PROJECT PORTFOLIO

VANSHATA JAISWAL

TRAINITY

JAN 2024 - FEB 2024

Professional Background

Dedicated and adaptable professional with a background in .NET support applications, transitioning into the field of data analysis/science. Proficient in Python, MS Excel, SQL, Power BI, Tableau, and possessing foundational knowledge in machine learning and statistics. Eager to leverage analytical skills and a passion for continuous learning to contribute effectively to data-driven decision-making processes.

Education:

Bachelor's Degree in Information Technology
Kalinga Institute of Industrial Technology, 2017-2021

Technical Skills:

- Programming: Python, SQL
- Data Analysis Tools: MS Excel (Data Cleaning, EDA, Data Analysis), Power BI (Beginner), Tableau
- Machine Learning: Beginner level proficiency
- Statistics: Proficient in foundational concepts

Soft Skills:

- Analytical Thinking: Ability to dissect complex problems and derive meaningful insights.
- Communication: Clear and concise articulation of technical concepts to stakeholders.
- Adaptability: Willingness to learn and adapt to new tools, technologies, and methodologies.
- Collaboration: Proven track record of working effectively in cross-functional teams to achieve common goals.
- Time Management: Efficient organization and prioritization of tasks to meet deadlines effectively.

TABLE OF CONTENTS

Data Analysis Process

Organizing a Party/ Hosting a get-together

1. **Plan:** Invitations and Culinary Considerations

-Extend invitations to guests for a dinner gathering via phone calls or text messages.
Anticipating the need for preparation in both beverages and food.

2. **Prepare:** Thorough Culinary Checklist

-Creating a detailed list categorizing items suitable for starters, main courses, and desserts.

3. **Process:** Inventory Check

-Evaluate the availability of necessary items at home.
Identify and list items that need to be purchased.

4. **Analyze:** Refining Culinary Choices

-Choosing food combinations that enhance the overall dining experience.
-Consider preferences for hot or cold drinks.
Selecting the choice of vegetarian or non-vegetarian food items.

5. **Share:** Collaborative Decision-Making

-Engage in a collective discussion with the homemaker, sharing insights gathered during the analysis
phase.

6. **Act:** Culinary Execution and Hospitality

-The homemaker takes actionable steps, ensuring a warm welcome to guests with a delightful array
of thoughtfully chosen food and beverages.

Project Link:
https://drive.google.com/file/d/1JIDvTDh2Xr84O5EZlyIPzAex15prmGqo/view?usp=drive_link

Instagram User Analytics

Project Description

- To know the profit or loss of any product used by the public, we need to analyse the product's or business's usage, sales and govern its security.
- This project helps in giving insights to the team for better growth and promotion of the digital product Instagram by drawing conclusions based on few criteria from the database using MySQL

Findings:

1. Loyal User Reward- The 5 oldest users of the Instagram.
2. Inactive User Engagement -The users who have never posted a single photo on Instagram.
3. Contest Winner Declaration - The user who gets the most likes on a single photo.
4. Hashtag Research - The top 5 most commonly used hashtags.
5. Ad Campaign Launch - Day of the week do most users register on.
6. User Engagement - Average user posts on Instagram.
7. Bots & Fake Accounts - Users (bots)who have liked every single picture on the site.

Approach

The database tables creation query existed in the documents. So, a database named ig_clone was created under which there were 7 tables.

1. Users
2. photos
3. comments
4. likes
5. follows
6. tags
7. photo_tag

After creation, the data was inserted into the respective tables. Now, the queries were executed based on the requirements of the company.

Insights

- Every table is linked with one another using foreign keys.
- Primary keys have also been used which are auto incremented in few tables while others have used a set of attributes for a primary key.
- We now know which user is linked to which photo and his number of likes or comments.
- There are total of 100 users using Instagram clone.
- There are 26 users who have never posted a photo. The management should understand user behaviour and Analyze the profiles of these users
- Is he a genuine user or not, his followers and followee, or when he had joined this social media (Instagram) and the tags that he created along with when was it made.
- We found that Zack_Kemmer93 with user id 52 has got highest number of likes
- We found that the most commonly used #hastags# are smile.
- We found that Sunday and Thursday are the days when most of the users were registered. The possible reasons could be holiday. However among Sunday and Thursday, it is more beneficial to launch Ads campaign on Sunday since most users would be having a holiday.
- We found that out of 100 users , there are 13 users who have liked all the 257 photos which any normal user would not be able to do and are therefore fake.

Project Link:
https://drive.google.com/file/d/1CsPDAHuS3sHfsnEL4AqPn9VIPkiopgW0/view?usp=drive_link

OPERATION ANALYTICS & INVESTIGATING METRICSPIKES

Project Description

- This project explains operations of a company for its better growth.
- It analysis various aspects of the company to work on areas of improvement.
- As a data analyst, we can deep dive into the database of the company and bring out some insights for the company's future.
- Investigating metric spike is also an important part of operation analytics. We must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement?
- Why have sales taken a dip? Etc. Questions like these must be answered daily and for that it's very important to investigate metric spike.

Findings

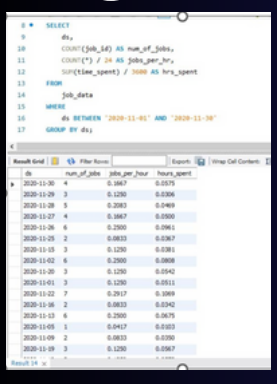
We are going to find few insights from the data like-

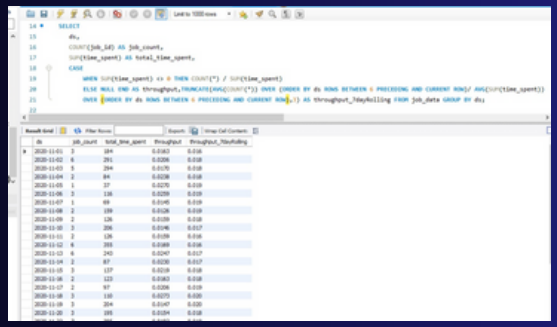
- Number of jobs reviewed.
- Calculating the percentage share of each language over a period
- Analyzing the throughput using window function
- User engagement in a week
- Engagement per device on a weekly basis and more.

