugatti coupe body type has the highest average MSRP, and the Chrysler coupe body style has the lowest average MSRP.
9. This graph shows that the MSRP for an automatic gearbox is more than for other gearbox types.
10. The three graphs above allow us to rapidly understand the distribution of fuel economy over time. We can see that in 2000, the passenger van had the lowest efficiency (14.5), while the greatest efficiency (4dr hatchback) in 2014 was 45.46.
11. Here, we can see that all three variables are associated with one another; as engine horsepower grows, MPG declines and MSRP rises, indicating that as engine horsepower rises, so doe’s fuel consumption.

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 opportunities 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.

The Problem

- Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).
- 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, …..)
- 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.)
- 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] and distribution of those 30 calls are as follows:

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

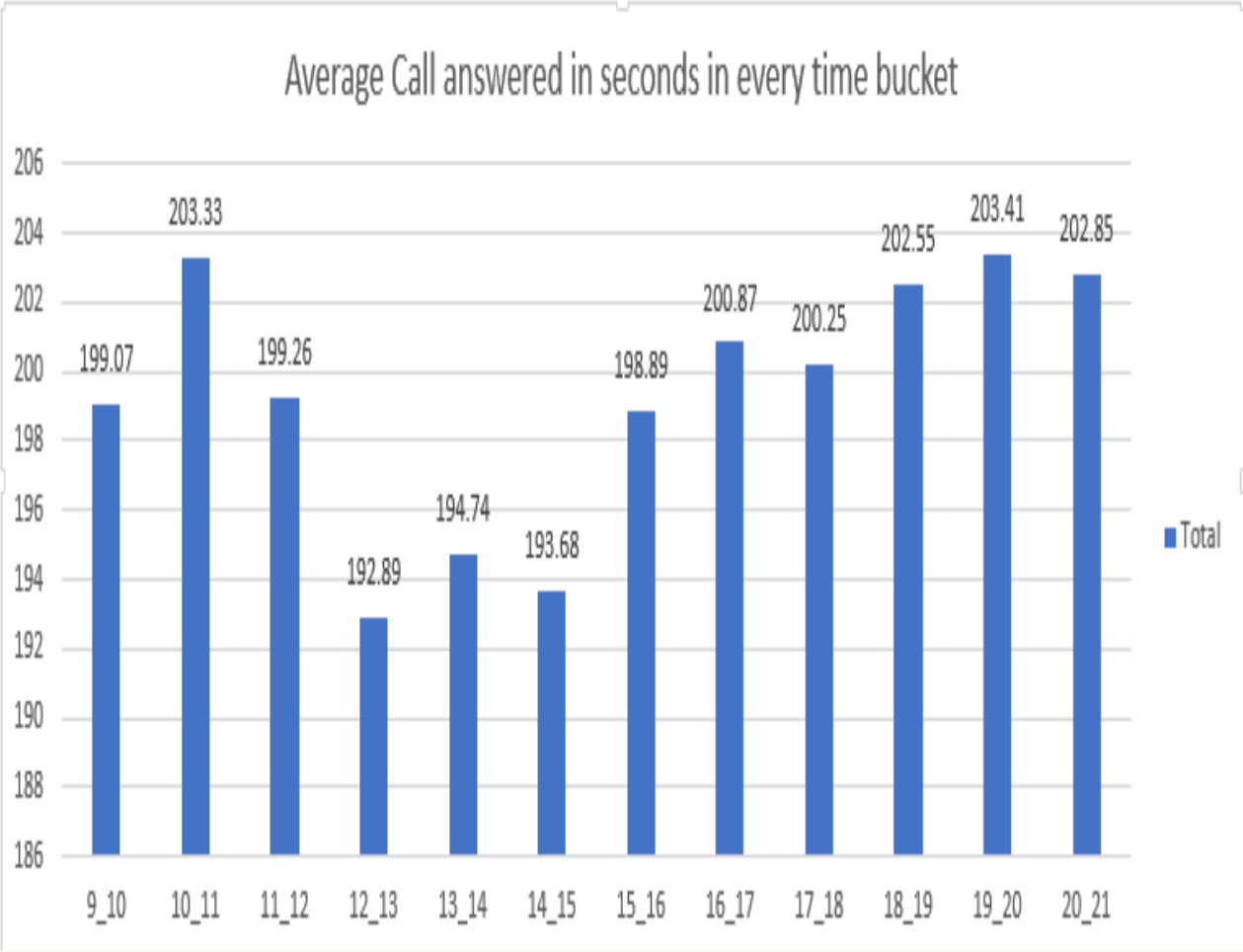
Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.

