DATA ANALYSIS PORT FOLIO

By -
Pooja Prakash

Professional Background-

Educational Qualification: Completed Bachelor of Technology (B.Tech) in Electronics and Communication Engineering with a GPA of 7.4.

- **Technical Skills & Certifications**:
  Acquired comprehensive skills in Data Analytics through various certified programs and hands-on learning in tools and technologies including:
- Python, SQL, Tableau, Power BI, Microsoft Excel
- Digital Marketing, including SEO, PPC, and content analytics.

Certifications include:

- Google Data Analytics Specialization
- Fundamentals of Digital Marketing (Google)
- Google Digital Marketing & E-Commerce Specialization
- The Data Analyst Course: Complete Data Analyst Bootcamp – Udemy
- SQL - MySQL for Data Analytics and Business Intelligence – Udemy
- **Industry Experience**:
- **Summer Internship – BSNL ALTTC (3 Months)**
  Gained practical exposure to telecommunications infrastructure and operations.
- **Data Visualization Job Simulation – Tata Consultancy Services (TCS)** via Forage – February 2025
  Simulated real-world business scenarios, including:
  - Designing insightful data visualizations for senior client leadership
  - Formulating strategic questions for stakeholder engagement
  - Presenting data-driven insights to support executive decision-making
- **Data Analytics Training Program – Trainity**
  Successfully completed **8 live industry-relevant projects**, focusing on end-to-end analytics processes including data cleaning, visualization, and interpretation.

Contents

1.Data Analytics Process-

Description-

- The purpose of this project was to make us realize that we use data analytics in our day-to-day life without even noticing it. Task given was to give example of a real-life situation where we can apply the principle of data analytics.

Design-

- Real-life scenario chosen – Buying my first car.
- When I will buy my first car I will have various options of different styles of vehicle, different brands, additional features and even more variation of models in the base model which I choose. So, I will decide which car to buy by doing a step-by-step analysis of all available options in my budget.

Steps to be taken-

- **Plan-** First I will decide I need to buy a car, which will be comfortable for a family trip and accommodate all four of my family members and will also give a good mileage.
- **Prepare-** I will check on various platforms the most popular brands and then decide a certain amount (say 15 lacs) above which I can't spend, also will look for the car that has top safety rating in my budget.
- **Process-** I will shortlist few models which I like, and which completes my requirements and fit under my budget, also have high popularity score high and safety ratings.
- **Analyze-** I will compare prices of selected models and feature variation they offer, reject few and finalize top two or three then test drive them to check overall performance.
- **Share-** Then I will share my findings with my friends and family ask

their opinion also let my car dealer know my final choices so they all can help me decide.

- **Act-** Finally I will buy my first car.

Conclusion-

Finally, we have seen how we can apply the six steps of data analytics, i.e., Plan, prepare, process, analyze, share and act in our everyday problems.

2.Instagram User Analytics-

Description-

- the goal of this project is to extract meaningful insights from the data, in such a manner that it can help the various teams at Instagram to decide their future strategy and plan of action to make sure the existing users are active on platform and also to be able to acquire new users.

The Problem-

Marketing Analysis-

- **Loyal User Reward:** The marketing team wants to find who are the most loyal user (who have used the platform for longest time) so that they can reward them. Find top five loyal users.
- **Inactive user engagement:** Team wants to send promotional call to encourage inactive users to start posting. We have to identify users who have never posted a single photo on Instagram.
- **Contest winner declaration-** A contest has been organized by the team where user with most likes on a single photo will win. Determine winner of contest.
- **Hashtag research-** A brand partner wants to know the most popular hashtags to use in their posts that will reach the most number of people. Identify and suggest the top five most commonly used hashtags.
- **Ad campaign launch-** The team wants to know the best day of the week to launch ad. Find out the day of the week when most users register on Instagram.Provide insights on when to schedule an ad campaign.

Investor Metrics:

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

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

Design-

Steps taken to load the data into the database

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

Software used - MySQL Workbench 8.0 CE

Findings I –

Identifying the Five oldest Instagram Users from the Database:

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

Findings II-

To determine which users have never uploaded a photo:

Result

Output/Results:-

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

Findings III –

To determine the user who received the highest number of likes on a

id	User_name	Photo_id	Image_url	likes
52	Zack_kemmer93	145	https://jarret.name/	48

Findings IV –

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

Tag id	Tag name	Tag count
21	smile	59
20	beach	42
17	party	39
13	fun	38
18	concert	24

Findings V-

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

Day	No of users registered
Thursday	16
Sunday	16

Finding VI-

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

- Average number of posts per user – 3.473
- Total photos divided by the total number of user – 2.57

Finding VII-

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

- There are thirteen accounts which appear to be bots or fake.

$ no.	User id	No of likes
1.	21	257
2.	71	257
3.	5	257
4.	66	257
5.	41	257
6.	14	257
7.	57	257
8.	24	257
9.	76	257
10.	75	257
11.	54	257
12.	91	257
13.	36	257

Analysis-

After completing my analysis of the given dataset my findings are as follows-

- **Inactive Users**: Analysis reveals that 26 users have registered on Instagram but have not engaged in any activity, such as posting photos, videos, or text updates. To encourage participation, the marketing team should consider reaching out to these users with reminders or promotional content.
- **Top Performer**: User *Zack_Kemmer93* (User ID: 52) stands out as the most liked contributor, having received 48 likes on a single photo (Photo ID: 145), making him the leading participant in terms of engagement.
- **Popular Hashtags**: The five most frequently used hashtags on the platform are:

1. #smile – 59
2. #beach – 42 occurrences
3. #party – 39 occurrences
4. #fun – 38 occurrences
5. #concert – 24 occurrences

6. **Optimal Days for User Registration**: Data indicates that Thursdays and Sundays are the most common days for user registrations, each accounting for 16 sign-ups. This suggests that launching advertising campaigns on these days could maximize reach and effectiveness.

- **Average Posts per User**: The platform has an average of 3 posts per user. By dividing the total number of photos by the total number of users, we find an average of approximately 2.57 posts per user.
- **Highly Active Users**: There are 30 users who have liked every photo available on the site. Such consistent activity may indicate automated behavior, suggesting these accounts could be bots.

Root Cause Analysis Using the 5 Whys Approach

- **Why did the marketing team seek information about the least active users?**

  ➔ Their goal was to identify and contact these users via email to understand the reasons behind their inactivity and encourage future engagement with the platform.
- **Why were they interested in the five most frequently used hashtags?**

  ➔ The team may be planning to implement specialized filters or features for posts using those popular hashtags, enhancing user experience and content discoverability.
- **Why investigate which day of the week has the highest new user registrations?**

  ➔ By identifying peak registration days, the marketing team can strategically schedule ad campaigns to coincide with high user activity, thereby maximizing brand exposure and potential revenue.
- **Why did investors want to understand the average number of posts per user?**

  ➔ User engagement is a critical indicator of a platform's value. Investors are interested in assessing whether Instagram's user base is both authentic and active, and this insight also helps the tech team scale infrastructure efficiently without affecting performance.
- **Why the interest in detecting bot and fake accounts?**

  ➔ Investors need to ensure that their investment is going toward a platform with genuine growth and not inflated metrics. Minimizing the presence of inauthentic users reduces risk and ensures a healthier, more trustworthy user ecosystem.

Conclusion-

leveraging customer data analysis is a fundamental strategy employed by Instagram and numerous other social media and commercial enterprises to extract meaningful insights from user behavior. This analytical approach enables businesses to identify and nurture high-value customers, thereby transforming them into long-term assets rather than potential liabilities.

Regularly conducting such analyses—be it weekly, monthly, quarterly, or annually—allows companies to adapt to evolving market dynamics and customer preferences. By segmenting their customer base, businesses can tailor their marketing strategies, optimize product offerings, and enhance customer experiences, leading to increased profitability and sustained growth.

Ultimately, the strategic application of customer segmentation and data analysis not only maximizes returns on investment but also ensures that resources are allocated efficiently, fostering a more personalized and effective engagement with the target audience.

3. Operation Analytics and Investigating Metric Spike:

Description-

Operational analytics is a process that focuses on examining and enhancing an organization's day-to-day activities. By analyzing both real-time and historical data, businesses can uncover inefficiencies, streamline processes, and make informed decisions to boost overall performance.

In a role akin to a Lead Data Analyst at a major tech company, one would collaborate closely with departments such as operations, customer support, and marketing. The objective is to transform raw data into actionable insights that drive strategic initiatives and operational improvements.

A critical aspect of this role involves investigating unexpected changes in key performance indicators (KPIs), such as sudden drops in user engagement or sales figures. This process requires a thorough examination to determine whether these anomalies stem from data quality issues, seasonal trends, or genuine shifts in user behavior.

By leveraging operational analytics, organizations can not only address immediate operational challenges but also lay the groundwork for long-term strategic planning and competitive advantage.

The Problem-

Case Study 1: Job Data Analysis

- **Jobs Reviewed Over Time**: Calculate the number of jobs reviewed per hour for each day in November 2020.We have to Write an SQL query to calculate the number of jobs reviewed per hour for eachday in November 2020.
- **Throughput Analysis:** Calculate the 7-day rolling average of throughput (number of events per second). Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.
- **Language Share Analysis:** Calculate the percentage share of each language in the last 30 days. Write an SQL query to calculate the percentage share of each language over the last 30 days**.**
- **Duplicate Rows Detection:** Identify duplicate rows in the data.

Your Task: Write an SQL query to display duplicate rows from the job_data table.

Case Study 2: Investigating Metric Spike-

- **Weekly User Engagement:** Measure the activeness of users on a weekly basis. Write an SQL query to calculate the weekly user engagement.
- **User Growth Analysis:** Analyze the growth of users over time for a product. Write an SQL query to calculate the user growth for the product.
- **Weekly Retention Analysis:** Analyze the retention of users on a weekly basis after signing up for a product. Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.
- **Weekly Engagement Per Device:** Measure the activeness of users on a weekly basis per device. Write an SQL query to calculate the weekly engagement per device.
- **Email Engagement Analysis:** Analyze how users are engaging with the email service. Write an SQL query to calculate the email engagement metrics.

Design-

Steps taken to load the data into the data base:

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

Software used for querying the results:-

--> MySQL Workbench 8.0 CE

Software used for analyzing using Bar plots:-

--> Microsoft Excel

Finding I –

Number of jobs reviewed per hour for each day in November 2020

Date	Jobs reviewed	Time spent
2020-11-25	1	45
2020-11-26	1	56
2020-11-27	1	104
2020-11-28	2	33
2020-11-29	1	20
2020-11-30	2	40

Finding II-

The 7-day rolling average

Date	Throughput	Rolling average
2020-11-25	0.0222	0.022200
2020-11-26	0.0179	0.020050
2020-11-27	0.0096	0.016567
2020-11-28	0.0606	0.027575
2020-11-29	0.0500	0.032060
2020-11-30	0.0500	0.035050

Finding III-

Percentage share of each language-

Language	Count	Percentage share
Persian	3	37.50
English	1	12.50
Arabic	1	12.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50

Finding IV-

Find number of duplicate rows-

There are no duplicate rows.