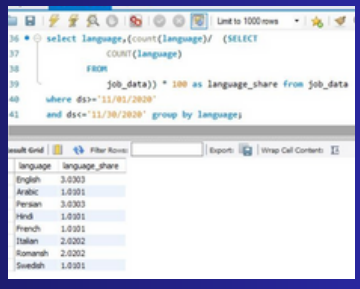
Approach

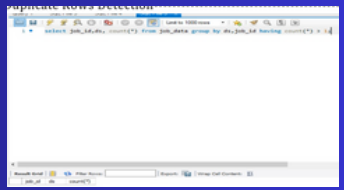
- I have created the needed database (Project3) and tables in MySQL workbench and
- imported the entire csv file in the database as tables so that we could analyze the
- data.
- We write MySQL queries to extract information that is needed for better analysis as
- the data is huge.
- The tables that are created:
    1. Job_Data
    2. Events
    3. Email_Events
    4. Users

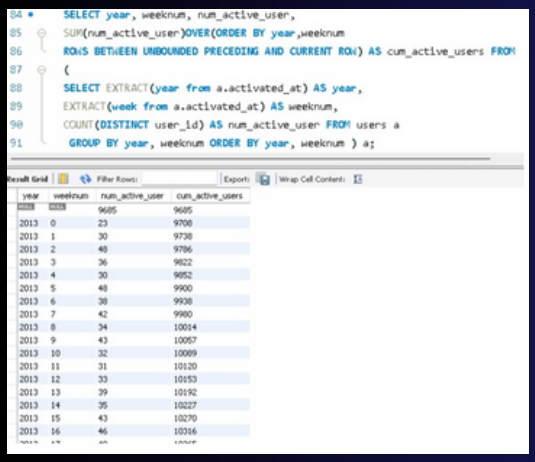
Insights

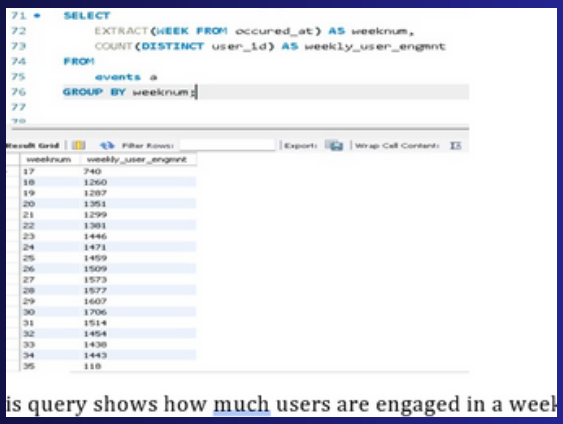




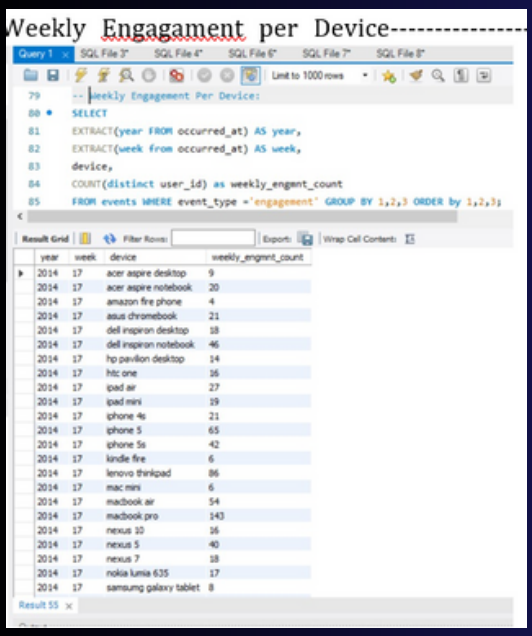


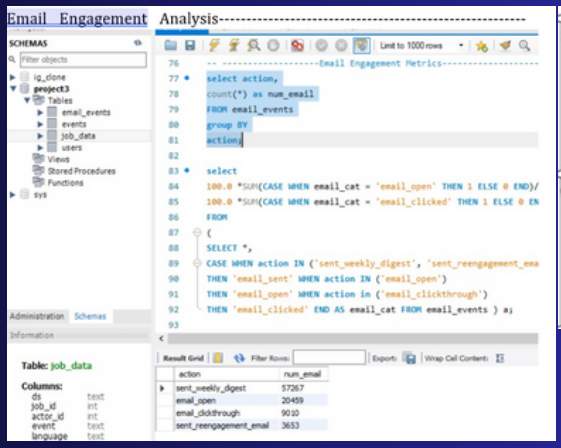


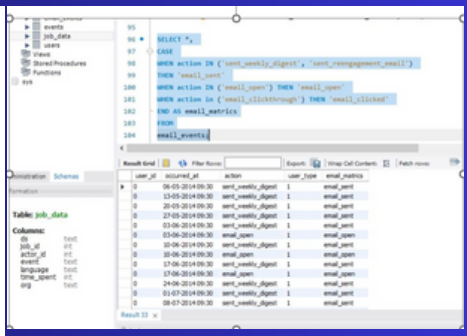




is query shows how much users are engaged in a week







- The number of distinct jobs reviewed per hour per day for November 2020 is 83%.
- We used the 7-day rolling average of throughput as it gives the average for all the days right from day 1 to day 7 whereas, daily metric gives the average for only that particular day itself.
- The percentage share of Persian and English language is the most.
- There is no duplicacy of rows.
- The overall count of weekly engagement per device used is the most for MacBook users and iPhone users.
- The email opening rate is around 34% and email clicking rate is around 15%.

Project Link:

https://drive.google.com/file/d/1thXXJ5C1zB2h6fUmibeESdmnZOm42vco/view?usp=drive_link

HIRING PROCESS ANALYTICS

Project Description

- Hiring process is very important for a company and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department.
- So, the task is to analyse the given data so that we could provide better insights and make this process easier.

Approach

- I am working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hiring and have asked me to answer certain questions making sense out of that data.
- For this we will clean the data, by clearing the missing data, detecting the outliers.
- We create charts and calculate few values for this purpose.

Findings

The following are found using the excel functions:

- Hiring analysis based on gender.
- Average Salary Analysis
- Distribution of salaries
- Department Analysis
- Position tier analysis

Insights

- Males are hired more than females employees.
- The least average salary is for Marketing Department i.e. 48349.37 and highest for General department
- Maximum % of people are in the operations department.
- Production, purchase and marketing have 5% people working
- The maximum number is 767 which is under 40800-50799 interval.
- Most employees were employed on c9 post and then in c5
- n10,n6,n7 and M7 have only 1 person each.

Project Link:

https://drive.google.com/drive/folders/1-Vm6iigafLSiONPMEXle6rQj2qAHZklF?usp=drive_link

IMDB ANALYSIS

Project Description

- This project is about conducting a comprehensive analysis of IMDb data to gain
- insights into the trends, patterns, and factors influencing movie ratings, profitability,
- and the performance of directors and genres.
- • By leveraging statistical and data visualization techniques, we aim to provide a
- deeper understanding of the dynamics within the film industry.