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 – I

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

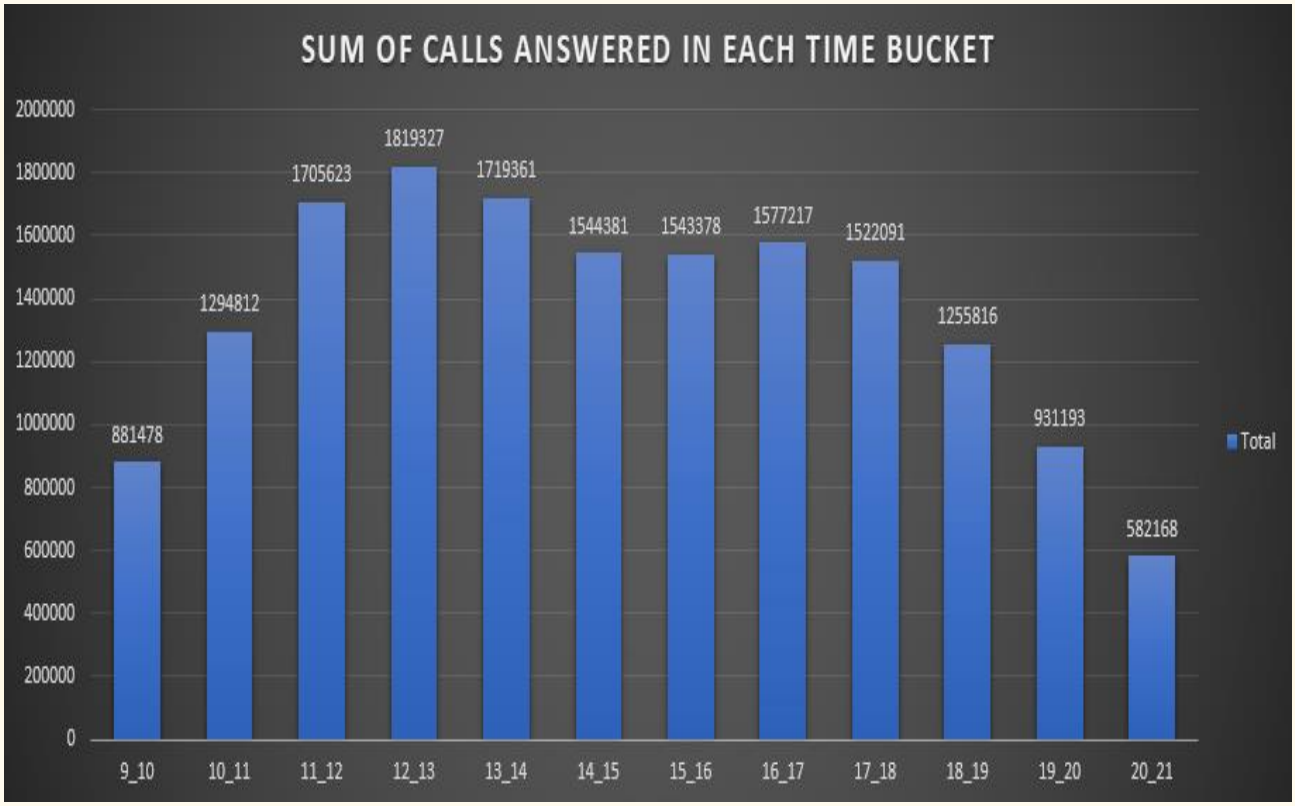


Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
9_10	199.0691057
10_11	203.3310302
11_12	199.2550234
12_13	192.8887829
13_14	194.7401744
14_15	193.6770755
15_16	198.8889175
16_17	200.8681864
17_18	200.2487831
18_19	202.5509677
19_20	203.4060725
20_21	202.845993
Grand Total	**198.6227745**

The time_bucket 19_20, or 7 PM to 8 PM, had the greatest average number of calls answered in seconds, at 203.4, according to the aforementioned column plot.

