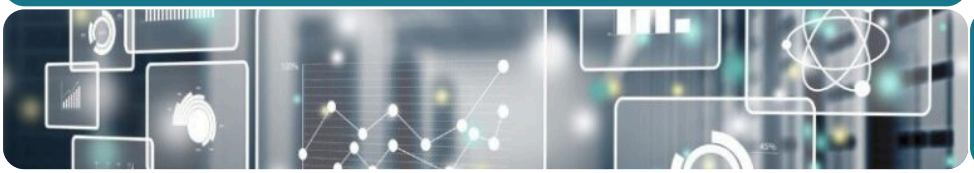
Data Analysis Portfolio



Sakhi Patel

Email: sakhipatel20@gmail.com

Professional Background

My name is Sakhi Patel, currently in my final year of pursuing a B.Tech in Information and Communication Technology, with a CGPA of 9.2 (till 6th semester). I have developed a strong foundation in various domains including Data Analysis, Python, AI, Data Structures and Algorithms (DSA), Database Management Systems (DBMS), Machine Learning, MySQL, and Excel.

Throughout my academic journey, I have actively engaged in several personal projects focusing on Data Analysis, Machine Learning, AI, and Python, which have honed my problem-solving skills and technical expertise. These projects have allowed me to apply theoretical knowledge to practical scenarios, enhancing my understanding of real-world applications.

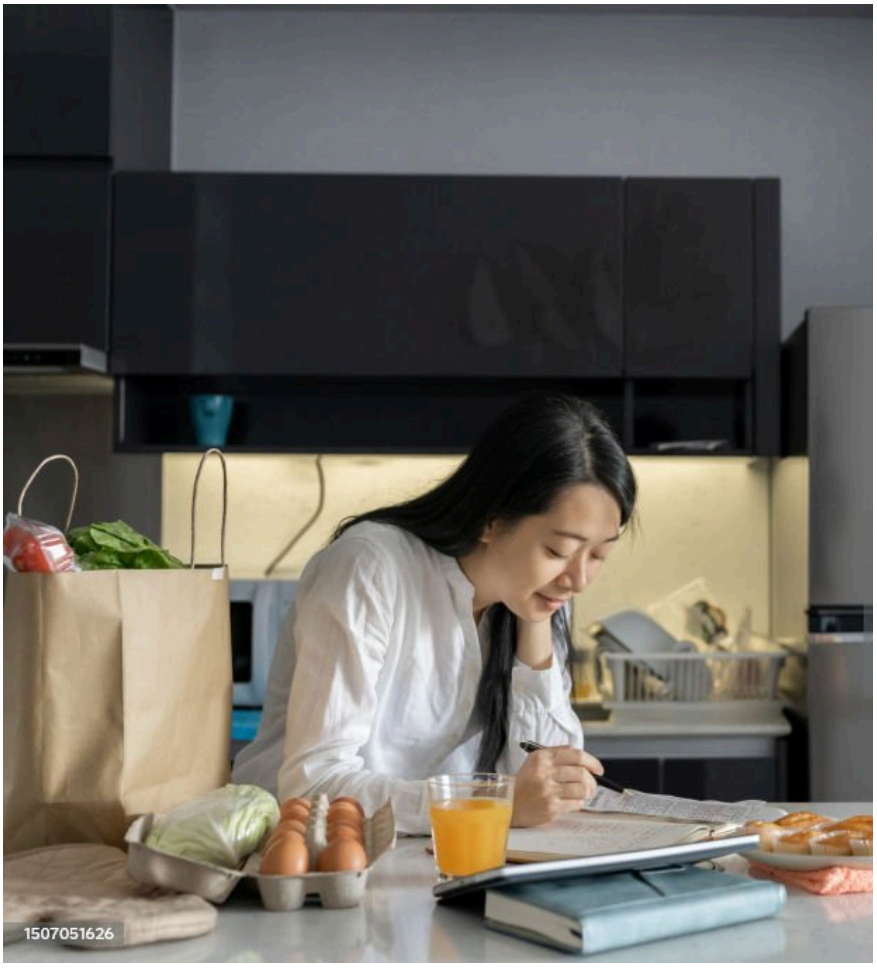
As a fresher, I am eager to embrace the challenges of the corporate world and gain firsthand experience in how the industry operates. I am confident in my flexibility and adaptability, which I believe are crucial for continuous learning and growth. My theoretical knowledge provides a solid base, and I am enthusiastic about leveraging it in practical settings. I am committed to putting in significant effort to learn and excel in a professional environment.

Table of Contents

Data Analytics Process

Description

To demonstrate how the principles of Data Analytics are applied in everyday life using the example of Meal Planning and Cooking. The project will illustrate each step of the Data Analytics Process—Plan, Prepare, Process, Analyze, Share, and Act—showing its practical application and importance.



Design

Scenario: Grocery shopping is a common activity where we unknowingly apply data analytics. This project will break down the process into analytical steps, providing a structured approach to decision-making.

1. **Plan**
   - Menu Creation: I decide on meals for the week based on preferences, nutritional needs, and available ingredients.

    - Inventory Check: I assess my pantry and perishables to see what needs restocking or using up.
2. **Prepare**
    - Grocery Shopping: I make a shopping list considering portion sizes, spoilage rates, and my budget.
    - Meal Prepping: I prepare ingredients ahead of time, like washing veggies or marinating proteins.
3. **Process**
    - Cooking Methods: I choose how to cook each dish based on time, flavor, and nutrition.
    - Recipe Execution: I follow recipes or improvise, adjusting seasonings and ingredients as needed.
4. **Analyze**
    - Nutritional Analysis: I check the nutritional content of my meals to ensure they're healthy.
    - Taste Testing: I taste the food as I cook to ensure it’s delicious.
5. **Share**
    - Family Preferences: I consider preferences and dietary restrictions when planning meals.
    - Recipe Sharing: I share successful recipes and cooking tips with others.
6. **Act**
    - Serve Meals: I present my cooked dishes nicely and serve them at the right time.
    - Gather Feedback: I ask for feedback to see how I did and what I can improve next time.

Conclusion

Through the lens of meal planning and cooking, it’s evident that Data Analytics plays a significant role in our daily lives, often without us realizing it. By following a structured approach—planning, preparing, processing, analyzing, sharing, and acting—we can make informed decisions that enhance our efficiency, health, and overall satisfaction. This project demonstrates that the principles of Data Analytics are not confined to professional environments but are integral to everyday tasks like meal preparation. By leveraging these principles, we can optimize our routines, reduce waste, and improve the quality of our meals. Ultimately, understanding and applying Data Analytics in everyday scenarios empowers us to make better, data-driven decisions in all aspects of life.

Instagram User Analytics

Description

This project uses SQL analysis to extract insights from Instagram user data. Key objectives include identifying loyal users, re-engaging inactive users, determining contest winners, researching popular hashtags, optimizing ad campaign timing, evaluating user engagement, and detecting potential bot accounts. These insights will help the product team improve Instagram's performance and user experience.

The Problem

A) Marketing:

1. Rewarding Most Loyal Users:
   - **Task:** Find the 5 oldest Instagram users from the database.
2. Remind Inactive Users:
   - **Task:** Identify users who have never posted a photo.
3. Declaring Contest Winner:
   - **Task:** Find the user with the most likes on a single photo and provide their details.
4. Hashtag Researching:
   - **Task:** Identify the top 5 most commonly used hashtags on the platform.
5. Launch AD Campaign:
   - **Task:** Determine which day of the week has the highest user registrations to schedule ad campaigns effectively.

B) Investor Metrics:

1. User Engagement:
   - **Task:** Calculate the average number of posts per user and the total number of photos divided by the total number of users.
2. Bots & Fake Accounts:
   - **Task:** Identify users who have liked every single photo on the site, as they may be bots or fake accounts.

Design

Steps to Load Data into the Database:

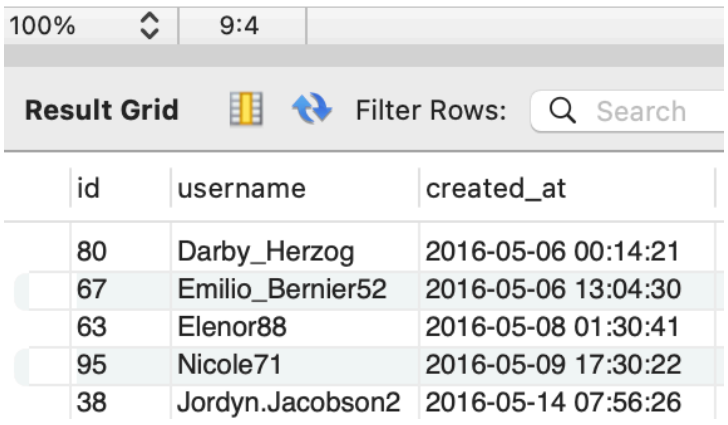
1. Create the Database:
   - Use MySQL to create a new database.
2. Add Tables and Column Names:
   - Define and create tables with relevant columns for storing data.
3. Insert Values:
   - Populate the tables with initial data values.
4. Query the Data:
   - Use querying tools to retrieve and analyze data from the database.
5. Software Used:
   - **MySQL Workbench 8.0 CE:** A graphical tool used for managing and querying MySQL databases.

Findings

I. Most Loyal Users:

- **Objective:** Identify the top 5 oldest Instagram users.
- **Approach:** Select the username and created_at columns from the users table. Order the results by created_at in ascending order and use the LIMIT function to display the top 5 oldest users.
- **Output/Result:** The usernames and registration dates of the top 5 oldest users on Instagram.

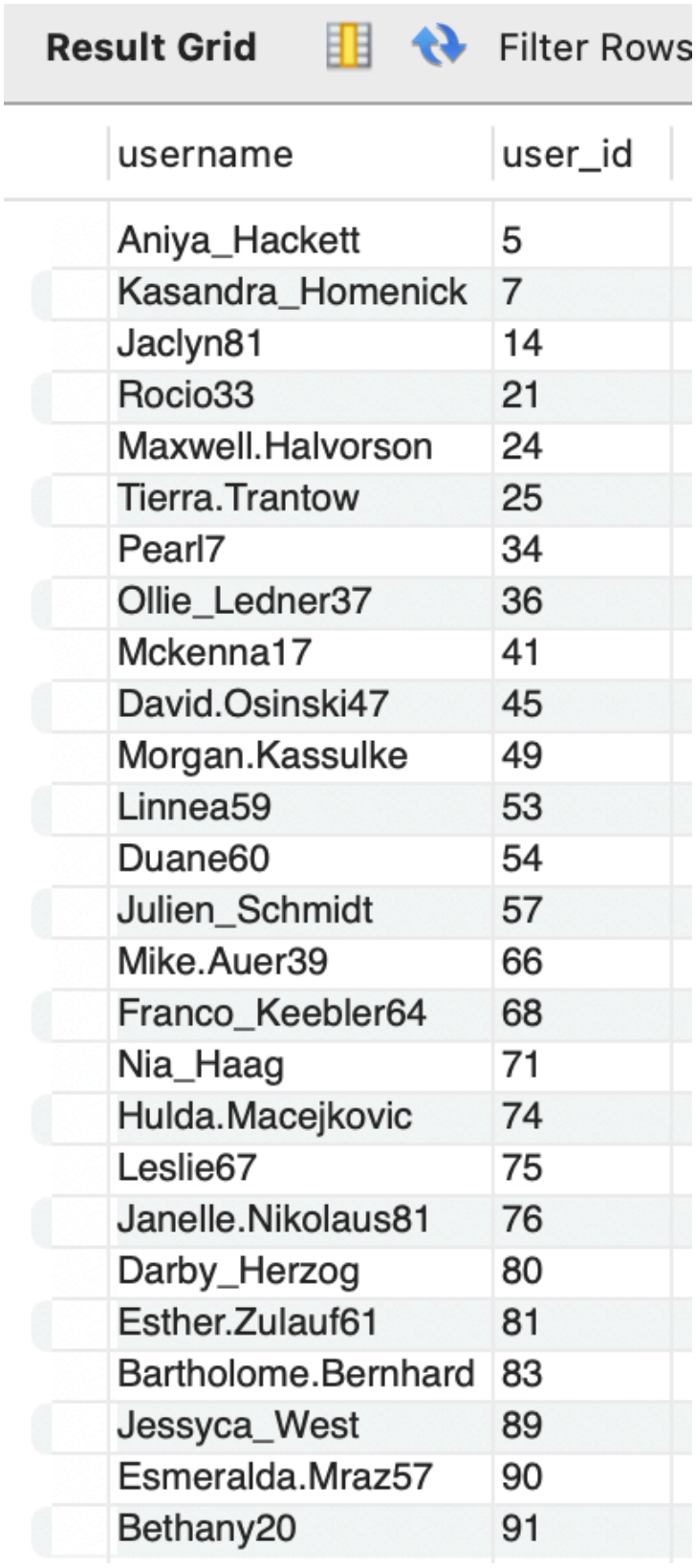
```
1 •   select *
2     from users
3     order by created_at ASC
4     limit 5;
```



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

II. Most Inactive Users:

- **Objective:** Find users who have never posted a photo.
- **Approach:** Select the username column from the users table. Perform a left join with the photos table on users.id and photos.user_id. Find users where photos.id is NULL.
- **Output/Result:** The list of users who have never posted any photos.

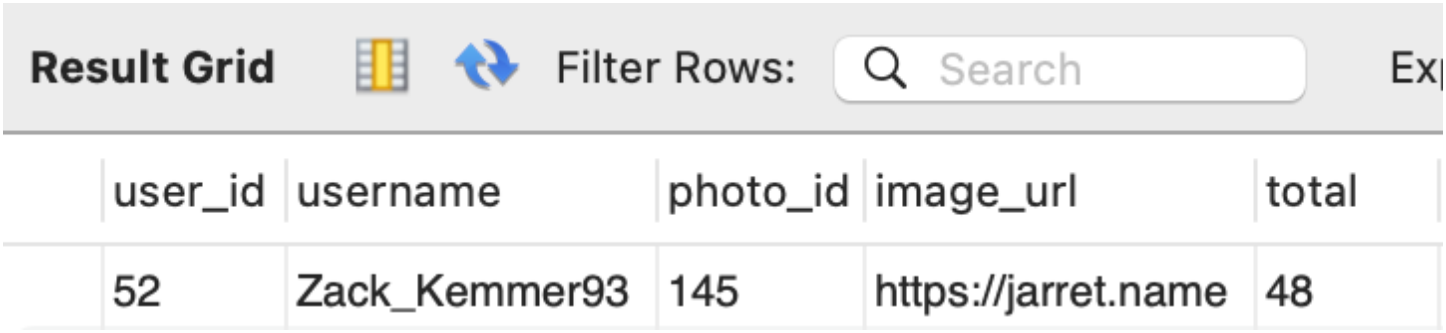


username	user_id
Aniya_Hackett	5
Kasandra_Homenick	7
Jaclyn81	14
Rocio33	21
Maxwell.Halvorson	24
Tierra.Trantow	25
Pearl7	34
Ollie_Ledner37	36
Mckenna17	41
David.Osinski47	45
Morgan.Kassulke	49
Linnea59	53
Duane60	54
Julien_Schmidt	57
Mike.Auer39	66
Franco_Keebler64	68
Nia_Haag	71
Hulda.Macejkovic	74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

III. Most Liked Photo:

- **Objective:** Identify the photo with the most likes and provide details.

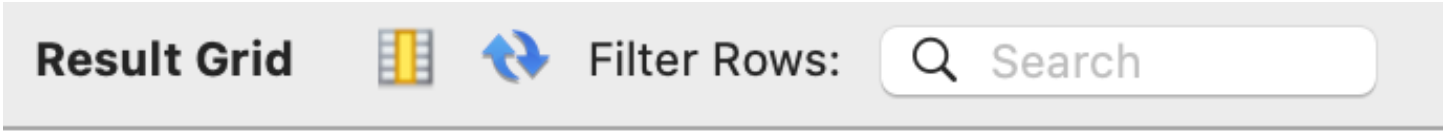
- **Approach:** Select users.username, photos.id, photos.image_url, and the count of likes. Inner join photos, likes, and users tables. Group by photos.id and sort by the count of likes in descending order. Use LIMIT to get the top liked photo.
- **Output/Result:** The username, photo ID, image URL, and total number of likes for the most liked photo.



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

IV. Top 5 Most Commonly Used Hashtags:

- **Objective:** Find the top 5 most commonly used hashtags on Instagram.
- **Approach:** Select tag_name from the tags table and count occurrences. Join tags with photo_tags on tag_id. Group by tag_name and sort by the count in descending order. Use LIMIT 5 to get the top 5 hashtags.
- **Output/Result:** The top 5 most commonly used hashtags on Instagram.

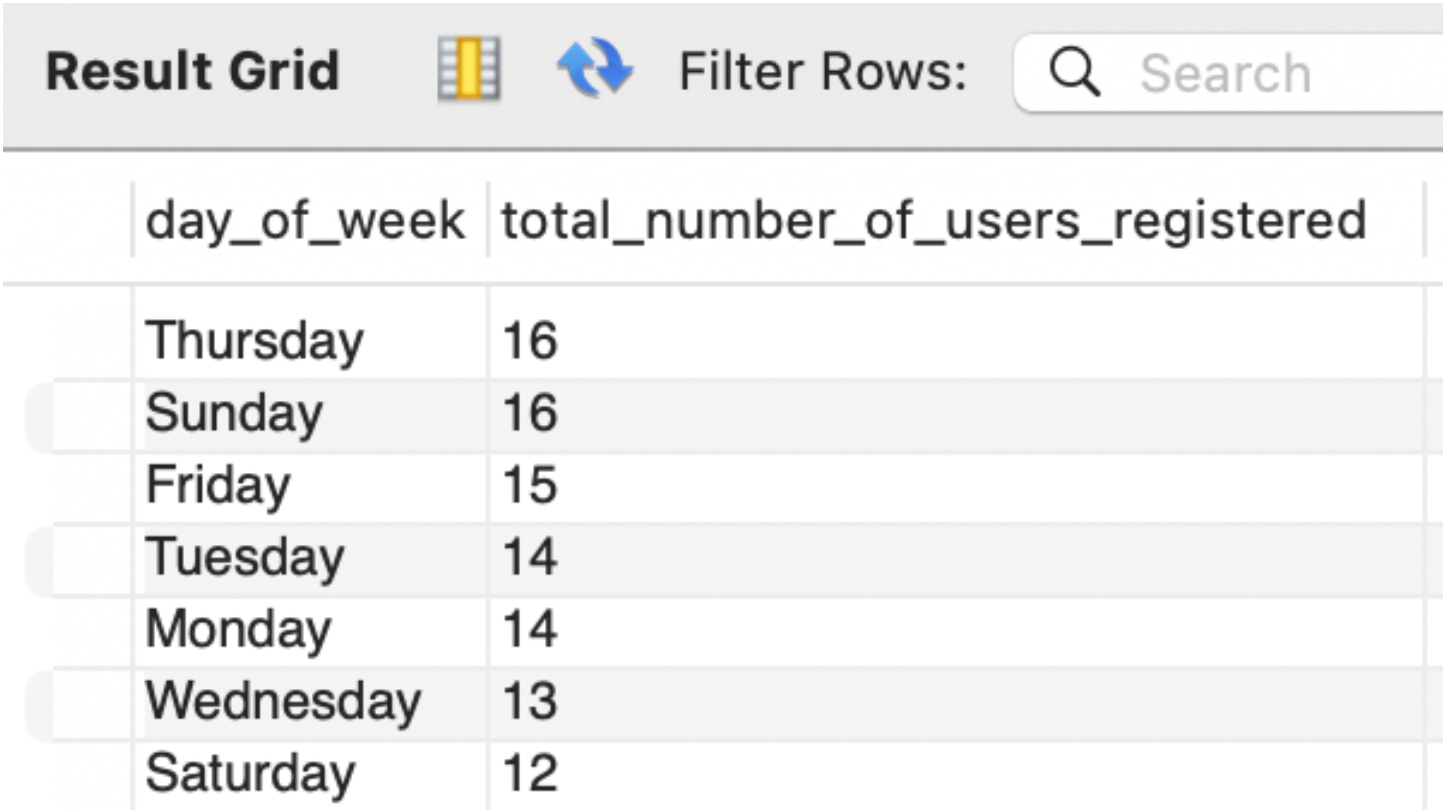


tag_name	total_number_of_times_tag_used_individually
smile	59
beach	42
party	39
fun	38
concert	24

V. Optimal Ad Campaign Day:

- **Objective:** Determine the best day to launch ad campaigns based on user registration data.
- **Approach:** Select the day of the week from the created_at column and count the number of users registered on each day. Group by the day of the week and sort by the total number of registrations in descending order.

- **Output/Result:** The day of the week with the highest user registrations, ideal for ad campaigns.


Result Grid Filter Rows: Search

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

VI. Average User Posts:

- **Objective:** Determine the average number of posts per user.
- **Approach:** Count the total number of photos in the photos table and the total number of users in the users table. Calculate the average by dividing the total number of photos by the number of users. Also, determine the frequency of posts by each user.
- **Output/Result:** The average number of posts per user and the total number of posts.

Result Grid   Filter R

user_id	user_post_count
1	5
2	4
3	4
4	3
6	5
8	4
9	4
10	3
11	5
12	4
13	5
15	4
16	4
17	3
18	1
19	2
20	1
22	1
23	12
26	5
27	1
28	4
29	8
30	2
31	1
32	4
33	5
35	2
37	1
38	2
39	1
40	1
42	3
43	5
44	4
46	4
47	5
48	1
50	3

51	5
52	5
55	1
56	1
58	8
59	10
60	2
61	1
62	2
63	4
64	5
65	5
67	3
69	1
70	1
72	5
73	1
77	6
78	5
79	1
82	2
84	2
85	2
86	9
87	4
88	11
92	3
93	2
94	1
95	2
96	3
97	2
98	1
99	3
100	2

VII. Bots and Fake Accounts:

- **Objective:** Identify potential bots or fake accounts.
- **Approach:** Select user_id from the photos table and username from the users table. Count the total number of likes from the likes table. Inner join users and likes on users.id and likes.user_id. Group by likes.user_id and identify users who have liked every photo on the site.
- **Output/Result:** Users (potential bots) who have liked every photo on Instagram.

Result Grid   Filter Rows:  Search

user_id	username	total_likes_per_user
5	Aniya_Hackett	257
14	Jaclyn81	257
21	Rocio33	257
24	Maxwell.Halvorson	257
36	Ollie_Ledner37	257
41	Mckenna17	257
54	Duane60	257
57	Julien_Schmidt	257
66	Mike.Auer39	257
71	Nia_Haag	257
75	Leslie67	257
76	Janelle.Nikolaus81	257
91	Bethany20	257

These findings leverage SQL queries to analyze Instagram data, providing insights into user behavior, engagement, and potential issues with account authenticity.

Analysis

Analysis Summary:

1. **Most Loyal Users:**
   - The top 5 oldest users have been identified based on their registration dates.
2. **Inactive Users:**
   - Out of 100 total users, 26 have never posted any content on Instagram (photos, videos, or text). The Marketing team should remind these inactive users.
3. **Contest Winner:**
   - Zack_Kemmer93 (user_id 52) won the contest with the most liked photo (photo_id 145) having 48 likes.
4. **Top 5 Hashtags:**

    - The most commonly used hashtags are:
        - #smile (59 uses)
        - #beach (42 uses)
        - #party (39 uses)
        - #fun (38 uses)
        - #concert (24 uses)
5. **Optimal AD Campaign Days:**
    - Most users registered on Thursdays and Sundays (16 users each). It is advisable to launch AD campaigns on these days.
6. **Average Posts Per User:**
    - There are 257 photos and 100 users, resulting in an average of 2.57 posts per user.
7. **Bots and Fake Accounts:**
    - 13 users have liked every photo on Instagram, suggesting these accounts might be bots or fake.

Root Cause Analysis Using the 5 Whys:

1. **Why did the Marketing team want to know the most inactive users?**
    - To reach out and understand why they are not using Instagram, and to encourage their activity.
2. **Why did the Marketing team want to know the top 5 hashtags used?**
    - To potentially add filtering features for these popular hashtags to enhance user experience and engagement.
3. **Why did the Marketing team want to know the day with the most new user registrations?**
    - To schedule more ads and promotional activities on these high-traffic days to maximize reach and profit.
4. **Why did the Investors want to know the average number of posts per user?**
    - To gauge user engagement and ensure the platform maintains an active and authentic user base, which is crucial for handling traffic and optimizing platform performance.
5. **Why did the Investors want to know the count of bots and fake accounts?**
    - To ensure their investment is in a credible and valuable asset, rather than a platform with potential issues from fake accounts.

Conclusion

In conclusion, the analysis of Instagram user data illustrates how social media and commercial firms leverage data analytics to gain valuable insights. By examining user behavior, engagement, and interactions, these firms can identify valuable customers and address issues such as inactivity or fake accounts. This process helps firms enhance user experience, optimize marketing strategies, and ensure a genuine and engaged user base. Such analysis is typically conducted on a weekly, monthly, quarterly, or yearly basis, depending on the firm's needs.

Operation Analytics and Investigating Metric Spike

Description

Operation Analytics involves analyzing end-to-end company operations to identify improvement areas. As a Data Analyst Lead, you work with various teams—operations, support, marketing—to derive insights from data. This analysis helps predict company growth or decline, enhances automation, and improves cross-functional collaboration.

Key tasks include investigating metric spikes to address issues like dips in engagement or sales. You are responsible for analyzing data sets and answering questions from different departments, ensuring the company operates efficiently and effectively.

The Problem

Case Study 1: Job Data

1. **Number of Jobs Reviewed per Hour per Day (November 2020):**
   - **Task:** Calculate the number of jobs reviewed per hour for each day in November 2020.
   - **Approach:** Aggregate the total number of jobs reviewed, then divide by the number of hours each day to get a per-hour rate.
2. **Throughput and Rolling Average:**
   - **Task:** Calculate the 7-day rolling average of throughput.
   - **Approach:** Compute the average number of events happening per second over a 7-day window.
   - **Metric Preference:** Choose between a daily metric or a 7-day rolling average based on the stability and variability of the data. A 7-day rolling average smooths out short-term fluctuations and provides a clearer view of trends.
3. **Percentage Share of Each Language (Last 30 Days):**
   - **Task:** Calculate the percentage share of each language used in content over the last 30 days.
   - **Approach:** Sum the total occurrences of each language and divide by the total number of occurrences to get the percentage share.
4. **Displaying Duplicate Rows:**
   - **Task:** Identify and display duplicate rows in the data.
   - **Approach:** Use SQL queries or data analysis tools to find rows with identical values and list them for review.