Approach

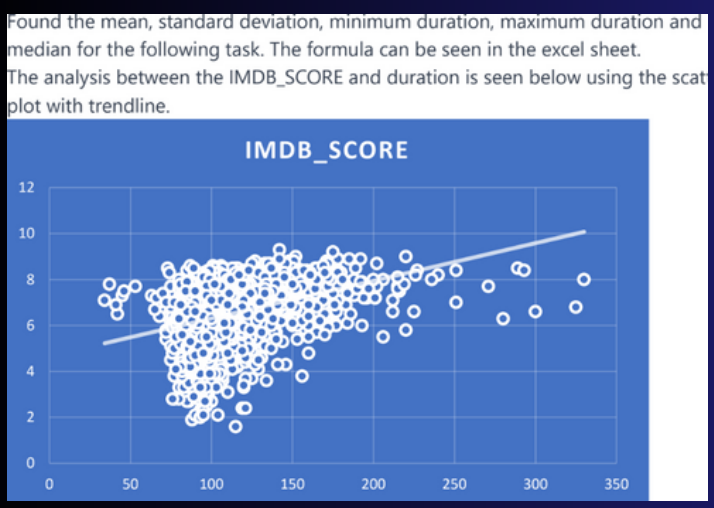
- We will clean the data by removing blanks, irregular patterns in any column and
- unwanted columns. Here we are handling missing and inconsistencies in data. Then
- use excel functions to gain insights from the data
- we used Five 'Whys' approach to determine its root cause by repeatedly asking the question “Why”. While asking Why is easy, what we're interested in is the answer. Each time we answer why the next time gets more difficult as we must think deeper behind the reasons for this. As we ask why, we may find that we have multiple answers for the same question.

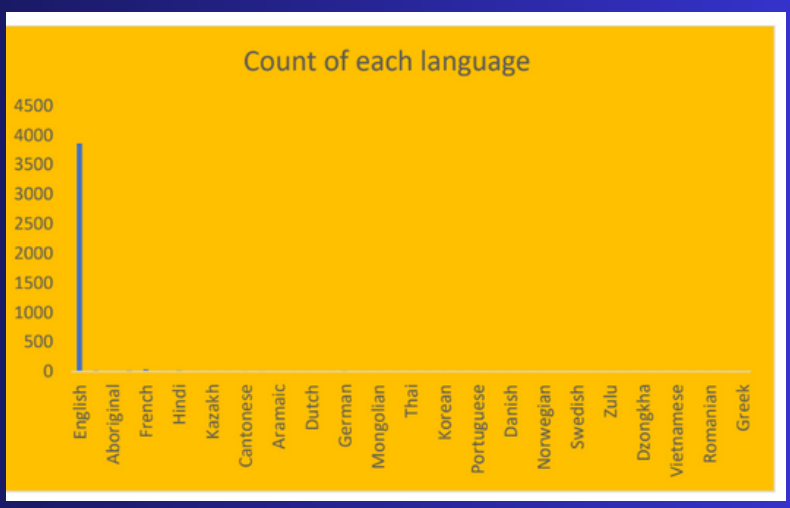
Findings

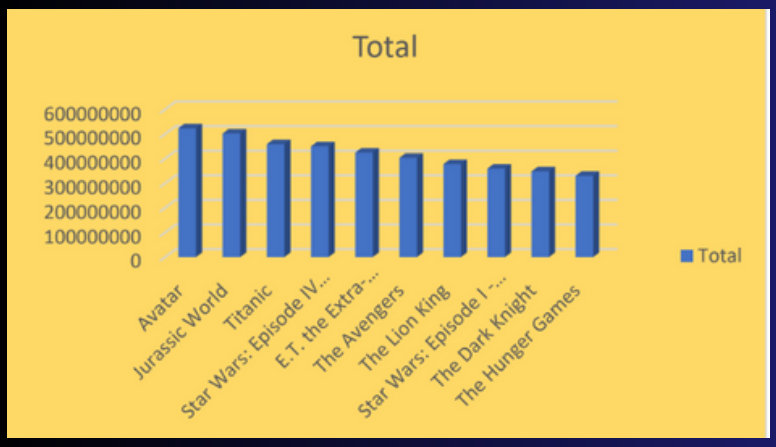
There are 5 tasks provided:
• Movie Genre Analysis
• Movie Duration Analysis
• Language Analysis
• Director Analysis
• Budget Analysis

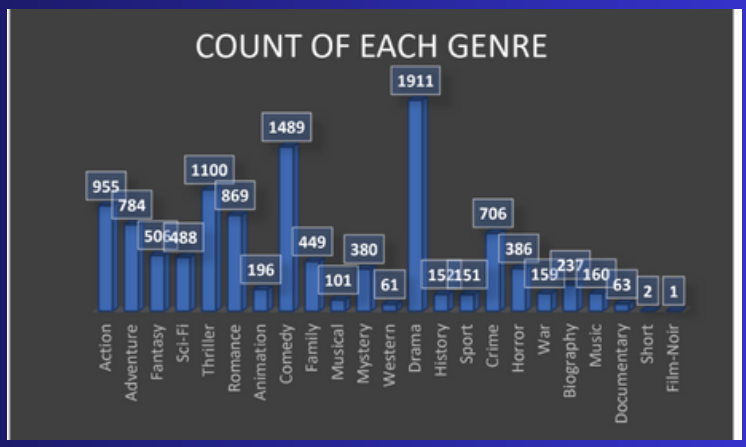
Insights

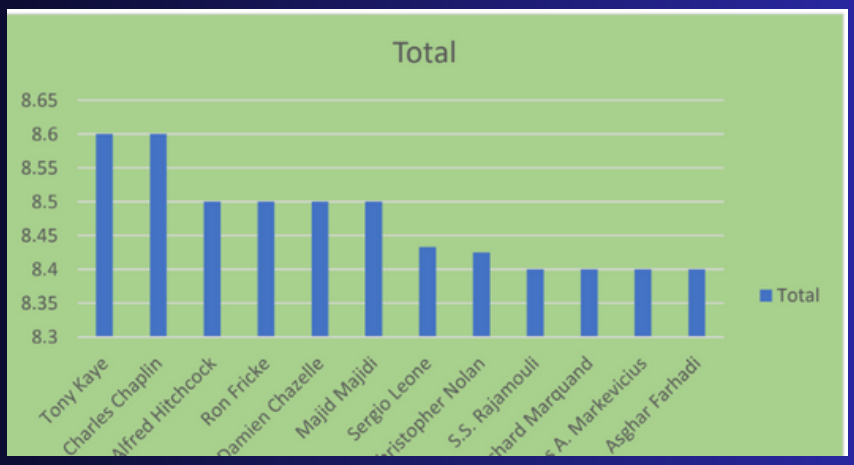
- The movie with the highest profit is ‘Avatar’.
- We analyze that the most popular genre is Drama.
- English is the most used language
- We found that movie with duration 100-150 minutes has maximum number of movies
- Tony Kaye and Charles Chaplin are the most famous directors











Project Link:
https://drive.google.com/drive/folders/1IyLd8PELJq4xyNZLN4WIJvK2jt86qpMa?usp=drive_link

BANK LOAN CASE STUDY

Project Description

- This project aims to give you an idea of applying EDA in a real business scenario.
- In this case study, apart from applying the techniques that you have learnt in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- When a client applies for a loan, there are 4 scenarios:
- 1. Approved
- 2. Cancelled
- 3. Refused
- 4. Unused offer

Approach

- The data is quite huge with around 30lakhs rows and 122 columns in application_data which is the
- main file. So we need to check for the missing data and outliers in the dataset. This csv file contains
- all the information of the client at the time of application. The data is regarding if the client has
- difficulty in paying the loan. I have also made use of pivot tables extensively.
- The second dataset is previous_application which contains information about the client's previous
- loan data. It tells us whether the previous application had been Approved, Cancelled, Refused or
- Unused offer.
- Columns_description.csv describes the meaning of each variable.