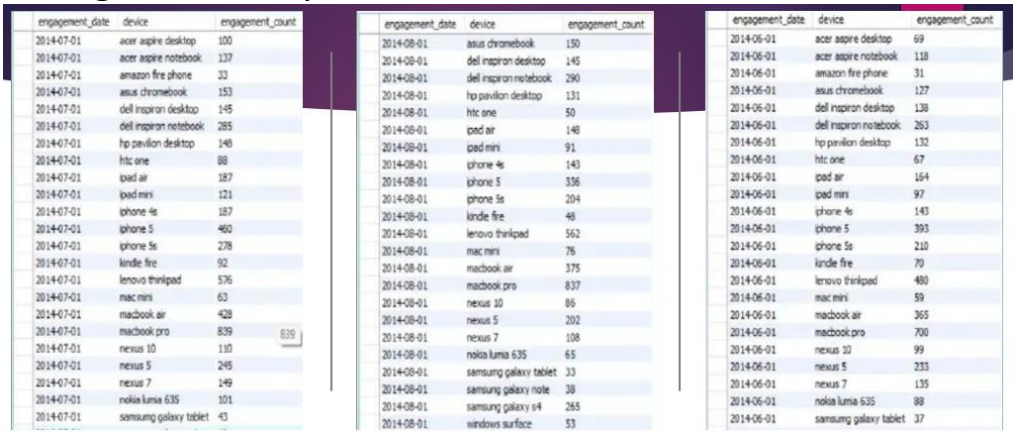
Finding V-

Weekly user engagement

Week	Engagement_count
18	701
19	1054
20	1094
21	1147
22	1113
23	1173
24	1219
25	1263
26	1249
27	1271
28	1355
29	1345
30	1363
31	1443
32	1266
33	1215
34	1203
35	1194

Finding VI-

User growth analysis



engagement_date	device	engagement_count
2014-07-01	acer aspire desktop	100
2014-07-01	acer aspire notebook	137
2014-07-01	amazon fire phone	33
2014-07-01	asus chromebook	153
2014-07-01	dell inspiron desktop	145
2014-07-01	dell inspiron notebook	285
2014-07-01	hp pavilion desktop	148
2014-07-01	htc one	88
2014-07-01	ipad air	187
2014-07-01	ipad mini	121
2014-07-01	iphone 4s	187
2014-07-01	iphone 5	460
2014-07-01	iphone 5s	278
2014-07-01	kindle fire	92
2014-07-01	lenovo thinkpad	576
2014-07-01	mac mini	63
2014-07-01	macbook air	428
2014-07-01	macbook pro	839
2014-07-01	nexus 10	110
2014-07-01	nexus 5	245
2014-07-01	nexus 7	149
2014-07-01	nokia lumia 635	101
2014-07-01	samsung galaxy tablet	43

engagement_date	device	engagement_count
2014-08-01	asus chromebook	150
2014-08-01	dell inspiron desktop	145
2014-08-01	dell inspiron notebook	290
2014-08-01	hp pavilion desktop	131
2014-08-01	htc one	50
2014-08-01	ipad air	148
2014-08-01	ipad mini	91
2014-08-01	iphone 4s	143
2014-08-01	iphone 5	336
2014-08-01	iphone 5s	204
2014-08-01	kindle fire	48
2014-08-01	lenovo thinkpad	562
2014-08-01	mac mini	76
2014-08-01	macbook air	375
2014-08-01	macbook pro	837
2014-08-01	nexus 10	86
2014-08-01	nexus 5	202
2014-08-01	nexus 7	108
2014-08-01	nokia lumia 635	65
2014-08-01	samsung galaxy tablet	33
2014-08-01	samsung galaxy note	38
2014-08-01	samsung galaxy s4	265
2014-08-01	windows surface	53

engagement_date	device	engagement_count
2014-06-01	acer aspire desktop	69
2014-06-01	acer aspire notebook	118
2014-06-01	amazon fire phone	31
2014-06-01	asus chromebook	127
2014-06-01	dell inspiron desktop	138
2014-06-01	dell inspiron notebook	263
2014-06-01	hp pavilion desktop	132
2014-06-01	htc one	67
2014-06-01	ipad air	164
2014-06-01	ipad mini	97
2014-06-01	iphone 4s	143
2014-06-01	iphone 5	393
2014-06-01	iphone 5s	210
2014-06-01	kindle fire	70
2014-06-01	lenovo thinkpad	480
2014-06-01	mac mini	59
2014-06-01	macbook air	365
2014-06-01	macbook pro	700
2014-06-01	nexus 10	99
2014-06-01	nexus 5	233
2014-06-01	nexus 7	135
2014-06-01	nokia lumia 635	88
2014-06-01	samsung galaxy tablet	37

Finding VII-

Weekly retention analysis-

Device	Avg_Week_Eng_Per_Dev
macbook pro	3148.33
lenovo thinkpad	2032.39
macbook air	1471.22
iphone 5	1420.89
dell inspiron notebook	1081.67
samsung galaxy s4	1024.22
nexus 5	907.33
iphone 5s	874.67
dell inspiron desktop	557.61
iphone 4s	528.11
asus chromebook	523.78
ipad air	518.94
acer aspire notebook	490.22
hp pavilion desktop	488.22
nexus 7	358.67
nokia lumia 635	308.28
ipad mini	306.94
acer aspire desktop	284.44
nexus 10	281.50
mac mini	245.00
htc one	234.89
kindle fire	224.44
windows surface	188.83
samsung galaxy note	146.78
amazon fire phone	118.67

Finding VIII-

Email engagement analysis

email_services	Email_Engagement_count
sent_weekly_digest	57267
email_open	20459
email_clickthrough	9010
sent_reengagement_email	3653

Analysis-

Root Cause Analysis Using the 5 Whys Framework

- **Why is job review data measured hourly each day?**
  ➔ To ensure that no job postings are overlooked, maintaining consistency and timely review across all time slots.
- **Why use a 7-day rolling average to calculate throughput instead of a daily average?**
  ➔ A 7-day rolling average provides a smoother and more comprehensive view by incorporating data from an entire week, unlike the daily average, which only reflects performance for a single day. This method helps in identifying broader trends and reduces volatility in the analysis.
- **Why is the share of the Persian language at 37.5% while all other languages combined make up only 12.5%?**
  ➔ This discrepancy might be due to either the presence of duplicate records involving Persian-language data or the actual presence of a larger number of unique users who prefer Persian.
- **Why should we detect and address duplicate records in a dataset?**
  ➔ Duplicate entries can distort analytical results, leading to incorrect conclusions and potentially poor business decisions. Removing them ensures data integrity and enhances the accuracy of insights.
- **Why does user engagement tend to be low initially but increase later?**
  ➔ When a new product or service is introduced, it typically gains awareness gradually. In the early stages, few users try it. Over time, as more people share positive experiences, engagement levels tend to rise — a pattern suggesting successful adoption in this case.

- **Why is weekly user retention important?**
  ➔ Weekly retention metrics help businesses re-engage users who start but don't complete the sign-up process. With targeted outreach and support, these users may convert into loyal customers.
- **Why is tracking weekly engagement per device significant?**
  ➔ This metric helps businesses identify which devices deliver the best user experiences and where improvements are needed, ensuring better device compatibility and higher satisfaction across platforms.
- **Why is email engagement an important metric?**
  ➔ Email interaction rates indicate the effectiveness of a company's communication strategy. For example, if only 33.58% of emails are opened and 14.79% of those result in clicks, it suggests the need for more compelling subject lines and better-crafted content to boost user interest and action.

Conclusion-

operational analytics and the investigation of metric anomalies are essential practices for businesses aiming to optimize performance and adapt to evolving market dynamics. Regular analysis—whether daily, weekly, monthly, or quarterly—enables organizations to identify inefficiencies, respond to trends, and make informed decisions that drive growth.

Equally important is the focus on email engagement strategies. Crafting compelling subject lines, personalizing content, and offering exclusive promotions can significantly enhance open and click-through rates. Implementing automated email sequences, such as welcome series or re-engagement campaigns, ensures consistent communication and fosters stronger customer relationships.

Additionally, establishing dedicated support channels for users who abandon the sign-up process can help address their concerns and convert them into loyal customers. By proactively reaching out and providing assistance, businesses can improve user experience and increase conversion rates.

By integrating these strategies—continuous operational analysis, targeted email marketing, and attentive customer support—businesses can

enhance efficiency, boost customer retention, and drive sustainable growth.

4.Hiring Process Analytics-

Description-

The hiring process plays a pivotal role in shaping a company's long-term success. By analyzing key metrics such as rejection rates, interview stages, job categories, and vacancy durations, organizations can gain valuable insights to refine their recruitment strategies. As a Data Analyst at a multinational corporation like Google, you're entrusted with examining historical hiring data to uncover patterns and inform decision-making. Utilizing data analytics in recruitment offers several advantages:

- Data helps identify the most effective sourcing channels and methods, leading to time and cost savings.
- **Reducing Bias**: Data-driven approaches can mitigate unconscious biases by standardizing evaluation criteria and highlighting discrepancies in hiring patterns.
- **Predictive Insights**: Employing predictive analytics enables forecasting candidate success and turnover rates, allowing for more informed hiring decisions.
- **Continuous Improvement**: Establishing feedback loops and monitoring key performance indicators (KPIs) facilitates ongoing refinement of the recruitment process.

In your role, you'll delve into the company's hiring data to identify trends, assess the effectiveness of various recruitment strategies, and provide actionable recommendations to optimize the hiring process. This data-centric approach ensures that the organization attracts and retains top talent, driving sustained growth and competitiveness.

The Problem-

- **Hiring Analysis:** The hiring process involves bringing new individuals into the organization for various roles. Determine the gender distribution of hires. How many males and females have been hired by the company?
- **Salary Analysis:** The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees. What is the average salary offered by this company? Use Excel functions to calculate this.
- **Salary Distribution:** Class intervals represent ranges of values, in this

case, salary ranges. The class interval is the difference between the upper and lower limits of a class. Create class intervals for the salaries in the company. This will help you understand the salary distribution.

- **Departmental Analysis**: Visualizing data through charts and plots is a crucial part of data analysis. Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.
- **Position Tier Analysis**: Different positions within a company often have different tiers or levels. Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

Design-

Prior to initiating the analysis, a systematic data preparation process was followed to ensure data integrity and reliability:

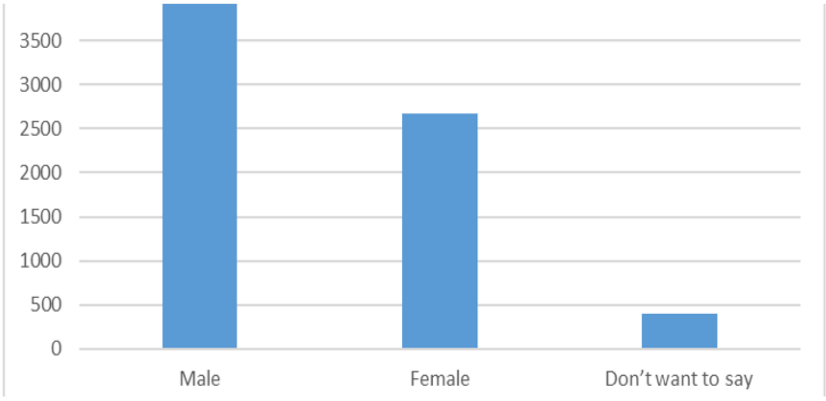
1. **Creating a Working Copy:**
   To preserve the original dataset, a duplicate was made. This approach allows for experimentation and modifications without compromising the raw data.
2. **Eliminating Non-Essential Columns:**
   Columns that were not pertinent to the analysis objectives were identified and removed. This streamlining facilitates a more focused and efficient analytical process.
3. **Identifying Missing Values:**
   The dataset was examined for empty cells and NULL entries, which can adversely affect analytical outcomes.
4. **Imputing Missing Numerical Data:**
   For numerical columns with missing values, imputation was performed using statistical measures:
   - **Mean Imputation:** Applied when data distribution is approximately normal.
   - **Median Imputation:** Utilized for skewed distributions to mitigate the impact of outliers.

Note: Imputation methods were chosen based on the distribution characteristics of each column.

5. **Handling Outliers:**
   Outliers were detected using statistical techniques such as the Interquartile Range (IQR) method. Identified outliers were then replaced with the median value of the respective column to minimize distortion in the analysis.
6. **Imputing Missing Categorical Data:**
   For categorical columns with missing entries, the most frequent category (mode) was used for imputation. This approach maintains the categorical distribution and integrity of the data.
7. **Removing Duplicate Records:**
   Duplicate rows, which can skew analysis and insights, were identified and removed to ensure each record in the dataset is unique.