Findings – II

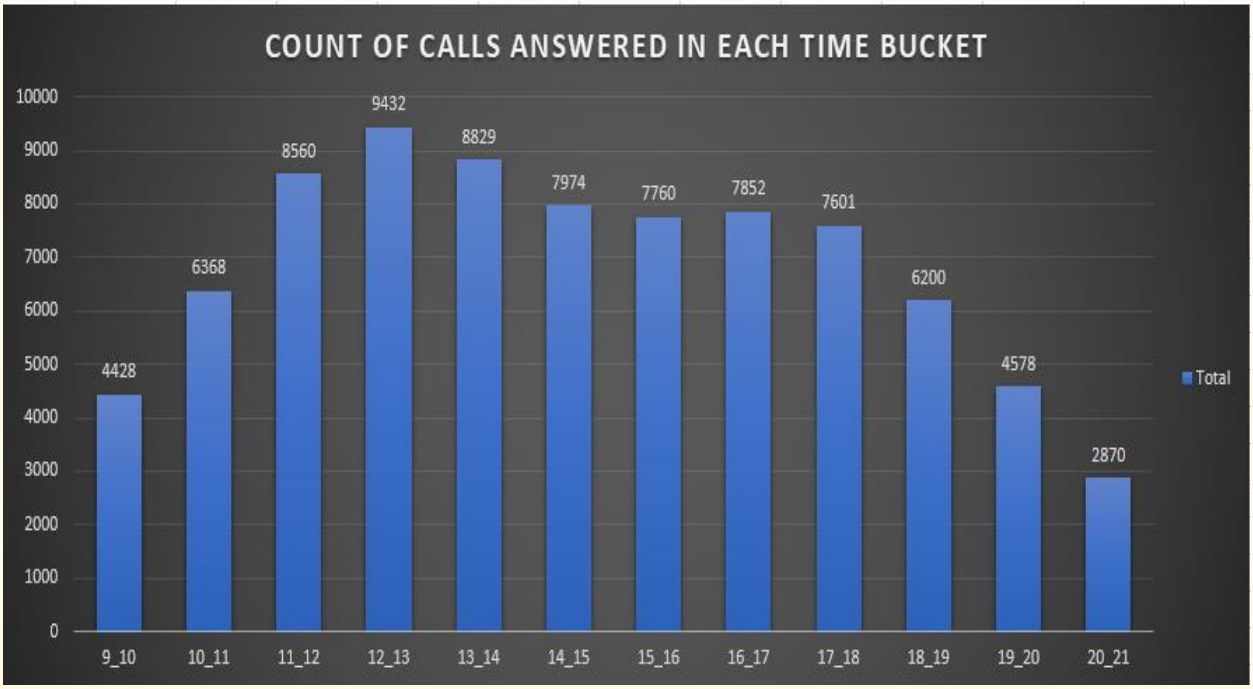
Call_Status	answered
Row Labels	**Sum of Call_Seconds (s)**
9_10	881478
10_11	1294812
11_12	1705623
12_13	1819327
13_14	1719361
14_15	1544381
15_16	1543378
16_17	1577217
17_18	1522091
18_19	1255816
19_20	931193
20_21	582168
Grand Total	**16376845**



The time_bucket 12_13, or from 12 to 1 PM, had the largest overall number of calls answered, with 1819327, according to the above column plot.