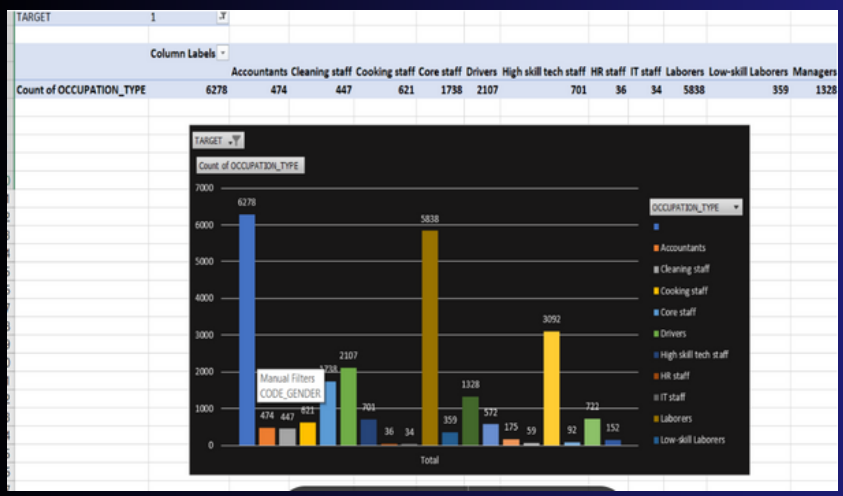
Findings

- Identify Missing Data and Dealing with it Appropriately.
- Identify Outliers in the Dataset.
- Determining if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- Determine the univariate, segmented univariate and bivariate analysis
- Identify Top Correlations for Different Scenarios

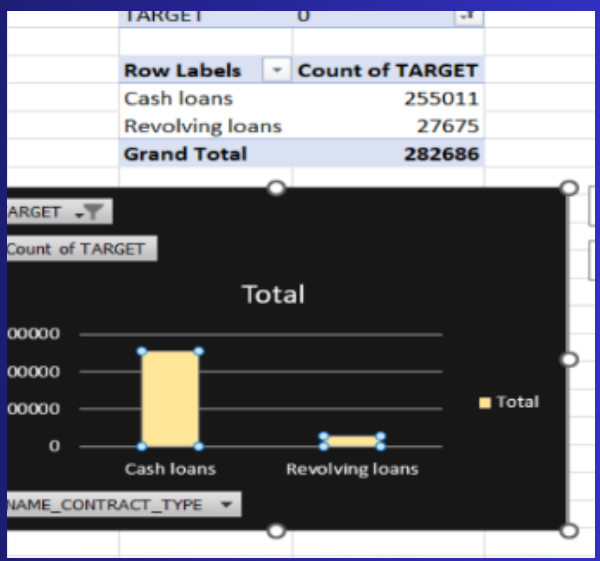
Insights

- As the age and experience increases, the chances of defaulting increases
- Educated clients tend to default less as compared to people who are less educated
- Corporate cleints are a safer choice as compared to labour class
- Male clients tend to default more
- As the age increases, the amount taken by the clients is higher.

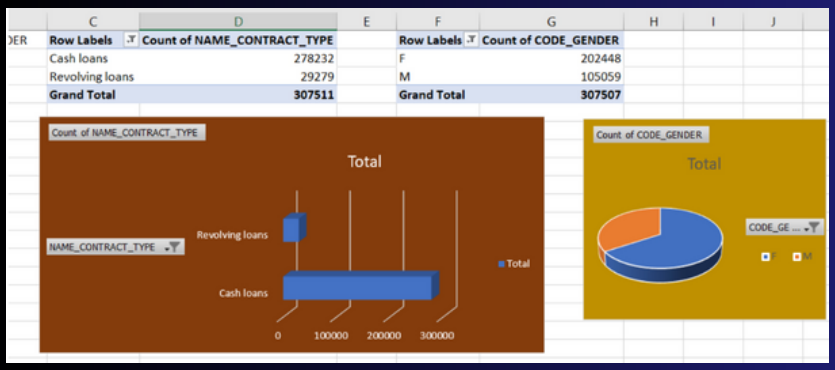
Bivariate



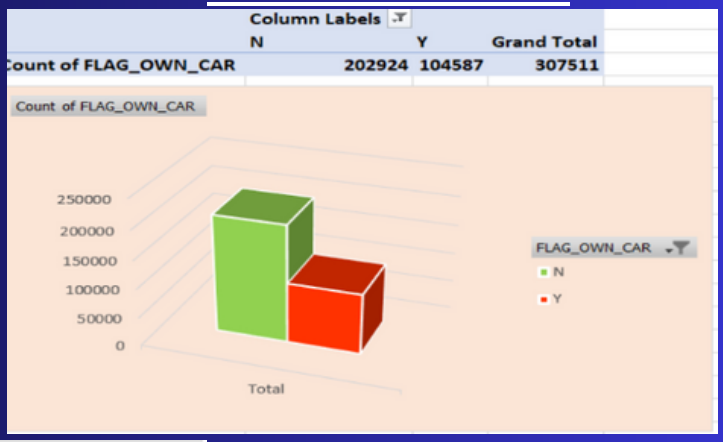
segmented univariate



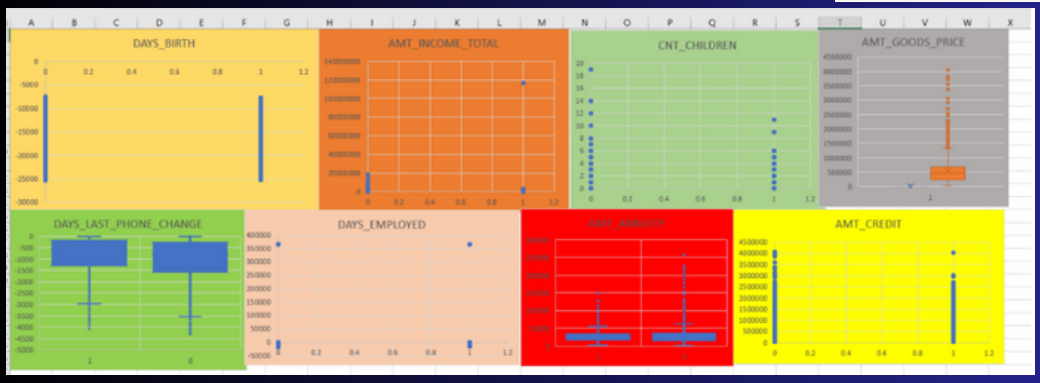
Univariate



Data Imbalance



Outliers



Project Link:
https://drive.google.com/drive/folders/1OAed94jNafE_5d5kUpenjoedTeOzDuR5?usp=drive_link

ANALYZING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

Project Description

- The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation.
- With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.
- Overall, this dataset could be a valuable resource for data analysts interested in exploring various aspects of the automotive industry and could provide insights that could informed decisions related to product development, marketing, and pricing.