Finding I –

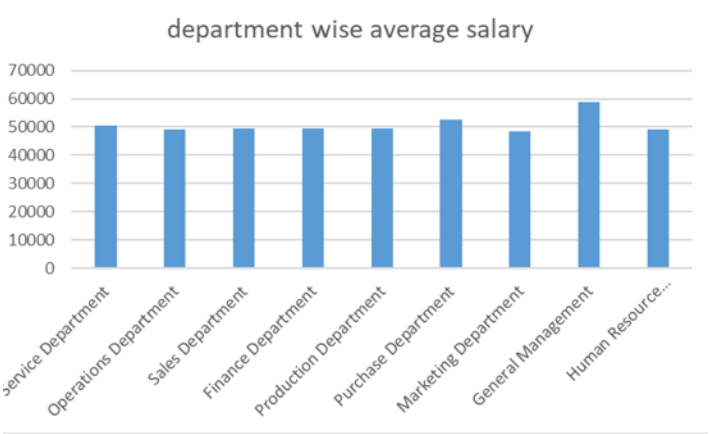
To determine gender distribution-



Finding II-

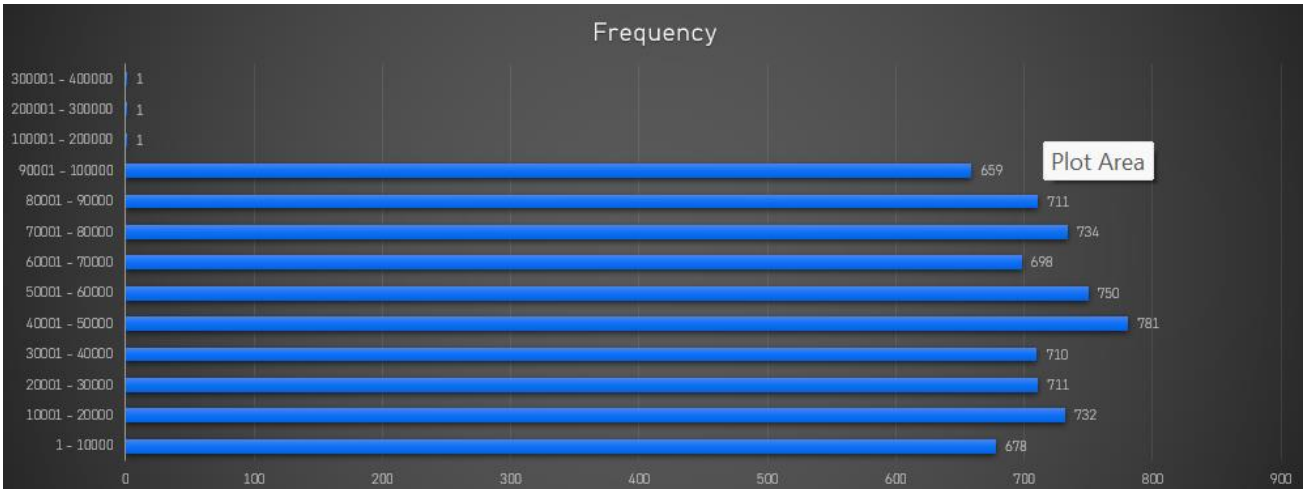
Salary analysis-

- Overall average salary is 49983.



Finding III-

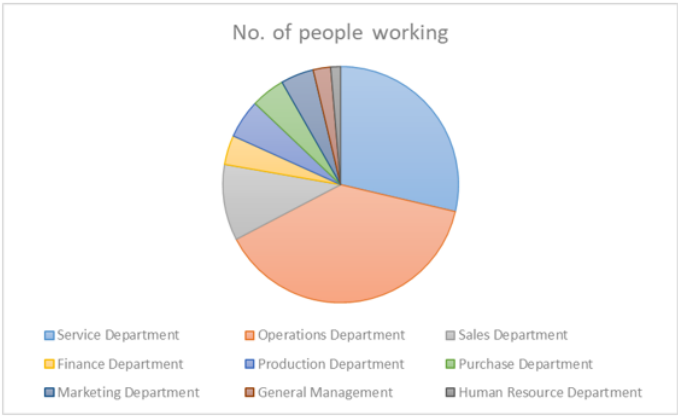
Create class intervals for salaries –



Finding IV –

Departmental analysis-

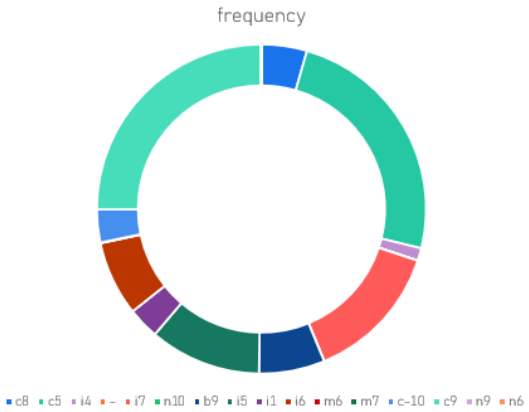
Departmental Analysis-



Department name	No of people working
Service Department	2055
Operations Department	2771
Sales Department	746
Finance Department	288
Production Department	380
Purchase Department	333
Marketing Department	325
General Management	172
Human Resource Department	97

Finding V –

Position tier analysis



Post	No. of employees	Average salary
c8	320	50701.46
c5	1747	50213.50
i4	88	48877.84
-	1	85914.00
i7	981	50065.36
n10	1	26990.00
b9	463	49666.76
i5	787	49391.93
i1	222	49943.94
i6	527	48839.25
m6	3	34521.33
m7	1	41402.00
c-10	232	51134.62
c9	1792	50201.19
n9	1	46219.00
n6	1	44700.00

We can see that c9 has highest number of of employees.

Analysis-

Using the whys approach to find the root cause of following-

1. Gender Disparity in Hiring

Observation: A significant imbalance exists between the number of male and female hires.

Analysis: This disparity may stem from entrenched gender stereotypes and biases prevalent in various regions. In some cultures, traditional roles and societal expectations limit women's participation in certain professions, leading to underrepresentation in the workforce. Additionally, unconscious biases during recruitment processes can inadvertently favor male candidates, further exacerbating the imbalance.

2. Salary Distribution Patterns

Observation: A smaller proportion of employees earn salaries exceeding ₹85,000, while a larger segment falls within the ₹35,000 to ₹60,000 range.

Analysis: Higher salary brackets are typically associated with specialized roles requiring extensive experience and expertise. Such positions are fewer in number and demand specific skill sets, justifying the elevated compensation. Conversely, most employees occupy roles that, while essential, are more standardized and thus fall within the mid-salary range.

This distribution reflects the hierarchical structure of organizations, where top-tier positions are limited, and mid-level roles are more abundant.

3. Predominance of the Operations Department

Observation: The Operations department has the highest number of employees.

Analysis: Operations serve as the backbone of an organization, ensuring the seamless execution of daily activities and processes. Given its central role in coordinating various functions, the department naturally requires a larger workforce to manage tasks ranging from logistics to quality control. The extensive scope and critical nature of operations necessitate a substantial staffing allocation to maintain efficiency and productivity.

By examining these patterns, organizations can identify areas for improvement, such as implementing unbiased recruitment practices, ensuring equitable compensation structures, and optimizing departmental resource allocation to foster a more balanced and inclusive workplace.

Conclusion-

Hiring process analytics serves as a critical tool for organizations aiming to align their recruitment strategies with future workforce requirements. By systematically analyzing hiring data on a monthly, quarterly, or annual basis, companies can make informed decisions regarding job openings, ensuring they meet evolving business demands.

Leveraging hiring process analytics fosters data-driven decision-making, promoting efficiency, equity, and strategic growth within the organization.

5. IMDB Movie Analysis –

Description-

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects**.**

The Problem –

- **Movie Genre Analysis**: Analyze the distribution of movie genres and their impact on the IMDB score. Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.
- **Movie Duration Analysis**: Analyze the distribution of movie durations and its impact on the IMDB score. Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
- **Language Analysis**: Situation: Examine the distribution of movies based on their language. Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- **Director Analysis**: Influence of directors on movie ratings. Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.
- **Budget Analysis:** Explore the relationship between movie budgets and their financial success. Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

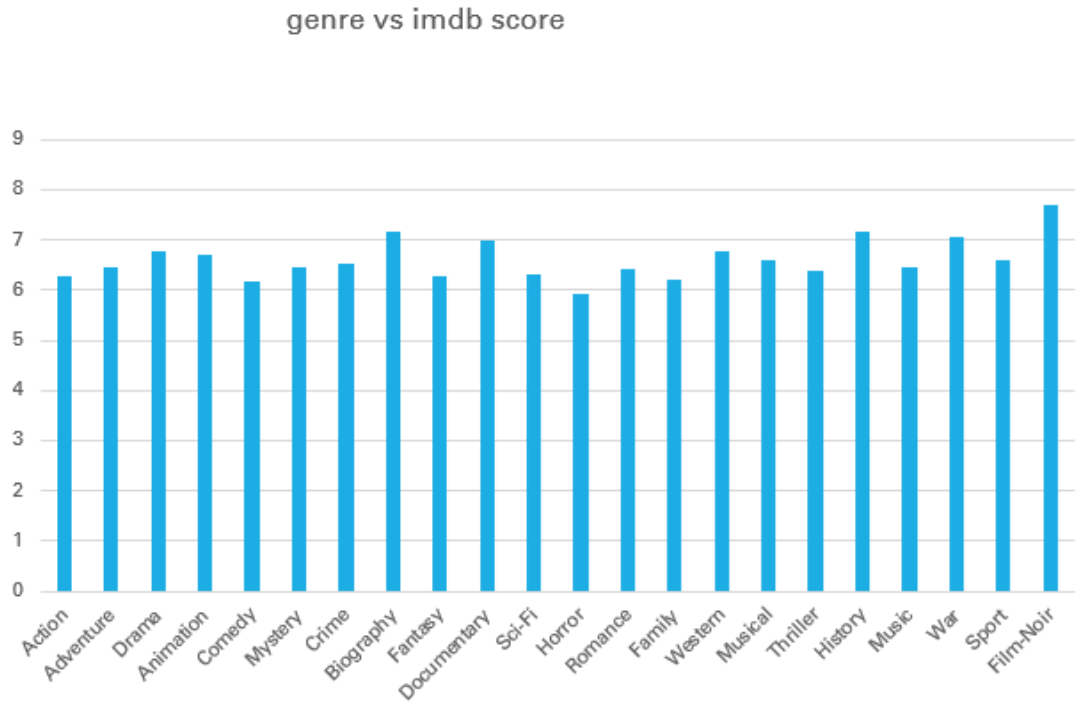
Design –

Before starting the analysis, I undertook a systematic data cleaning process to ensure the dataset's integrity and relevance:

1. **Creating a Backup of the Original Dataset**:
   To safeguard the original data, I created a duplicate copy. This approach allows for experimentation and modifications without compromising the raw data.
2. **Eliminating Irrelevant Columns**:
   I identified and removed columns that were not pertinent to the analysis objectives.
3. **Handling Missing Values**: I scanned the dataset for missing or null values. Rows containing such values were removed to maintain data consistency and reliability.
4. **Removing Duplicate Entries**: To ensure each record's uniqueness and prevent skewed analysis, I identified and eliminated duplicate rows from the dataset.
5. **Tools used** – Microsoft excel.