Case Study 2: Investigating Metric Spikes

1. **Weekly User Engagement:**
   - **Task:** Calculate the weekly user engagement.
   - **Approach:** Aggregate user activity data on a weekly basis to measure engagement levels.
2. **User Growth:**
   - **Task:** Calculate the growth in users over time for a product.
   - **Approach:** Track the number of new users added over specific time periods and compute growth rates.
3. **Weekly Retention of User Sign-Up Cohort:**
   - **Task:** Calculate the weekly retention rate for users who signed up for a product.
   - **Approach:** Measure the percentage of users who continue using the product each week after their initial sign-up.
4. **Weekly Engagement per Device:**
   - **Task:** Calculate the weekly engagement metrics for each device type.
   - **Approach:** Analyze user engagement data segmented by device type on a weekly basis.
5. **Email Engagement Metrics:**
   - **Task:** Calculate metrics related to user engagement with email services.
   - **Approach:** Track metrics such as open rates, click-through rates, and response rates to assess email engagement.

Findings

Job Data Findings

Finding I: Number of Jobs Reviewed per Hour per Day (November 2020)

- **Steps:**
  1. Use job_id data from the job_data table.
  2. Divide the total number of job_ids by the product of 30 days and 24 hours to get the average number of jobs reviewed per hour per day.
- **Output/Result:**

Result Grid Filter Rows:

number_of_jobs_reviewed_per_day
0.0083

Finding II: 7-Day Rolling Average of Throughput

- **Steps:**
  1. Count job_id occurrences and order by date of interview.
  2. Use the ROW function to include the current row and the previous 6 rows.
  3. Compute the average of the jobs reviewed over this 7-day window.
- **Output/Result:**

Result Grid Filter Rows: Search Export:

date_of_review	jobs_reviewed	throughput_7_rolling_average
11/25/2020	1	1.0000
11/26/2020	1	1.0000
11/27/2020	1	1.0000
11/28/2020	2	1.2500
11/29/2020	1	1.2000
11/30/2020	2	1.3333

Finding III: Percentage Share of Each Language (Last 30 Days)

- **Steps:**
  1. Calculate the total number of occurrences of each language.
  2. Divide by the total number of rows to get the percentage share.
  3. Group by language for distinct and non-distinct counts.
- **Output/Result:**

Result Grid Filter Rows: Search Export:

job_id	language	total_of_each_language	percentage_share_of_each_distinct_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

Finding IV: Viewing Duplicate Rows

- **Steps:**
  1. Identify the column(s) to check for duplicates.
  2. Use the ROW_NUMBER function to assign row numbers.
  3. Filter rows with row numbers greater than 1 to identify duplicates.
- **Output/Result:**

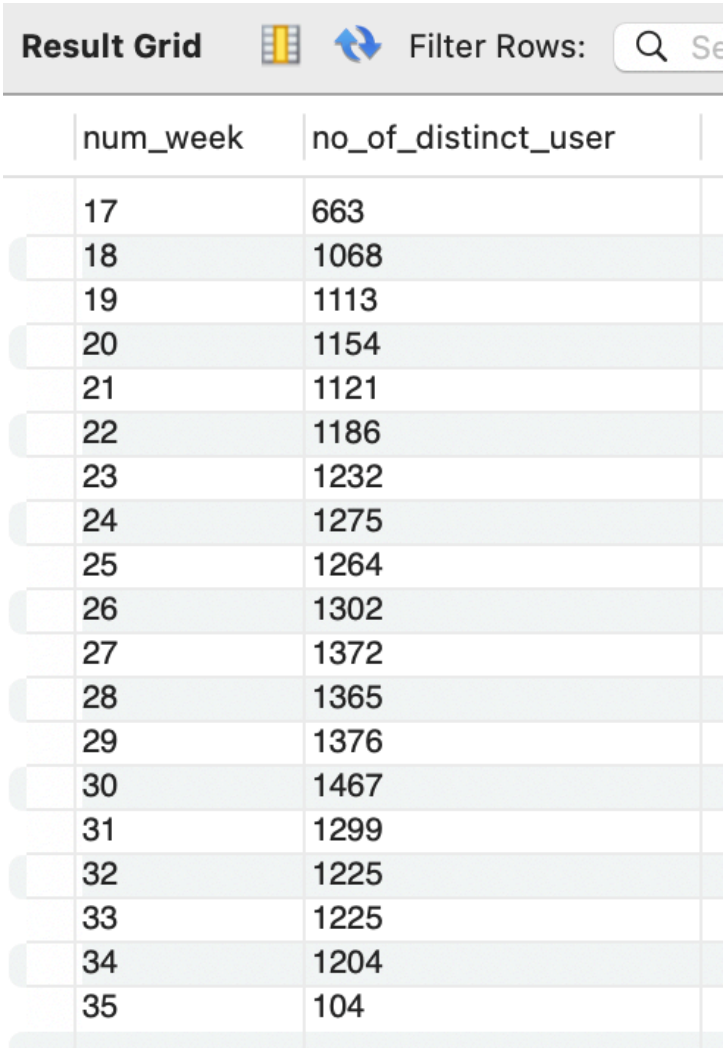
Result Grid Filter Rows: Search Export:

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

Investigating Metric Spike Findings

Finding I: Weekly User Engagement

- **Steps:**
  1. Extract the week from the occurred_at column using EXTRACT and WEEK functions.
  2. Count distinct user_ids per week.
  3. Group by week.
- **Output/Result:**



Result Grid Filter Rows:

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

Finding II: User Growth (Active Users per Week)

- **Steps:**
  1. Extract year and week from the occurred_at column using EXTRACT, YEAR, and WEEK functions.
  2. Group by year and week number.
  3. Order the results by year and week.
  4. Calculate cumulative active users using SUM, OVER, and ROW functions.
- **Output/Result:**

year_num	week_num	num_active_users	cum_active_users
2013	0	23	23
2013	1	30	53
2013	2	48	101
2013	3	36	137
2013	4	30	167
2013	5	48	215
2013	6	38	253
2013	7	42	295
2013	8	34	329
2013	9	43	372
2013	10	32	404
2013	11	31	435
2013	12	33	468
2013	13	39	507
2013	14	35	542
2013	15	43	585
2013	16	46	631
2013	17	49	680
2013	18	44	724
2013	19	57	781
2013	20	39	820
2013	21	49	869
2013	22	54	923
2013	23	50	973
2013	24	45	1018
2013	25	57	1075
2013	26	56	1131
2013	27	52	1183
2013	28	72	1255
2013	29	67	1322
2013	30	67	1389
2013	31	67	1456
2013	32	71	1527
2013	33	73	1600
2013	34	78	1678
2013	35	63	1741
2013	36	72	1813
2013	37	85	1898
2013	38	90	1988
2013	39	84	2072
2013	40	87	2159
2013	41	73	2232
2013	42	99	2331
2013	43	89	2420
2013	44	96	2516

year_num	week_num	num_active_users	cum_active_users
2013	44	96	2516
2013	45	91	2607
2013	46	88	2695
2013	47	102	2797
2013	48	97	2894
2013	49	116	3010
2013	50	124	3134
2013	51	102	3236
2013	52	47	3283
2014	0	83	3366
2014	1	126	3492
2014	2	109	3601
2014	3	113	3714
2014	4	130	3844
2014	5	133	3977
2014	6	135	4112
2014	7	125	4237
2014	8	129	4366
2014	9	133	4499
2014	10	154	4653
2014	11	130	4783
2014	12	148	4931
2014	13	167	5098
2014	14	162	5260
2014	15	164	5424
2014	16	179	5603
2014	17	170	5773
2014	18	163	5936
2014	19	185	6121
2014	20	176	6297
2014	21	183	6480
2014	22	196	6676
2014	23	196	6872
2014	24	229	7101
2014	25	207	7308
2014	26	201	7509
2014	27	222	7731
2014	28	215	7946
2014	29	221	8167
2014	30	238	8405
2014	31	193	8598
2014	32	245	8843
2014	33	261	9104
2014	34	259	9363
2014	35	18	9381

Finding III: Weekly Retention of Sign-Up Cohort

- **Steps:**
  1. Extract the week from occurred_at.
  2. Filter rows for event_type = 'signup_flow' and event_name = 'complete_signup'.
  3. Optionally specify a week number.
  4. Left join with engagement events on user_id.
  5. Group and order by user_id.
- **Output/Result:**

user_id	count(d.user_id)	per_week_retention
11768	1	0
11770	1	0
11775	2	1
11778	3	0
11779	5	1
11780	2	1
11785	1	0
11787	3	1
11791	2	1
11793	6	1
11795	2	1
11798	6	1
11799	10	1

Finding IV: Weekly User Engagement per Device

- **Steps:**
  1. Extract year_num and week_num from occurred_at.
  2. Filter rows where event_type = 'engagement'.
  3. Group and order by year_num, week_num, and device.
- **Output/Result:**

year_num	week_num	device	no_of_users
2014	17	acer aspire desktop	9
2014	17	acer aspire notebook	20
2014	17	amazon fire phone	4
2014	17	asus chromebook	21
2014	17	dell inspiron desktop	18
2014	17	dell inspiron notebook	46
2014	17	hp pavilion desktop	14
2014	17	htc one	16
2014	17	ipad air	27
2014	17	ipad mini	19
2014	17	iphone 4s	21
2014	17	iphone 5	65
2014	17	iphone 5s	42
2014	17	kindle fire	6

Finding V: Email Engagement Metrics

- **Steps:**
  1. Categorize actions (email_sent, email_opened, email_clicked) using CASE, WHEN, THEN functions.
  2. Calculate the email opening rate as (email_opened / email_sent) * 100.
  3. Calculate the email clicking rate as (email_clicked / email_sent) * 100.
- **Output/Result:**

email_opening_rate	email_clicking_rate
33.58339	14.78989

Analysis

Job Data Insights

Number of Jobs Reviewed per Day

- Distinct jobs reviewed per day: 0.0083
- Non-distinct jobs reviewed per day: 0.0111

7-Day Rolling Average Throughput (for Nov 25-30, 2020)

- Distinct and non-distinct jobs: 1, 1, 1, 1.25, 1.2, 1.3333 respectively

Percentage Share of Each Language

- Arabic: 12.5%
- English: 12.5%
- French: 12.5%
- Hindi: 12.5%
- Italian: 12.5%
- Persian: 37.5%

Duplicates

- Two duplicate rows with job_id = 23 and language = Persian

Insights Using the 5 Whys Approach

1. **Difference in Distinct vs. Non-Distinct Jobs Reviewed per Day:**
   - **Why?** May be due to repeated values or duplicate rows in the dataset.
2. **Preference for 7-Day Rolling Average over Daily Metric Average:**
   - **Why?** The 7-day rolling average provides a smoother trend over a week, offering a more comprehensive view compared to daily metrics which only reflect one day's data.
3. **Percentage Share Discrepancy for Persian Language:**
   - **Why?** Potential reasons include duplicate rows with Persian language or a higher actual number of unique users speaking Persian.
4. **Importance of Identifying Duplicate Rows:**
   - **Why?** Duplicates can skew analysis results, leading to incorrect business decisions and potential financial losses.

Investigating Metric Spike Insights

Weekly User Engagement

- Highest engagement recorded in week 31: 1685

Total Active Users (Weeks 1, 2013 to Week 35, 2014)

- Total active users: 9381

Email Engagement Metrics

- Email opening rate: 33.5833%
- Email clicking rate: 14.7898%

Insights Using the 5 Whys Approach

1. **Increase in Weekly User Engagement:**
   - **Why?** Initial low engagement is typical for new products. Increased engagement suggests users found the product/service valuable.
2. **Importance of Weekly Retention:**
   - **Why?** Weekly retention helps firms follow up with users who have completed sign-up but may need encouragement to become active users.
3. **Significance of Weekly Engagement per Device:**
   - **Why?** Helps firms identify which devices perform well and which need improvements based on user feedback.
4. **Role of Email Engagement:**
   - **Why?** Email engagement rates (opening and clicking) guide firms in refining email strategies. Lower rates indicate the need for more compelling email content and targeted offers.

Conclusion

In conclusion, Operation Analytics and Investigating Metric Spikes are essential for understanding and improving business performance. These analyses should be conducted on a daily, weekly, monthly, quarterly, or yearly basis, depending on the specific needs of the firm.

Additionally, firms should prioritize email engagement by crafting compelling subject lines and offering attractive discounts or coupons to enhance customer interaction and retention.

It is also beneficial for firms to have a dedicated team or department to address issues faced by users who abandoned the sign-up process. By effectively guiding these potential customers, firms can increase conversion rates and build a more robust customer base.

Hiring Process Analytics

Description

The hiring process is crucial for any company, especially for multinational corporations (MNCs). Understanding trends such as the number of rejections, interviews conducted, types of jobs, and vacancies is essential for refining hiring strategies. This creates a significant role for Data Analysts, who can analyze these trends to provide valuable insights.

As a Lead Data Analyst at a prominent MNC like Google, you are tasked with examining historical hiring data. Your goal is to analyze this data and answer specific questions to help the hiring department make informed decisions.

The Problem

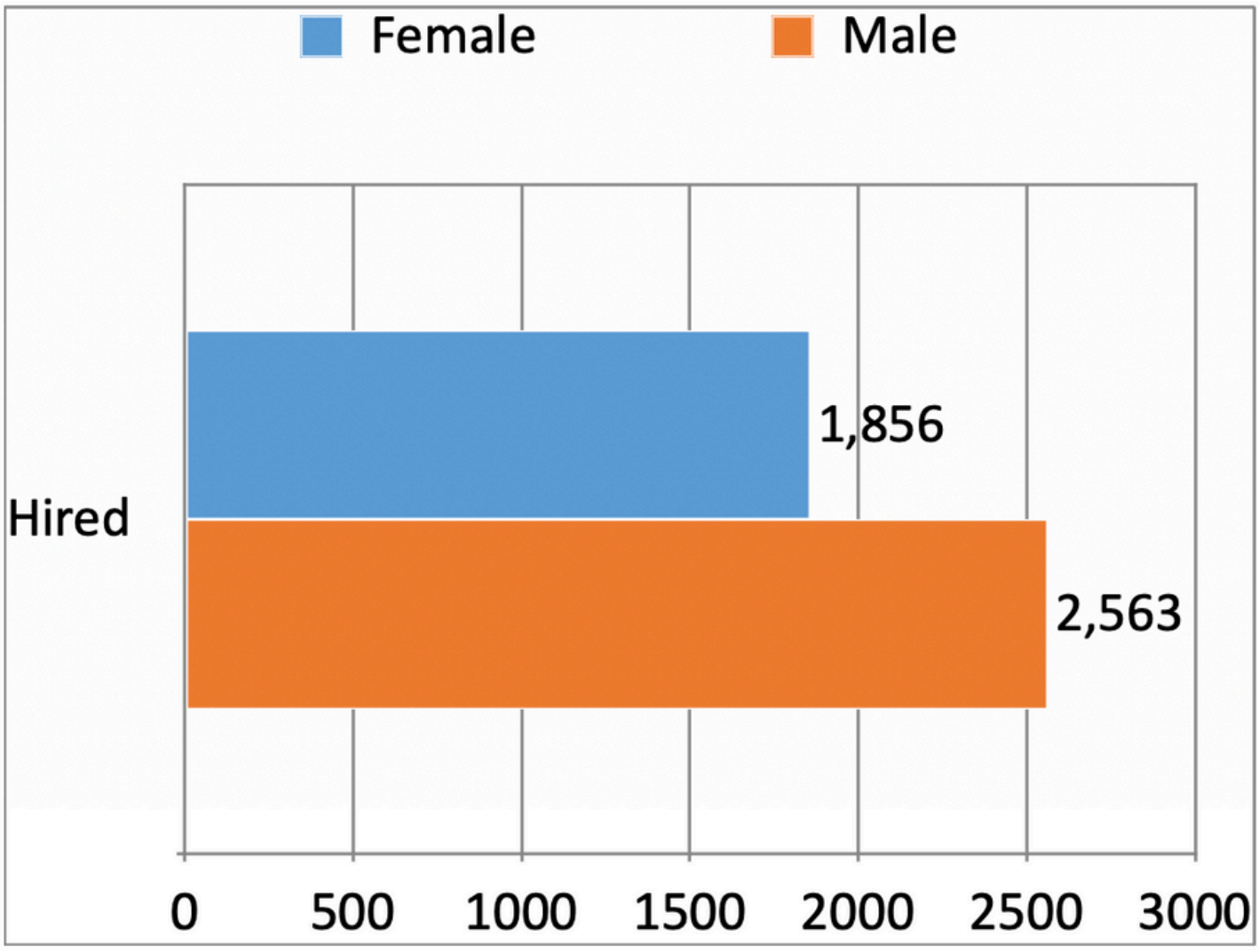
1. **Data Preparation:**
    - Copy Data: Created a copy of the raw data to prevent any changes to the original dataset.
    - Handle Missing Values:Checked for blank spaces and NULL values:
        - Numerical Data: Imputed blank and NULL cells with the mean of the column (if no outliers) or median (if outliers present).
        - Categorical Data: Filled blank cells with the most frequent category.
    - Outliers: Identified and replaced outliers with the median value of the column.
    - Duplicates: Removed any duplicate rows.
    - Irrelevant Data: Eliminated columns that were not necessary for the analysis.
2. **Analysis:**
    - Gender Hiring Analysis: Calculate the number of males and females hired.
    - Average Salary: Compute the average salary offered to employees.
    - Class Intervals: Create class intervals for salary distribution.
    - Charts and Plots:
        - Proportion of Departmental Employees: Use pie charts or bar graphs to visualize the distribution of employees across different departments.
        - Post Tiers Representation: Use charts or graphs to represent different post tiers within the company.
3. **Software Used:** Microsoft Excel

Findings

I. Gender Distribution:

- Males hired: 2,563
- Females hired: 1,855

Column D	Female	Male	Grand Total
Column C	Column A (Count All)		
Hired	1856	2563	4419
Grand Total	1856	2563	4419



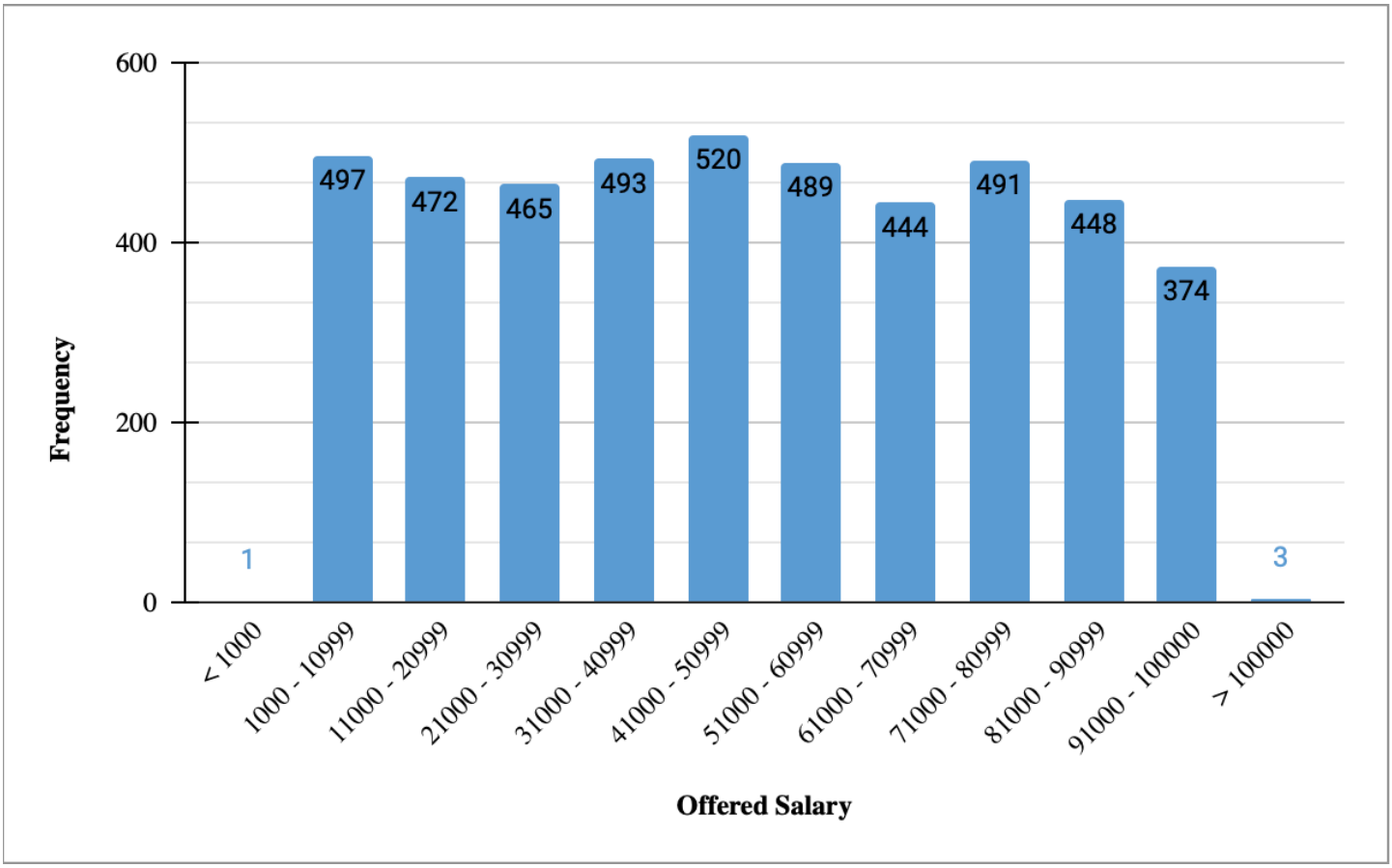
II. Average Salary:

- Average salary offered (after removing outliers): **49,982.13**

Column G (Average)
49892.134738897

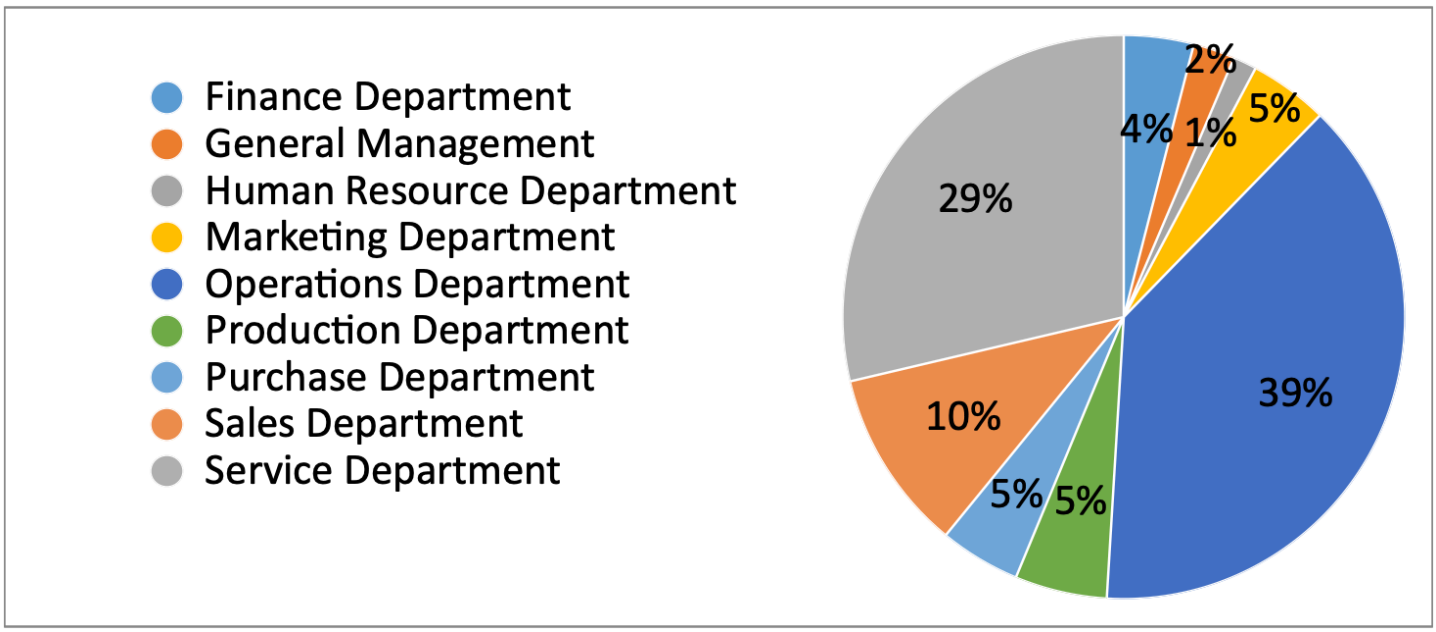
III. Salary Range Insights:

- Highest number of posts (hired and rejected): 770 in the salary range of 41,000 to 50,999.
- Highest number of hired posts: 520 in the salary range of 41,000 to 50,999.