Approach

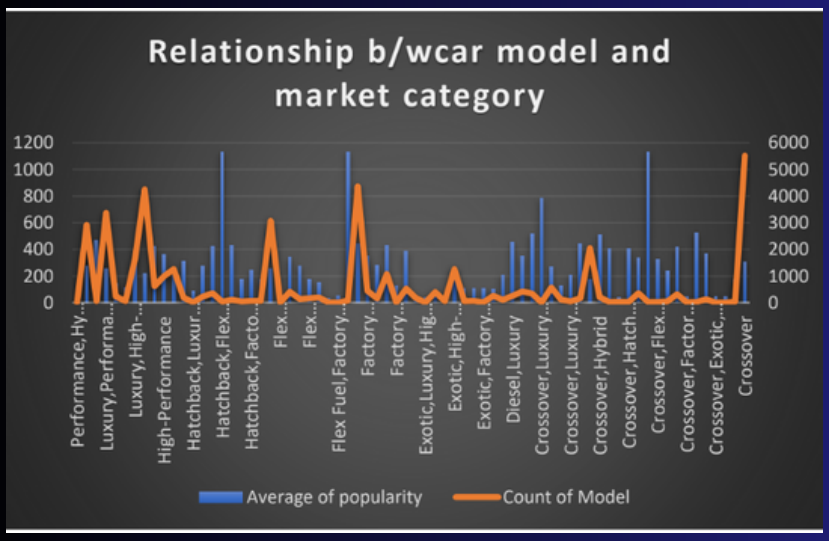
- For analytics, I have used MS-excel to perform various functions like pivot table, graph,
- regression. I will also use power BI to analyse the charts. I have also created interactive
- dashboards using slicers. I have also cleaned the data by removing blanks and duplicate data.

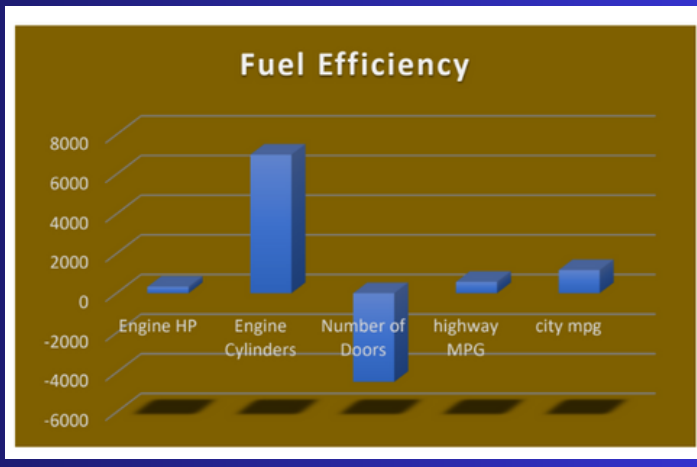
Findings

- How does the popularity of a car model vary across different market categories?
- What is the relationship between a car's engine power and its price?
- Which car features are most important in determining a car's price?
- How does the average price of a car vary across different
- manufacturers?
- What is the relationship between fuel efficiency and the number of
- cylinders in a car's engine?
- How does the distribution of car prices vary by brand and body style?
- Which car brands have the highest and lowest average MSRPs and how does this vary by body style?
- How do the different features such as transmission type affect
- the MSRP, and how does this vary by body style?
- How does the fuel efficiency of cars vary across different body
- styles and model years?
- How do the car's horsepower, MPG, and price vary across
- different Brands?

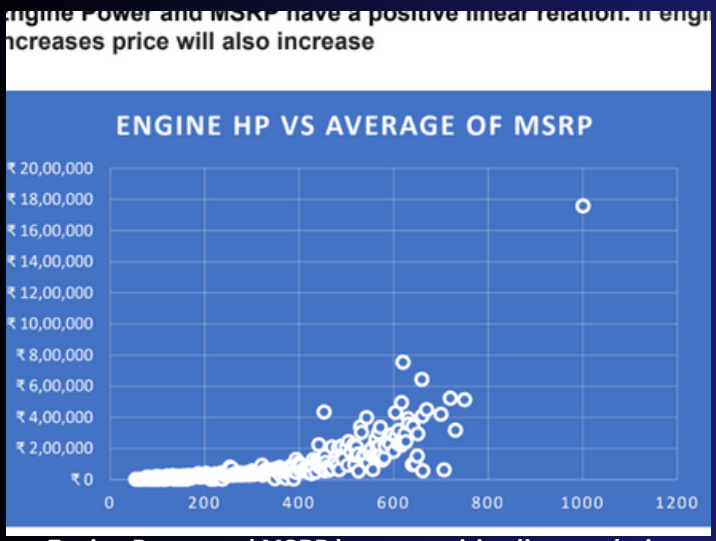
Insights

- Crossover category has the greatest number of models whereas the popularity average is more for 3 market categories with 5657 number.

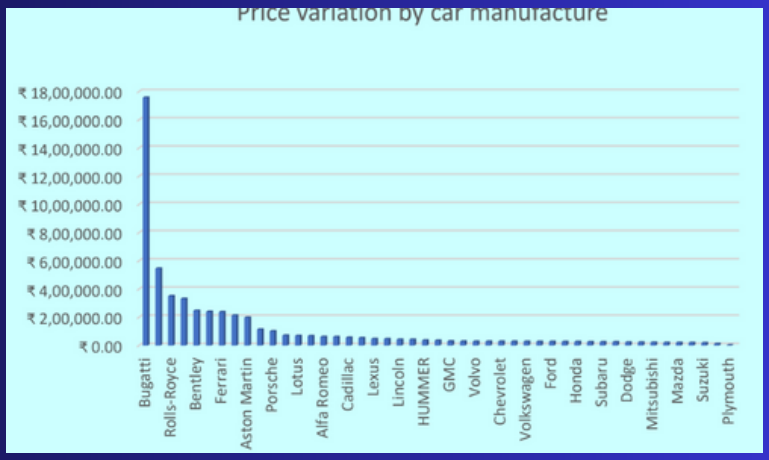




Engine Cylinder has the highest positive coefficient in determining MSRP.