Finding I –

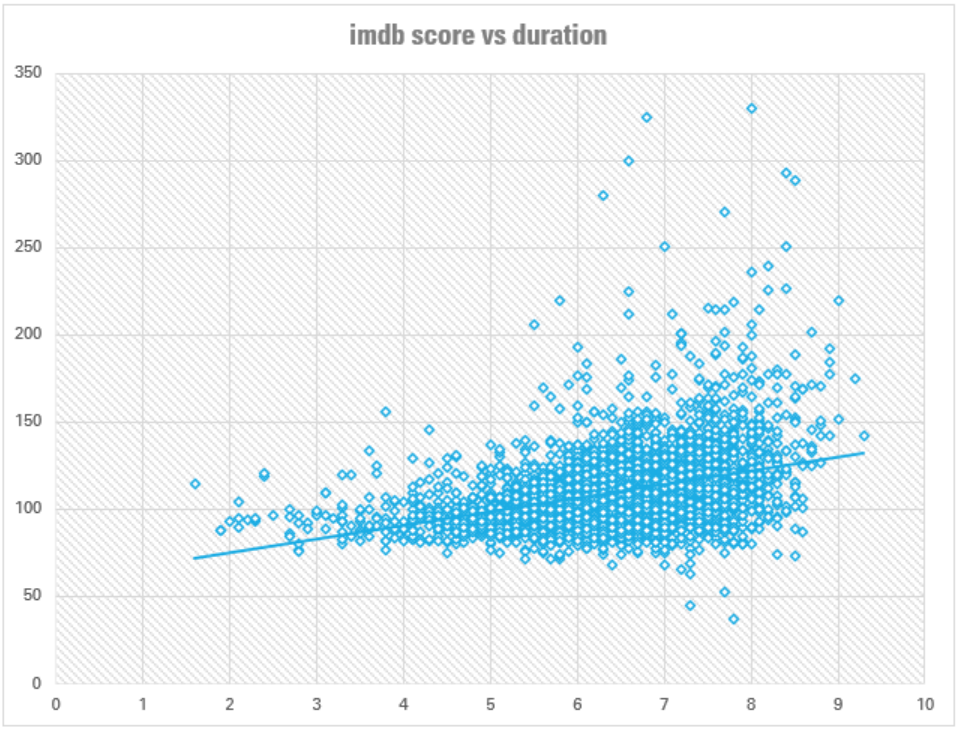
Movie Genre analysis –



Most common genre is drama.

Finding II –

Movie duration analysis –

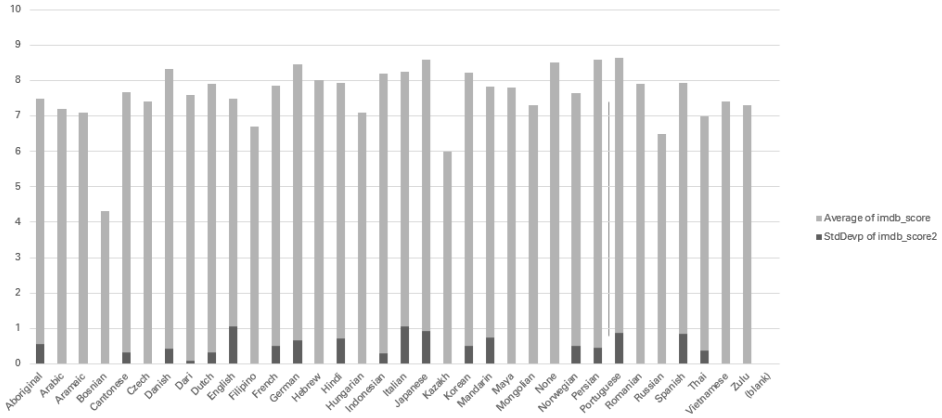


	duration	imdb_score
mean	110.258	6.4652822
stdeviation	22.6437	1.0561061
median	106	6.6

Movies with longer duration have better score.

Finding III –

Language Analysis:



top 10 languages	count
English	3598
French	34
Spanish	23
Mandarin	15
Japanese	10
German	10
Italian	7
Cantonese	7
Portuguese	5
Korean	5

English is the most common language.

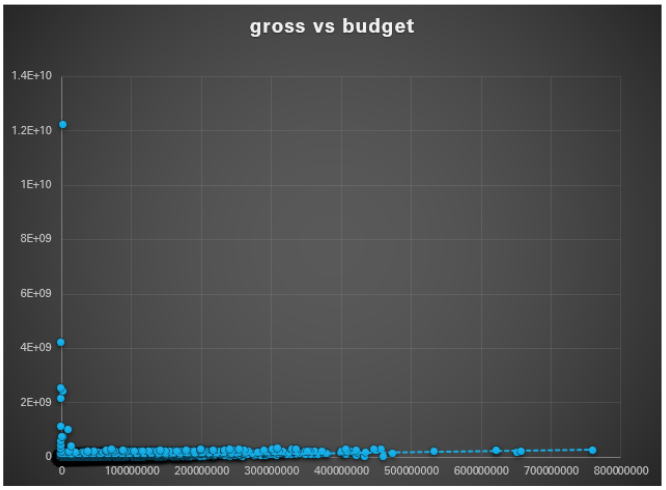
Finding IV –

Director Analysis-

top 10 directors	imdb score	percentile
Akira Kurosawa	8.7	0.99400
Tony Kaye	8.6	0.99200
Charles Chaplin	8.6	0.99200
Ron Fricke	8.5	0.98700
Majid Majidi	8.5	0.98700
Alfred Hitchcock	8.5	0.98700
Damien Chazelle	8.5	0.98700
Sergio Leone	8.433333333	0.98700
Christopher Nolan	8.425	0.98700
Richard Marquand	8.4	0.98300

Finding V –

Budget Analysis –



Budget analysis-

correlation coefficient	0.0994964
highest profit margin	523505847

Movie with highest margin is Avatar.

Analysis –

Applying the why approach –

🎭 1. Popularity of Drama and Comedy Genres

Observation: Drama and comedy films consistently rank among the most popular genres worldwide.

Analysis: The widespread appeal of drama and comedy can be attributed to their ability to resonate with audiences on both emotional and psychological levels.

- **Drama**: Drama films often delve into profound themes, offering viewers a chance to reflect on human experiences and emotions. Individuals who appreciate drama tend to seek meaningful narratives that provide insight into the human condition.
- **Comedy**: Comedy serves as a form of escapism, allowing audiences to experience joy and laughter, which can be therapeutic, especially during stressful times. The genre's emphasis on humour and light-heartedness makes it universally appealing.

In regions experiencing high levels of stress or societal pressures, such as India, the preference for light-hearted comedies has been linked to the population's desire for relief and entertainment that offers a temporary escape from daily challenges.

⏱️ 2. Movie Duration vs. IMDb Ratings

Observation: Longer movies do not necessarily receive higher IMDb ratings.

Analysis: While one might assume that longer films provide more content and, therefore, might be rated higher, audience preferences suggest otherwise. Data indicates that movies with runtimes between 100 to 120 minutes often receive favorable ratings, whereas those around 90 minutes may receive lower ratings. This trend suggests that audiences appreciate films that are concise yet comprehensive, delivering impactful narratives without unnecessary prolongation.

🌐 3. Prevalence of English in Films

Observation: English is the most commonly used language in films globally.

Analysis: The dominance of English in cinema can be attributed to several factors:

- **Global Lingua Franca**: English serves as a common language across many countries, facilitating broader audience reach.
- **Economic Influence**: The United States, being a major hub for film production, predominantly produces English-language films, which are then distributed worldwide.
- **Marketability**: English-language films often have higher marketability in international markets, leading to increased investments and returns.

In countries like India, the integration of English into cinema reflects societal shifts and the language's association with modernity and global appeal.

💰 4. Disparity Between IMDb Ratings and Box Office Profits

Observation: The highest-rated films on IMDb are not always the most profitable at the box office.

Analysis: Several factors contribute to this discrepancy:

- **Audience Reach**: IMDb ratings are based on user reviews, which may not represent the broader movie-going audience.
- **Marketing and Distribution**: A film's profitability heavily depends on its marketing strategy and distribution channels, which can influence box office performance regardless of critical acclaim.
- **Genre and Content**: Films with niche appeal or complex themes might receive high ratings from critics and certain audience segments but may not attract mass audiences necessary for high box office returns.

Therefore, while IMDb ratings reflect the opinions of a segment of viewers, box office profits are influenced by a combination of factors, including marketing, distribution, and broader audience appeal.

Conclusion-

Analyzing IMDb data has become an integral part of the film industry's decision-making process. Beyond filmmakers, this analysis is invaluable to investors, stakeholders, and theater owners who seek insights into audience preferences and market trends.

Such analyses are not just academic exercises; they play a pivotal role during both the pre-production and post-production phases of filmmaking.

In the pre-production stage, understanding trends can guide script development, casting choices, and marketing strategies. Post-production analysis helps assess a film's performance and informs future projects. It's important to note that a high IMDb rating doesn't always equate to box office success. Box office profits are primarily driven by ticket sales, which can be influenced by factors like marketing, release timing, and genre appeal. For instance, genres like comedy and drama often attract wider audiences seeking entertainment and emotional connection, whereas action or horror films may cater to more niche markets. Therefore, directors and production teams should incorporate comprehensive data analysis into their planning processes. By doing so, they can align their creative vision with audience expectations, optimize resource allocation, and enhance the potential for both critical acclaim and financial success.

6.Bank loan Case Study –

Description –

Imagine you are a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who do not have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

The Problem –

- Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis. Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
- Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset. Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models. Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes. Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.
- Identify Top Correlations for Different Scenarios: Understanding the

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

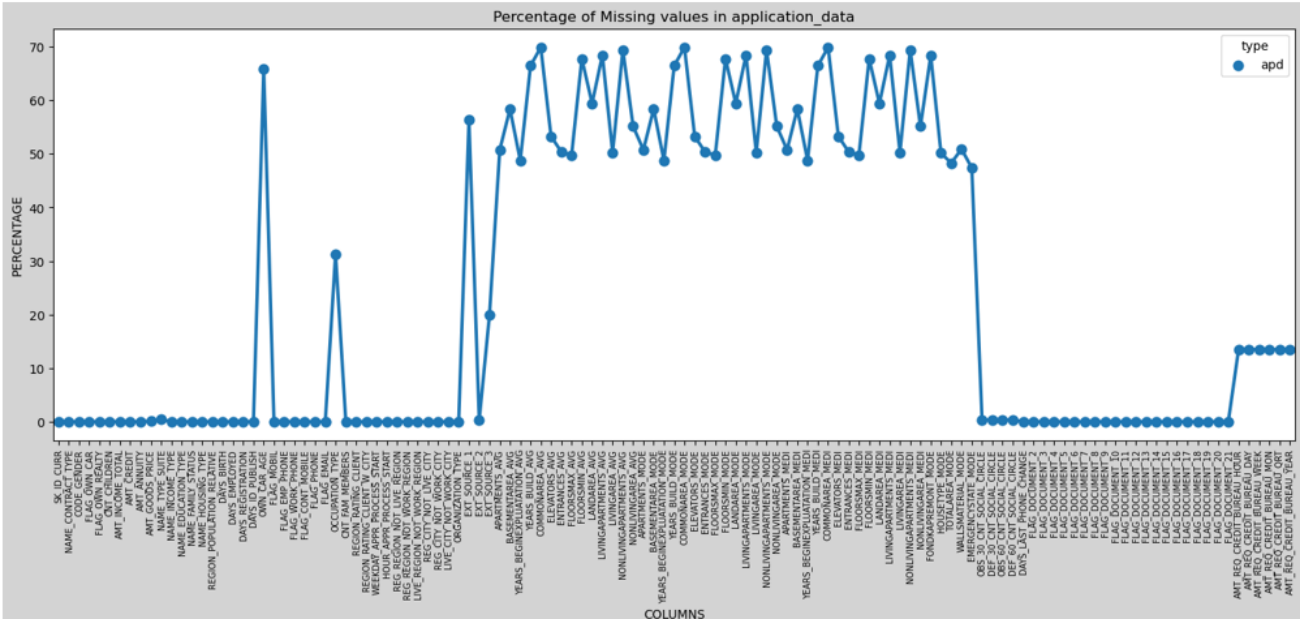
Design –

Prior to initiating the analysis, the following preparatory steps were undertaken to ensure data integrity and relevance:

- **Data Duplication for Preservation**: Created a separate copy of the original dataset to perform analyses, ensuring that the primary data remains unaltered.
- **Elimination of High Null-Value Columns**: Identified and removed columns where 50% or more of the entries were null, as such extensive missing data could compromise the reliability of the analysis.
- **Handling Missing Values in Numerical Columns**:
  For numerical columns with missing values, imputation was performed using either the mean or median, depending on the data distribution. The median was preferred in cases where the data exhibited skewness or contained outliers, as it provides a more robust central tendency measure in such scenarios.
  Outlier Treatment:
- Outliers were identified using appropriate statistical methods. To mitigate their impact, these outlier values were replaced with the median of the respective column, ensuring that the central tendency of the data remained unaffected.
- **Handling Missing Values in Categorical Columns**:
  Categorical columns with missing entries were addressed by imputing the mode, i.e., the most frequently occurring value in each column. This approach maintains the distributional characteristics of the categorical variables.
- **Removal of Duplicate Entries**:To prevent data redundancy and potential biases in analysis, duplicate rows within the dataset were identified and removed.