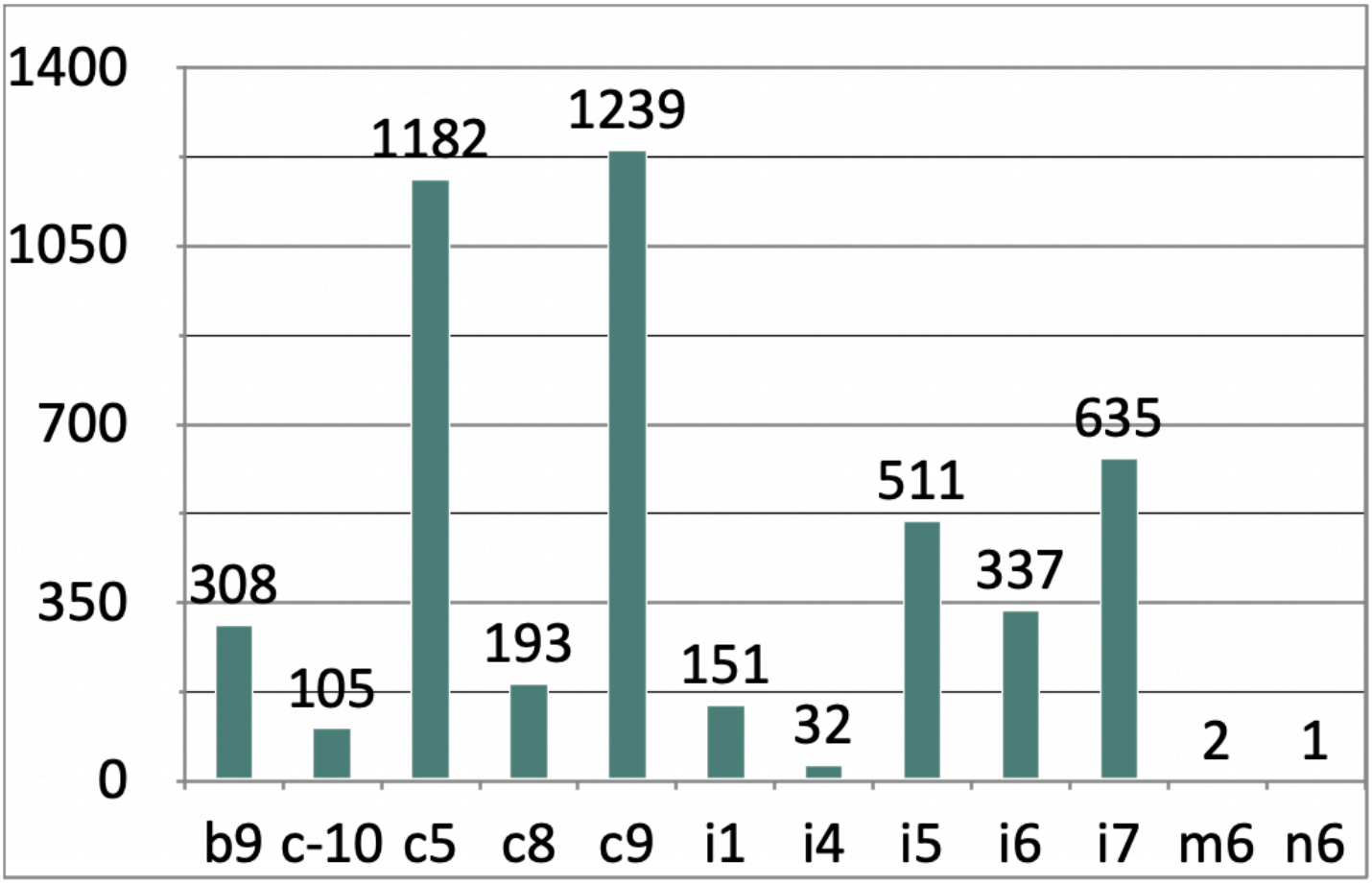
IV. Departmental Distribution:

- Highest number of employees were in the Operations Department: 1,843, accounting for approximately 39% of the total workforce.



V. Post Tiers:

- The post c9 had the highest number of openings: 1,792, representing 25% of the total job openings in the company.



Analysis

Insights Using the Why's Approach

1. Gender Disparity in Hiring:

- Why is there a significant difference in the number of males and females hired?
  - Reason: This disparity may be due to varying gender equality standards across different regions. In some regions, such as certain Gulf countries, parts of Africa, and specific Asian countries, gender equality in the workplace has not yet been fully achieved, leading to a higher number of males being hired compared to females.

2. Salary Distribution:

- Why are there fewer employees with salaries above 85,000 and more with salaries between 35,000 and 60,000?
  - Reason: High salaries are typically offered for specialized positions requiring extensive experience and expertise, which are less common. In contrast, salaries in the range of 35,000 to 60,000 are more typical for employees with several years of experience, reflecting incremental increases based on performance and tenure within the company.

3. High Number of Employees in Operations Department:

- Why does the Operations Department have the highest number of employees?
  - Reason: The Operations Department serves as a central hub, managing execution tasks and supporting various functions across the organization. Its broad responsibilities and essential role in daily operations contribute to a higher workload and consequently, a larger workforce compared to other departments.

Conclusion

In conclusion, Hiring Process Analytics is crucial for companies to effectively plan their recruitment and workforce management. It provides valuable insights that assist in determining future job openings and workforce needs.

- Frequency of Analysis: Hiring Process Analytics is typically performed on a monthly, quarterly, or yearly basis, depending on the company's requirements and policies.
- Departmental Workforce Distribution: The Operations Department often has the highest number of employees due to its central role in executing and managing various tasks across the organization.
- Salary Distribution: Employees with high salary packages usually possess specialized skills and extensive experience, which justifies their compensation compared to other employees.
- Benefits of Hiring Process Analytics: This analysis helps companies set appropriate salaries for new hires, assess departmental workforce needs, and make informed decisions regarding appraisals and increments for current employees.

IMDB Movie Analysis

Description

For this IMDB Movie Analysis project, you'll explore factors influencing movie success based on IMDB ratings. Start by analyzing genre distribution and its effect on ratings, calculating statistics like mean and median for each genre. Then, examine how movie duration impacts ratings, using scatter plots to visualize the relationship. Next, assess the influence of language on ratings, determining the average ratings for different languages. Investigate the role of directors by finding those with the highest average ratings and using percentiles to gauge their impact. Finally, analyze the correlation between movie budgets and gross earnings, identifying films with the highest profit margins. Data cleaning is essential to handle missing values and remove duplicates. Use the Five 'Whys' approach to dig deeper into trends and provide actionable insights in your report with visualizations.

The Problem

1. Movies with Highest Profit:

- **Task**: Identify the movies with the highest profit.
- **Steps**:
  1. Create a profit column as the difference between gross and budget.
  2. Sort by the profit column in descending order.
  3. Plot profit (y-axis) vs. budget (x-axis) to observe outliers.

2. Top 250 Movies:

- **Task**: Find the IMDb Top 250 movies.
- **Steps**:
  1. Create an IMDb_Top_250 column for the top 250 movies by imdb_score.
  2. Ensure these movies have more than 25,000 user votes (num_voted_users).
  3. Add a Rank column from 1 to 250.
  4. Extract non-English movies into Top_Foreign_Lang_Film.

3. Best Directors:

- **Task**: Identify the best directors based on IMDb scores.
- **Steps**:
  1. Group by director_name.
  2. Find the top 10 directors with the highest mean imdb_score.

3. Sort alphabetically in case of a tie and store in top10director.

4. Popular Genres:

- **Task**: Identify popular genres.
- **Steps**:
  - Analyze genres similarly to previous steps.

5. Actor Analysis and Charts:

- **Task**: Identify critic-favorite and audience-favorite actors.
- **Steps**:
  1. Create columns Meryl_Streep, Leo_Caprio, and Brad_Pitt for movies starring each actor.
  2. Combine these into a Combined column.
  3. Group by actor_1_name and calculate mean num_critic_for_reviews and num_users_for_review.
  4. Identify actors with the highest means.
  5. Create a decade column (e.g., 1920s, 1930s).
  6. Group by decade and sum user votes, storing results in df_by_decade.
  7. Use a bar chart to observe changes in user votes over decades.

Findings

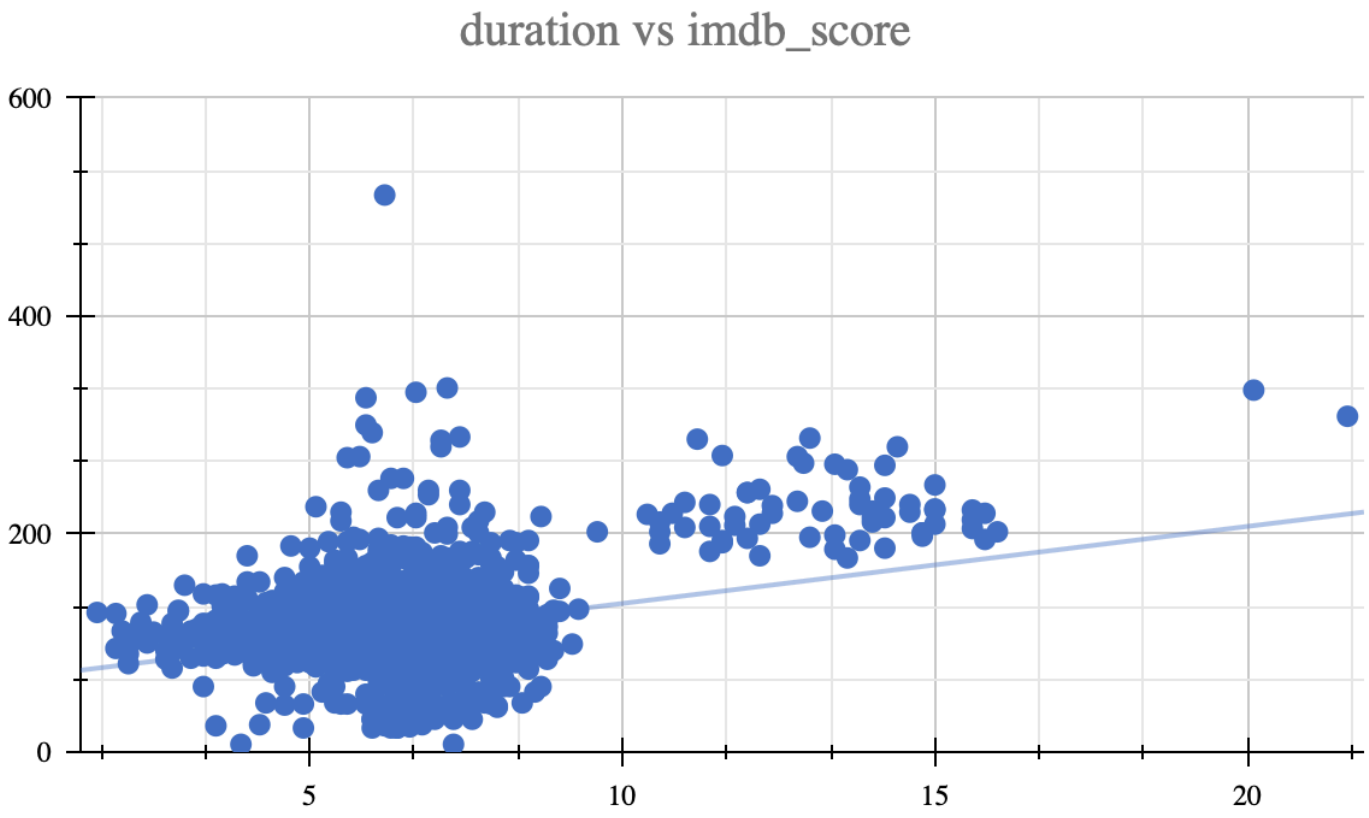
I: Movie Genre Analysis

- Analyzed the distribution of movie genres in the dataset.
- Used Excel's COUNTIF function to identify the most common genres.
- Calculated descriptive statistics (mean, median, mode, range, variance, and standard deviation) for IMDB scores within each genre.
- Findings revealed how different genres impact movie ratings.

genres	COUNTA of genres	SUM of imdb_score	AVERAGE of imd	MAX of imdb_sc	MIN of imdb_sc	MEDIAN of imdb	VARP of imdb_s	STDEVP of imdb
Action	6	35.5	5.916666667	8.1	3.7	5.8	1.978055556	1.40643363
Action|Adventure	9	61.7	6.855555556	8.5	4.8	6.7	1.355802469	1.164389312
Action|Adventure|Animation|Comedy|Crime|Family|Fantasy	1	6.2	6.2	6.2	6.2	6.2	0	0
Action|Adventure|Animation|Comedy|Drama|Family|Sci-Fi	2	15.9	7.95	8	7.9	7.95	0.0025	0.05
Action|Adventure|Animation|Comedy|Family	5	35.8	7.16	7.6	6.7	7.2	0.0904	0.3006659276
Action|Adventure|Animation|Comedy|Family|Fantasy	4	29.3	7.325	7.9	6.7	7.35	0.281875	0.5309190145
Action|Adventure|Animation|Comedy|Family|Fantasy|Sci-Fi	2	11.4	5.7	6.3	5.1	5.7	0.36	0.6
Action|Adventure|Animation|Comedy|Family|Sci-Fi	3	17.9	5.966666667	6.5	5.4	6	0.2022222222	0.4496912521
Action|Adventure|Animation|Comedy|Fantasy	1	7.2	7.2	7.2	7.2	7.2	0	0
Action|Adventure|Animation|Comedy|Fantasy|Sci-Fi	1	6.9	6.9	6.9	6.9	6.9	0	0
Action|Adventure|Animation|Drama|Mystery|Sci-Fi|Thriller	1	7.1	7.1	7.1	7.1	7.1	0	0
Action|Adventure|Animation|Family	1	8	8	8	8	8	0	0
Action|Adventure|Animation|Family|Fantasy	1	7	7	7	7	7	0	0
Action|Adventure|Animation|Family|Fantasy|Sci-Fi	1	6.8	6.8	6.8	6.8	6.8	0	0
Action|Adventure|Animation|Family|Sci-Fi	2	12.5	6.25	6.6	5.9	6.25	0.1225	0.35
Action|Adventure|Animation|Family|Sci-Fi|Thriller	1	6.9	6.9	6.9	6.9	6.9	0	0
Action|Adventure|Animation|Fantasy	1	6.3	6.3	6.3	6.3	6.3	0	0
Action|Adventure|Animation|Fantasy|Romance|Sci-Fi	1	6.4	6.4	6.4	6.4	6.4	0	0
Action|Adventure|Biography|Drama|History|Romance|War	1	5.5	5.5	5.5	5.5	5.5	0	0
Action|Adventure|Biography|Drama|History|Thriller	1	7	7	7	7	7	0	0
Action|Adventure|Comedy	10	59.6	5.96	7.3	5.1	5.8	0.3364	0.58
Action|Adventure|Comedy|Crime	6	36.4	6.066666667	7.6	4.4	6.45	1.245555556	1.116044603
Action|Adventure|Comedy|Crime|Family|Romance|Thriller	1	5	5	5	5	5	0	0
Action|Adventure|Comedy|Crime|Music|Mystery	1	6.3	6.3	6.3	6.3	6.3	0	0
Action|Adventure|Comedy|Crime|Mystery|Thriller	1	6.9	6.9	6.9	6.9	6.9	0	0
Action|Adventure|Comedy|Crime|Thriller	1	5.5	5.5	5.5	5.5	5.5	0	0
Action|Adventure|Comedy|Drama|Music|Sci-Fi	1	6.7	6.7	6.7	6.7	6.7	0	0
Action|Adventure|Comedy|Drama|Thriller	1	5.2	5.2	5.2	5.2	5.2	0	0
Action|Adventure|Comedy|Drama|War	1	7.1	7.1	7.1	7.1	7.1	0	0

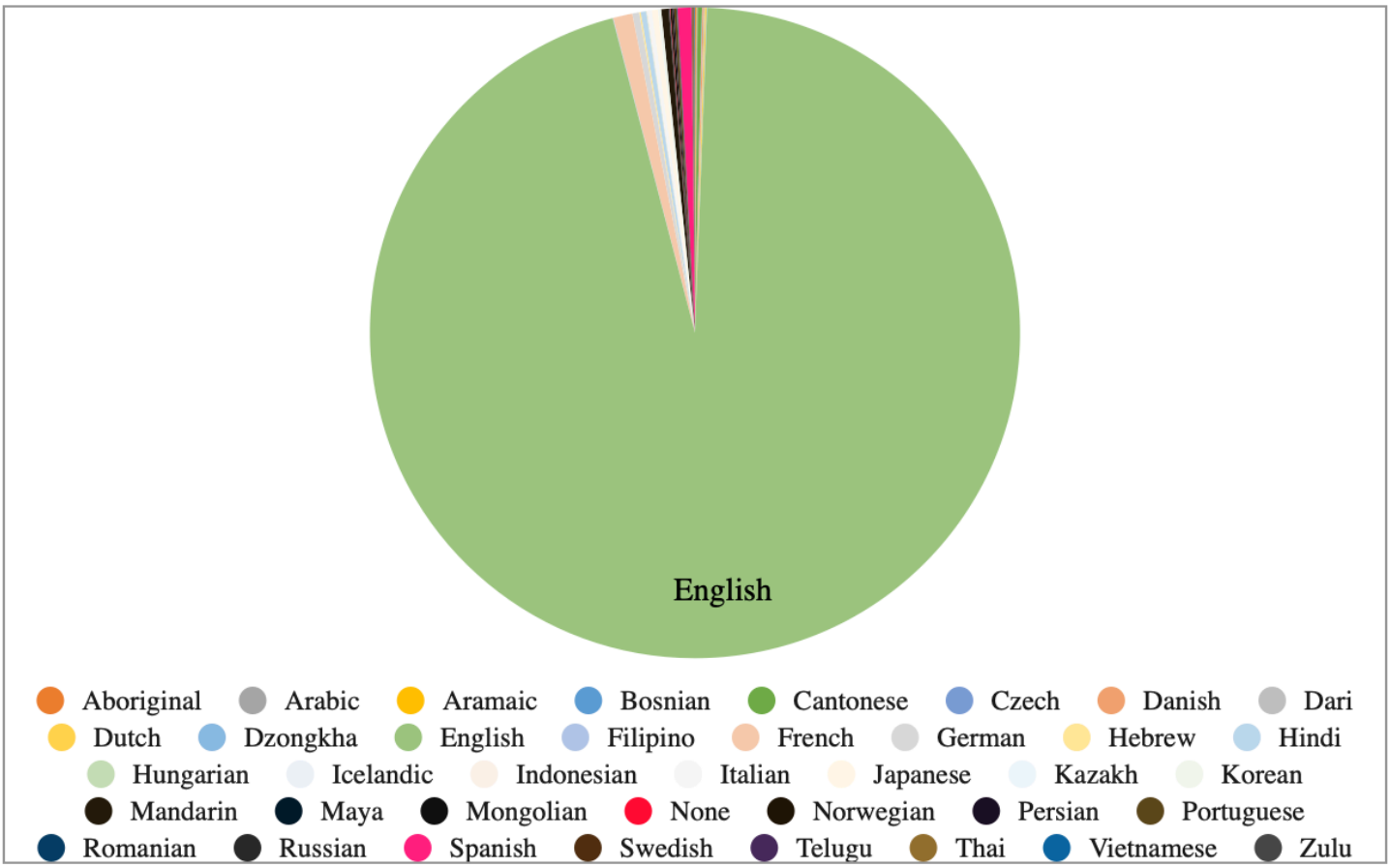
II: Movie Duration Analysis

- Examined the distribution of movie durations and their relationship with IMDB scores.
- Computed descriptive statistics (mean, median, and standard deviation) for movie durations.
- Created a scatter plot to visualize the relationship between movie duration and IMDB score, including a trendline to assess direction and strength.
- Insights on how movie length affects ratings were derived.



III: Language Analysis

- Investigated the distribution of movies by language.
- Used Excel's COUNTIF function to determine the most common languages.
- Calculated descriptive statistics (mean, median, and standard deviation) for IMDB scores by language.
- Analyzed how language influences movie ratings.



IV: Director Analysis

- Identified top directors based on their average IMDB scores.
- Created a pivot table to compute the average IMDB score for each director.
- Used Excel's PERCENTILE function to determine the top directors.
- Compared these top directors' scores to the overall distribution to understand their impact on movie success.

1	director_name	AVERAGE of imdb_score
2	Tony Kaye	8.6
3	Charles Chaplin	8.6
4	Ron Fricke	8.5
5	Majid Majidi	8.5
6	Damien Chazelle	8.5
7	Alfred Hitchcock	8.5
8	Sergio Leone	8.433333333
9	Christopher Nolan	8.425
10	S.S. Rajamouli	8.4
11	Richard Marquand	8.4

V: Budget Analysis

- Explored the relationship between movie budgets and their financial success.
- Calculated the correlation coefficient between budgets and gross earnings using Excel's CORREL function.
- Determined the profit margin (gross earnings - budget) for each movie.
- Identified movies with the highest profit margins to understand financial success.

MAX	Movie
523505847	AvatarÂ

Analysis

1. Why is it that the most-rated IMDB movie and the highest profit movie are not the same?

- **Insight**: IMDB ratings are influenced by the opinions of users who actively rate movies on the platform, which may not be representative of the broader audience. In contrast, profit is determined by the total box office revenue, which reflects the movie's commercial success across a global audience. Therefore, a movie might be highly rated by a niche group of critics or enthusiasts without necessarily being a top commercial success.

2. Why were there more votes during the decade 2001-2010?

- **Insight**: This period saw significant advancements in technology, including improvements in computer graphics and an increase in the production of films. Additionally, the proliferation of the internet and changes in film production regulations contributed to a higher volume of movie releases and consequently more voting activity on IMDB.

3. Why are only English-language movies in the top 5 ranked movies on IMDB?

- **Insight**: English-language movies, predominantly from the USA, benefited from a strong domestic film industry and robust economic conditions during this period. The influence of Hollywood and its global reach likely contributed to these films dominating IMDB rankings. Additionally, the high financial investment and marketing in English-language films increased their visibility and appeal.

4. Why do Drama and Comedy genres have the highest popularity?

- **Insight**: People often seek entertainment and relaxation, which Drama and Comedy genres provide. Comedy offers a way to relieve stress and enjoy a lighter experience, while Drama can offer deep emotional engagement. The preference for these genres reflects a desire for escapism and emotional relief from everyday stresses.

5. Why were there more votes in the decade 2001-2010 compared to 2011-2020, despite advancements in graphics and animation during 2011-2020?

- **Insight**: The 2001-2010 decade saw a surge in movie production and audience engagement. Technological advancements, including the introduction of VPNs and increased internet access, led to greater film consumption and voting. However, the rise of piracy and illegal distribution during the 2011-2020 period may have discouraged theatrical viewership and voting, leading to fewer votes despite technological progress.

Conclusion

In conclusion, IMDB Movie Analysis is essential for movie makers, investors, stakeholders, and theatre owners. This analysis is valuable both during pre-production and post-production phases. It's important to note that a high IMDB rating does not always equate to the highest profit, as profit is primarily driven by global ticket sales. Audiences often gravitate towards Comedy and Drama genres due to their need for relaxation, while Action and Horror genres may be less preferred. Directors and production teams should consider these insights to optimize their pre-production strategies and better align with audience preferences.

Bank Loan Case Study

Description

Loan-providing companies often face challenges in approving loans due to applicants' insufficient or non-existent credit histories. This situation can be exploited by some consumers who then default on their loans. Suppose you work for a consumer finance company specializing in lending various types of loans to urban customers. To address this, you need to employ Exploratory Data Analysis (EDA) to uncover patterns in the data. This will help ensure that applicants who are capable of repaying the loan are not rejected.

When the company receives a loan application, it must decide on loan approval based on the applicant's profile. There are two types of risks associated with the company's decision:

1. If the applicant is likely to repay the loan, not approving the loan results in a loss of business for the company.
2. If the applicant is not likely to repay the loan (i.e., is likely to default), approving the loan may lead to a financial loss for the company.

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

1. **Approved**: The company approves the loan application.
2. **Cancelled**: The client cancels the application during the approval process, either due to changing their mind or receiving worse pricing due to higher risk.
3. **Refused**: The company rejects the loan application because the client does not meet their requirements.
4. **Unused Offer**: The loan offer is cancelled by the client at different stages of the process.

The Problem

This case study aims to apply Exploratory Data Analysis (EDA) in a real business scenario, particularly within a consumer finance company specializing in urban loans. The goal is to identify patterns indicating loan repayment difficulties to minimize financial risk. Using EDA, the company wants to uncover key variables driving loan defaults to make informed decisions about loan approvals, rejections, and interest rate adjustments.

Steps:

1. **Copy the Raw Data**: Create a copy of the original dataset.
2. **Handle Missing Data**:

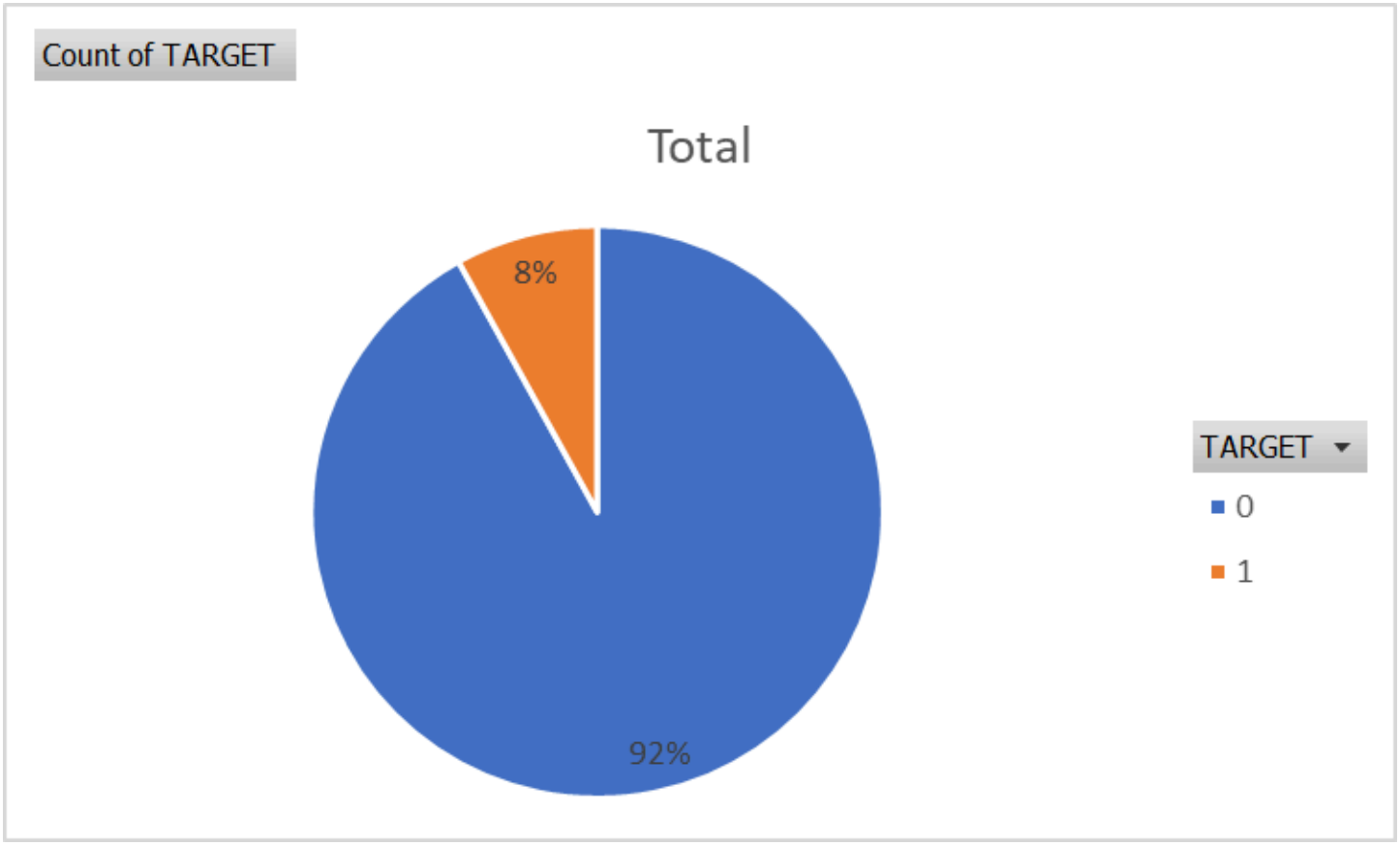
- Drop columns with more than 50% null values:
  - OWN_CAR_AGE, EXT_SOURCE_1, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BUILD_AVG, COMMON_AREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG, FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG, NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, 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.
- Replace null values in columns with less than 50% null data with mean, median, or mode as appropriate.