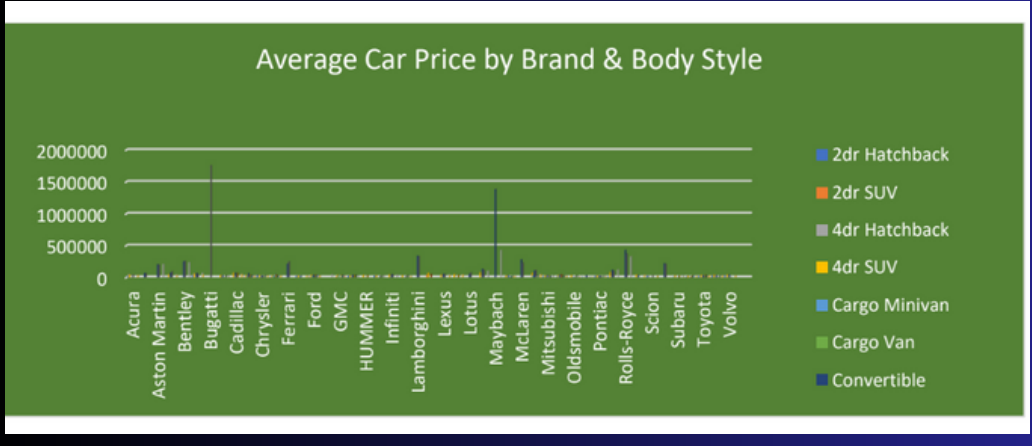


- **Engine Power and MSRP have a positive linear relation. If engine power increases price will also increase.**

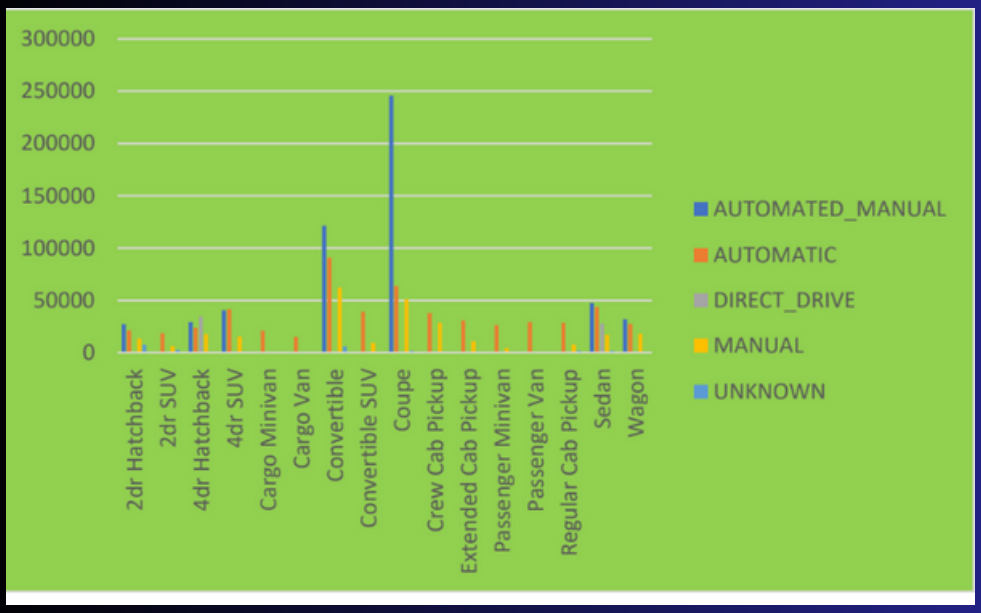


Buggati is the most expensive car sold

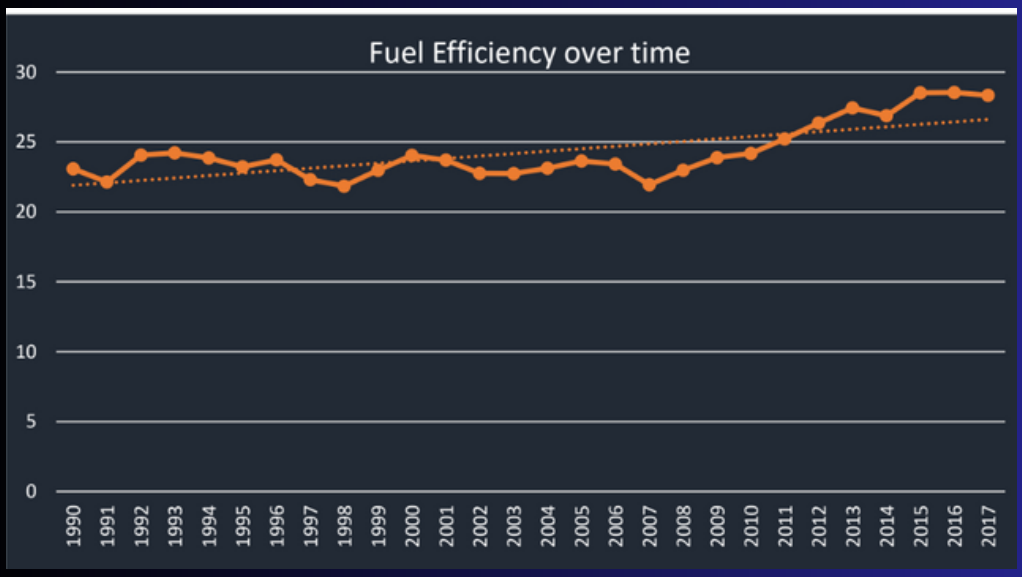
- **Chevrolet has the highest price distribution.**



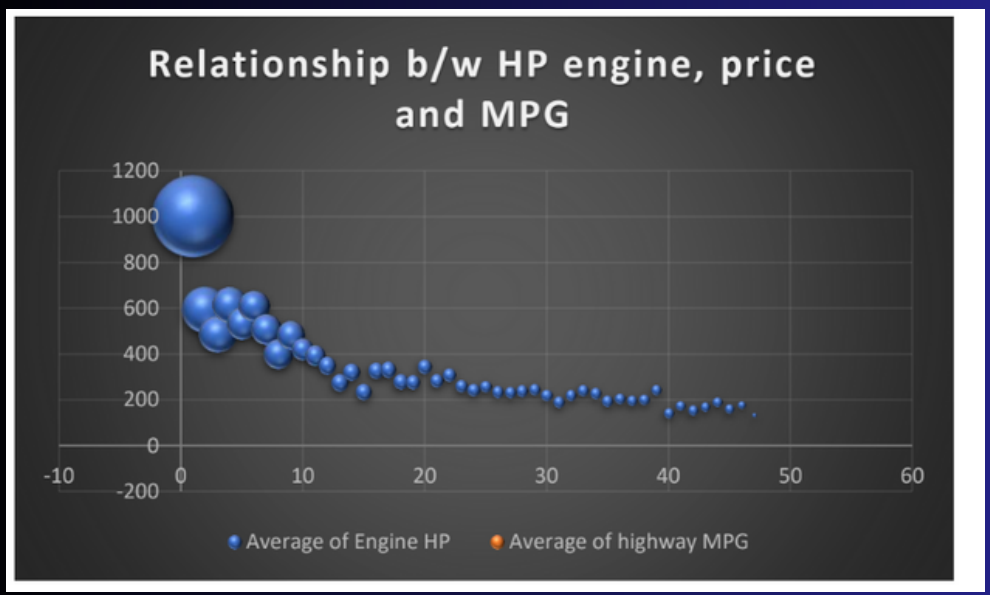
- **Bugatti has the highest MSRP and Plymouth has the lowest Average MSRP**



- **AutomatedManual is the most expensive category and the most popular also.**



- **Over the year fuel efficiency is increasing at a slow speed.**



- **If engine HP increases, highway mpg will decrease, and the price will also increase.**

Project Link: https://drive.google.com/drive/folders/1IyLd8PELJq4xyNZLN4WIJvK2jt86qpMa?usp=drive_link

ABC CALL VOLUME TREND ANALYSIS

Project Description

- The project analyzes a dataset of an insurance company's Customer Care team for 23 days.
- The dataset includes information about agents, call durations, and statuses. Data includes
- Agent_Name, Agent_ID, Queue_Time , Time, Time_Bucket, Duration, Call_Seconds, call
- status (Abandon, answered, transferred).

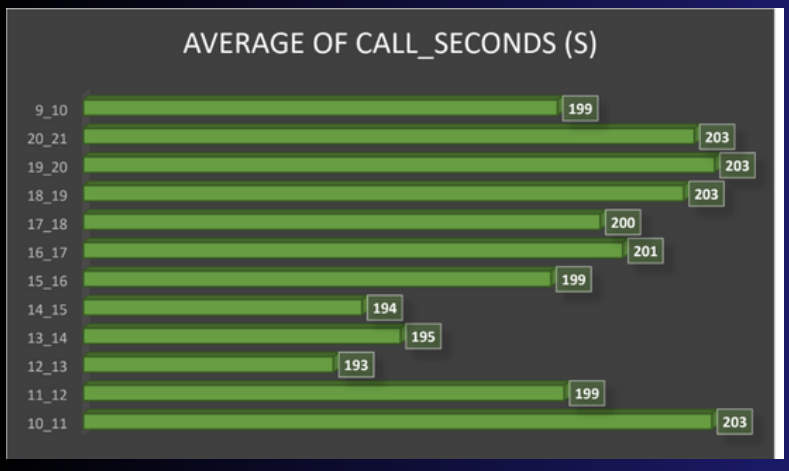
Approach

- There are 13 columns and 117989 rows. We are not planning on deleting extra columns as
- They do not tend to disturb our analysis.
- I have created the charts and done a few calculations to understand the data and draw some insights.

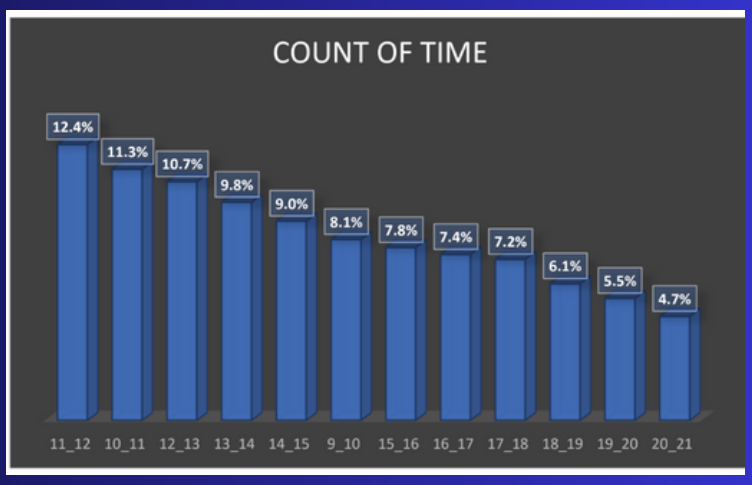
Findings

- The average call time duration for all incoming calls received by agents (in each Time_Bucket).
- The total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time].
- Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.
- Propose a manpower plan required during each time bucket in a day[9 pm to 9 am].
- Maximum Abandon rate assumption would be same 10%.

Insights

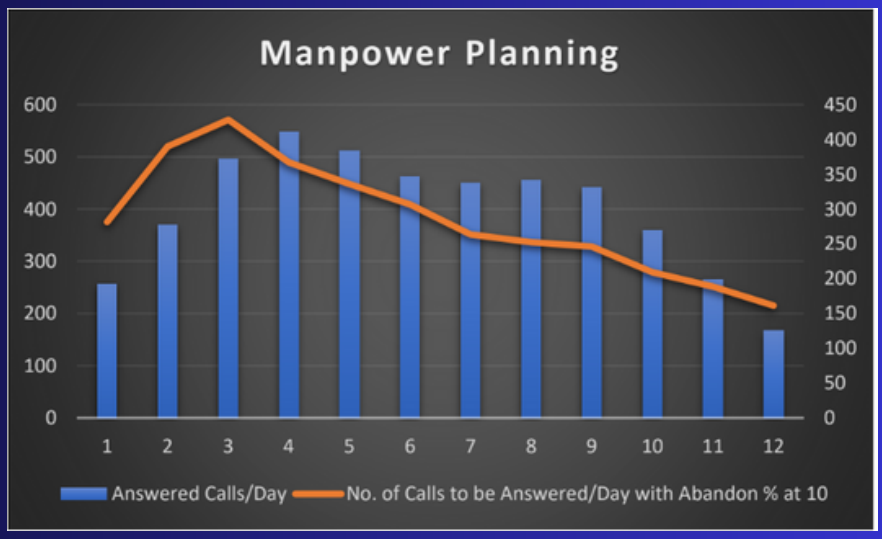


Average call time is 199 seconds. I have created a pivot table and added a filter so that the data can be filtered according to the call status.



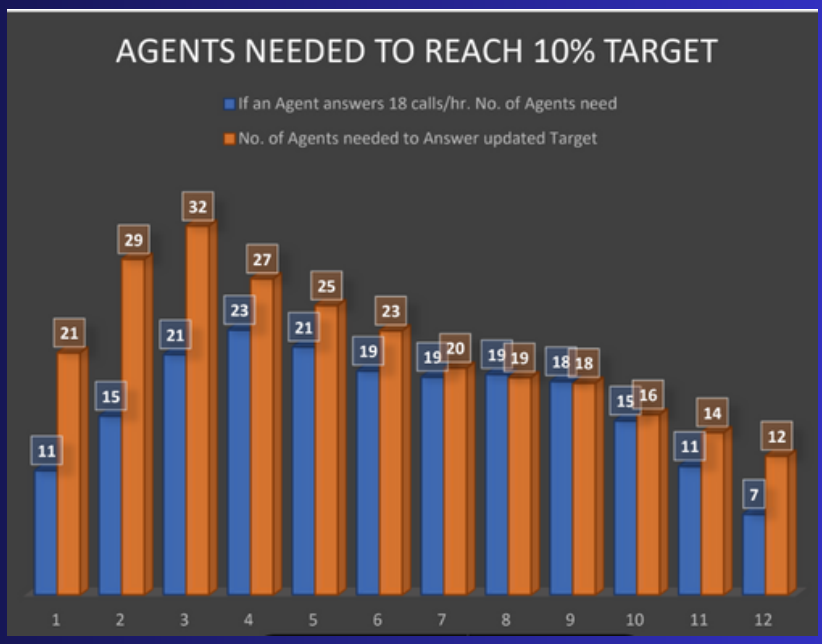
Most number of calls are in 11_12 interval that is 12.4% of the total number of calls.

Time_Bucket	Answered Calls/Day	If an Agent answers 18 calls/hr. No. of Agents need	No. of Calls to be Answered/Day with Abandon % at 10
09_10	193	11	375
10_11	277	15	520
11_12	372	21	571
12_13	410	23	489
13_14	384	21	448
14_15	347	19	409
15_16	337	19	351
16_17	341	19	336
17_18	330	18	328
18_19	270	15	279
19_20	199	11	251
20_21	125	7	215



If an Agent answers 18 calls/hr. No. of Agents need	No. of Agents needed to Answer updated Target
11	21
15	29
21	32
23	27
21	25
19	23
19	20
19	19
18	18
15	16
11	14
7	12

Here we see the comparison of current calls Vs the updated call numbers per day for each time bucket as per the new Abandon Rate. Based on which we'd be calculating the number of Agents needed in each time bucket.



As you can see, the current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%.

Here we see the comparison of current calls Vs the updated call numbers per day for each time bucket as per the new Abandon Rate. Based on which we'd be calculating the number of Agents needed in each time bucket.

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)

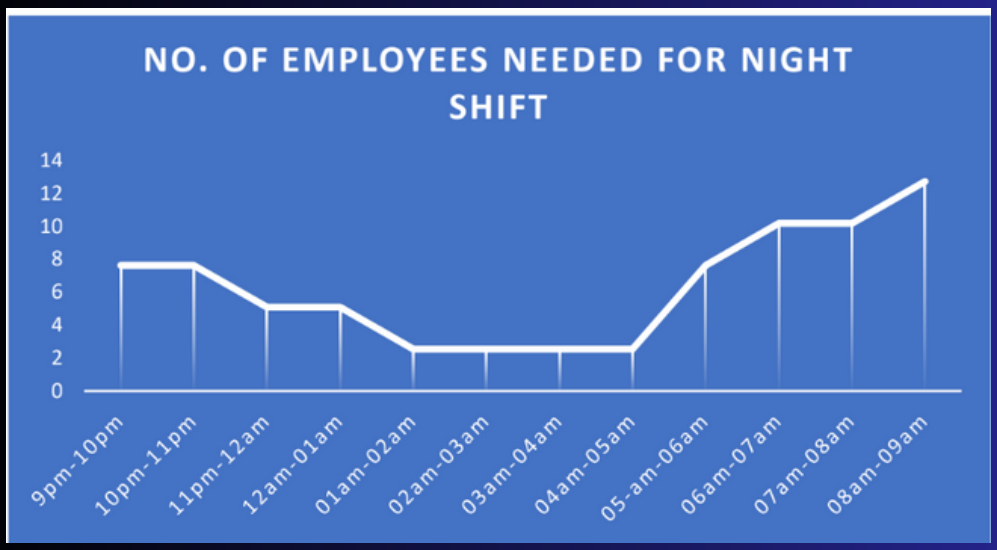
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

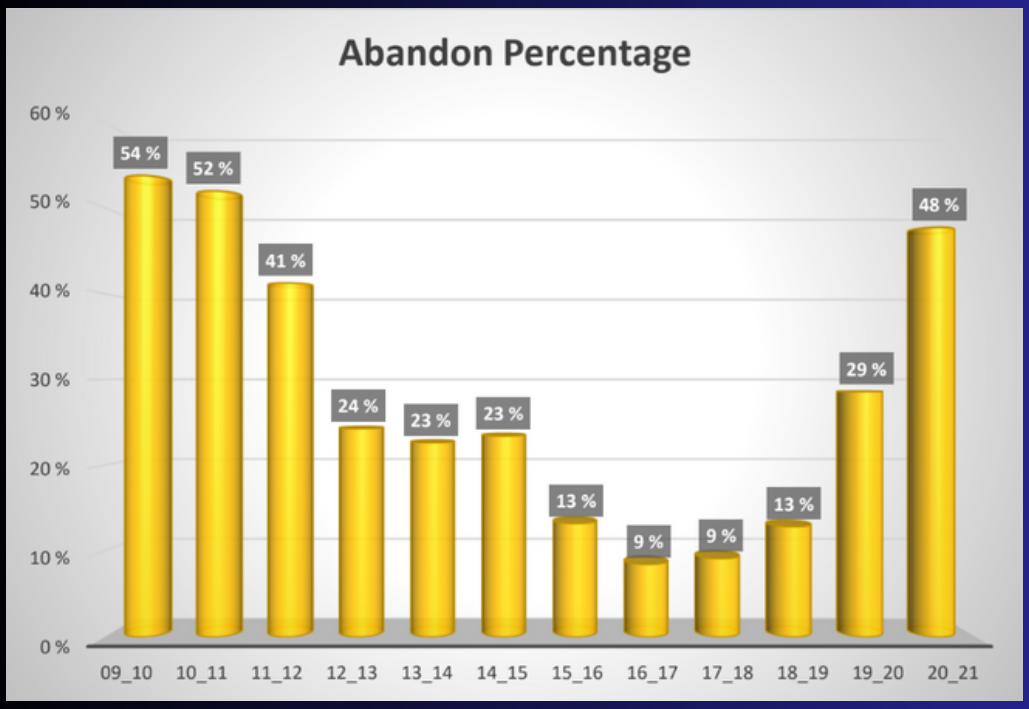
Time_Slot	Distribution of 30 calls	Percentage Distribution
9pm-10pm	3	0.1000
10pm-11pm	3	0.1000
11pm-12am	2	0.0667
12am-01am	2	0.0667
01am-02am	1	0.0333
02am-03am	1	0.0333
03am-04am	1	0.0333
04am-05am	1	0.0333
05-am-06am	3	0.1000
06am-07am	4	0.1333
07am-08am	4	0.1333
08am-09am	5	0.1667
Total	30	1

Time_Bucket	No. of Calls to be Answered/Day with Abandon % at 10
09_10	375
10_11	520
11_12	571
12_13	489
13_14	448
14_15	409
15_16	351
16_17	336
17_18	328
18_19	279
19_20	251
20_21	215
Total Calls on an average/day	4573
Total Calls on an average/night	1372

This is from previous question.

Let's say customers also call this ABC insurance company in night but didn't get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that a customer made during 9 Am to 9 Pm, customer also made 30 calls in night between interval [9 Pm to 9 Am].





- Total Calls answered with 10% abandon rate on an average/day:4573
- 30% of calls received during the day are received during night
- Total Calls answered on an average/night: 4573*0.3 = 1372
- Multiplying the Total Calls with percentage distribution, we get the total calls expected/time slot.
- As we know the number of calls an Agent can answer/hr. is 18, dividing the No. of Calls by 18, we get the number of agents required.

- Percentage distribution Calculations: Distribution of 30 calls/30 * 100 gives percentage distribution

Project link: https://drive.google.com/drive/folders/1YjhiK-jBT_MTy3lVoxJWmCUXa5tM9ide?usp=drive_link