Finding I –

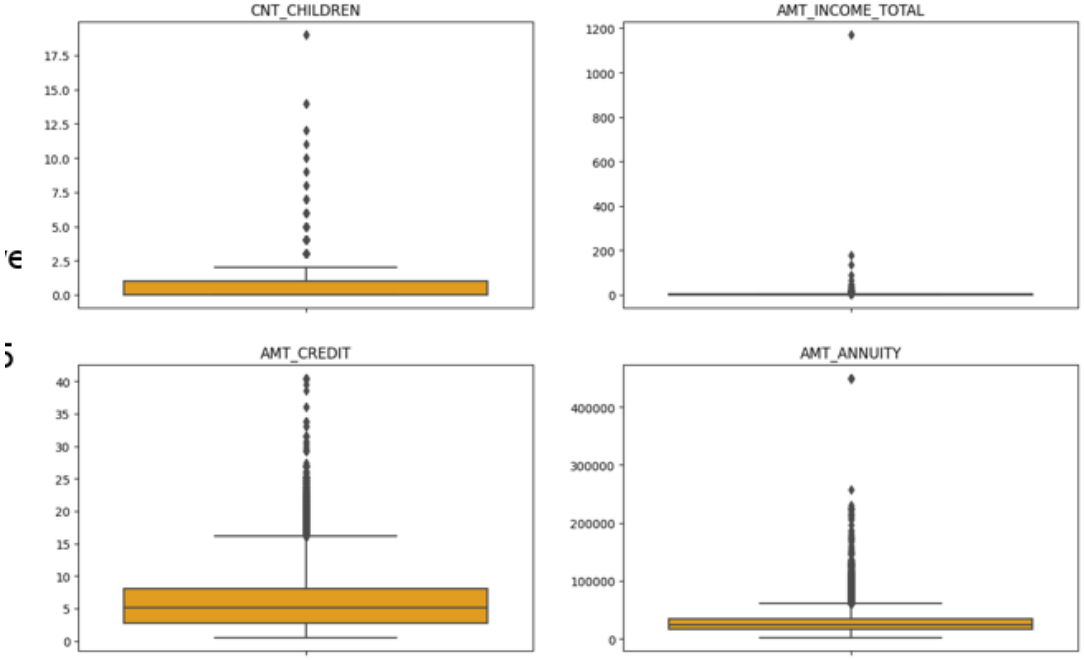
Handling missing data-

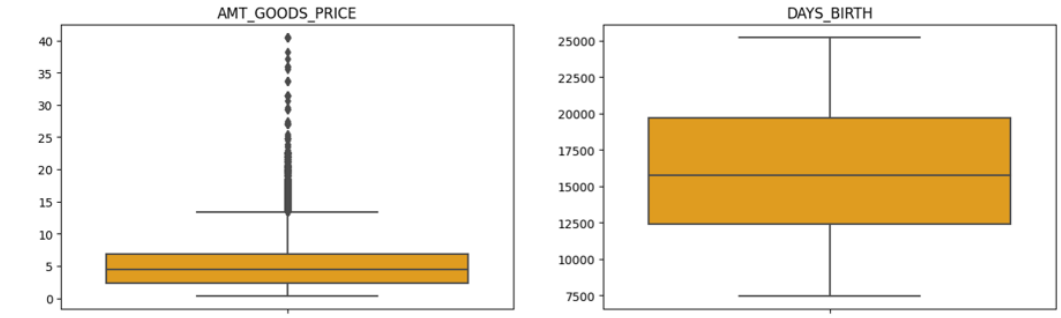


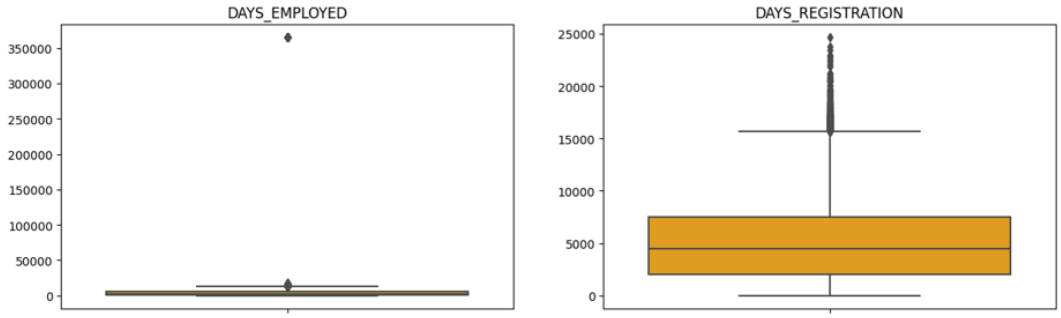
- We dropped the columns that have null values greater than 40%.
- After that we replaced null values of numerical column with median and null values of categorical columns with mode.

Finding II –

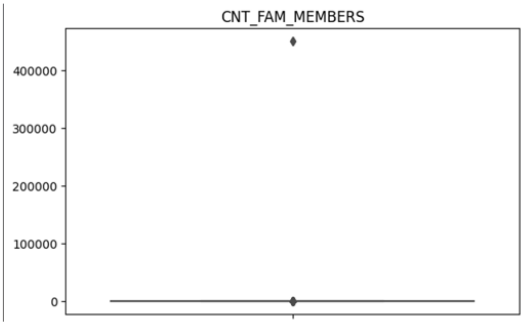
Handling outliers-





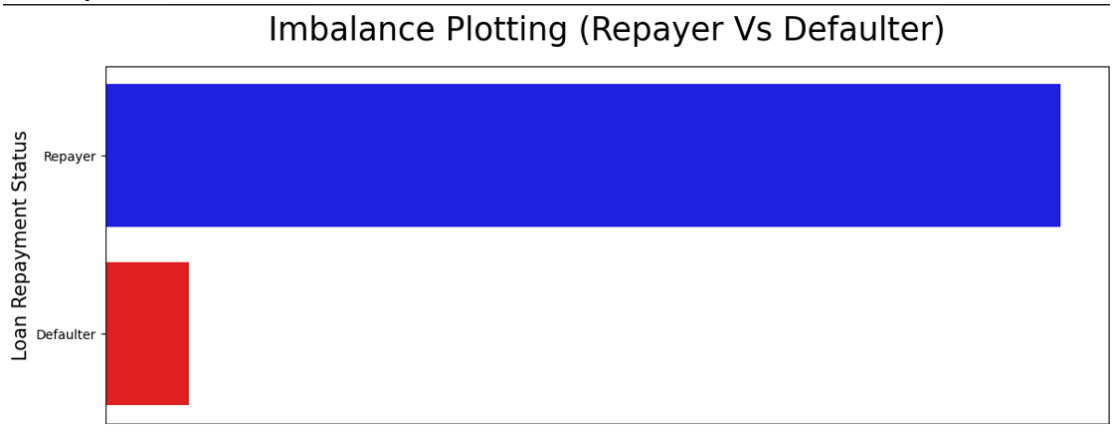


- As we can see the graph of family members count shows value greater than 40000 which is not possible.



Finding III-

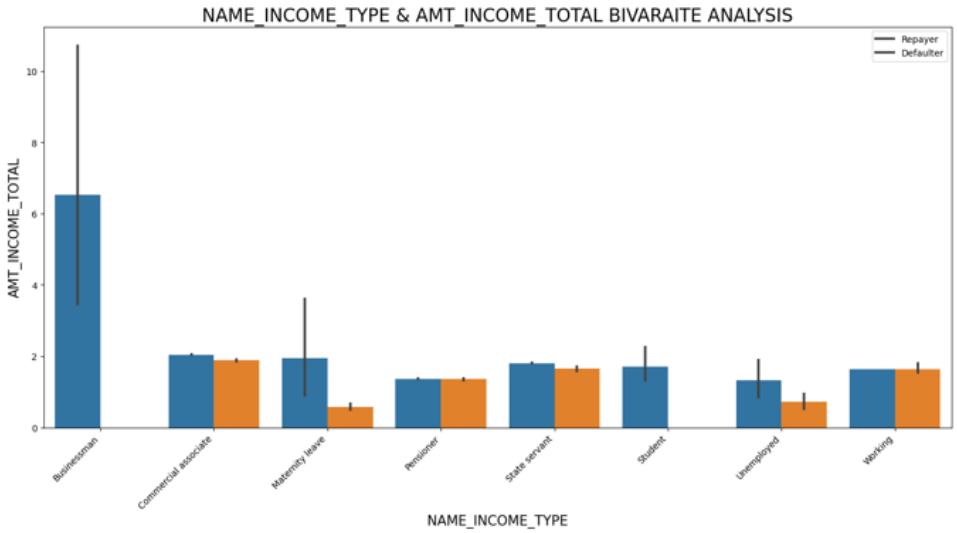
Analyse data imbalance –



Repayer percentage 91.93 and defaulter – 8.07%.

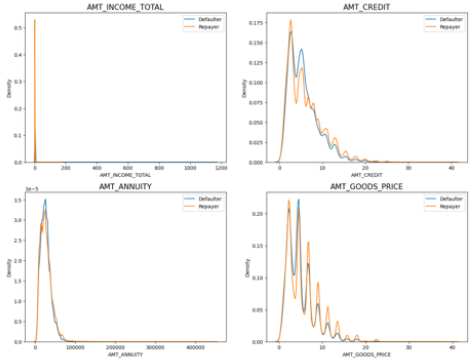
Finding IV –

Categorical variable analysis –



Finding V –

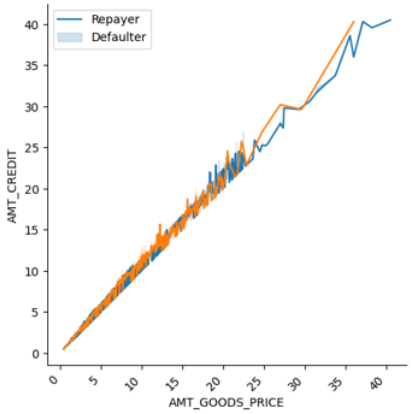
Numerical variable analysis –



- The graphs show Amount and loan repayment status.
- Highest number of loans- goods priced below 10lakhs
- Common annuity amount- below 50k
- Common amount for credit loan – less than 10 lakhs
- But as repayer and defaulter overlap in map no decision can be made.