3. **Drop Irrelevant Columns**:
   - FLAG_MOBILE, FLAG_EMPLOY_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL, CNT_FAMILY_MEMBERS, REGION_RATING_CLIENT, REGION_RATING_CLIENT_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 to FLAG_DOC_21.
4. **Impute Missing Values**:
   - OCCUPATION_TYPE: Replace blanks with 'Laborers'.
   - AMT_ANNUITY: Replace blanks with the median (24,903).
   - AMT_GOODS_PRICE: Replace blanks with the median (450,000).
   - NAME_TYPE_SUITE: Replace blanks with 'Unaccompanied'.
   - ORGANIZATION_TYPE: Replace blanks with 'Business Entity Type 3'.
5. **Previous Application Dataset**:
   - Drop irrelevant columns: 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.
   - Remove rows with 'XNA' and 'XAP' in NAME_TYPE_SUITE.

Findings

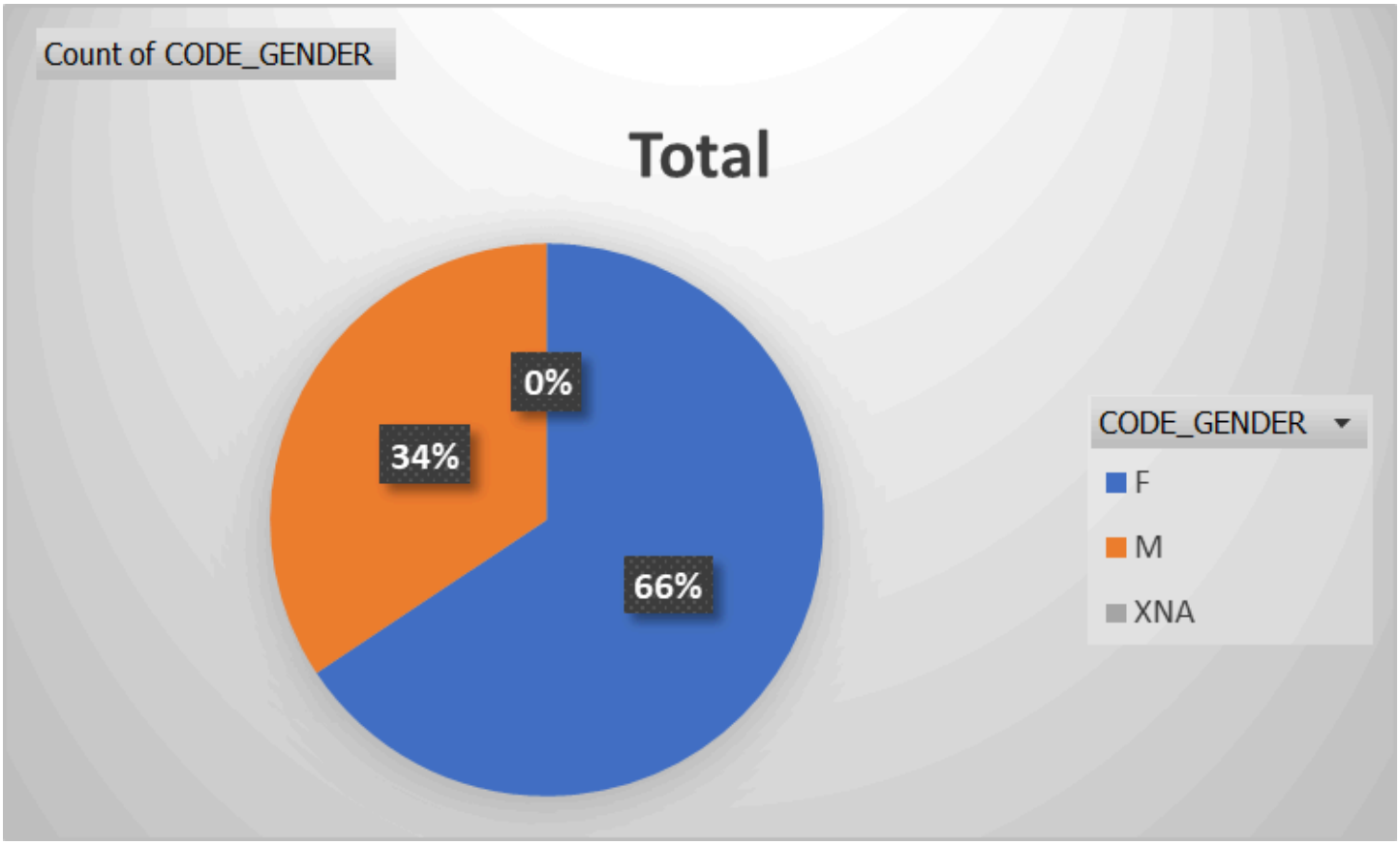
I. **Target Variable Analysis**:

- Approximately 92% of clients had no payment issues, while 8% experienced problems.



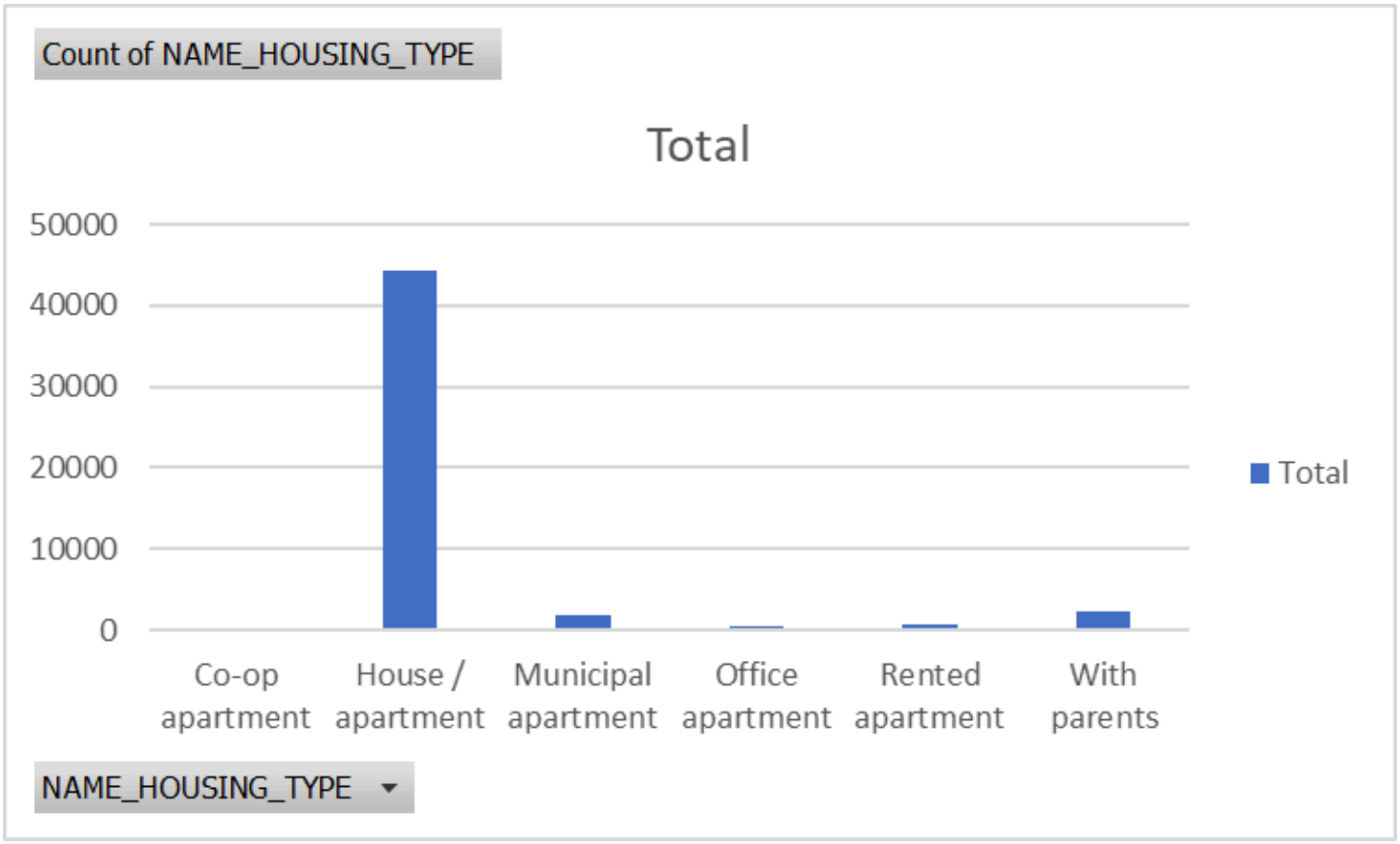
II. **Gender Distribution**:

- Around 66% of clients are female, and 34% are male. No applicants are classified as XNA.



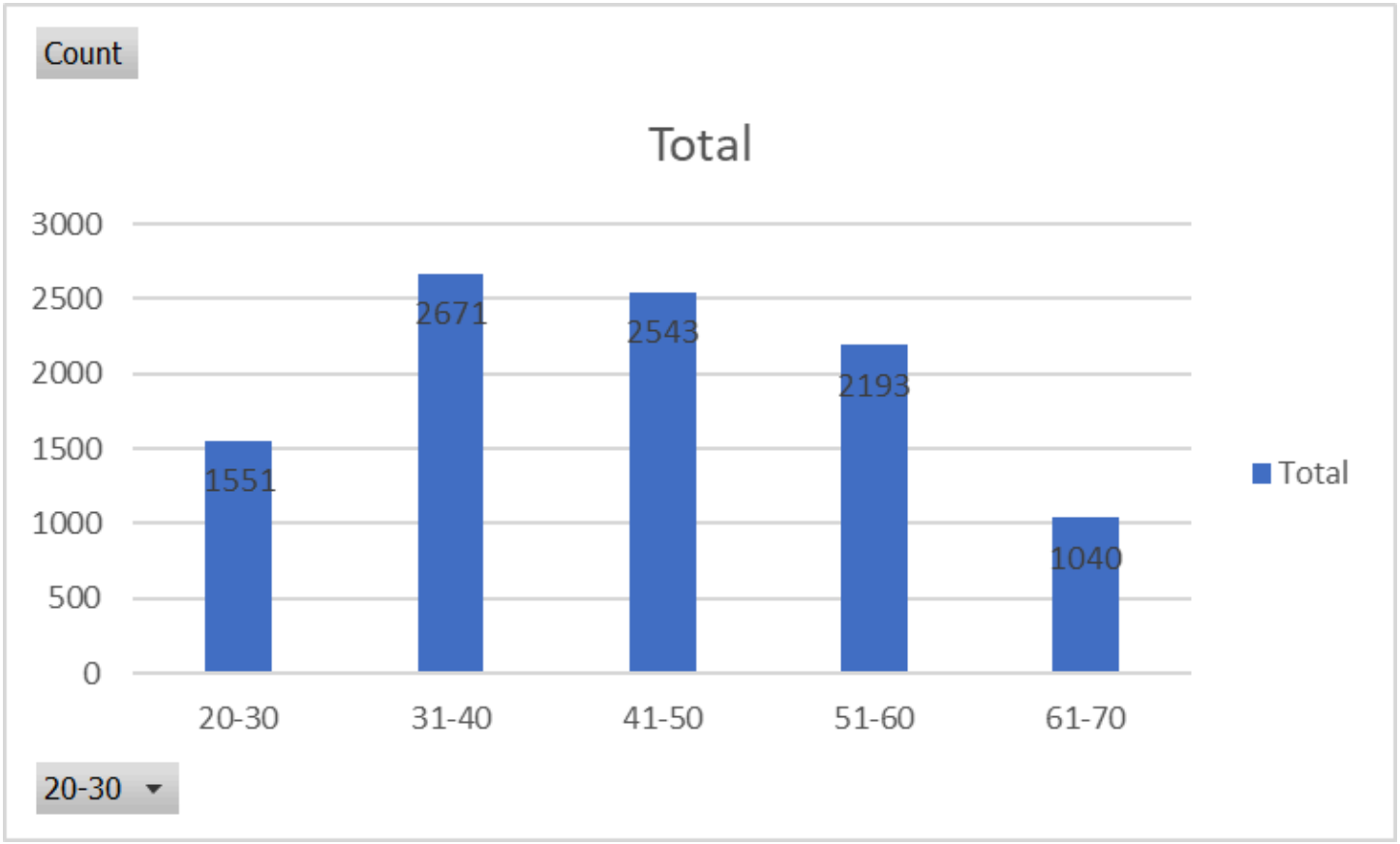
III. **Housing Situation**:

- The bank should target clients without their own apartments, such as those in co-op, municipal, rented apartments, or living with parents.



IV. **Age Group**:

- Most applicants are aged 31-40.

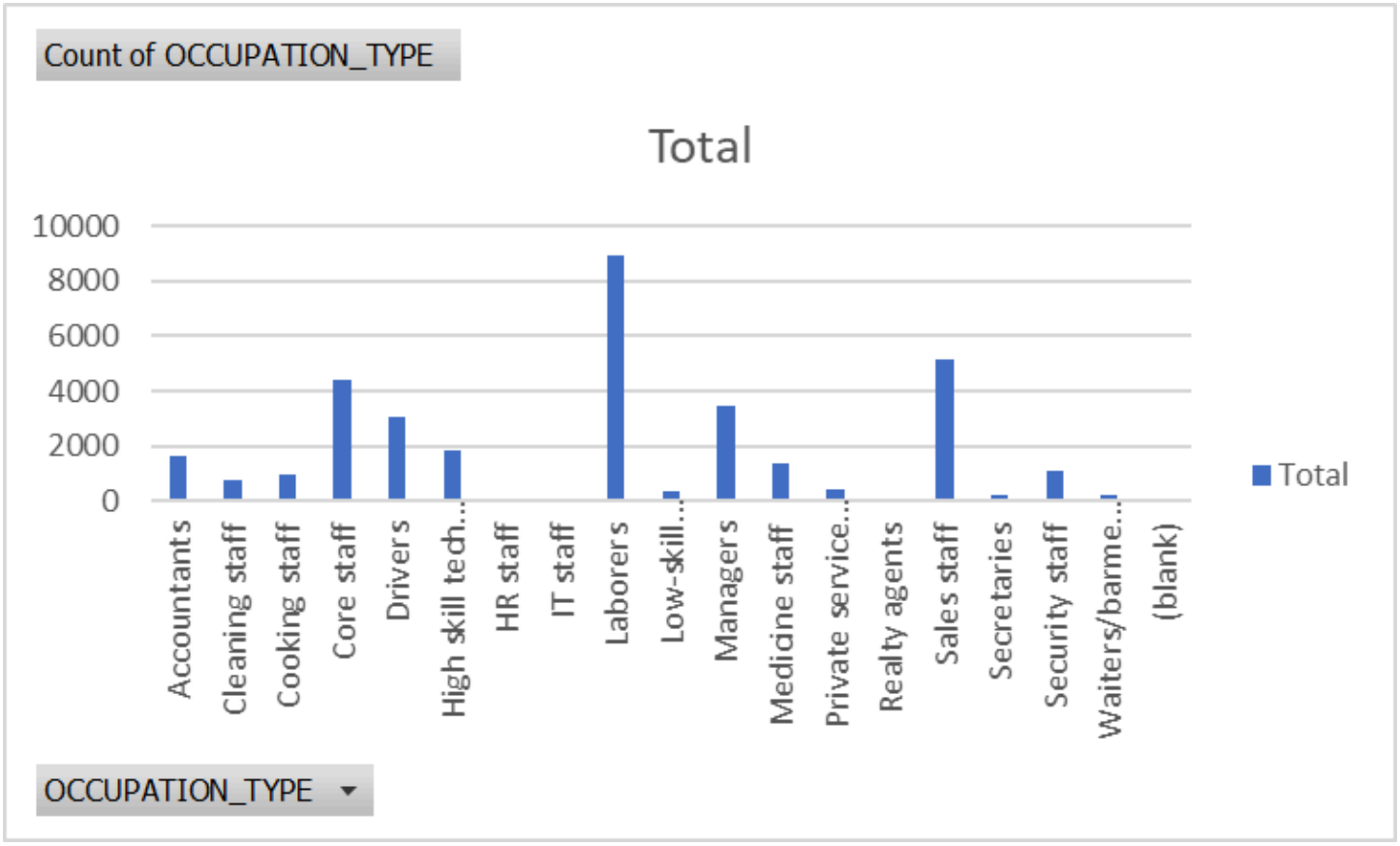


V. **Age Group Payment Behavior**:

- The 31-40 age group has the highest number of both successful payments and payment issues.

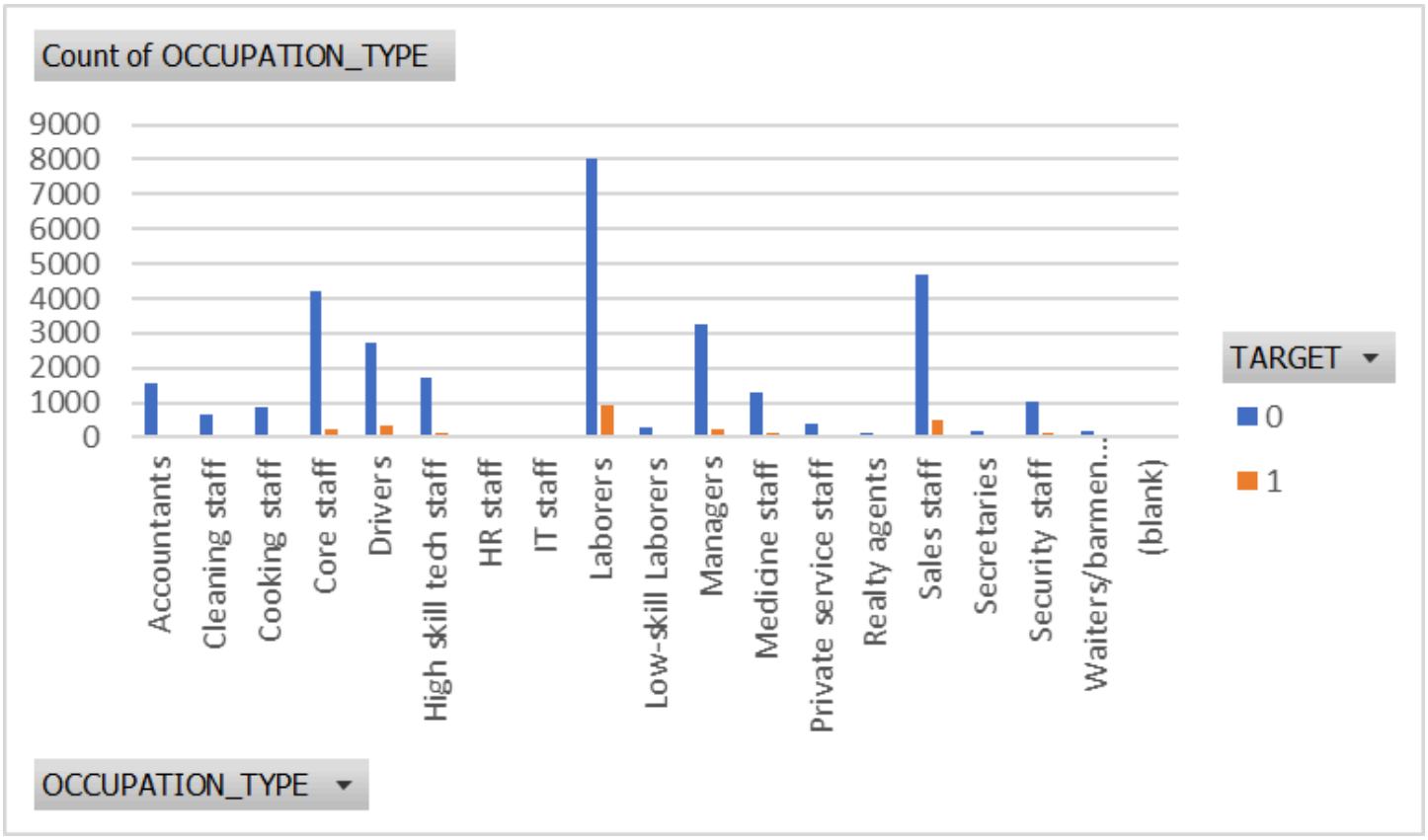
VI. **Income and Payment Issues**:

- Low-income clients have the highest count of no payment issues, while medium-income clients have the highest count of payment issues.



VII. **Occupation and Payment Issues**:

- Clients in the 'Laborers' occupation have the highest number of both successful payments and payment issues.

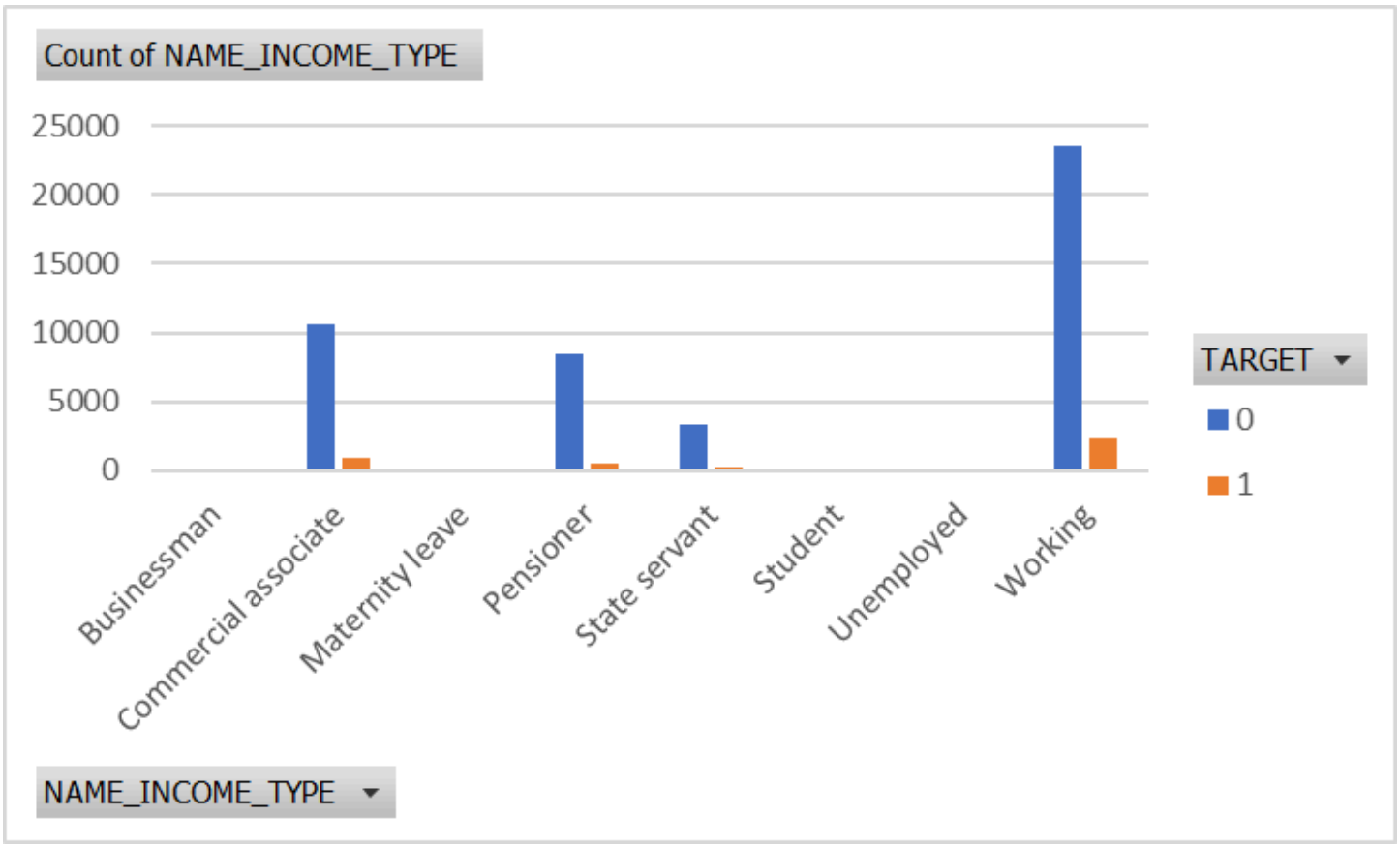


VIII. **Income Type and Payment Issues**:

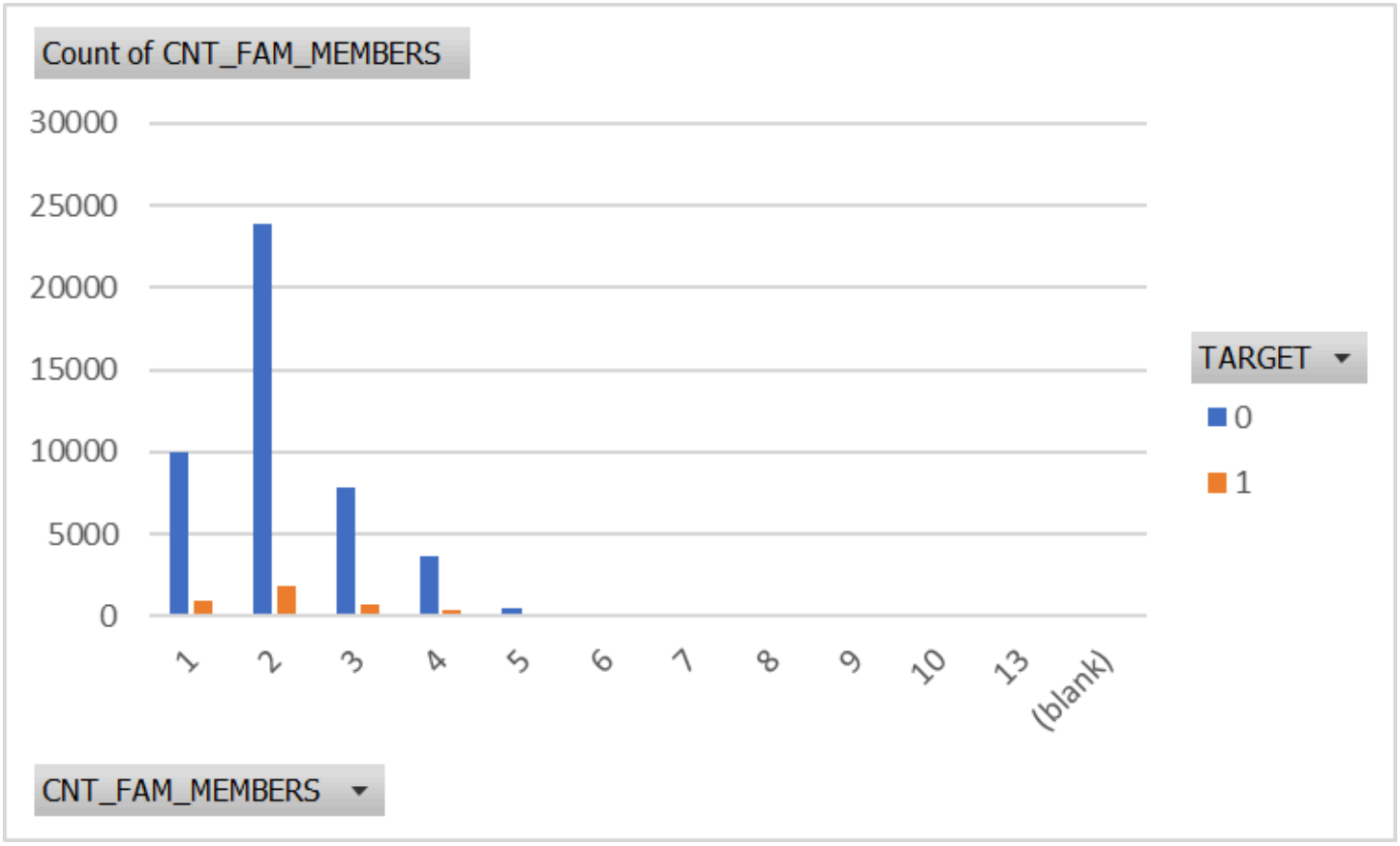
- Clients with 'WORKING' income type have the highest count of both successful payments and payment issues.