Findings – III

Call_Status	answered
Row Labels	**Count of Call_Seconds (s)**
9_10	4428
10_11	6368
11_12	8560
12_13	9432
13_14	8829
14_15	7974
15_16	7760
16_17	7852
17_18	7601
18_19	6200
19_20	4578
20_21	2870
Grand Total	**82452**

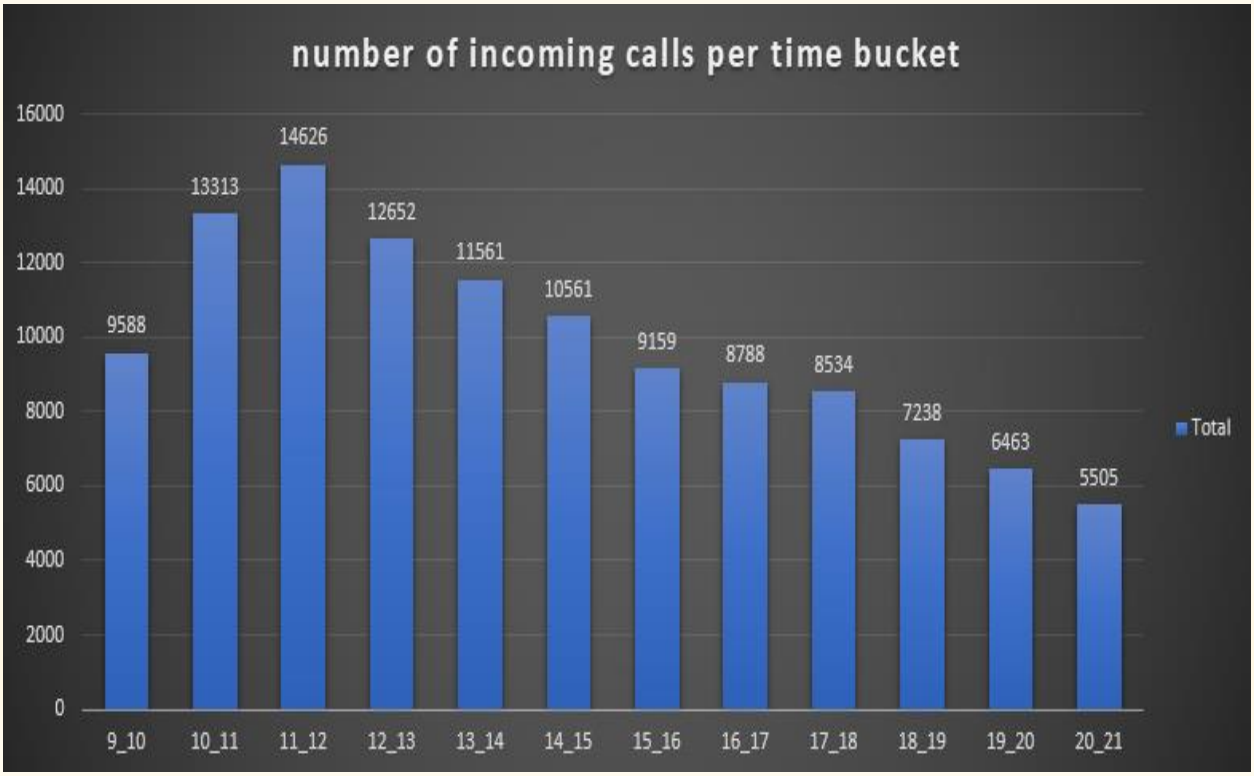


The time_bucket 12-13, or 12 PM to 1 PM, had the greatest total of calls answered, with 9432, according to the above column plot.

Findings – IV

2. 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, …..).