Finding VI –

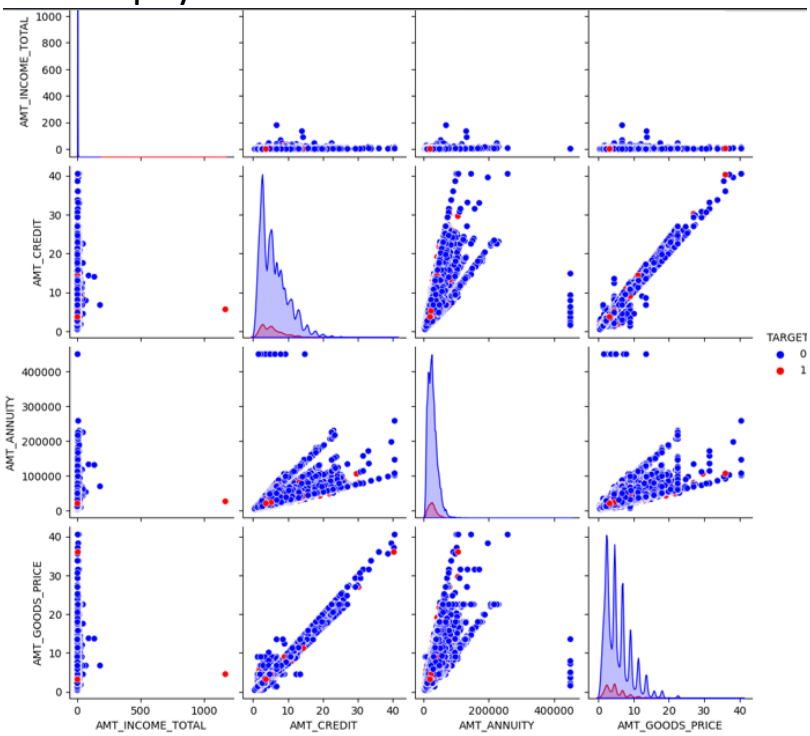
Numerical bivariate analysis –



When credit amount crosses 30lakhs there is an increase in defaulters.

Finding VII –

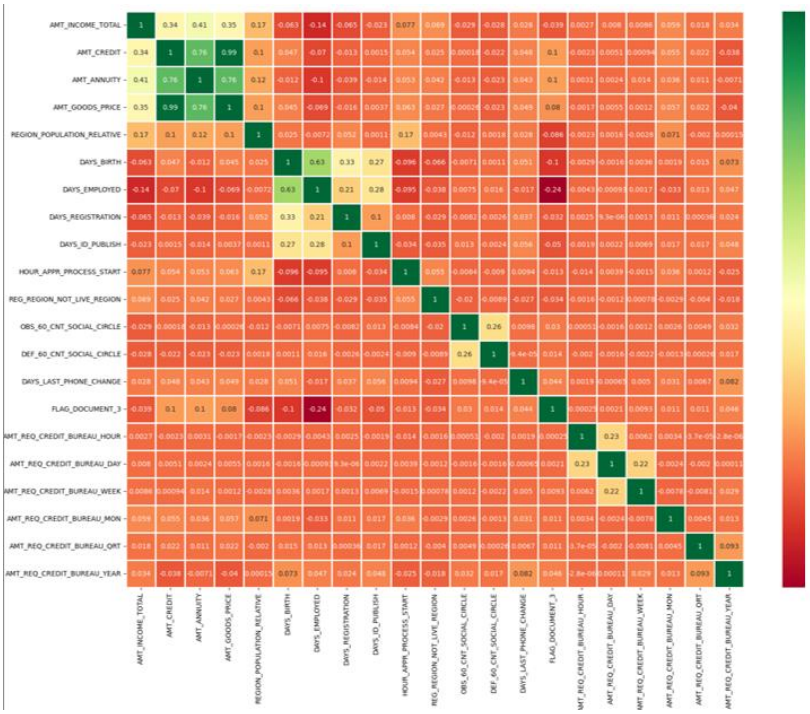
Loan repayment status –



High correlation between price of goods and loan amount.

Finding VIII –

Correlation for different scenarios –



Credit amount is highly correlated with annuity, total income and price of good.

Analysis –

Root Cause Analysis Using the "Why" Approach

1. Why Should Banks Consider Clients Residing in Non-Standard Housing Types, Despite Higher Non-Default Rates Among House/Apartment Dwellers?

While clients living in traditional houses or apartments exhibit higher non-default rates, it's essential for banks to also focus on individuals residing in non-standard housing types such as municipal apartments, rented accommodations, or those living with family. These clients often aspire to own a home, driven by factors like the declining joint family system in India and the desire for personal space. Targeting this segment can open new avenues for home loan disbursements, as these individuals are more likely to seek financing to transition into homeownership.

2. Low-income female borrowers exhibit notably lower default rates.

Conclusion –

Key Insights from Loan Repayment Analysis

1. **High Repayment Rates Among Clients**: The majority of clients demonstrate consistent loan repayment behavior, indicating a generally reliable borrower base.
2. **Gender Dynamics in Lending and Defaults**: While banks tend to extend more loans to female clients, data suggests that male borrowers have slightly lower default rates. This trend may be influenced by factors such as income disparities and varying financial obligations between genders.
3. **Impact of Age and Experience**: An increase in age and professional experience correlates with a decrease in default likelihood. Older, more experienced clients often have established financial stability, reducing the risk of loan delinquency.
4. **Prevalence of Cash Loans**: Cash loans constitute a significant portion of the bank's lending portfolio, reflecting a common preference among clients for immediate liquidity solutions.
5. **Educational Attainment and Default Rates**: Clients with higher educational qualifications tend to have lower default rates compared to those with only secondary education. This may be attributed to

better financial literacy and more stable employment opportunities among the educated demographic.

6. **Family Size and Loan Acquisition**: There is an observable trend where clients with more children are less likely to take on new loans, possibly due to increased financial responsibilities and risk aversion associated with larger family sizes.
7. **Unemployment and Credit Risk**: Unemployed clients represent a higher credit risk, often associated with larger loan amounts and increased default rates. This underscores the importance of thorough risk assessment when considering loan applications from unemployed individuals.
8. **Age-Related Loan Amounts and Default Rates**: As clients age, they tend to take out larger loans; however, their default rates decrease. This suggests that older clients, despite higher loan amounts, are more reliable in repayments, making them less risky and potentially more profitable for the bank.

7.Analyzing the Impact of Car Features on Price and Profitability –

Description –

The dataset includes variables such as car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP). The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. It is important to know the impact of car features on price and profitability in the automotive industry. The purpose is to analyze the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

The Problem –

Tasks: Analysis

• Insight Required: How does the popularity of a car model vary across different market categories?

Task 1.A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

• Insight Required: What is the relationship between a car's engine power and its price?

Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

•Insight Required: Which car features are most important in determining a car's price?

Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.
Insight Required: How does the average price of a car vary across different manufacturers?
Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.
Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.
• Insight Required: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?
Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.
Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.
Building the Dashboard:
• Task 1: How does the distribution of car prices vary by brand and body style?
• Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?
• Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?
• Task 4: How does the fuel efficiency of cars vary across different body styles and model years?
• Task 5: How does the car's horsepower, MPG, and price vary across different Brands?

Design –

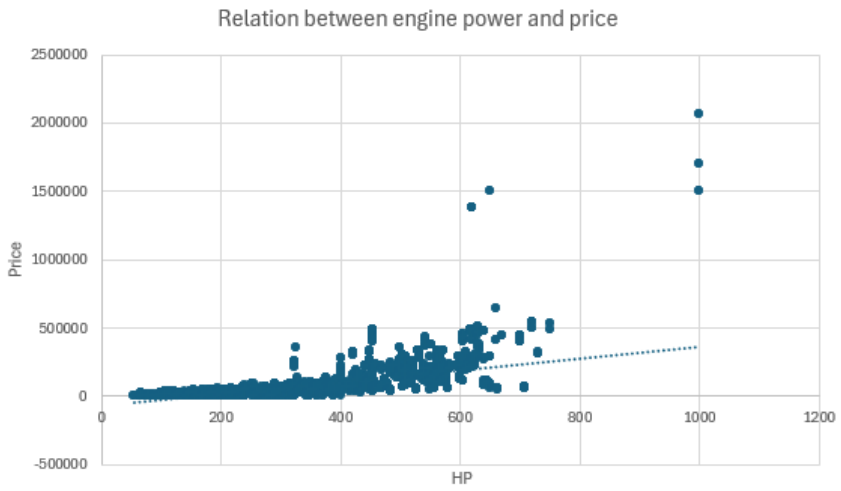
Before initiating the analysis, the following data preparation steps were undertaken using Microsoft Excel:

1. **Creation of a Working Copy**: To preserve the integrity of the original dataset, a duplicate was created for analysis purposes. This approach ensures that any modifications do not affect the source data.
2. **Elimination of Irrelevant Columns**: Columns deemed non-essential for the analysis were identified and removed, streamlining the dataset for more efficient processing.
3. **Handling Missing Data**: Rows containing blank spaces or NULL values were systematically removed to maintain data consistency and reliability.
4. **Removal of Duplicate Entries**: To ensure data accuracy, duplicate rows were identified and eliminated from the dataset.

Finding I –
How does the popularity of a car model vary across different market categories?

Row Labels	Average of Popularity	Count of Model
Crossover	1494.261415	1993
Diesel	2190.353234	201
Exotic	908.4887064	487
Factory Tuner	1735.825944	609
Flex Fuel	2102.40102	1177
Hatchback	1370.208022	1072
High-Performance	1539.069343	1370
Hybrid	1755.839763	337
Luxury	1238.607575	3221
N/A	1657.725223	3370
Performance	1412.930256	1950
(blank)		
Grand Total	**1513.44847**	**15787**

Finding II –
What is the relationship between a car's engine power and its price?

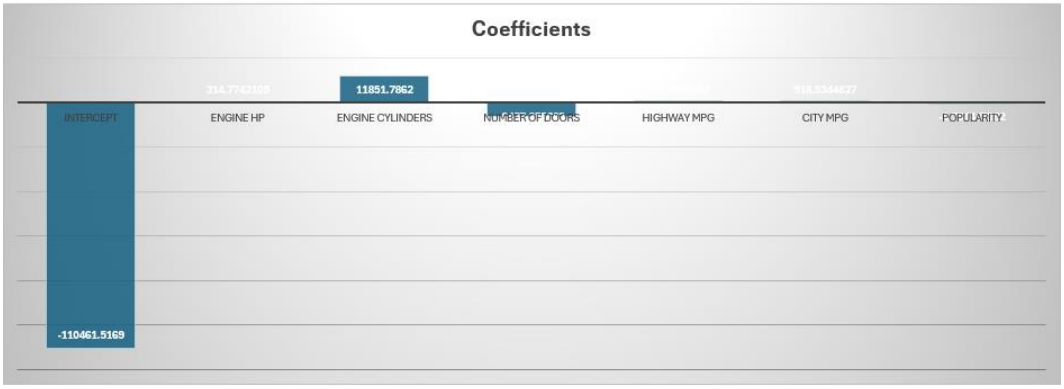


higher engine power means high price as it will involve complex design additional elements.