IX. **Total Income Range**:

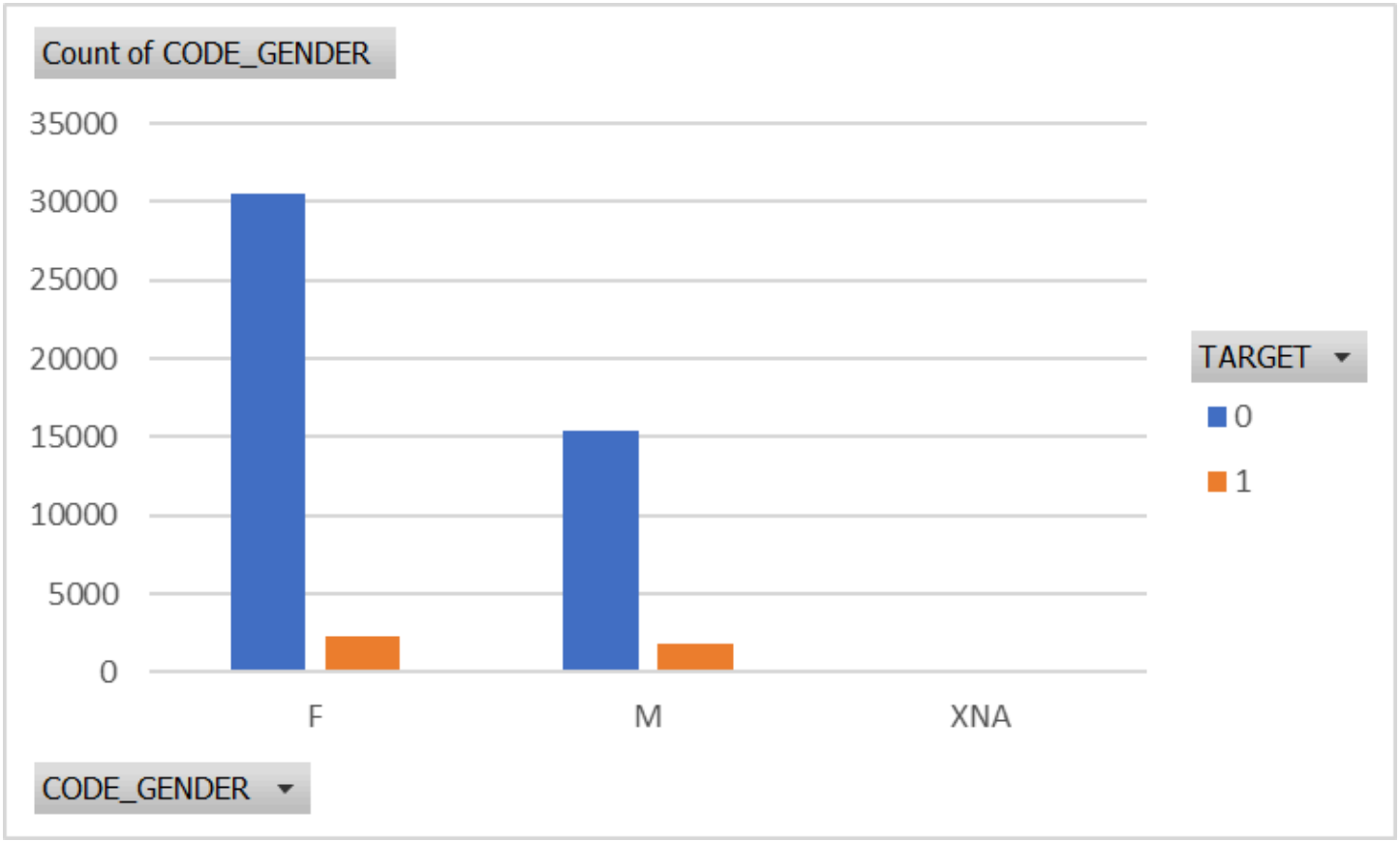
- Low-income clients have the highest count of both no payment issues and payment issues.



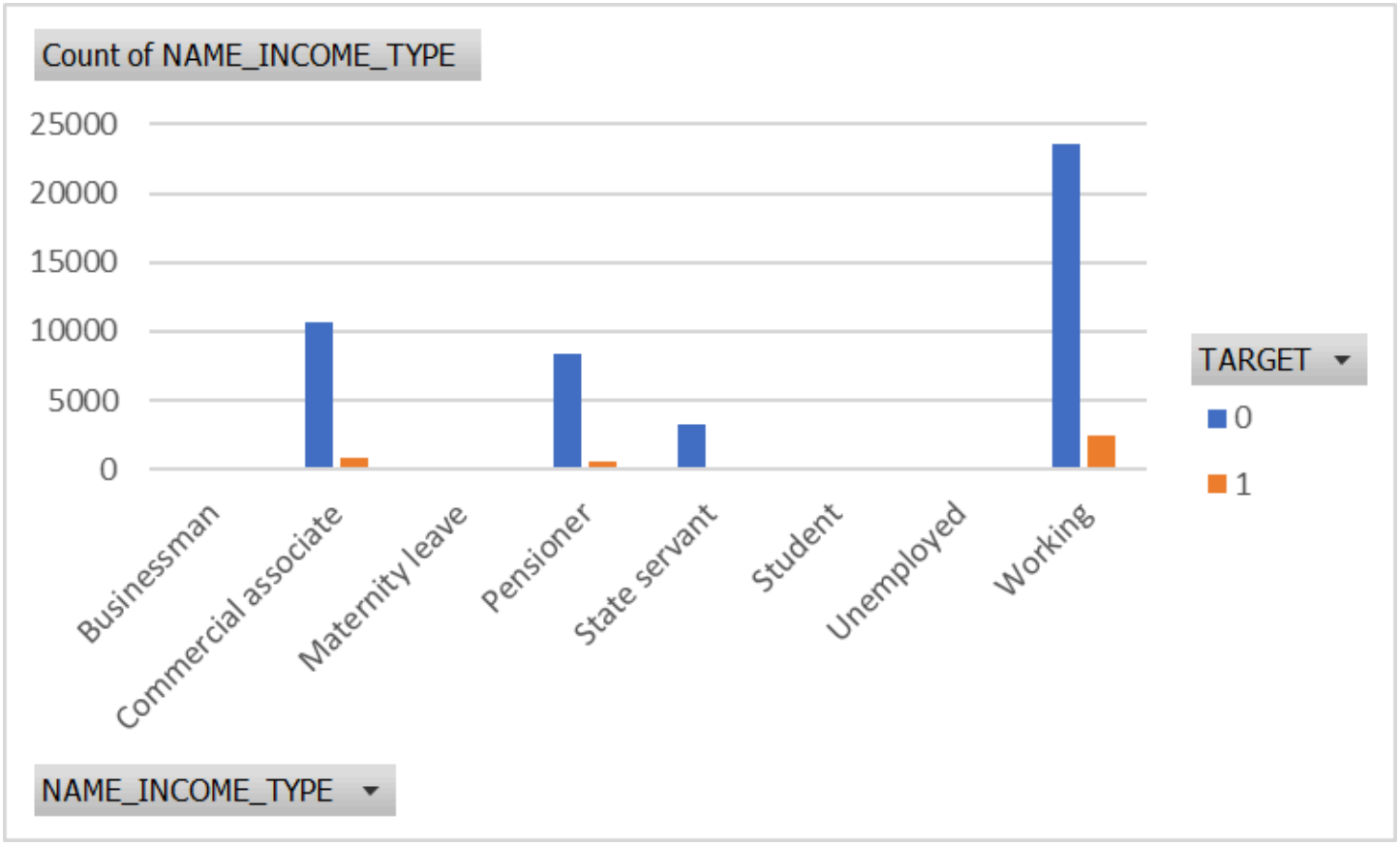
X. **Family Members**: - Clients with two family members have the highest count of both no payment issues and payment issues.



XI. **Gender and Default**: - Female clients have the highest number of non-defaulters.



XII. **Income Type and Non-Defaulters**: - Clients with 'WORKING' income type have the highest count of non-defaulters.

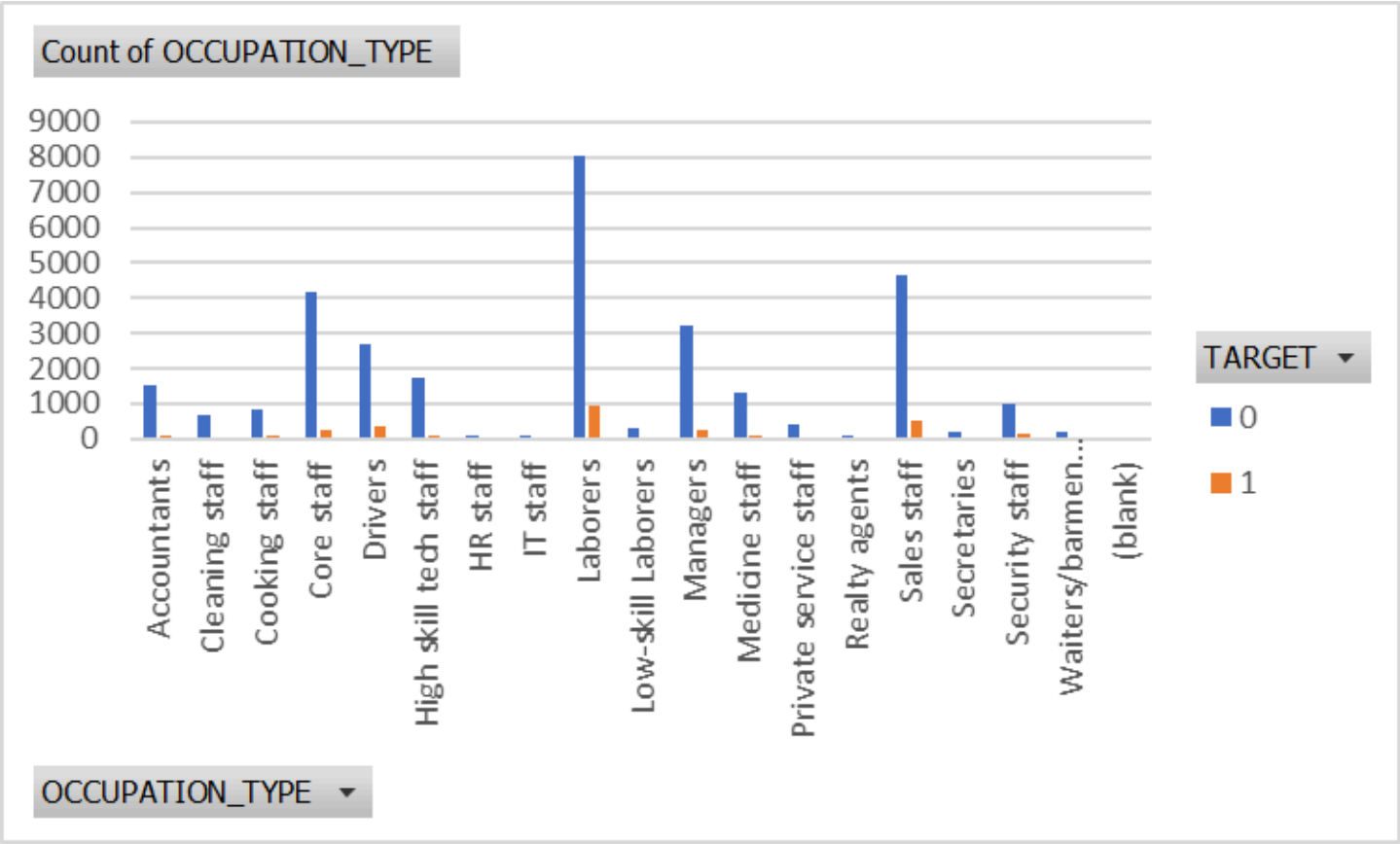


XIII. **Education Level and Non-Defaulters**: - Clients with 'SECONDARY/SECONDARY SPECIAL' education have the highest count of non-defaulters.

XIV. **Family Status and Non-Defaulters**: - Married clients have the highest count of non-defaulters.

XV. **Housing Type and Non-Defaulters**: - Clients with 'House/Apartment' housing type have the highest count of non-defaulters.

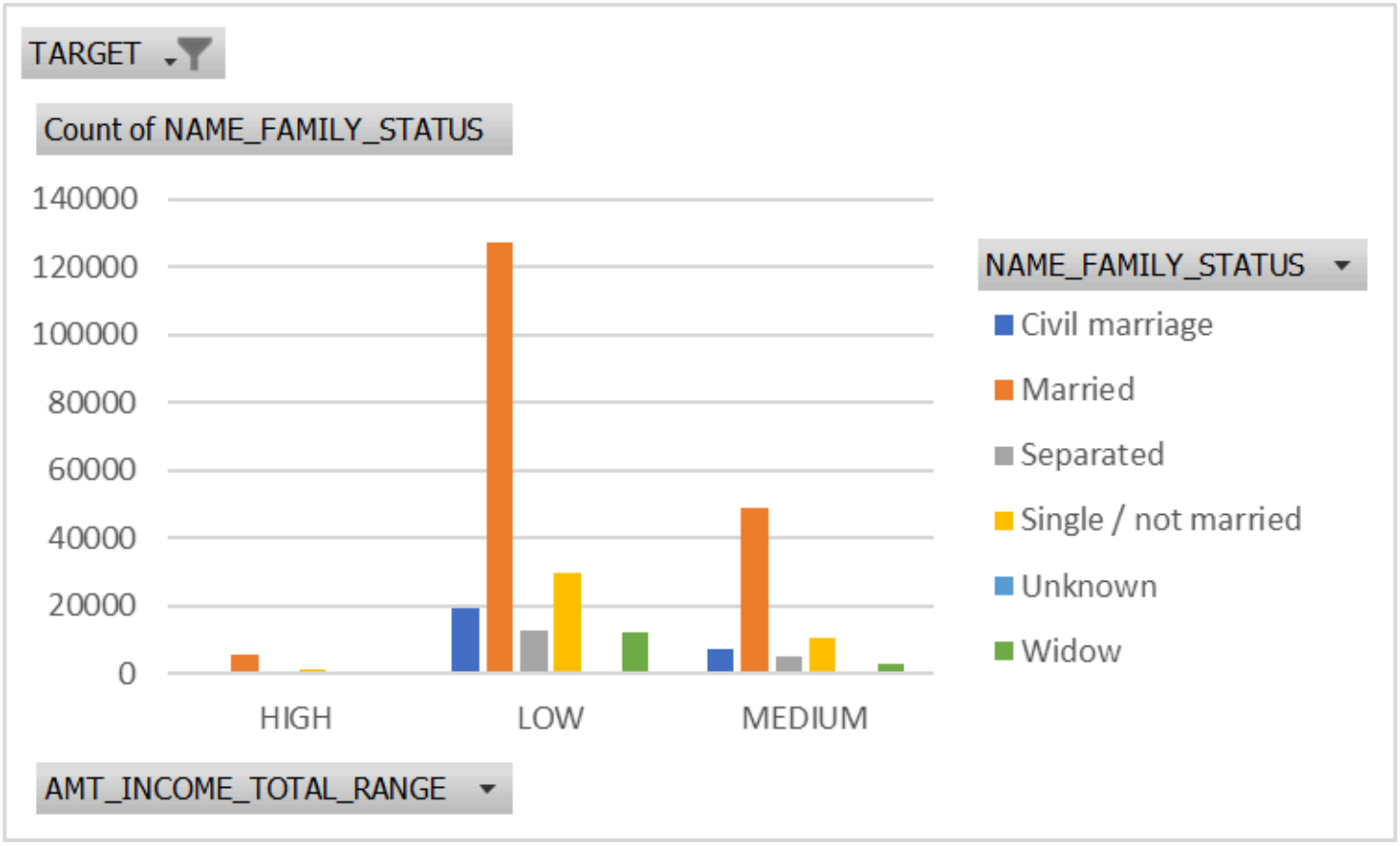
XVI. **Occupation Type and Non-Defaulters**: - 'Laborers' have the highest count of non-defaulters.



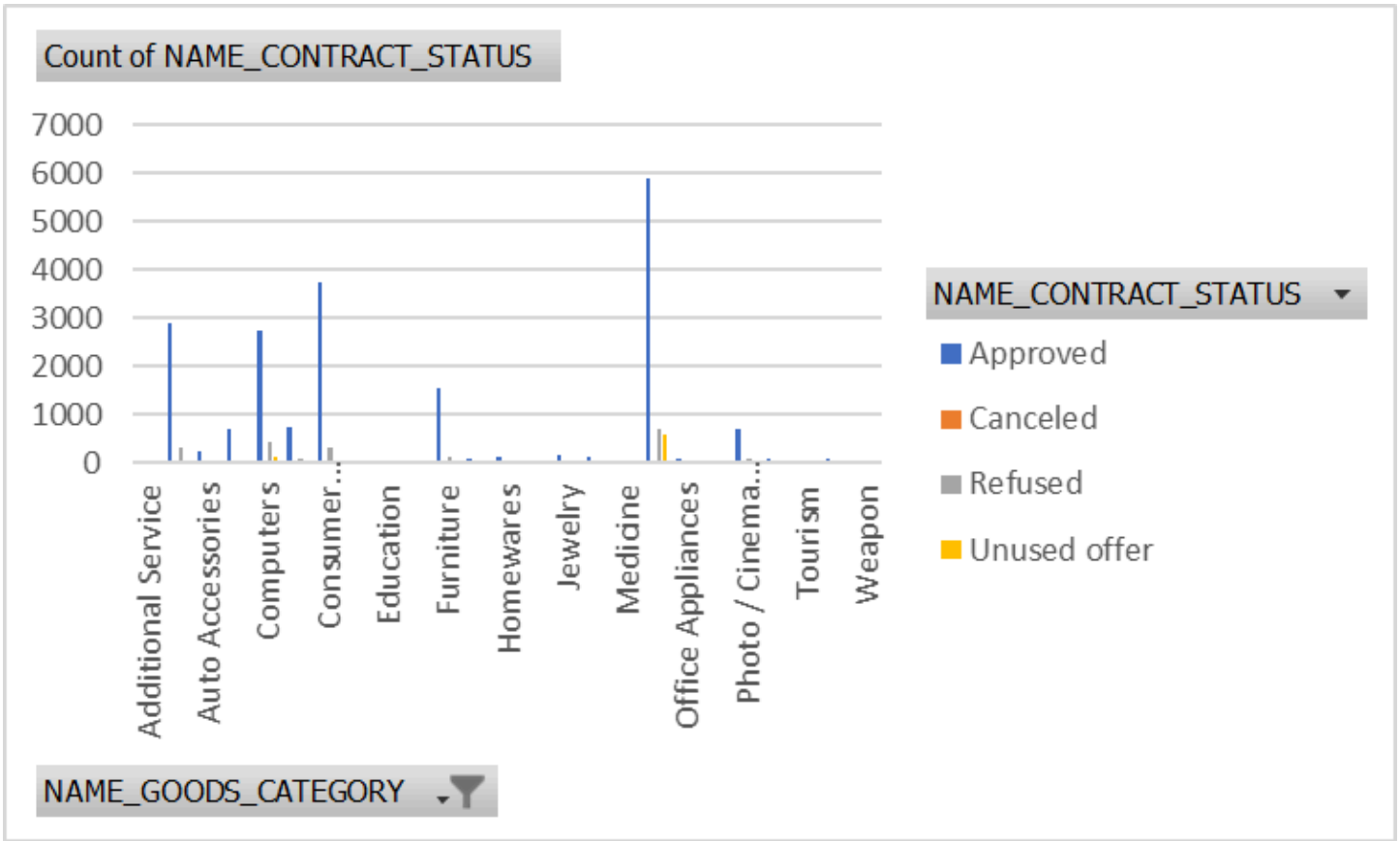
XVII. **Gender and Income Group**: - Females in the low-income group have the highest number of clients with no payment issues and also the highest number of clients with payment issues.

XVIII. **Credit Amount and Education Level**: - Clients with low credit amounts and 'Secondary/Secondary Special' education have the highest count of no payment issues. - Clients with medium credit amounts and 'Secondary/Secondary Special' education have the highest count of payment issues.

XIX. **Total Income and Family Status**: - Married clients in the low-income range have the highest count of both no payment issues and payment issues.



XX. **Loan Approval**: - Repairs work has the highest count of approved loans based on contract status.



Analysis

1. **Why is the target variable so important?**
   - The target variable indicates whether the client had payment issues (1) or did not have payment issues (0). It is crucial because it helps the bank decide on interest rate adjustments for loans. With 92% of clients having no payment issues, the bank's credit score is strong, indicating few or no non-performing accounts.
2. **Why is the proportion of female clients higher than male clients?**
   - In countries like India, there are government laws supporting women who want to start their own businesses or services, offering them loans at lower interest rates. Additionally, some people use the names of their retired or household mothers or wives to get concessions on interest rates for home loans.
3. **Why should the bank prefer clients from other housing types, even though House/Apartment clients have the highest proportion of non-defaulters?**
   - People in other housing categories, such as Municipal Apartments, Co-op Apartments, Rented Apartments, or living with parents, are often seeking their own homes. The trend in India is moving away from joint family systems, with future generations preferring their own smaller homes.
4. **Why should the bank opt for working-class clients over state-government class clients, despite the latter's benefits and regular salaries?**
   - State government employees receive significant housing allowances and may even get apartments for the duration of their employment. Working-class clients, however, do not enjoy such benefits and are more likely to seek home loans to purchase their own houses.
5. **Why should the bank not prioritize approving loans for 'Laborers' occupation type clients, despite their high count of non-defaulters?**
   - Laborers typically take personal loans for purposes like marriage or house repairs, which are smaller in amount and come with lower interest rates. These loans yield less profit for the bank compared to larger loans like home or car loans.
6. **Why do females in the low-income group have the lowest count of defaulters?**
   - Females in this group usually take small loans to start their own businesses or services like catering or parlors. They often benefit from government schemes aimed at supporting such endeavors, leading to a lower default rate.

Conclusion

In conclusion, the analysis reveals the following insights:

1. **Default Rates**:
   - Approximately 8% of clients are defaulters (target = 1).
   - Around 92% of clients are non-defaulters (target = 0).
2. **Gender-Based Lending**:
   - The bank generally lends more to female clients, as their count in the defaulter’s list is lower compared to male clients.

    - The bank can consider increasing loans to male clients if their creditworthiness is satisfactory.
3. **Employment Class**:
    - Clients from the working class tend to repay their loans on time, followed by clients who are commercial associates.
4. **Education Status**:
    - Clients with secondary/higher secondary education or more tend to repay their loans on time.
    - The bank can prefer lending to clients with such education levels.
5. **Age Group**:
    - Clients aged 31-40 have the highest repayment rate, followed by those aged 41-60.
6. **Credit Amount Range**:
    - Clients with a low credit amount range tend to repay their loans on time compared to those with high and medium credit ranges.
7. **Housing Type**:
    - Clients living with their parents tend to repay their loans more quickly than those with other housing types.
    - The bank can prioritize lending to clients who live with their parents.
8. **Loan Purpose**:
    - Clients taking loans for purchasing a new home (home loans), a new car (car loans), or those who are state servants tend to repay their loans on time.
    - The bank should prefer clients with these backgrounds.
9. **Repairs Purpose**:
    - The bank should be cautious when lending to clients who take loans for repairs, as they have a high count of defaulters.