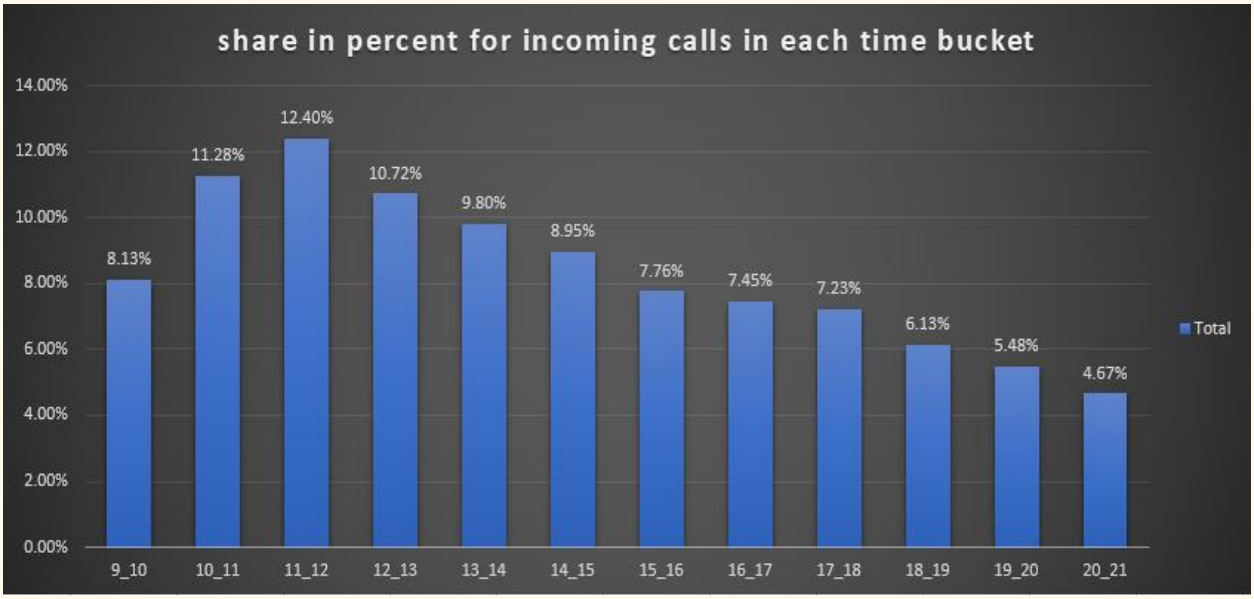
Row Labels	Count of Customer_Phone_No
9_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
Grand Total	**117988**



The time bucket 11_12, or 11 AM to 12 PM, has the greatest count for the total number of incoming calls, with a count of 14626, according to the above column plot.

Findings – V

Row Labels	Count of Time
9_10	8.13%
10_11	11.28%
11_12	12.40%
12_13	10.72%
13_14	9.80%
14_15	8.95%
15_16	7.76%
16_17	7.45%
17_18	7.23%
18_19	6.13%
19_20	5.48%
20_21	4.67%
Grand Total	100.00%



The time bucket 11_12, or 11 AM to 12 PM, has the biggest percentage of incoming calls (12.40%), according to the aforementioned column plot.

Findings – VI

3. 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.)

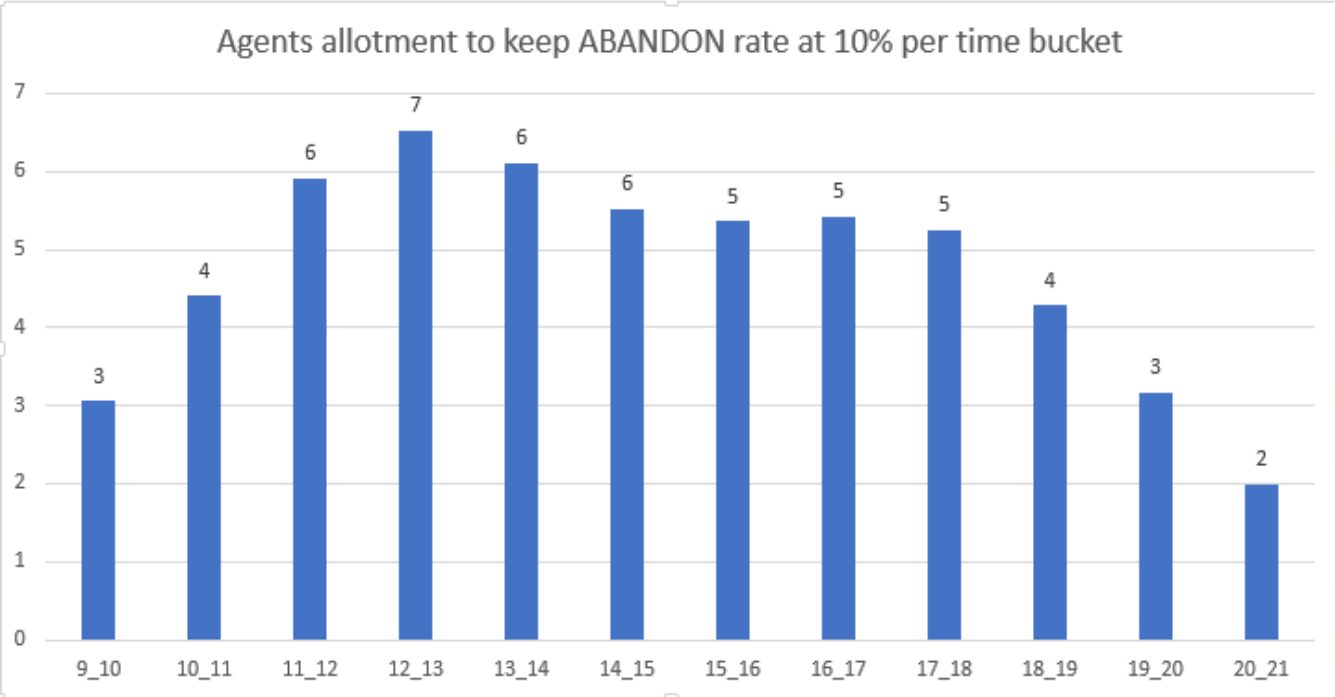
The data indicates that the current abandon rate is about 30%.A manpower plan, or the new average number of employees working each day, has to be proposed.

Count of Call_Status	Column Labels			
Row Labels	abandon	answered	transfer	Grand Total
Jan				
1-Jan	684	3883	77	4644
2-Jan	356	2935	60	3351
3-Jan	599	4079	111	4789
4-Jan	595	4404	114	5113
5-Jan	536	4140	114	4790
6-Jan	991	3875	85	4951
7-Jan	1319	3587	42	4948
8-Jan	1103	3519	50	4672
9-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 in daily basis	1495.782609	3584.87	49.261	5129.913
% Avg calls in daly basis	29.16%	69.88%	0.96%	

- We may infer from the preceding study that the average number of calls answered by each agent in each time bucket is 198.6.
- The abandon rate has to be decreased by 30% (actual) - 10% (desired) = 20%. Specifically, we must raise the call responded rate by 70% (current) + 20% (change), which equals 90%.
- In order to lower the abandon rate to 10%, we must answer 90% of all incoming calls.
- 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
- Time needed to answer 90% of incoming calls is calculated as follows: 5130 * 198.6 * 0.9 / 3600 = 254.7001826
- Therefore, the new average number of agents working each day is 255 divided by the actual number of hours each agent works (on a customer contact), or 4.5, which is 56.67. This results in 57 agents working each day .
- In order to achieve a 10% abandon rate, 57 agents must be working each day.

Findings – VII

Call_Status	answered	
Row Labels	**Count of Customer_Phone_No**	**Agents allotment**
9_10	4428	3
10_11	6368	4
11_12	8560	6
12_13	9432	7
13_14	8829	6
14_15	7974	6
15_16	7760	5
16_17	7852	5
17_18	7601	5
18_19	6200	4
19_20	4578	3
20_21	2870	2
Grand Total	**82452**	**57**



The distribution of the manpower plan by time bucket to maintain a 10% abandon rate, or a 90% call response rate.

The following observations were made based on the assumptions made:-

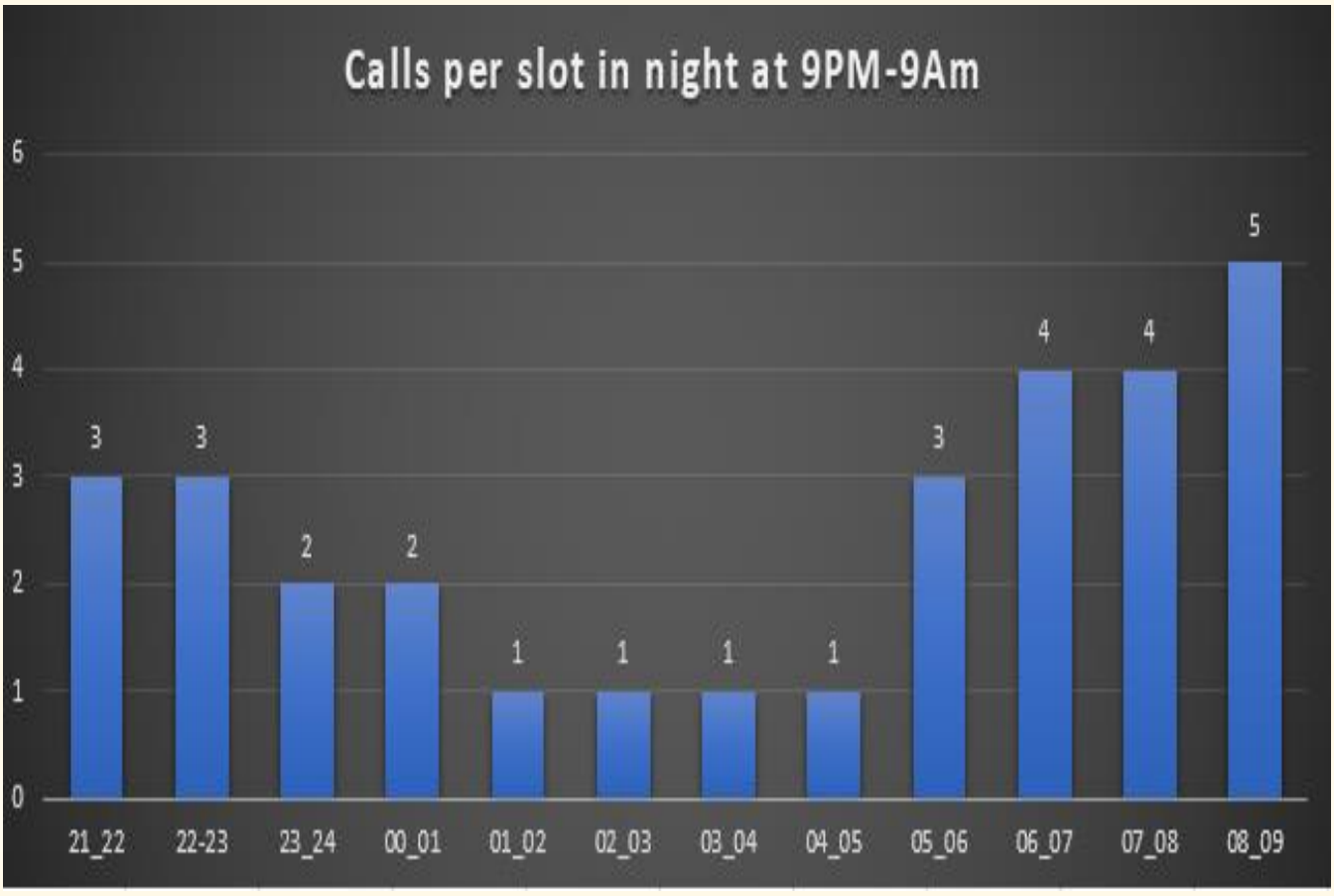
- An agent works for 9 hours in a day.
- Out of that total, 1.5 hours are taken for lunch and coffee/tea breaks. That leaves 9 minus 1.5 hours, or 7.5 hours. Of those 7.5 hours, an agent is on consumer calls for only 60% of the time, or 0.6 * 7.5 hours. This means that an agent spends only 4.5 hours per day out of a total of 7.5 hours on consumer calls.
- 6 days a week are worked by an agency.
- In a month of 30 days, there are 6 days per week; there are 4 weeks in a month of 30 days; 7 days per week means a total of 28 days, of which 4 days are unplanned leave; the number of days an agent spends on the floor is 20 * 7/28, or 5 days; the number of days left is 28 - 4 = 24; there are 4 Sundays in a month of 30; the number of days an agent can work is 24 - 4, or 20; as

Finding-VIII

4. 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] and distribution of those 30 calls are as follows:

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

- Keeping the abandon rate at 10%, or keeping the answered rate at 90%, we must now distribute the total manpower available for each time bucket, starting from 9 AM to 9 PM and then from 9 PM to 9 AM.
- There are 30 night calls for every 100 day calls, therefore there will be 1539 night calls for every 5130 day calls (5130*30/100).In all, there are 5130 day calls and 1539 night calls.
- Consequently, 1539 more working hours will be required to maintain a 90% response rate. * 198.6 (average number of answered calls per second) * 0.9 / 3600 (number of seconds in an hour) = 76.41135.
- Therefore, the organisation needs more employees to handle nighttime calls (76.41135/4.5 = 16.98 == 17).
- We thus require an extra 17 agents to answer night calls, bringing the total number of agents working each day to 57 (day call answer 90%) + 17 (night call answer 90%), or 74 agents.
- Therefore, we require 74 agents per day to handle customer calls from day and night while maintaining an answered rate of 90% and an abandon rate of 10%.



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 agents that work in the 19_20, 20_21 time buckets must wait and work in the 21_22, and 22_23 time buckets as well because we only have 17 agents available at night.
- Additionally, agents that work during the time buckets 9_10 and 10_11 may be requested to work during the time buckets 7_8 and 8_9.
- To keep the abandon rate at 10%, the agents who operate in time buckets 1–2, 2–3, 3–4, and 4–5 might be instructed to work in time buckets 6–7, 7–8, and 8–9.

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 junor 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:-

1. According to the aforementioned column plot, the time_bucket 19_20, or 7 PM to 8 PM, had the highest average number of calls answered in seconds, at 203.4.
2. According to the aforementioned column diagram, the time_bucket 12_13, or from 12 to 1 PM, had the highest total number of calls answered (1819327).
3. According to the above column plot, the time_bucket 12-13, or 12 PM to 1 PM, had the most calls answered overall with 9432.

4. According to the aforementioned column plot, the time bucket 11_12, or 11 AM to 12 PM, has the highest count for the overall number of incoming calls, with a count of 14626.
5. In order to achieve a 10% abandon rate, 57 agents must be working each day.
6. According to the aforementioned column plot, the time bucket 11_12, or 11 AM to 12 PM, has the highest percentage of incoming calls (12.40%).
7. We only have 17 agents accessible at night, thus the agents that work in the 19_20 and 20_21 time buckets must wait and work in the 21_22 and 22_23 time buckets as well.
8. Additionally, it is possible to require that agents who work in the time buckets 9_10 and 10_11 also work in the time buckets 7_8 and 8_9.
9. The agents who work in time buckets 1-2, 2-3, 3-4, and 4-5 may be told to work in time buckets 6-7, 7-8, and 8-9 in order to maintain the abandon rate at 10%.

Appendix

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

Data Analytics Trainee Assignement

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

Data Analytics Trainee Task 2

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

Data Analytics Trainee Task - 3

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

Data Analytics Trainee Task - 4

- **IMDB Movie Analysis-**
  - Link for the shared PDF on Google Drive:

Data Analytics Trainee Task - 5

Google drive Folder link with all Analysis file

- **Bank Loan Case Study:-**
  - Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task 6

Google Drive folder link with all files

- **Analyzing the Impact of Car Features on Price and Profitability**
  - Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task - 7

Google drive folder link with all files

- **ABC Call Volume Trend Analysis:-**
  - Link for the shared file on Google Drive:

Trainity Data Analytics Trainee Task 8

Google drive folder link with all files

- Link to LinkedIn Profile:-

My Linkedin Profile

- Link to GitHub Portfolio:-

Nagendra Pratap Singh. Github