Finding III –

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

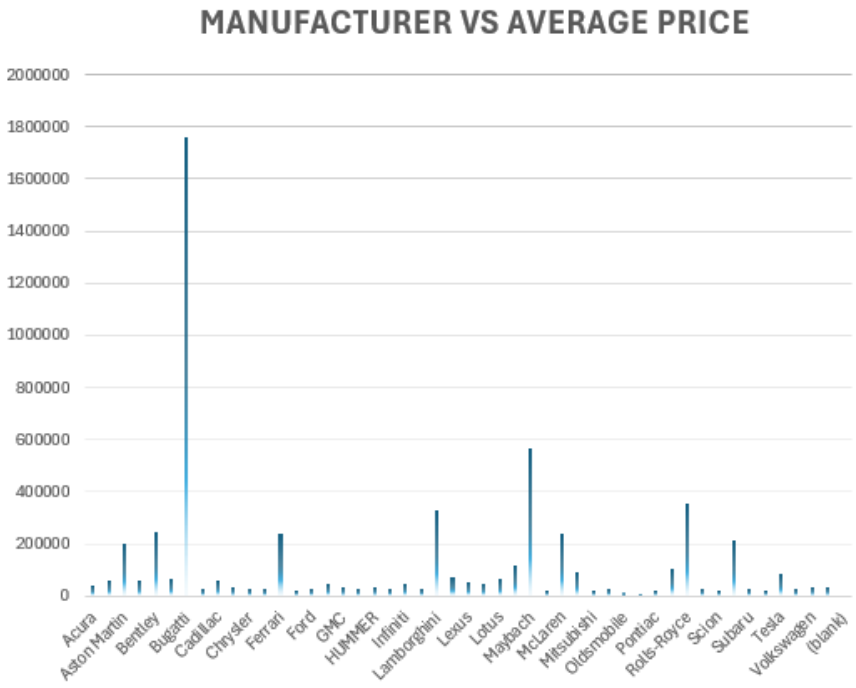
I	J	K	L	M	N	O	P	Q	R
	SUMMARY OUTPUT								
	Regression Statistics								
	Multiple R	0.71128281							
	R Square	0.50592324							
	Adjusted R Square	0.50573538							
	Standard Error	55314.9184							
	Observations	15787							
	ANOVA								
		df	*SS*	*MS*	*F*	*Significance F*			
	Regression	6	4.94404E+13	8.2401E+12	2693.05951	0			
	Residual	15780	4.82827E+13	3059740200					
	Total	15786	9.77231E+13						
		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
	Intercept	-110461.517	3688.271819	-29.949397	9.948E-192	-117690.951	-103232.082	-117690.951	-103232.082
	Engine HP	314.774211	5.904913415	53.3071679	0	303.1999055	326.348516	303.199905	326.348516
	Engine Cylinders	11851.7862	424.8850612	27.8940996	3.481E-167	11018.9629	12684.6095	11018.9629	12684.6095
	Number of Doors	-6427.35568	501.9271086	-12.8053567	2.3481E-37	-7411.1902	-5443.52116	-7411.1902	-5443.52116
	highway MPG	670.515917	122.0655233	5.4930819	4.0109E-08	431.2535354	909.778298	431.253535	909.778298
	city mpg	918.534483	114.5148455	8.02109525	1.1212E-15	694.0722931	1142.99667	694.072293	1142.99667
	Popularity	-4.79444728	0.30989392	-15.4712531	1.3423E-53	-5.4018748	-4.18701977	-5.4018748	-4.18701977



Engine cylinders have the highest coefficient which means that they have very strong relationship with engine power.

Finding IV –

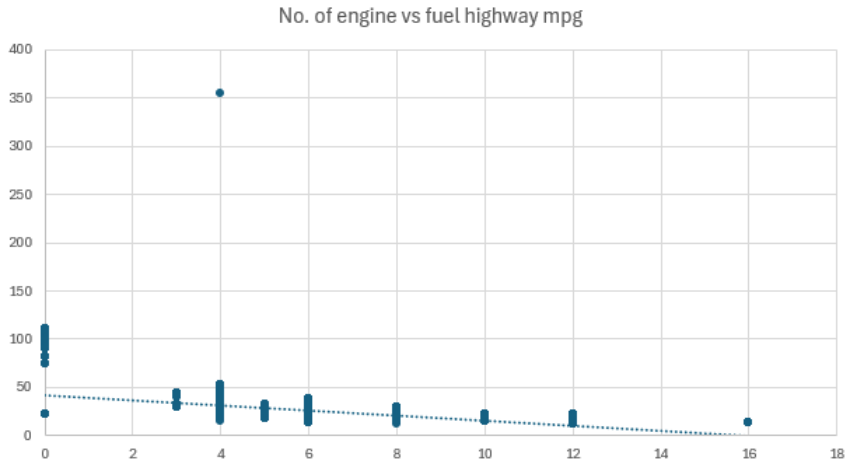
How does the average price of a car vary across different manufacturers?



We can see that luxury brands such as – Bugatti, Maybach, Rolls-Royce, Ferrari are highest priced.

Finding V –

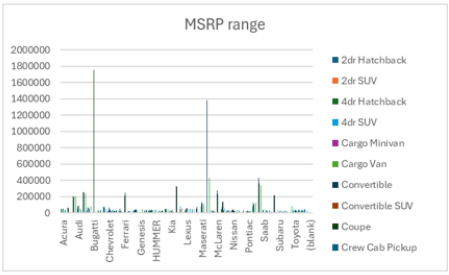
What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

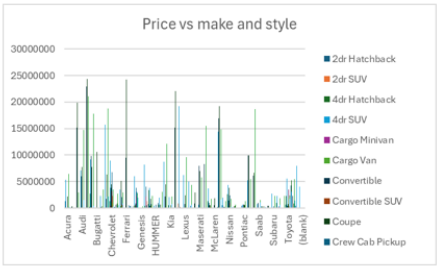


We can see that this is a negative slope therefore negative relation.

Finding VI –

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

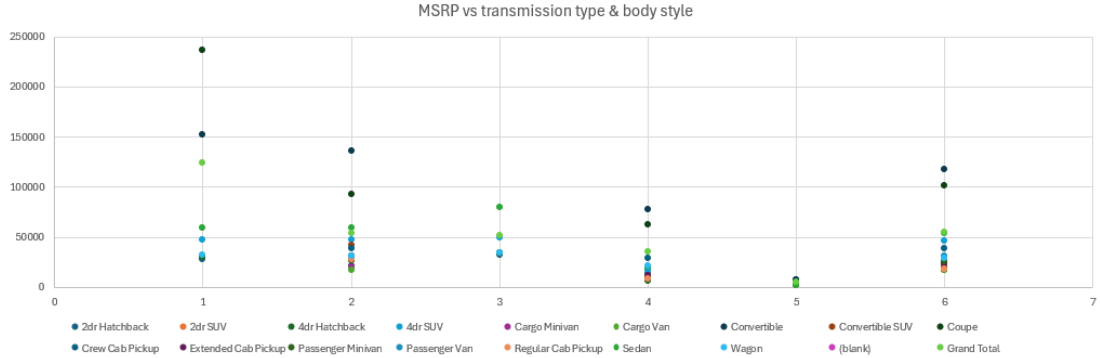




- The sum of MSRP is highest for Ferrari.
- Car model dating back to 1990-91 has lowest MSRP sum, although it may be because prices were not so high back then.

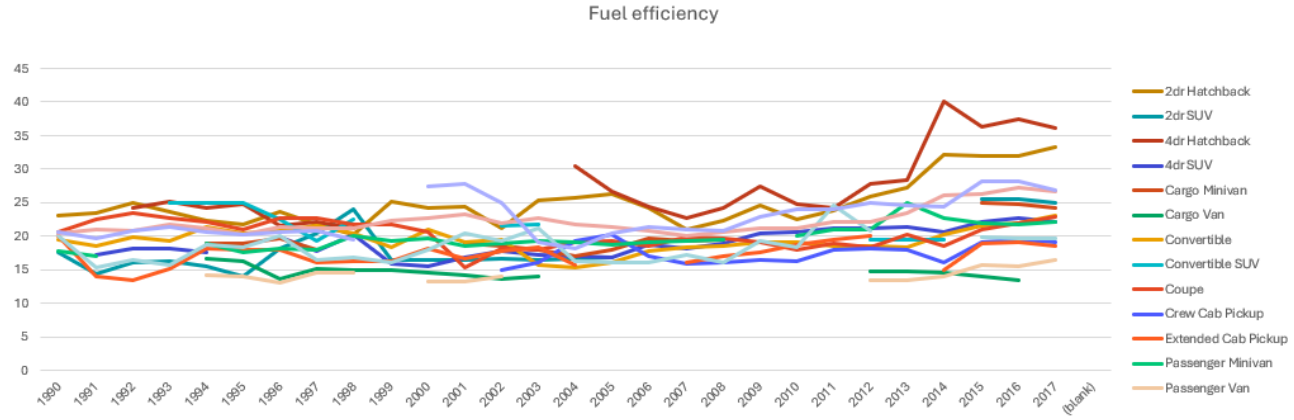
Finding VII –

How do the different feature affect the MSRP, and how does this vary by body style?



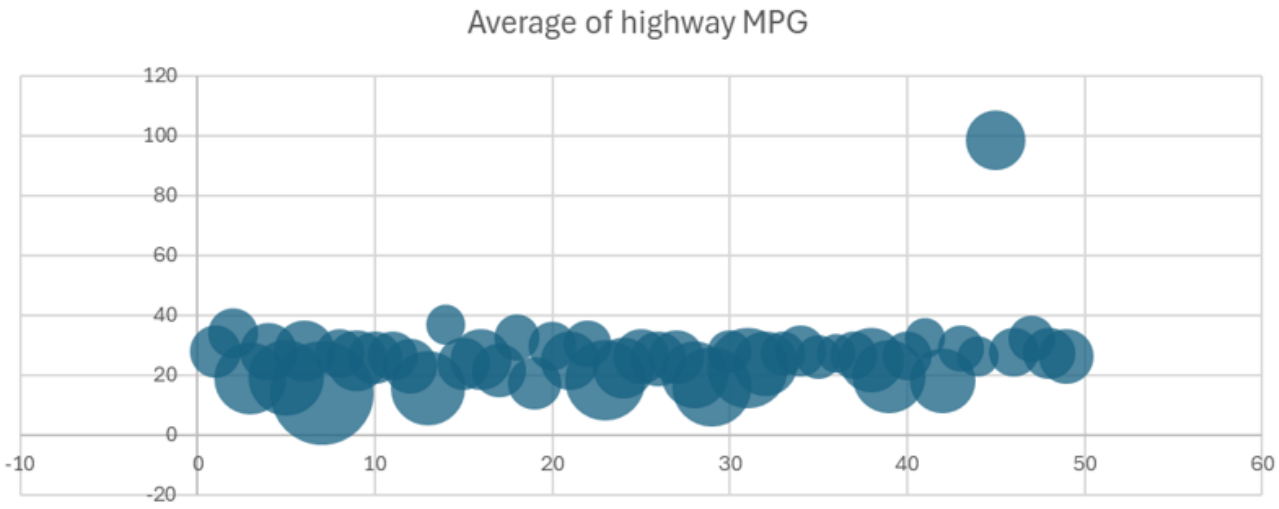
Finding VIII –

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



Finding IX –

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



Conclusion –

Key Insights from Car Market Analysis

1. **Popular Market Segments**: Vehicles featuring Flex Fuel and Diesel engines, particularly in Hatchback, Crossover, and Performance categories, dominate the market due to their versatility and efficiency.
2. **Engine Power and Pricing Correlation**: There's a strong positive correlation between engine power and vehicle price, indicating that as engine capacity increases, so does the cost of the car.

3. **Significance of Engine Cylinders**: The number of engine cylinders is a crucial determinant in pricing, with more cylinders often leading to higher performance and, consequently, higher prices.
4. **Brand Price Extremes**: Among various brands, Bugatti stands out with the highest average vehicle price, while Plymouth offers models at the lower end of the pricing spectrum.
5. **Cylinders and Fuel Efficiency**: An increase in the number of cylinders typically results in decreased highway miles per gallon (MPG), highlighting a trade-off between performance and fuel economy.
6. **Chevrolet's Pricing by Body Style**: Chevrolet exhibits the highest price distribution across different body styles, reflecting a diverse range of offerings catering to various market segments.
7. **Costliest Transmission and Body Style Combination**: Vehicles equipped with automated manual transmissions and a coupe body style tend to be among the most expensive, combining performance with sleek design.
8. **Fuel-Efficient Body Style**: Station wagons are recognized for their superior fuel efficiency, making them an economical choice for consumers seeking practicality without compromising on mileage.
9. **Engine Horsepower Impact**: As engine horsepower increases, there's a general trend of decreasing highway MPG, while the vehicle's price tends to rise, indicating a balance between performance and efficiency.