These insights can guide the bank in refining its lending strategies to minimize defaults and optimize loan repayments.

Analyzing the Impact of Car Features on Price and Profitability

Description

The automotive industry is rapidly evolving with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. Manufacturers need to understand the factors that drive consumer demand to optimize pricing and product development decisions. This project aims to analyze the relationship between car features, market categories, pricing, and profitability using a dataset of car models and their specifications.

The Problem

The key business problem is how a car manufacturer can optimize pricing and product development decisions to maximize profitability while satisfying consumer demand. Specifically, the analysis aims to:

- Identify which car features and market categories are most popular and profitable.
- Determine the relationship between car features and pricing.
- Develop strategies to balance consumer demand with profitability.

Design

The project is structured into several stages:

1. **Data Cleaning and Preprocessing**
   - **Handling Missing Values**: Identify and impute or remove missing data.
   - **Removing Duplicates**: Ensure each car model is unique.
   - **Converting Data Types**: Ensure numerical and categorical data are correctly formatted.
   - **Feature Engineering**: Create new variables or modify existing ones for better analysis.
2. **Descriptive Statistics**
   - Summarize numerical data (e.g., mean, median, standard deviation) and categorical data (e.g., frequency counts).
3. **Pattern Recognition**
   - Use visualizations (e.g., heatmaps, correlation matrices) to identify relationships between variables.
4. **Segmentation Analysis**
   - Group cars into segments based on key characteristics (e.g., market category, engine type).
5. **Visualizations**

- Create charts and graphs to highlight trends, patterns, and outliers.

6. **Regression Analysis**
   - Perform multiple regression analysis to identify the most influential variables on car price.
7. **Interactive Dashboard**
   - Develop an interactive Excel dashboard with filters and slicers for dynamic data exploration.

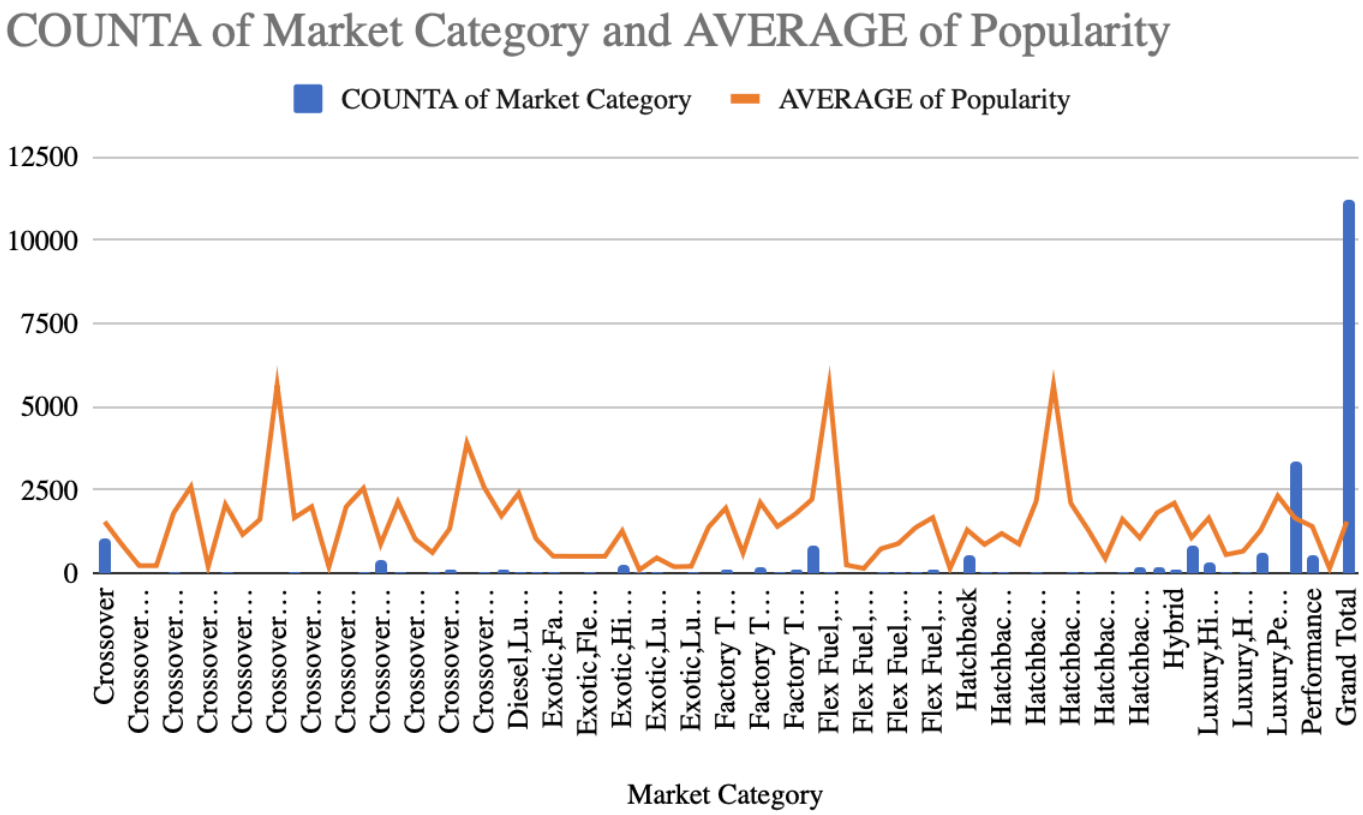
Findings

I. Popularity Across Market Categories

- **Task 1.A**: Pivot table showing the number of car models in each market category and their popularity scores.

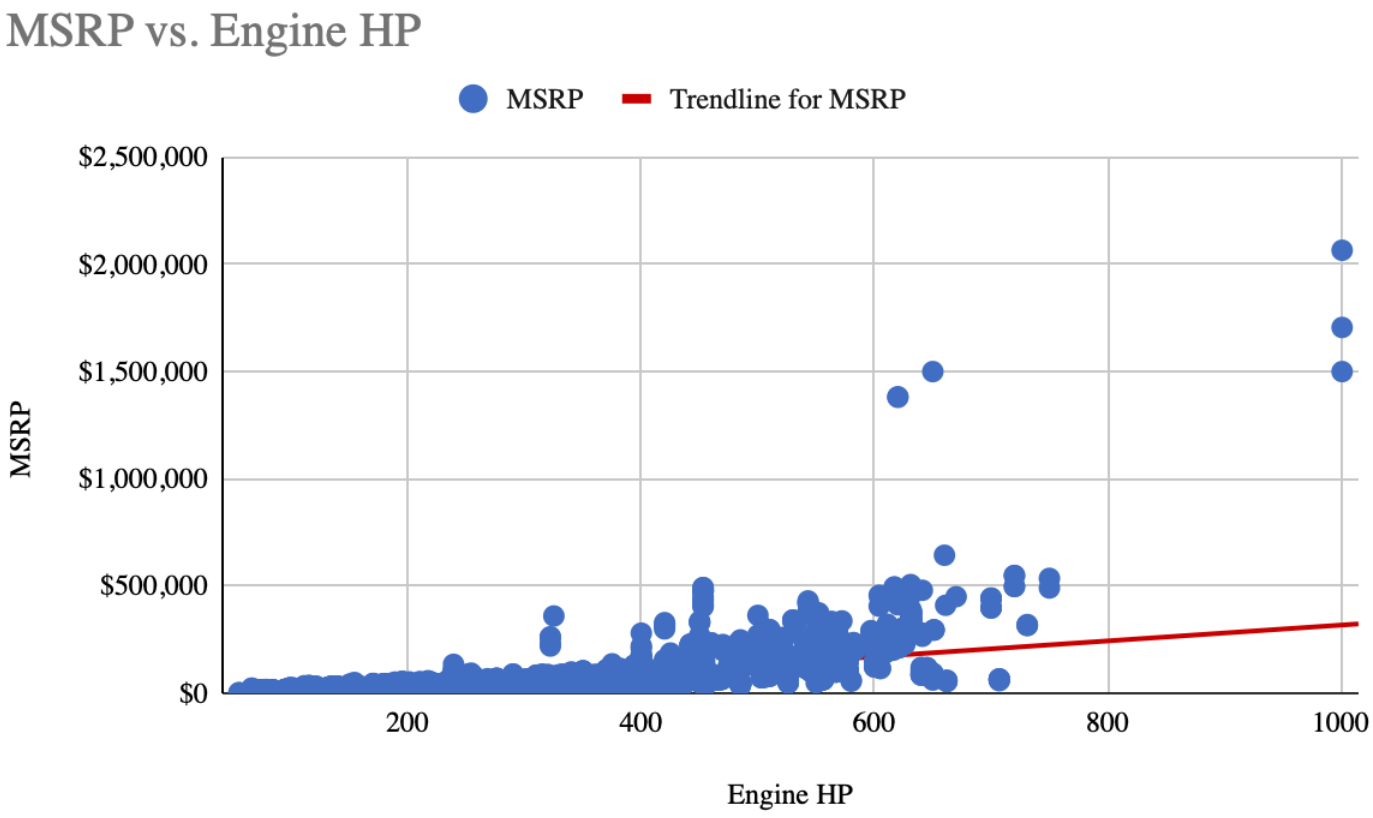
MARKET CATEGORY	COUNT OF MARKET CATEGORY	AVERAGE OF POPULARITY
Crossover	1075	1556.17
Crossover, Diesel	7	873.00
Crossover, Exotic, Luxury, High-Performance	1	238.00
Crossover, Exotic, Luxury, Performance	1	238.00
Crossover, Factory Tuner, Luxury, High-Performance	26	1823.46
Crossover, Factory Tuner,Luxury,Performance	5	2607.40
Crossover,Factory Tuner,Performance	4	210.00
Crossover,Flex Fuel	64	2073.75
Crossover,Flex Fuel,Luxury	10	1173.20

- **Task 1.B**: Combo chart visualizing the relationship between market category and popularity.



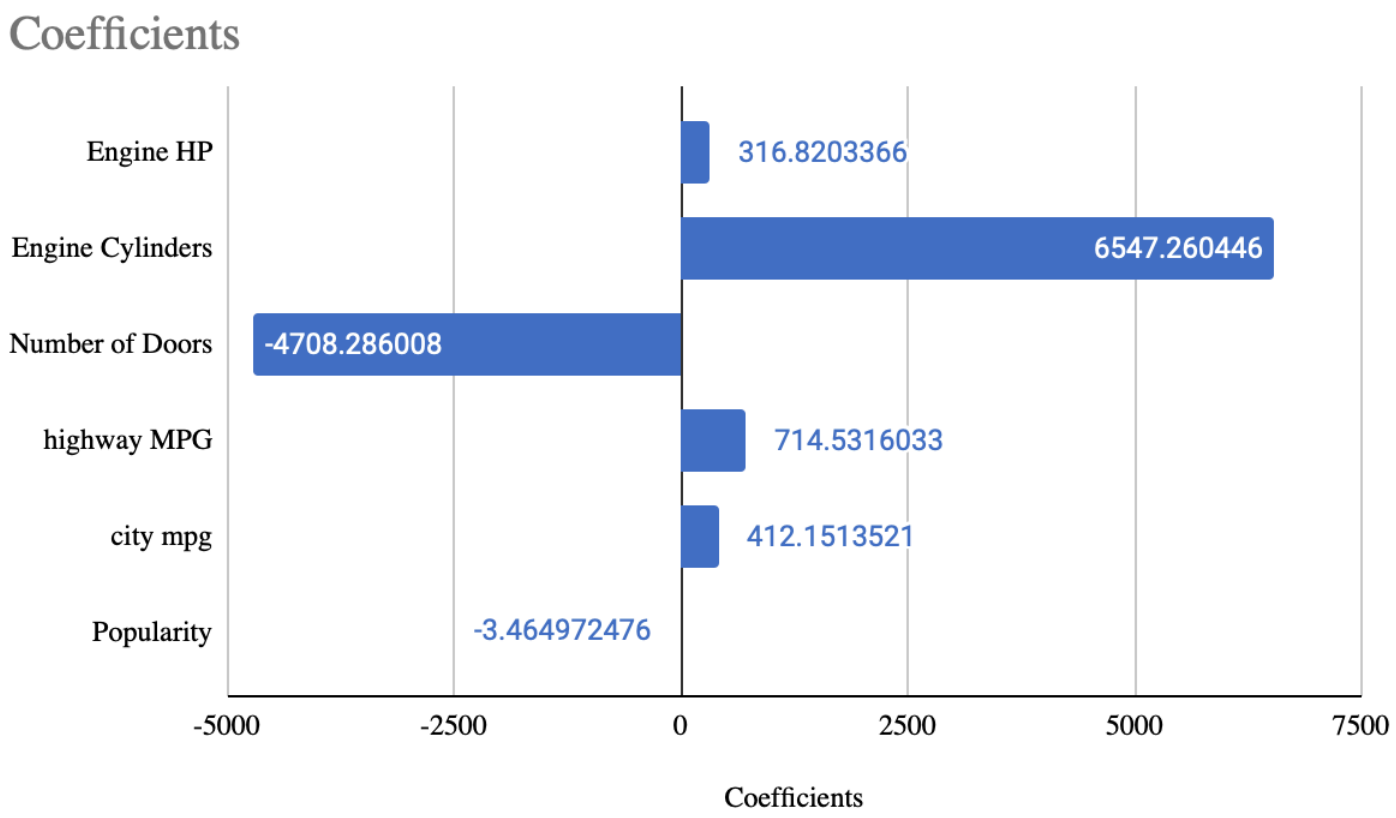
II. Engine Power vs. Price

- **Task 2**: Scatter chart plotting engine power against price with a trendline.



III. Important Car Features for Price Determination

- **Task 3**: Regression analysis to identify variables that influence car price and a bar chart showing their coefficients.

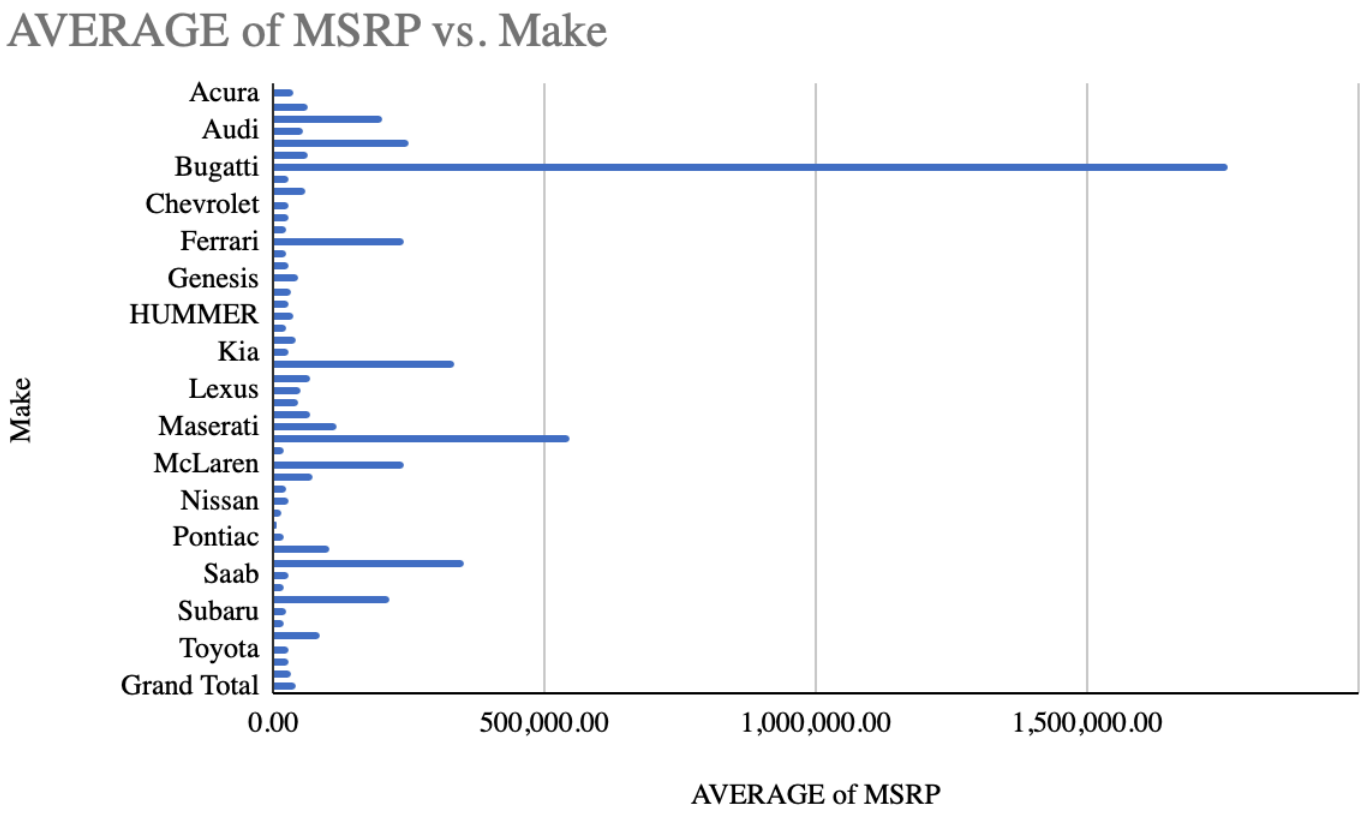


IV. Average Price by Manufacturer

- **Task 4.A**: Pivot table showing average car prices for each manufacturer.

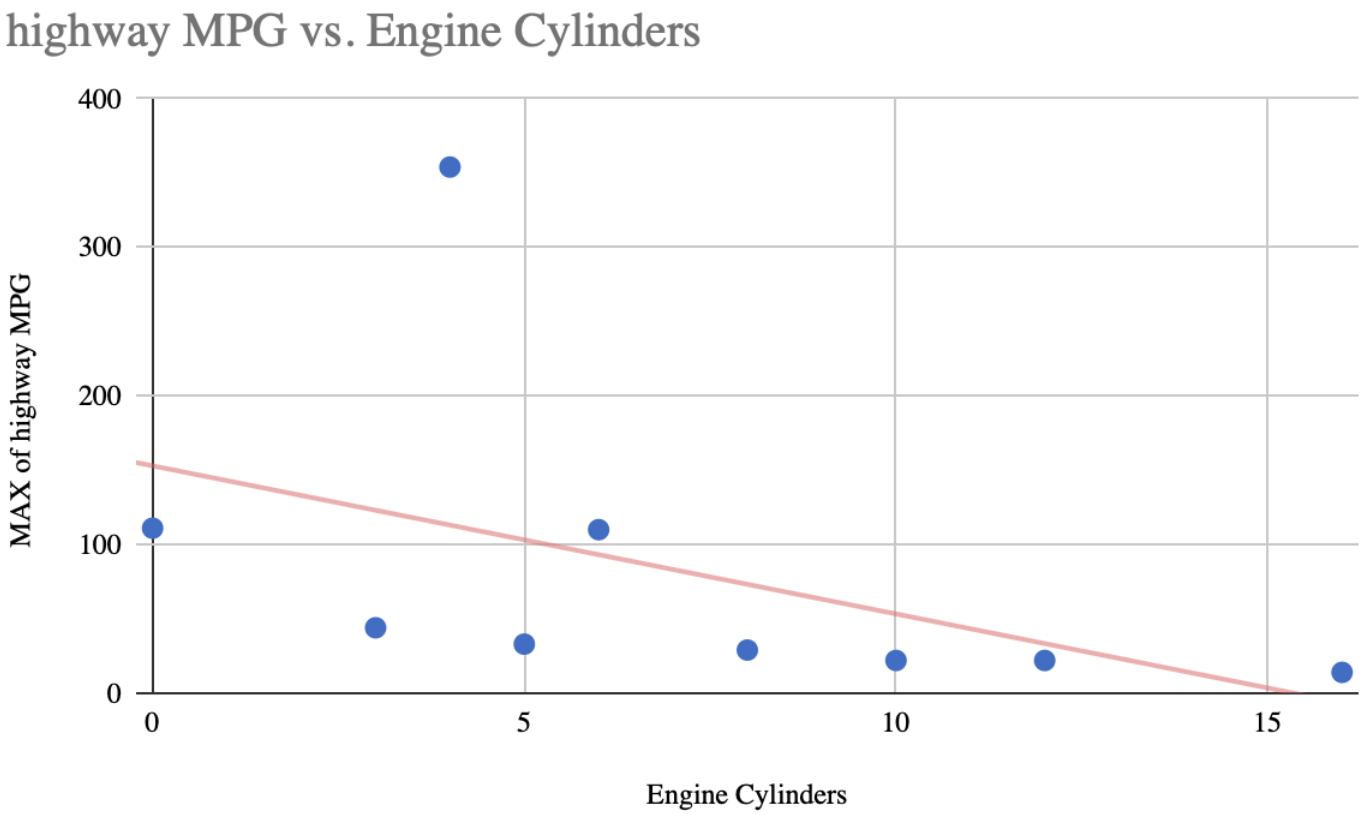
Make	
Acura	35,087.49
Alfa Romeo	61,600.00
Aston Martin	198,123.46
Audi	54,574.12
Bentley	247,169.32
BMW	62,162.56
Bugatti	1,757,223.67
Buick	29,034.19
Cadillac	56,368.27

- **Task 4.B**: Bar chart visualizing the relationship between manufacturer and average price.



V. Fuel Efficiency and Engine Cylinders

- **Task 5.A**: Scatter plot with number of cylinders on the x-axis and highway MPG on the y-axis, with a trendline.



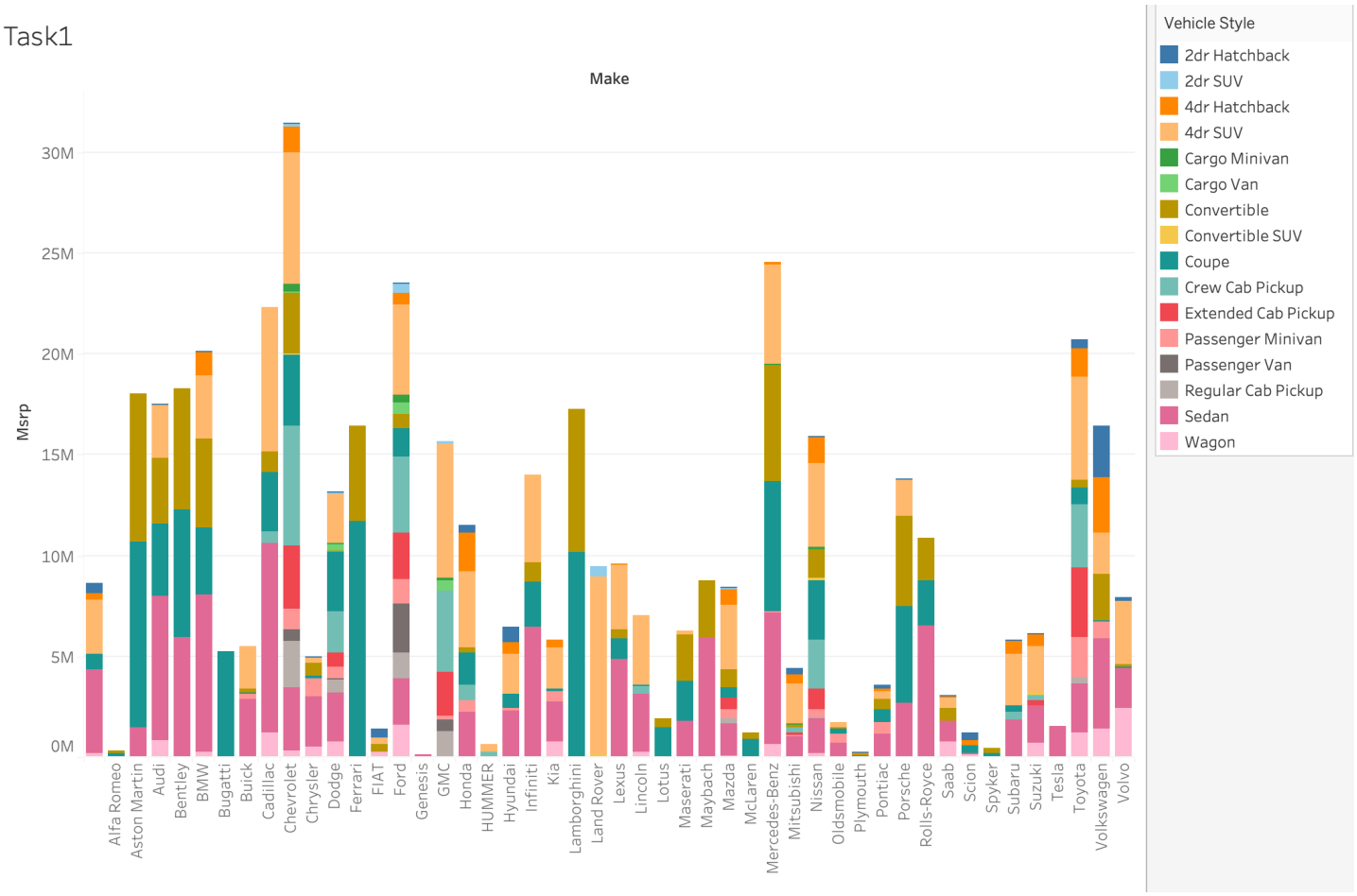
- **Task 5.B**: Calculation of the correlation coefficient between cylinders and highway MPG.

-0.5962460188

Building the Dashboard

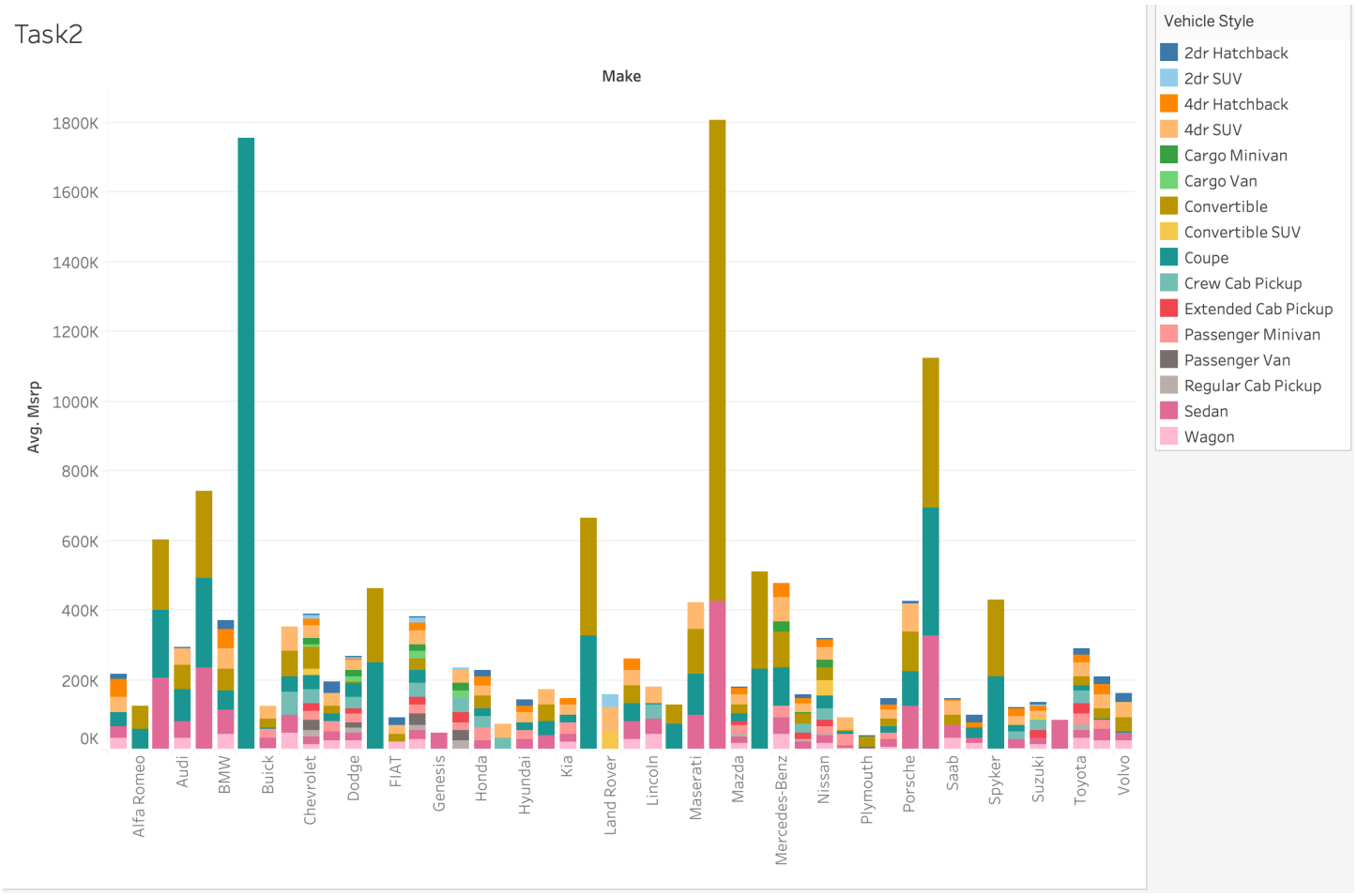
I. Distribution of Car Prices by Brand and Body Style

- **Task 1**: Use a stacked column chart to show the distribution of car prices by brand and body style. Utilize filters and slicers for interactivity.
  - **Insight**: Helps understand price distribution and identify trends based on brand and body style, aiding in targeted marketing and pricing strategies.



II. Car Brands' Average MSRPs by Body Style

- **Task 2**: Create a clustered column chart to compare the average MSRPs across different car brands and body styles.
  - **Insight**: Identifies brands with the highest and lowest average MSRPs, informing brand positioning and product development strategies.

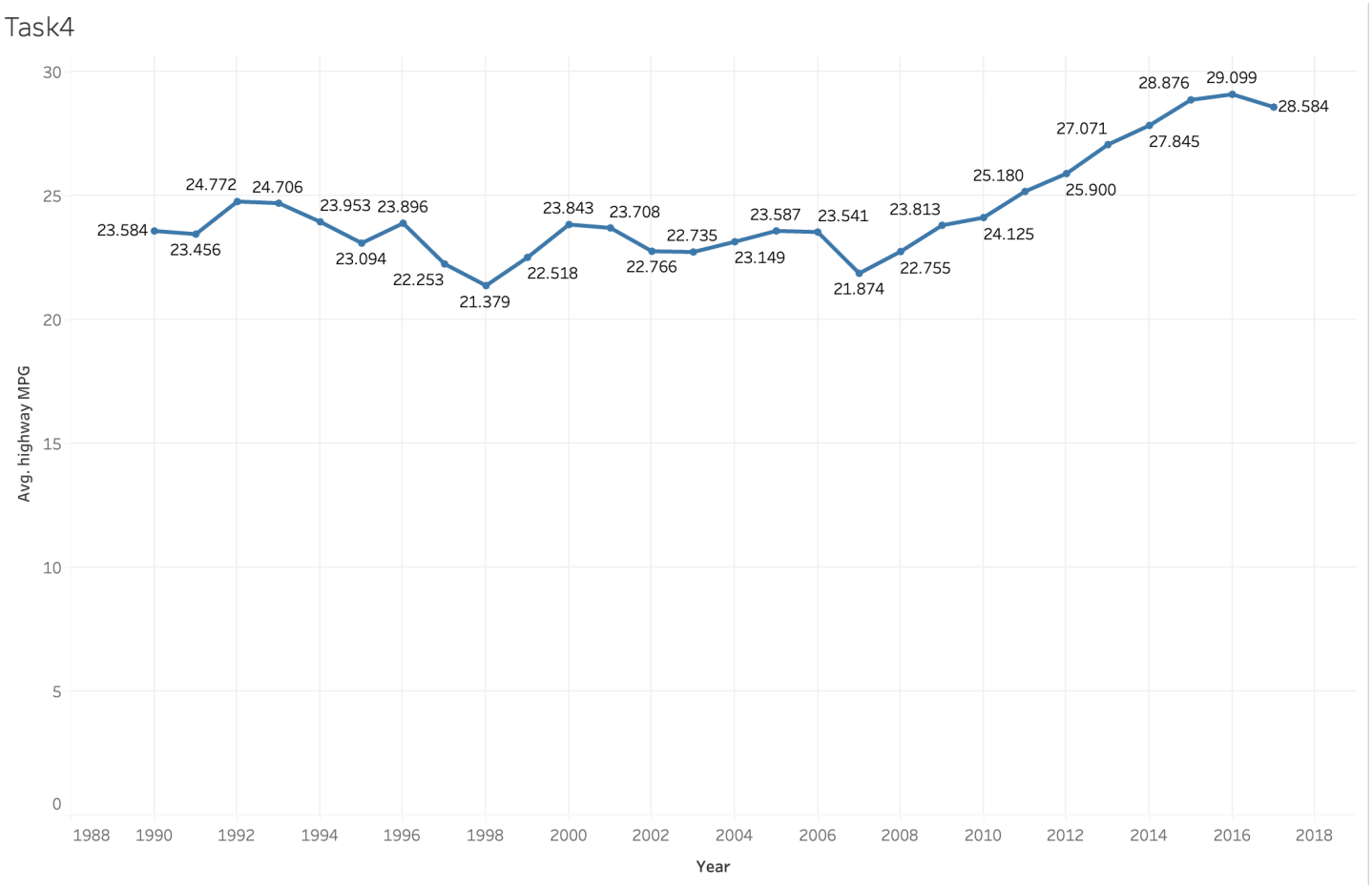


III. Effect of Transmission Type on MSRP by Body Style

- **Task 3**: Use a scatter plot to visualize the relationship between MSRP and transmission type, with different symbols for each body style.
  - **Insight**: Reveals how transmission type influences MSRP and varies by body style, guiding feature prioritization in new models.

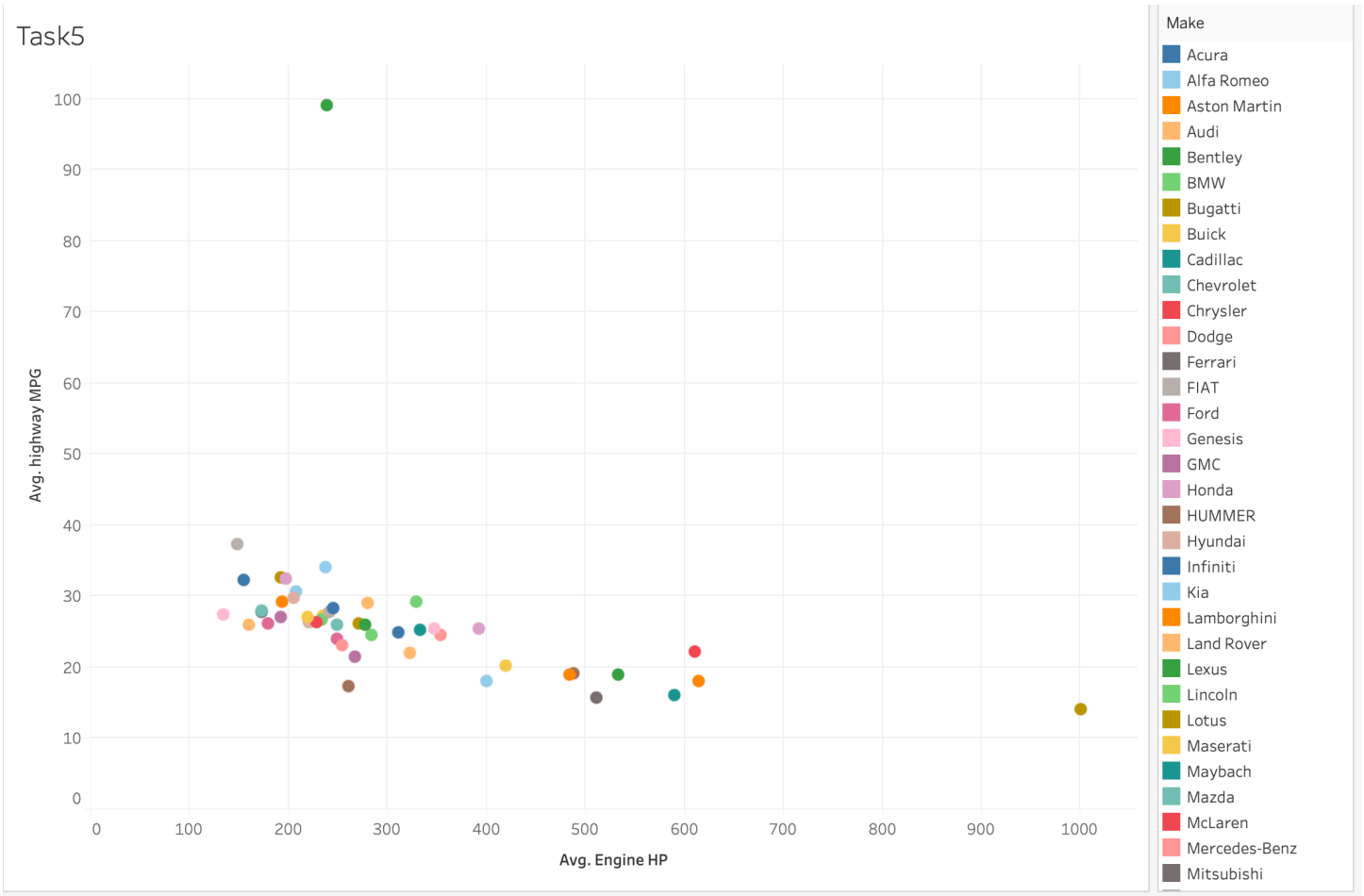
IV. Fuel Efficiency Across Body Styles and Model Years

- **Task 4**: Create a line chart to show the trend of fuel efficiency (MPG) over time for each body style.
  - **Insight**: Trends in fuel efficiency improvements across body styles and model years, informing future product development focused on sustainability.



V. Horsepower, MPG, and Price Across Brands

- **Task 5**: Use a bubble chart to visualize the relationship between horsepower, MPG, and price across different car brands.
  - **Insight**: Identifies the trade-offs between performance (horsepower), fuel efficiency (MPG), and price across brands, aiding in competitive analysis and product positioning.



Analysis

1. **Popularity Analysis**
   - The pivot table and combo chart show the most and least popular market categories.
   - Insights into which market categories are preferred by consumers.
2. **Engine Power vs. Price Relationship**
   - The scatter plot and trendline indicate a positive correlation between engine power and price.
   - Higher horsepower generally leads to higher prices.
3. **Regression Analysis for Price Determination**
   - Key features affecting car price include engine power, market category, and fuel efficiency.
   - The bar chart visualizes the relative importance of each feature.
4. **Manufacturer and Average Price Relationship**
   - The pivot table and bar chart reveal manufacturers with the highest and lowest average car prices.
   - Helps identify manufacturers targeting premium vs. budget segments.
5. **Fuel Efficiency and Engine Cylinders Relationship**
   - The scatter plot and correlation coefficient show an inverse relationship between the number of cylinders and highway MPG.
   - More cylinders generally lead to lower fuel efficiency.

Conclusion

The analysis provides valuable insights into how car features influence pricing and profitability. Key findings include the importance of engine power, market category, and fuel efficiency in determining car prices. Popular market categories and manufacturers' average pricing strategies were also identified. The interactive dashboard enables stakeholders to explore these insights dynamically, aiding in optimizing pricing and product development decisions to maximize profitability while meeting consumer demand.

ABC Call Volume Trend Analysis

Description

A Customer Experience (CX) team plays a vital role in analyzing customer feedback and data to provide valuable insights across an organization. These teams typically manage various responsibilities, including:

- **CX Programs**: Developing and implementing strategies to enhance the overall customer experience.
- **Digital Customer Experience**: Optimizing digital interactions and touchpoints.
- **Design and Processes**: Crafting and refining processes and designs to improve customer interactions.
- **Internal Communications**: Facilitating effective communication within the organization about customer experience initiatives.
- **Voice of the Customer (VoC)**: Capturing and acting on customer feedback to drive improvements.
- **User Experiences**: Enhancing the overall user journey across all touchpoints.
- **Customer Experience Management**: Overseeing the entire customer experience strategy.
- **Journey Mapping**: Visualizing and analyzing the customer journey to identify improvement areas.
- **Nurturing Customer Interactions**: Cultivating positive and meaningful interactions with customers.
- **Customer Success**: Ensuring customers achieve their desired outcomes with your product or service.
- **Customer Support**: Providing assistance and resolving issues to enhance customer satisfaction.
- **Handling Customer Data**: Managing and analyzing customer data to gain insights and drive decisions.
- **Learning About the Customer Journey**: Continuously understanding and improving the customer journey.

Today, AI-driven tools are revolutionizing customer experience management, including:

- **Interactive Voice Response (IVR)**: Automating customer interactions through voice prompts and responses.
- **Robotic Process Automation (RPA)**: Streamlining repetitive tasks to improve efficiency.
- **Predictive Analytics**: Anticipating customer needs and behaviors to proactively address issues.
- **Intelligent Routing**: Directing customer inquiries to the most appropriate support channels or representatives.

In the realm of customer service, there are abundant opportunities for various roles, including:

- **Email Support**: Handling customer inquiries and issues via email.
- **Inbound Support**: Managing incoming calls from current or potential customers to resolve their concerns and enhance their experience.
- **Outbound Support**: Proactively reaching out to customers for follow-ups, surveys, or promotional purposes.
- **Social Media Support**: Engaging with customers and addressing their needs through social media platforms.

Inbound customer support focuses on managing incoming voice calls from existing or prospective customers. This approach aims to attract, engage, and delight customers by solving their problems and ensuring they achieve success with your product or service. By providing exceptional inbound support, you can transform customers into loyal advocates and drive business growth.

The Problem

To address the problem of calculating the average call duration, visualizing call volume, and proposing a manpower plan to reduce the abandonment rate, we need to break down the problem into several steps. Let's work through these steps methodically:

1. Average Call Duration Calculation

- **Calculate Average Call Duration:**
  - For each time bucket, calculate the average duration of incoming calls.
  - Formula:

$$\text{Average Call Duration} = \frac{\text{Total Duration of Calls}}{\text{Number of Calls}}$$

2. Visualizing Call Volume

- **Create Charts/Graphs for Call Volume:**
  - Use a time bucket interval (e.g., 1-2 PM, 2-3 PM, etc.).
  - Plot the number of calls received during each time bucket.
  - Use bar charts or line graphs to represent the number of calls versus time buckets.

3. Manpower Planning to Reduce Abandon Rate

Assumptions:

- Current abandon rate: 30%
- Target abandon rate: 10%
- Agent work details:
  - Work for 6 days a week

    - 4 unplanned leaves per month
    - 9 hours of total work, 1.5 hours for breaks, 7.5 hours net working time
    - 60% of 7.5 hours (i.e., 4.5 hours) is spent on calls
- Total days in a month: 30

Step-by-Step Calculation:

I. Determine Call Volume During Day and Night:

- Assume a total of 100 calls are received from 9 AM to 9 PM.
- Each night (9 PM to 9 AM), there are 30 calls distributed.

II. Calculate Required Number of Agents During Each Time Bucket:

a. **Calculate Calls per Hour:**
   - Distribute the total calls equally across the hours in each time bucket.
b. **Calculate Calls per Time Bucket:**
   - Let's say the number of calls per bucket is Callsbucket.
c. **Calculate Required Number of Agents:**
   - Given the target abandonment rate is 10%, 90% of the calls need to be answered.
   - Total number of calls that need to be answered during each time bucket is Callsbucket×0.90.
   - Required number of agents Agentsbucket can be calculated as:

$$\text{Agents}_{\text{bucket}} = \frac{\text{Calls}_{\text{bucket}} \times 0.90}{4.5}$$

III. Adjust for Agent Availability:

- Factor in that agents work 6 days a week and have 4 unplanned leaves per month. Adjust calculations accordingly.

IV. Manpower Plan for 24-Hour Coverage:

- Divide the 30 calls during the night into buckets, e.g., 9 PM-10 PM, 10 PM-11 PM, etc.
- Apply the same calculation method to determine the number of agents required during night hours.
- Ensure sufficient coverage to meet the same 10% abandonment rate target.

Example Calculation

Let's assume you have the following data for call volume in time buckets:

- **Time Bucket**: 1-2 PM
    - Number of Calls: 150
- **Average Call Duration Calculation:**
    - Total Duration of Calls: 10,000 minutes.
    - Average Call Duration: 10,000/150≈66.67 minutes.
- **Required Number of Agents Calculation:**

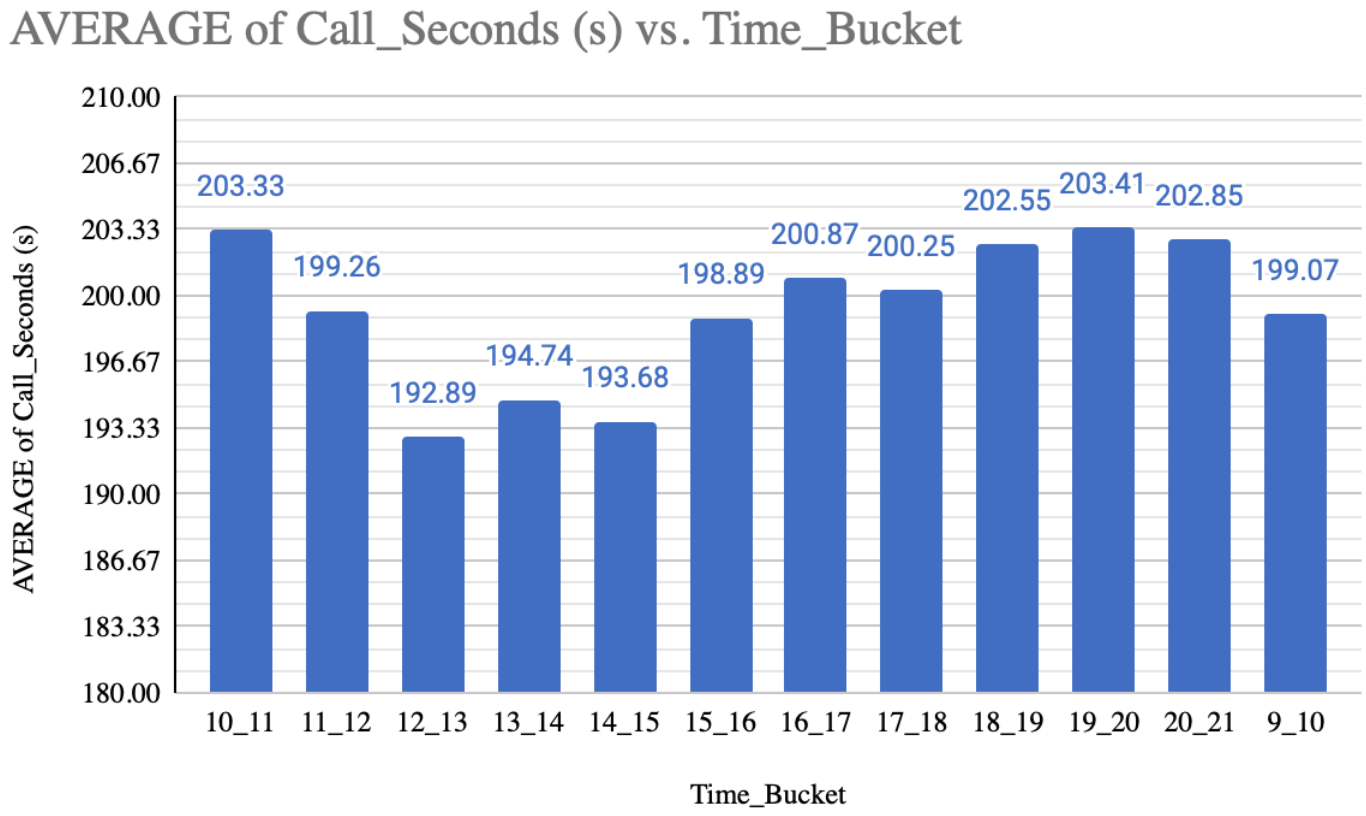
- Calls to be answered = 150 calls × 0.90 = 135 calls.
- Number of agents required = 135/4.5≈30 agents.

- **Adjusting for Availability:**
  - Assuming agents work 26 days (after accounting for unplanned leaves):
  - Effective working hours per agent per month = 26 days × 7.5 hours/day = 195 hours
  - If each agent handles 4.5 hours of calls per day, calculate the total required coverage and ensure staffing levels are adequate.

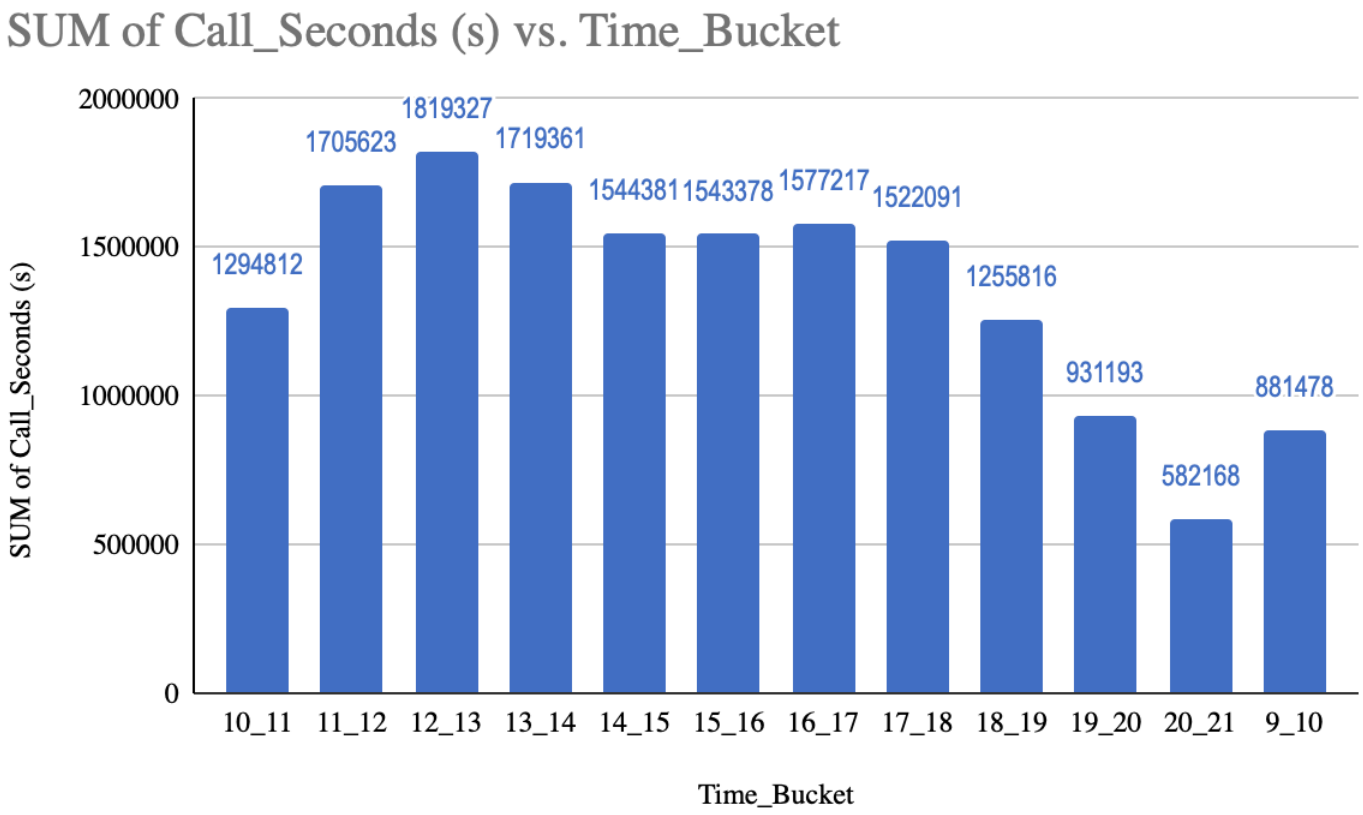
By following this approach, you can create a detailed manpower plan to meet the desired abandonment rate and ensure sufficient coverage during both day and night shifts.

Findings

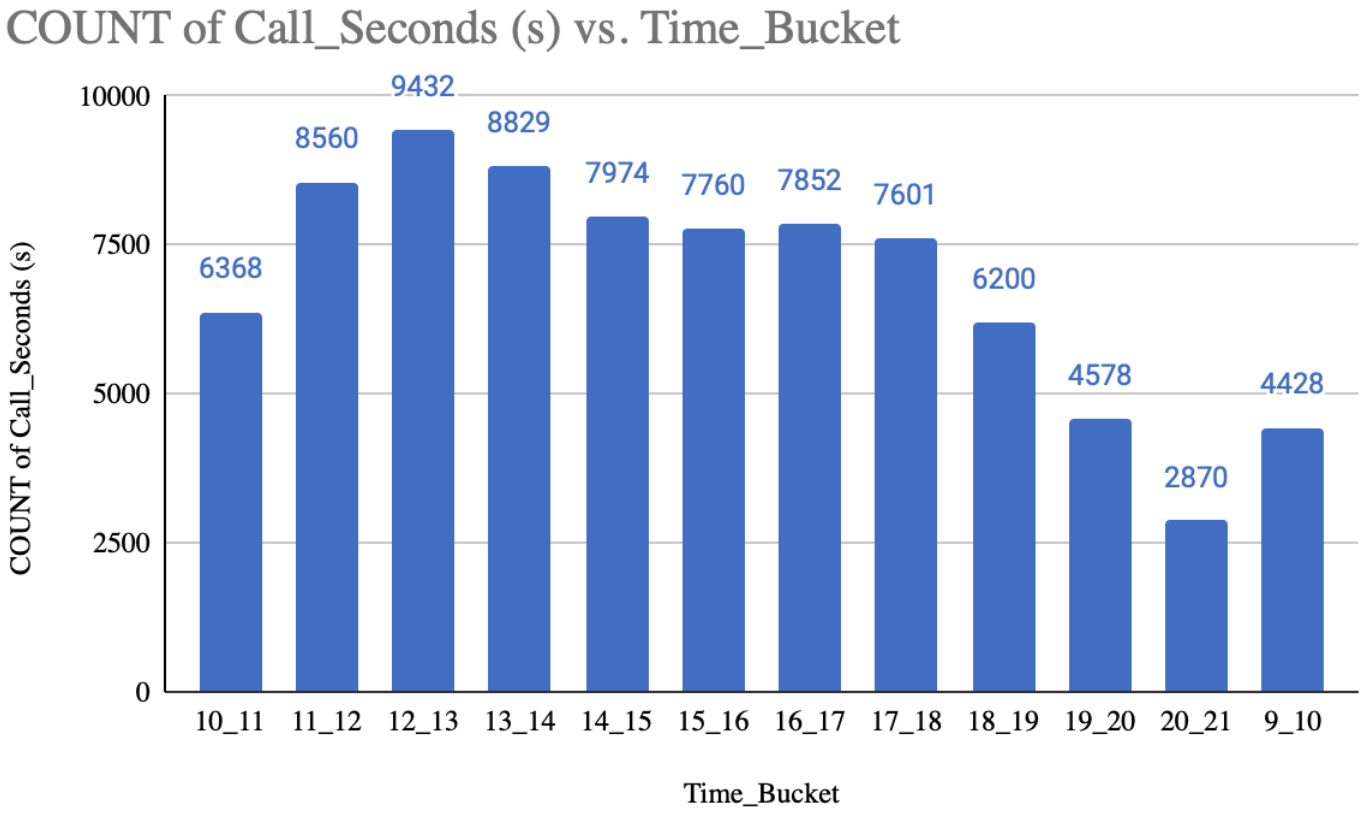
Finding I: The time bucket from 7 PM to 8 PM (19:00-20:00) has the highest average call duration of 203.4 seconds.



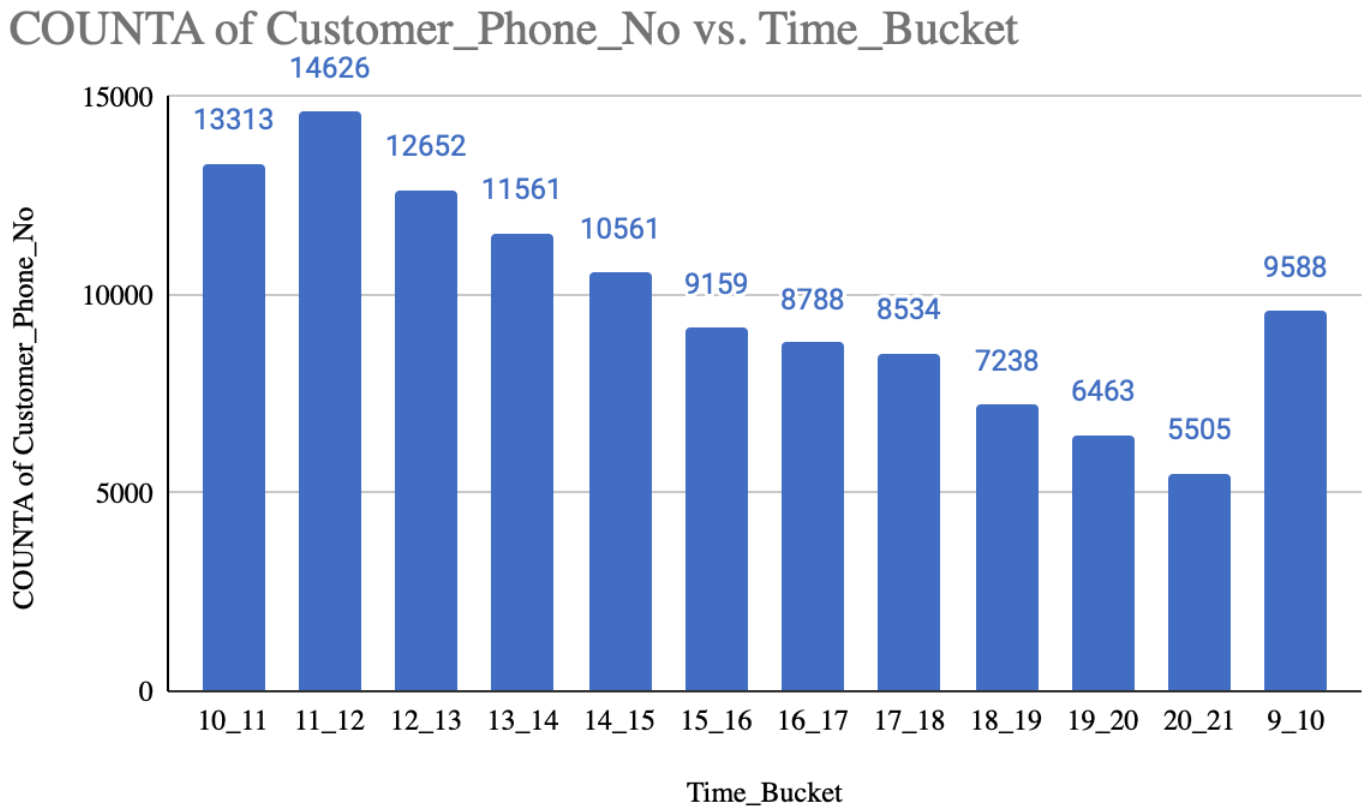
Finding II: The time bucket from 12 PM to 1 PM (12:00-13:00) had the highest total number of calls answered, totaling 1,819,327 calls.



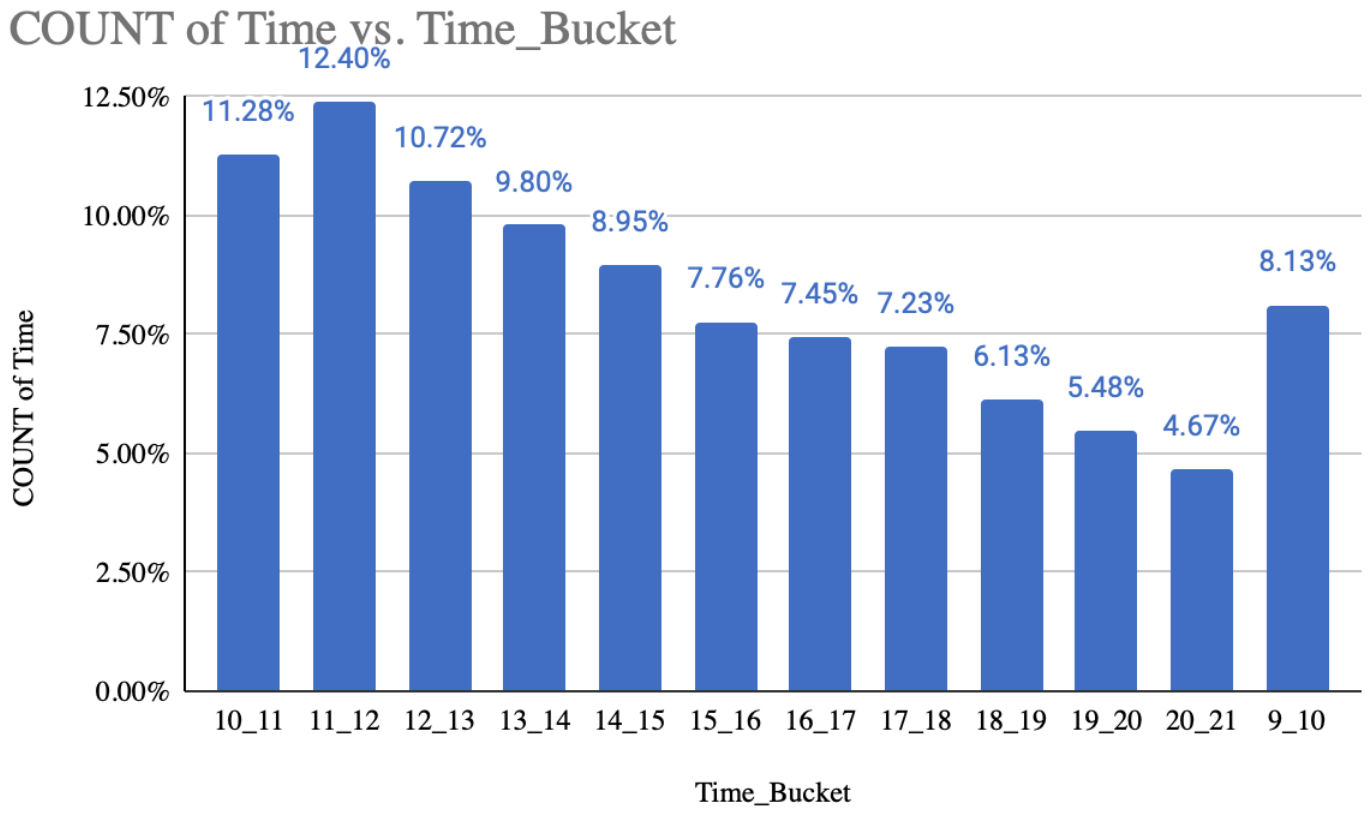
Finding III: The same time bucket (12 PM to 1 PM) had the highest count of calls answered, which was 9,432.



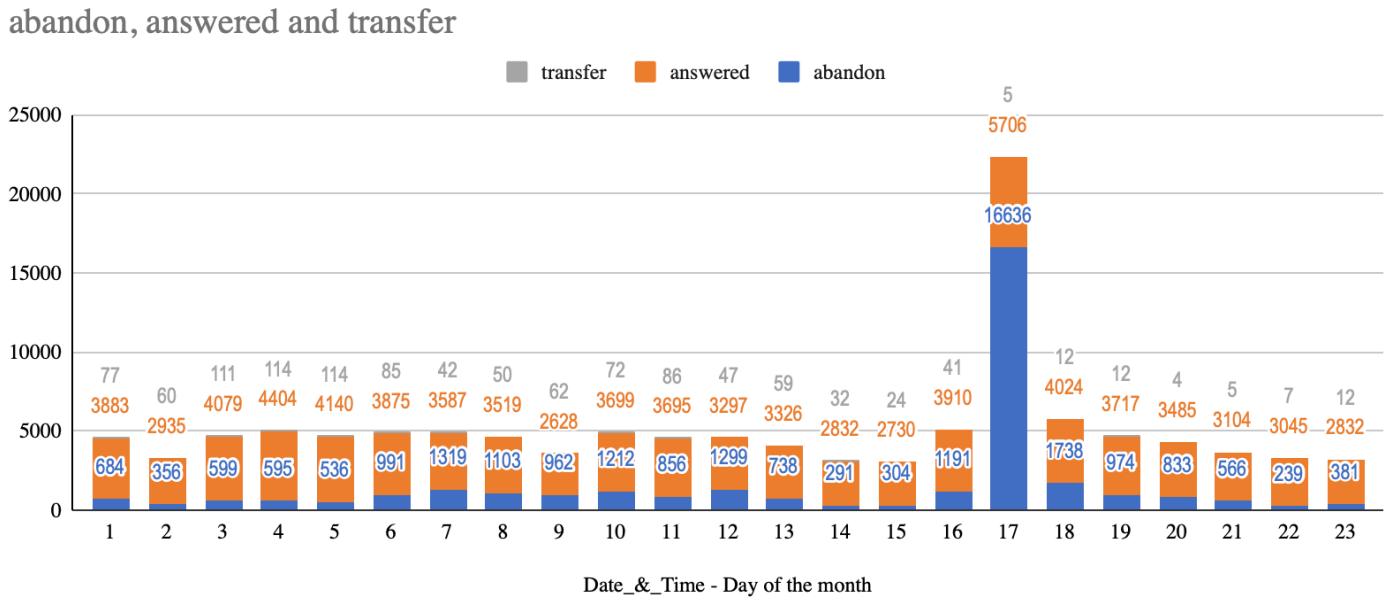
Finding IV: The time bucket from 11 AM to 12 PM (11:00-12:00) had the highest count of total incoming calls, totaling 14,626 calls.



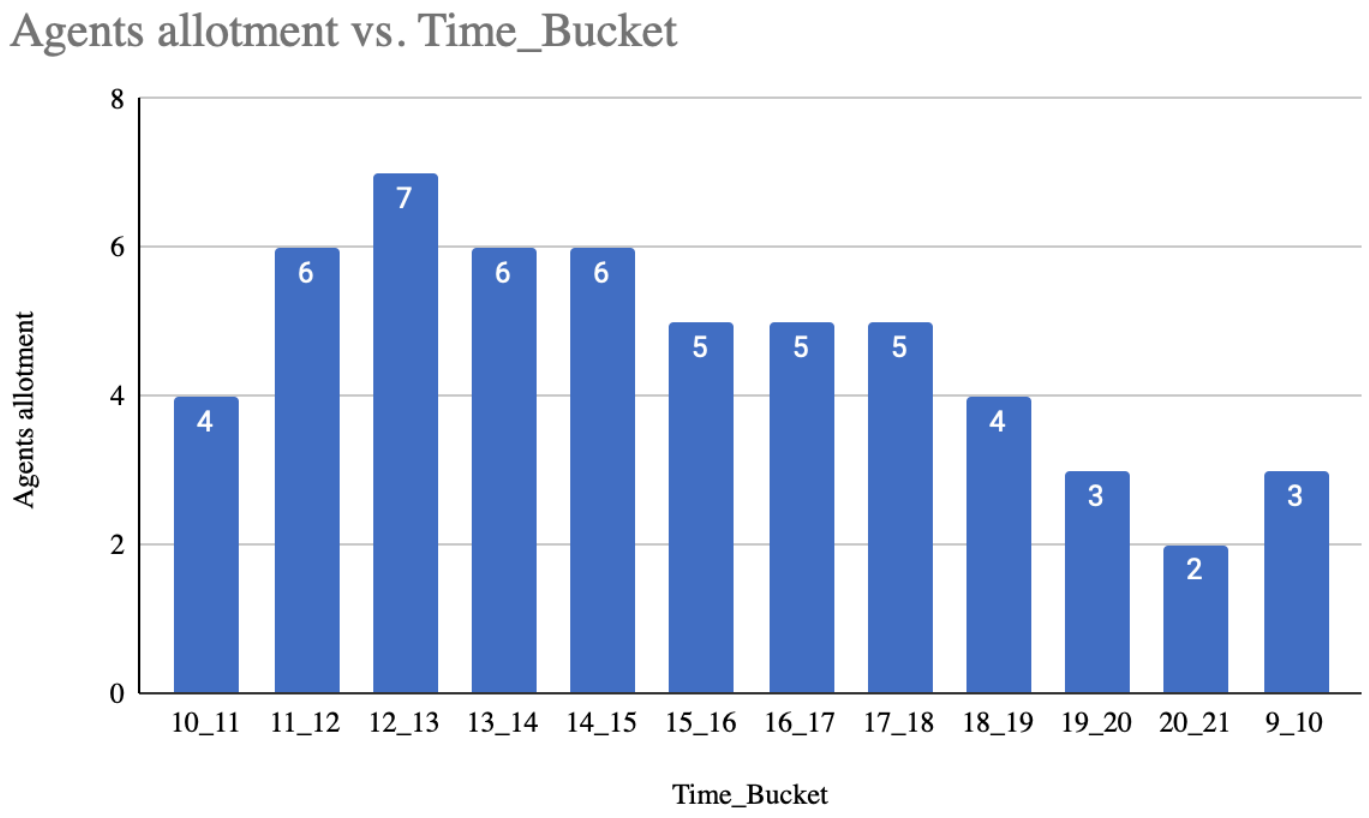
Finding V: The time bucket from 11 AM to 12 PM (11:00-12:00) had the highest share of incoming calls, accounting for 12.40%.



Finding VI: The current abandon rate is approximately 30%.



Finding VII: Desired distribution of night calls to maintain an abandonment rate of 10%.



Proposed Manpower Plan:

Daytime Coverage (9 AM to 9 PM)

1. **Determine Peak Call Times:**
   - Highest call volume and share: 11 AM to 12 PM and 12 PM to 1 PM.
2. **Manpower Calculation:**
   - Use the same methodology as described previously to calculate the required number of agents for each time bucket, focusing on high-volume periods (11 AM to 1 PM) to ensure adequate coverage.

Nighttime Coverage (9 PM to 9 AM)

1. **Distribution of Night Calls:**
   - Total night calls: 30 calls per 100 calls made during daytime.
   - Ensure distribution aligns with call patterns observed in previous time buckets.
2. **Manpower Adjustment:**
   - Utilize agents from high-volume periods during the day (e.g., 7 PM to 8 PM) for nighttime coverage.
   - Agents from time buckets 9 AM to 10 AM and 10 AM to 11 AM can be reassigned to cover time buckets like 7 PM to 8 PM and 8 PM to 9 PM during the night.
   - Similarly, agents working from 1 PM to 2 PM, 2 PM to 3 PM, etc., can be utilized during the early evening hours and late night hours.
3. **Agent Utilization Strategy:**
   - **9 PM to 10 PM:** Agents from 7 PM to 8 PM.

    - **10 PM to 11 PM:** Agents from 8 PM to 9 PM.
    - **11 PM to 12 AM:** Agents from 9 PM to 10 PM.
    - **12 AM to 1 AM:** Agents from 10 PM to 11 PM.
    - **1 AM to 2 AM:** Agents from 11 PM to 12 AM.
    - Continue this distribution pattern until 9 AM.
4. **Adjust for Abandon Rate:**
    - Ensure at least 90% of calls are answered by adjusting agent numbers based on historical call volume and average call duration.
    - Factor in the 10% abandonment rate target and adjust agent schedules accordingly to maintain this rate.

Summary of Recommendations:

- **Daytime:** Focus on high-volume periods, especially between 11 AM and 1 PM, and adjust agent numbers to handle peak loads.
- **Nighttime:** Utilize a strategic approach to distribute agents from peak daytime periods to ensure sufficient coverage during night hours.

By implementing these strategies, you can effectively manage call volumes, reduce the abandonment rate to 10%, and improve overall customer experience.

Analysis

1. **Higher Average Call Count in Specific Time Buckets (10-11 AM, 6-9 PM):**

- **Reason:** Customers are often office workers who call during transitions, either while heading to work or after returning home. These times are convenient for them to address concerns as they have some free time. Calls during these periods tend to be quick and straightforward, which might contribute to a higher average call count in these buckets.

2. **Highest Number of Incoming Calls vs. Average Answered Calls (11-12 AM):**

- **Reason:** The high volume of incoming calls during this period suggests that the call center might be understaffed relative to the demand. The inability to handle all queries efficiently during this peak time results in a lower average number of calls answered.

3. **Decrease in Incoming Calls After 12 PM:**

- **Reason:** Customers generally prefer to resolve their issues on the same day, so they tend to call before noon. This behavior is driven by the desire for timely resolution, leading to peak call volumes before noon and a decrease afterwards as fewer new issues arise later in the day.

4. **Proportion of Monthly Transfer Rate vs. Answered and Abandoned Rates:**

- **Reason:** Dedicated toll-free numbers and skilled personnel at the call center help resolve most queries efficiently. The low transfer rate indicates that most issues are resolved at

the initial contact. Abandonments occur due to high call volumes or insufficient staff, while transfers happen only for complex issues beyond the initial team's expertise.

5. **Challenges in Exact Distribution of Night Agents:**

- **Reason:** With a fixed number of 17 agents for night shifts, a precise analytical distribution is difficult. Practical considerations such as agent availability, commute issues, and transportation facilities complicate the ability to provide an exact distribution. Hence, a flexible, non-analytical approach is used where agents from peak periods are reassigned to cover night shifts, and factors like agent comfort and logistics are considered.

Recommendations:

1. **Address Staffing Shortages:**
   - **For Peak Times:** Increase staffing during high call volume periods (11 AM - 12 PM) to handle the high number of incoming calls and improve the average answered call count.
   - **For Night Shifts:** Implement a flexible staffing strategy where agents from busy periods are reassigned to night shifts, while considering their personal circumstances and logistics.
2. **Optimize Call Handling:**
   - **During High Call Periods:** Enhance training for agents to handle queries more efficiently and reduce the need for transfers.
   - **For Night Coverage:** Ensure that agents are comfortable and have access to reliable transportation to improve night shift coverage.
3. **Improve Customer Experience:**
   - **Address Abandonment:** Focus on reducing abandonment rates by optimizing staffing levels during peak hours and ensuring adequate coverage during off-peak times.

By addressing these insights and implementing targeted strategies, you can improve both the efficiency of call handling and overall customer satisfaction.

Conclusion

Based on the analysis, the following conclusions and recommendations can be drawn:

1. **Current Situation Analysis:**
   - The average number of calls answered per agent per time bucket is 198.6.
   - To achieve a reduction in the abandon rate from 30% to 10%, we need to increase the call answered rate to 90% of incoming calls.
2. **Required Call Handling Capacity:**
   - **Total Average Calls Incoming Per Day:** 5,130
   - **Desired Answer Rate:** 90% (0.9)
   - **Average Calls Answered Per Second:** 198.6
   - **Seconds Per Hour:** 3,600
3. **Time Required to Answer 90% of Incoming Calls:**
4. Time Required=5,130×198.6×0.93,600≈257.2 hours

5. **Agents Required Per Day:**
6. Total Agents Required=257.2/4.5≈57.1 agents
7. To maintain an abandonment rate of 10%, approximately **57 agents** are required per day.
8. **Agent Availability Calculation:**
    - **Daily Working Hours per Agent:** 9 hours
    - **Break Time:** 1.5 hours
    - **Effective Working Hours:** 7.5 hours
    - **Occupied Hours per Day on Calls:** 4.5 hours (60% of 7.5 hours)
    - **Working Days per Month:** 20 days (excluding unplanned leaves and official holidays)
9. Each agent is available for approximately **20 days per month**. To cover the daily requirement of 57 agents, the total number of agents should be:
10. Total Agents Required=57×3020=85.5 agents
11. Therefore, to meet the demand and reduce the abandonment rate, approximately **86 agents** should be available.
12. **Night Shift Consideration:**
    - **Total Night Calls per Day:** 5130×30100=1,539 calls
    - **Additional Hours Required for Night Calls:** Additional Hours=1,539×198.6×0.93,600≈76.4 hours
13. **Additional Agents for Night Calls:**
14. Additional Agents Required=76.44.5≈17 agents.
15. Hence, an additional **17 agents** are required for night shifts.
16. **Total Agents Required Per Day:**
    - **Daytime Agents:** 57 agents
    - **Nighttime Agents:** 17 agents
    - **Total Agents Required:** 57 + 17 = **74 agents**
17. To ensure that the abandon rate is reduced to 10% and maintain service levels both during the day and night, **74 agents per day** are needed.

Summary

To effectively handle the incoming call volume, maintain a 90% call answered rate, and achieve a 10% abandonment rate, the company needs to allocate approximately 74 agents daily. This includes 57 agents for daytime operations and an additional 17 agents for night shifts. By adopting this staffing strategy, the company can enhance its customer service efficiency and improve overall customer satisfaction.

Appendix

Task2:
Instagram User Analytics

https://drive.google.com/drive/folders/1pNDIeIw4pI37mOBP_sfhmYbA8W-jsHc0?usp=share_link

Task3:
Operation & Metric Analytics

https://drive.google.com/drive/folders/1FxvGkB-GySTdR2It-rifg_bcynUQ6exi?usp=sharing

Task4:
Hiring Process Analytics

https://drive.google.com/drive/folders/1K882OStPzEMCGb34EzAmpqRNlapjMjZp?usp=sharing

Task5:
IMDB Movie Analysis

https://drive.google.com/drive/folders/11Ms3ZPTLuR7B7gW1aPvrU1W7UGfVX37G?usp=sharing

Task6:

Bank Loan Case Study

https://drive.google.com/drive/folders/1vl1mfn-VlkrNTsA-477bpC7gwAe22zQw?usp=sharing

Task7:

Bank Loan Case Study

https://drive.google.com/drive/folders/1SUtK2LY_32ahO9lVyJp_9orY7W0ppxuu?usp=sharing

Task8:

ABC Call Volume Trend Analysis

https://drive.google.com/drive/folders/1ERl2pzDmsVVZDac_Q5cq5eviUiMrOYCw?usp=sharing