ABC Call Volume Trend Analysis –

Description –

A Customer Experience (CX) team analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, various types of support, including email, inbound, outbound, and social media support, among others. There are several AI-powered tools like include Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing are being used to enhance customer experience. Inbound customer support, which is the focus of this project, involves handling incoming calls from existing or prospective customers. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business. We have dataset that contains information about the inbound calls received by a company named ABC that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

The Problem –

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

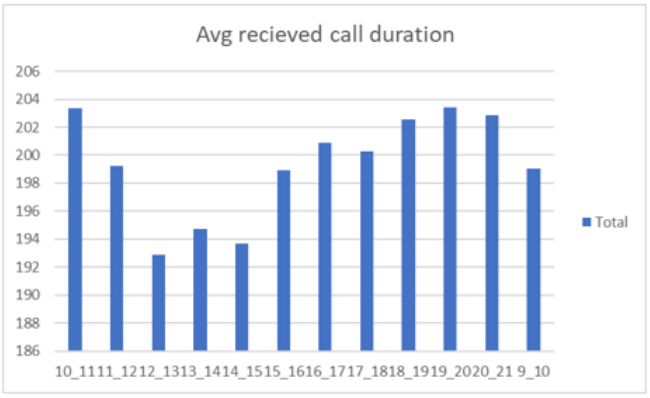
- Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets. Can you create a chart or graph that shows the number of calls received in each time bucket?
- Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are

answered. What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

• Night Shift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

Finding I –

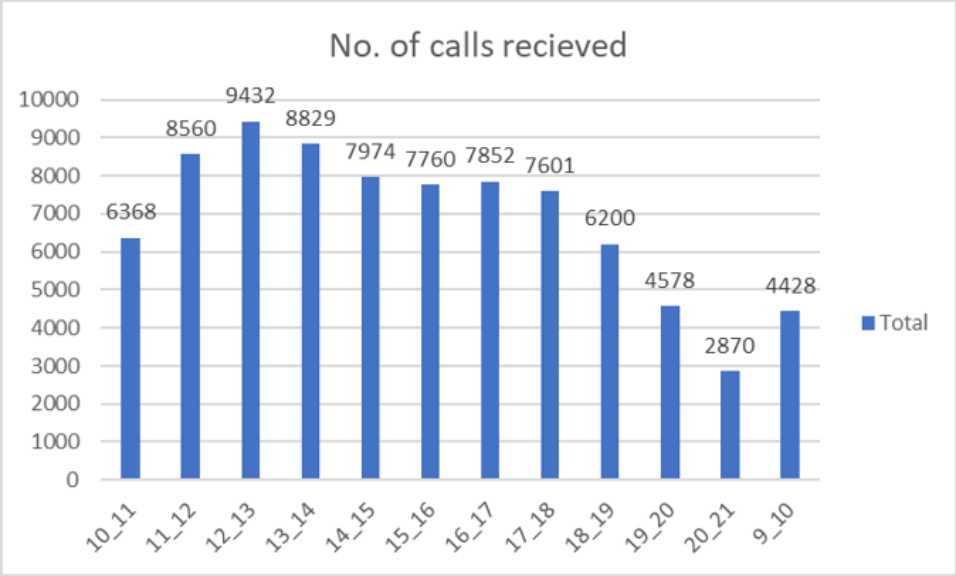
Average call duration-



Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
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
9_10	199.0691057
Grand Total	**198.6227745**

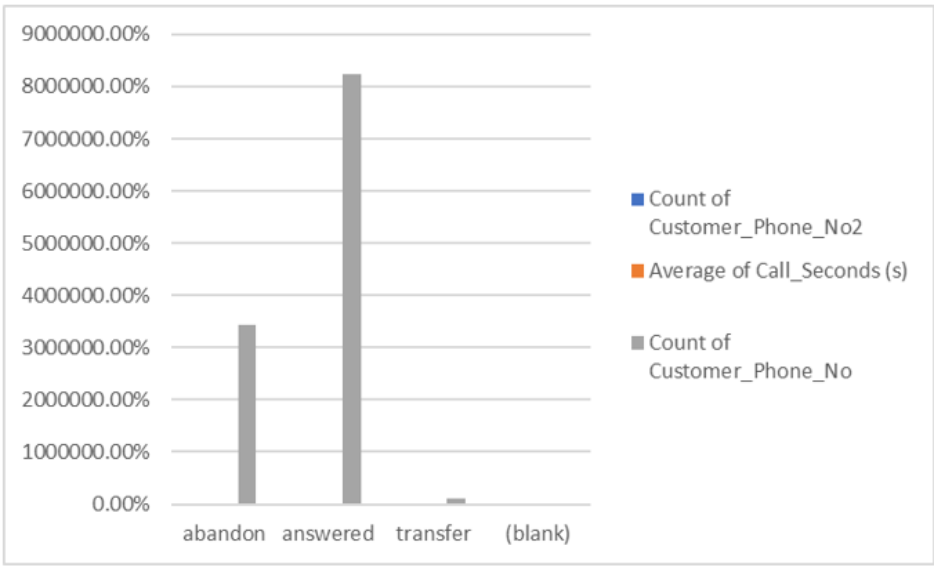
Finding II –

Call Volume Analysis –



Finding III –

Manpower Planning-



We can see that nearly 30% of calls are abandoned.

Total hours	9
Total working hours	7.5
Total hours actually working	60% of 7.5 = 4.5 hours
Total call hours during day	676664/3600 = 188
Total agents worked during day	188/4.5 = 42
If 42 agents do 70% work then 90% will be done by-	(42/70)*90 = 54 agents.

Grand Total	676664

Number of agent required to turn abandoned call percentage to 10% is 54.

Row Labels	Percentage of call seconds	Count of Call_Seconds	Time distribution	No. of agents required
10_11	11.28%	13313	0.11	6
11_12	12.40%	14626	0.12	7
12_13	10.72%	12652	0.11	6
13_14	9.80%	11561	0.10	5
14_15	8.95%	10561	0.09	5
15_16	7.76%	9159	0.08	4
16_17	7.45%	8788	0.07	4
17_18	7.23%	8534	0.07	4
18_19	6.13%	7238	0.06	3
19_20	5.48%	6463	0.05	3
20_21	4.67%	5505	0.05	3
9_10	8.13%	9588	0.08	4
(blank)	0.00%		0	0
Grand Total	**100.00%**	**117988**	**1**	**54**

Finding IV –

Night Shift Manpower Planning-

Time bucket	Call distribution (given)	Time distribution (Call dis/30)	No. of agents required
9_10	3	0.1	2
10_11	3	0.1	2
11_12	2	0.07	1
12_1	2	0.07	1
1_2	1	0.03	1
2_3	1	0.03	1
3_4	1	0.03	1
4_5	1	0.03	1
5_6	3	0.1	2
6_7	4	0.13	2
7_8	4	0.13	2
8_9	5	0.17	3
Total	**30**	**1**	**17**

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

Analysis –

Root Cause Analysis Using the "Why's" Approach

1. Elevated Call Answer Rates During Specific Time Buckets (10–11 AM, 6–9 PM):

The increased number of answered calls during the 10–11 AM slot can be attributed to customers, particularly working professionals, making calls during their commute or upon arriving at their workplace before commencing their duties. Similarly, the 6–9 PM window aligns with post-work hours when individuals are more likely to have free time to address personal matters, including contacting customer service. These periods reflect times when customers are most available and inclined to seek assistance for minor issues that can be resolved promptly.

2. Discrepancy Between High Incoming Calls and Answered Calls in the 11–12 PM Time Bucket:

Despite a surge in incoming calls between 11 AM and 12 PM, the number of answered calls does not correspondingly peak. This disparity may stem from insufficient staffing during this period, leading to longer wait times and a higher likelihood of missed calls. Understaffing during peak hours is a common challenge in call centers, often resulting in decreased service

quality and customer satisfaction.

Challenges in Defining Exact Night Shift Agent Distribution to Maintain a 10% Abandonment Rate

Determining a precise allocation of call center agents during nighttime hours (9 PM to 9 AM) to sustain a 10% call abandonment rate presents several challenges, even when the total number of available agents is known.

1. **Dynamic Call Volume and Agent Availability**: Nighttime call volumes can fluctuate unpredictably, making it difficult to assign agents strictly based on predefined schedules. A rigid analytical approach may not accommodate sudden surges in call traffic or unforeseen agent absences.
2. **Flexible Staffing Requirements**: To address varying call volumes, agents working in adjacent time slots (e.g., 7–8 PM or 8–9 AM) may need to extend their shifts or adjust their working hours. This flexibility helps in covering peak periods and maintaining service levels but complicates the creation of a fixed staffing model.
3. **Employee Well-being and Logistical Considerations**: Factors such as the distance between an agent's residence and the workplace, availability of safe transportation during late hours, and personal health considerations play a crucial role in scheduling. Ensuring agents' safety and well-being may necessitate deviations from an analytically optimal distribution.
4. **Regulatory and Organizational Policies**: Compliance with labor laws, organizational policies on maximum working hours, and mandatory rest periods further restrict the ability to define exact staffing patterns solely based on analytical models.

Given these complexities, a hybrid approach that combines analytical tools with managerial discretion and real-time adjustments is essential for effective night shift staffing in call centers.

Conclusion –

Comprehensive Analysis and Strategic Recommendations for Call Centre

Operations

1. **Implementing a Three-Shift System for 24/7 Customer Support**: To ensure continuous customer service availability, it's advisable to structure the workforce into three distinct shifts, covering morning, afternoon, and night hours. This approach facilitates round-the-clock assistance, catering to customer inquiries at any time.
2. **Average Call Duration Insights**: The analysis indicates that the average duration of calls handled by agents is approximately 198.62 seconds (or about 3.3 minutes). This metric is crucial for assessing agent efficiency and customer engagement levels.
3. **Peak Call Duration Timeframes**: Further examination reveals that the longest average call durations occur during the 10–11 AM and 7–8 PM time slots. These periods may correspond with customers addressing issues before starting their day and after typical work hours, respectively.
4. **Shortest Call Duration Period**: Conversely, the shortest average call durations are observed between 12–1 PM, possibly due to customers making brief inquiries during their lunch breaks.
5. **Highest Call Volume Interval**: The time slot from 12 PM to 1 PM experiences the highest volume of incoming calls, indicating a need for increased staffing during this peak period to maintain service levels.
6. **Lowest Call Volume Interval**: The analysis also identifies that the fewest calls are answered between 8 PM and 9 PM, suggesting a potential opportunity to reallocate resources during this quieter period.
7. **Evening Call Volume Trends**: Overall, evening hours witness a decline in incoming calls. Therefore, the company might consider optimizing workforce allocation by reducing the number of agents scheduled during these times, thereby improving operational efficiency.
8. **Staffing Requirements to Maintain Abandonment Rate**: To achieve a target call abandonment rate of 10%, it's estimated that a total of 54 agents are required. This staffing level should be strategically distributed across various shifts to meet customer demand

effectively.

9. **Night Shift Staffing Strategies**: For the night shift (9 PM to 9 AM), the company can either hire 17 dedicated agents or reassign someday-shift employees to cover these hours. This flexibility ensures adequate coverage during all operational periods.

By implementing these strategic recommendations, the company can enhance customer service efficiency, optimize workforce utilization, and ensure consistent availability to address customer needs effectively.

Appendix

Data Analytics Process: -
---> Link for the shared PDF on Google Drive:
Instagram User Analytics: -
----> Link for the shared file on Google Drive:
Operation Analytics and Investigating Metric Spike Analysis: -
-----> Link for the shared file on Google Drive:
Hiring Process Analytics: -
----> Link for shared PDF on google drive:
IMDB Movie Analysis-
---> Link for the shared PDF on Google Drive:
Bank Loan Case Study: -
----> Link for the shared file on Google Drive:
Analyzing the Impact of Car Features on Price and Profitability: -
-----> Link for the shared file on Google Drive:
ABC Call Volume Trend Analysis: -
-----> Link for the shared file on Google Drive: