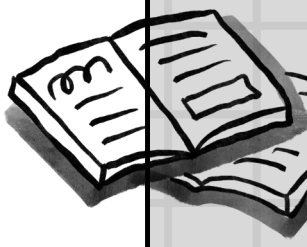
Data Analysis Portfolio

- Hitisha Soni

Professional Background



Currently a final year student, pursuing Bachelor of Computer Application (BCA), secured 8.8 AGPA in 2nd Year. Skilled in programming languages (Python, Java), data visualization tools (Power BI, Tableau), and database management (MySQL), Front-end (HTML, CSS, SASS/SCSS, Bootstrap). Strong communication skills with a commitment to delivering high-quality, data-driven solutions.

I have also published a research paper titled- "An elaborate study on cybersecurity: a review on cyber-threat and its protection" in a journal and have worked on several projects related to web development, data analysis.

Detail-oriented and results-driven Data Analyst with a strong foundation in computer applications. Proven expertise in extracting, analyzing, and interpreting complex data sets, identifying trends, to drive business decision-making. Possess a keen ability to translate data into actionable insights, along with exceptional technical and communication skills.

As a Fresher, My enthusiasm for problem-solving, adaptability, and a strong work ethic make me an ideal candidate for entry-level role in data analytics, web development or related areas. I am committed to continuous learning and growth and look forward to making a positive impact in the IT industry.

Table of Contents

ABC Call Volume Trend

Description:-

In this project, you'll be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. You'll be provided with a dataset that spans 23 days.

A Customer Experience (CX) team plays a crucial role in a company. They 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, among others.

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.

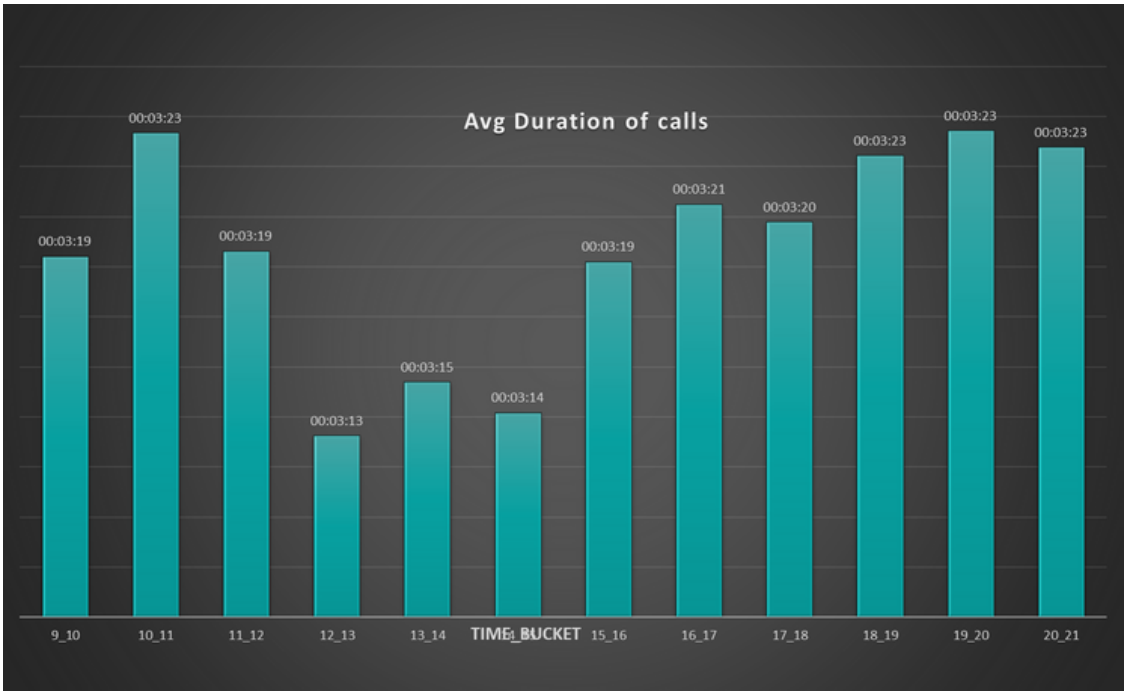
Tech-Stack Used:-

ABC Call Volume Trend

Data Analytics Tasks:-

1). 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?

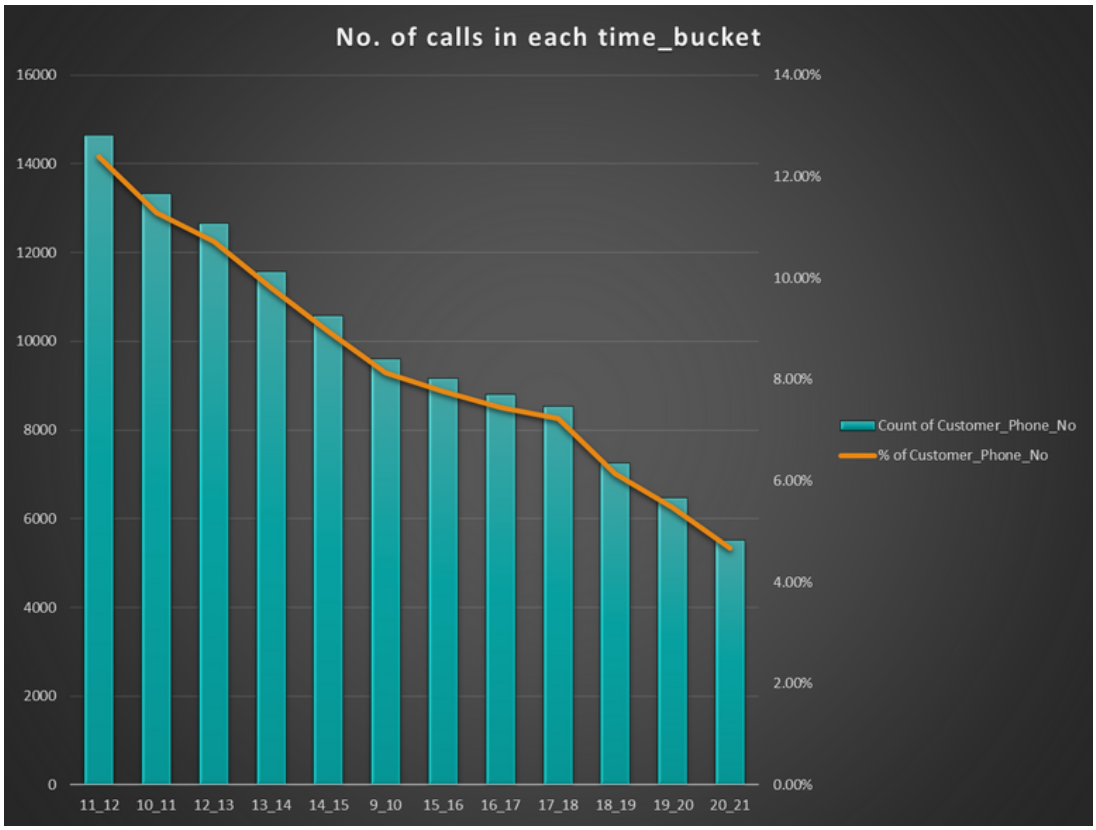


ABC Call Volume Trend

Data Analytics Tasks:-

2). 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 (e.g., 1-2, 2-3, etc.).

Your Task: Can you create a chart or graph that shows the number of calls received in each time bucket?



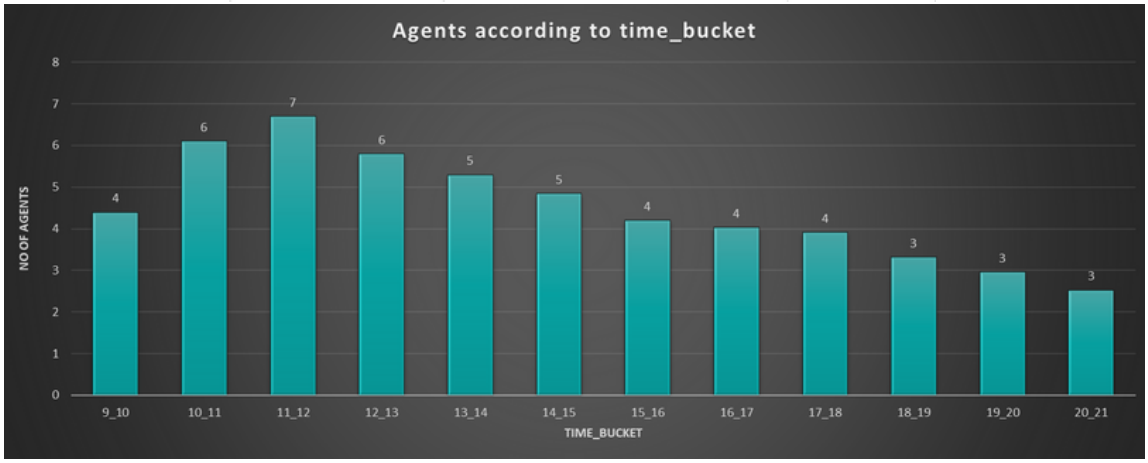
ABC Call Volume Trend

Data Analytics Tasks:-

3). 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.

Your Task: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

Time_bucket	Count of Customer_Phone_No	No. of agent
10_11	11%	6
11_12	12%	7
12_13	11%	6
13_14	10%	5
14_15	9%	5
15_16	8%	4
16_17	7%	4
17_18	7%	4
18_19	6%	3
19_20	5%	3
20_21	5%	3
9_10	8%	4
Grand Total	100.00%	54



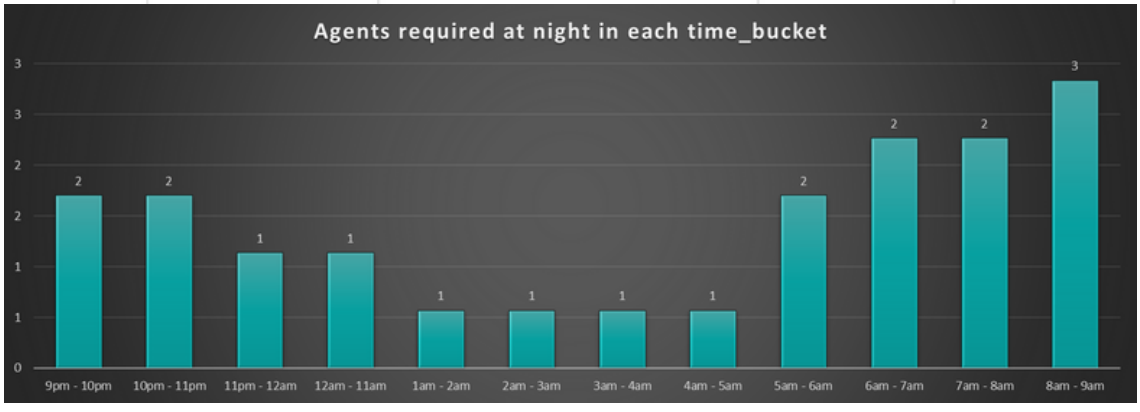
ABC Call Volume Trend

Data Analytics Tasks:-

4). 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. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am. The distribution of these 30 calls is as follows:

Your Task: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

Time	calls distribution 30 out of 100	Agents at night
9pm - 10pm	3	2
10pm - 11pm	3	2
11pm - 12am	2	1
12am - 11am	2	1
1am - 2am	1	1
2am - 3am	1	1
3am - 4am	1	1
4am - 5am	1	1
5am - 6am	3	2
6am - 7am	4	2
7am - 8am	4	2
8am - 9am	5	3
Grand Total	30	17



ABC Call Volume Trend

Results & Insights:-

1. Average duration of calls in grand total is 03:19(hh:mm:ss) or 199 seconds.
2. Highest no. of calls were between 11am-12pm, that is, 14626 which is 12.40% of grand total.
3. The minimum number of agents required in Day in each time bucket to reduce the abandon rate to 10% is 54.
4. The minimum number of agents required in Night is 17.

Requirements for 10% abandon rate for a day	
Agents at day	54
Agents at night	17
Total	71

Impact of Car Features

Description:-

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation.

With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars.

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

Tech-Stack Used:-

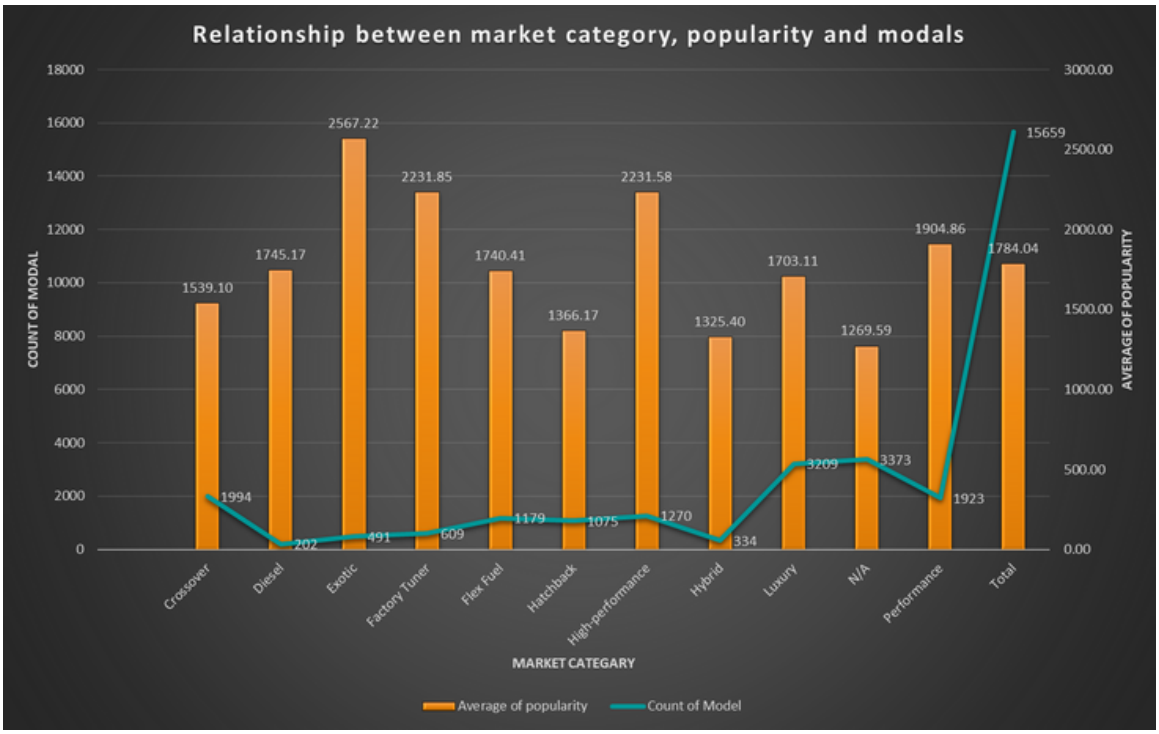
Impact of Car Features

Data Analytics Tasks:-

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.

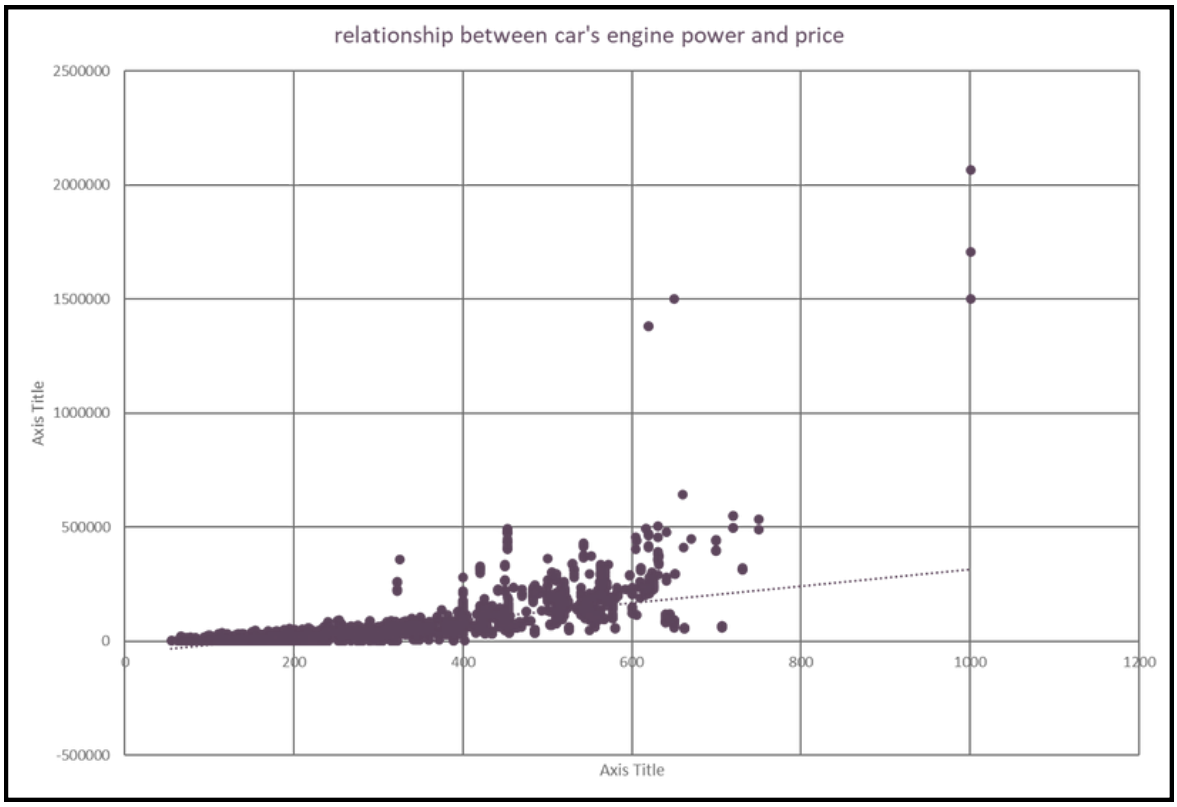


Impact of Car Features

Data Analytics Tasks:-

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.

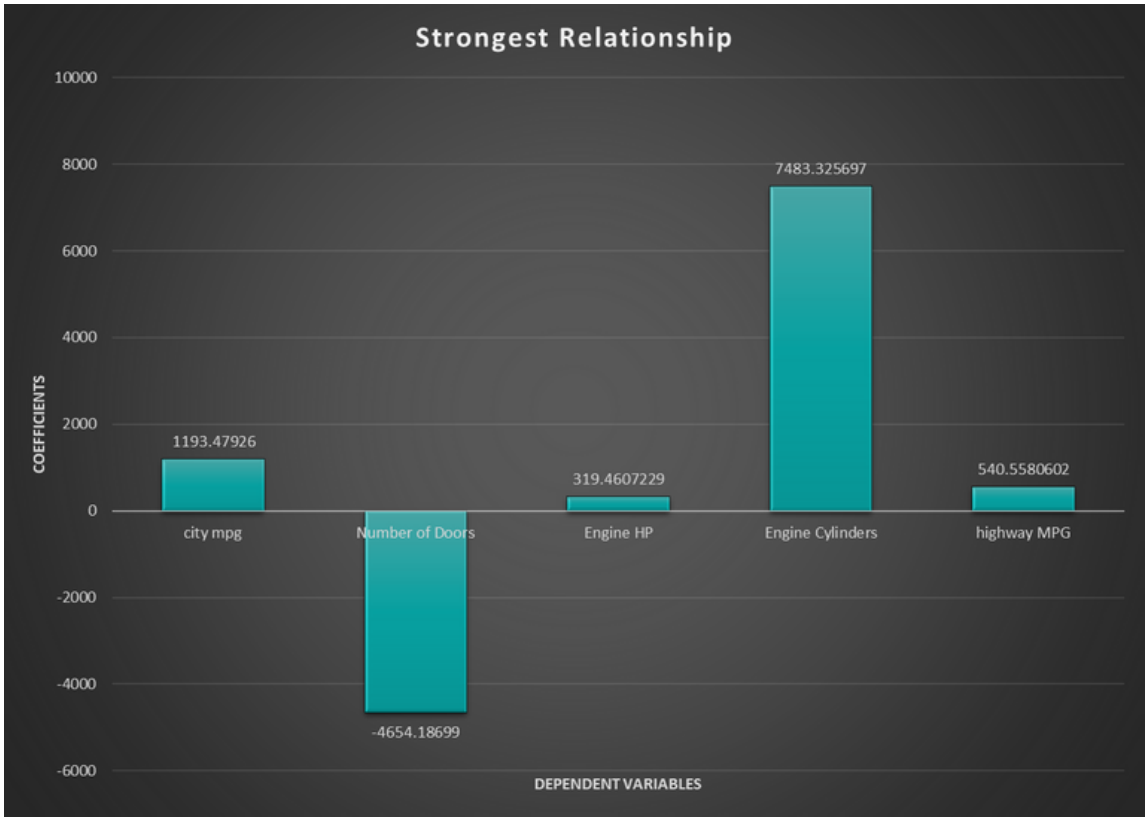


Impact of Car Features

Data Analytics Tasks:-

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.



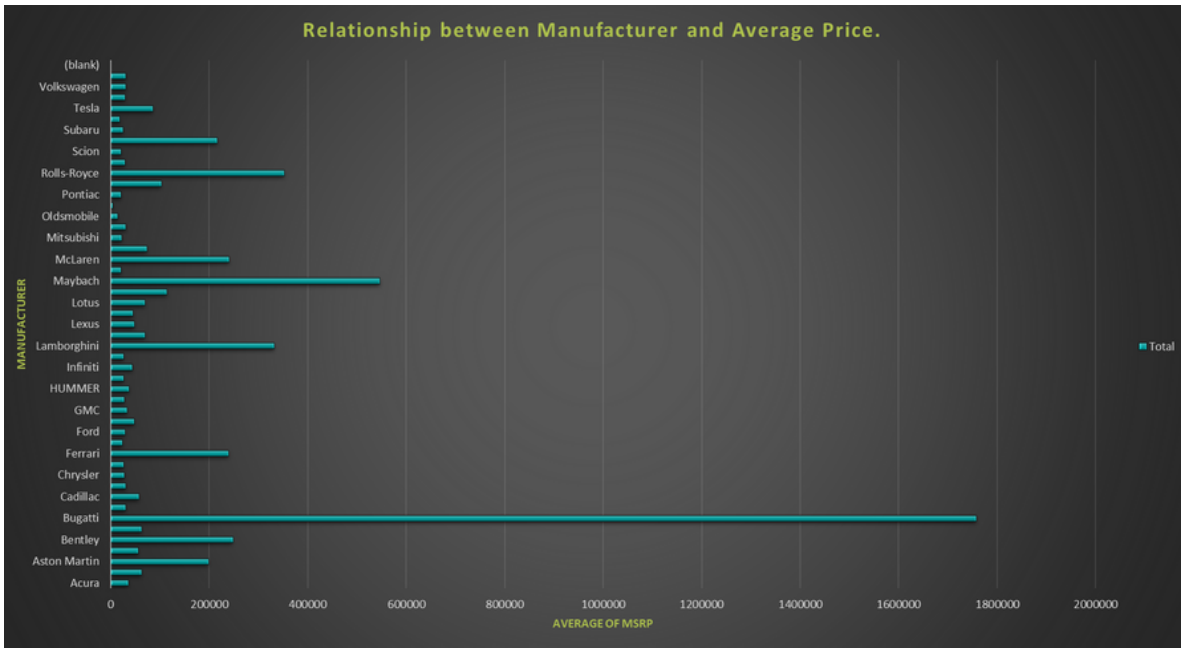
Impact of Car Features

Data Analytics Tasks:-

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.



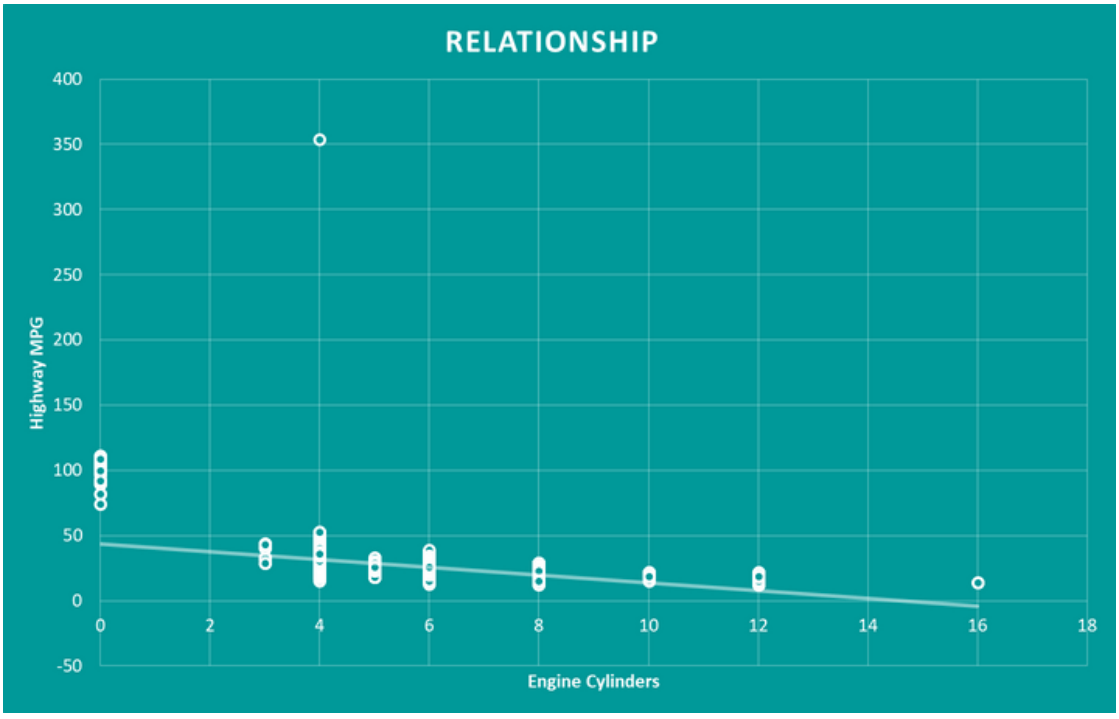
Impact of Car Features

Data Analytics Tasks:-

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.

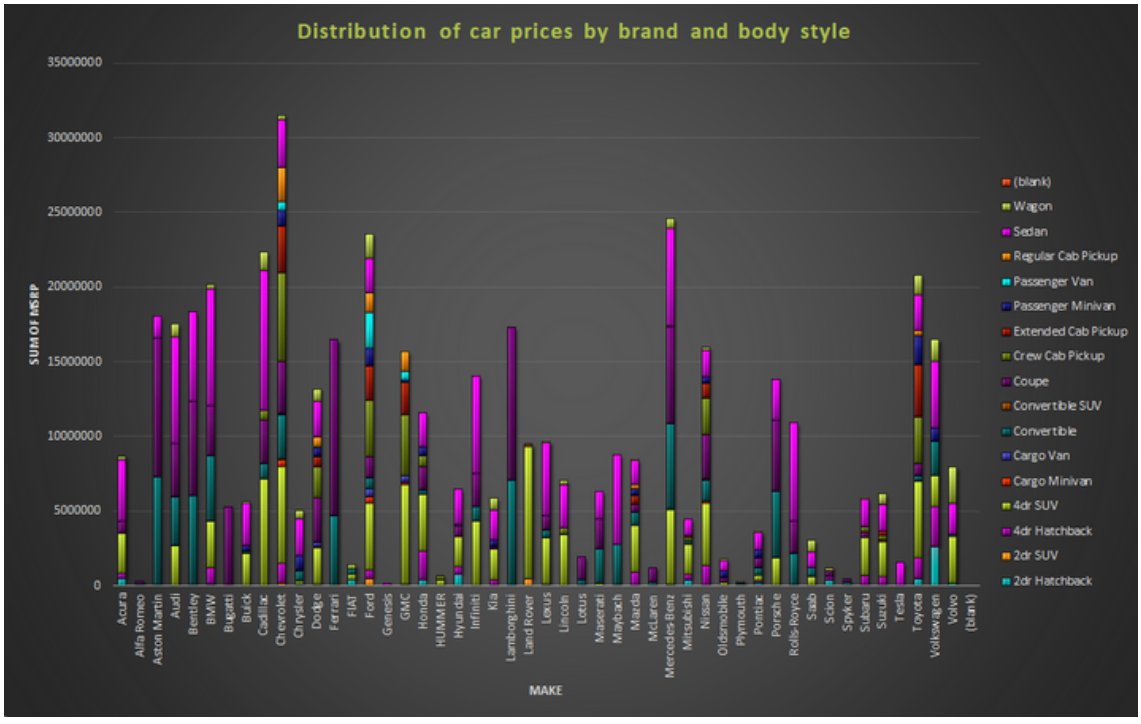


Impact of Car Features

Dashboard Tasks:-

Task 1: How does the distribution of car prices vary by brand and body style?

Hints: Stacked column chart to show the distribution of car prices by brand and body style. Use filters and slicers to make the chart interactive. Calculate the total MSRP for each brand and body style using SUMIF or Pivot Tables.

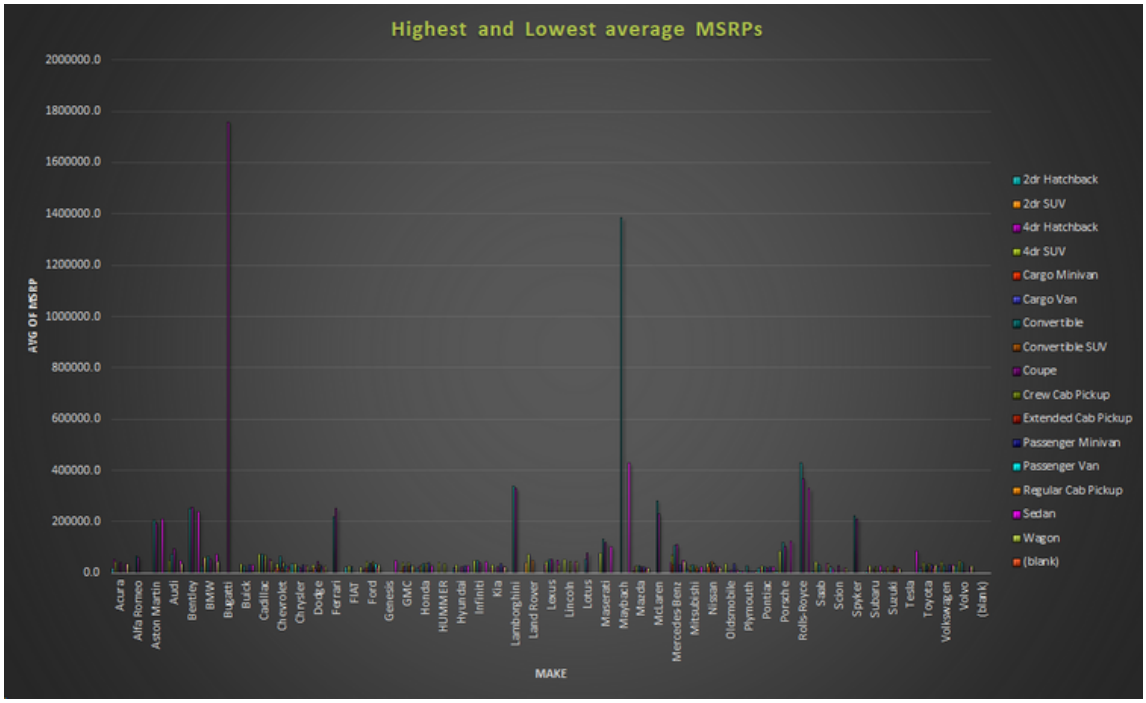


Impact of Car Features

Dashboard Tasks:-

Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

Hints: Clustered column chart to compare the average MSRPs across different car brands and body styles. Calculate the average MSRP for each brand and body style using AVERAGEIF or Pivot Tables.

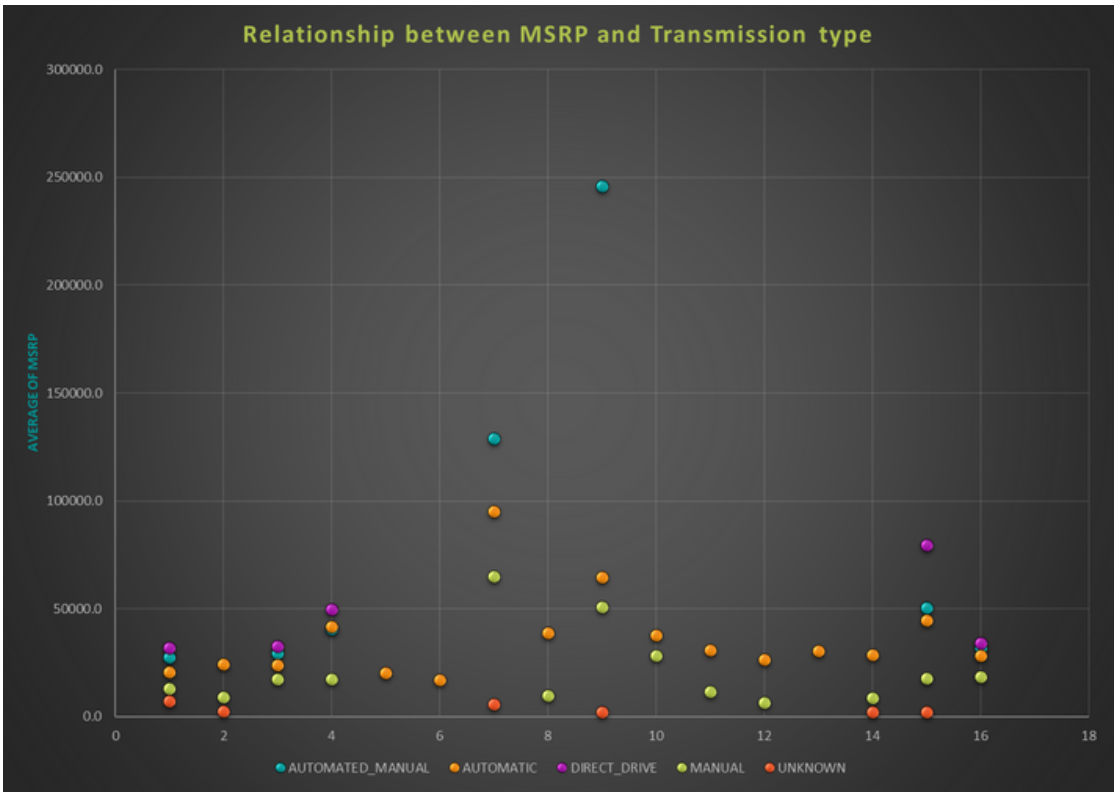


Impact of Car Features

Dashboard Tasks:-

Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

Hints: Scatter plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. Calculate the average MSRP for each combination of transmission type and body style using AVERAGEIFS or Pivot Tables.

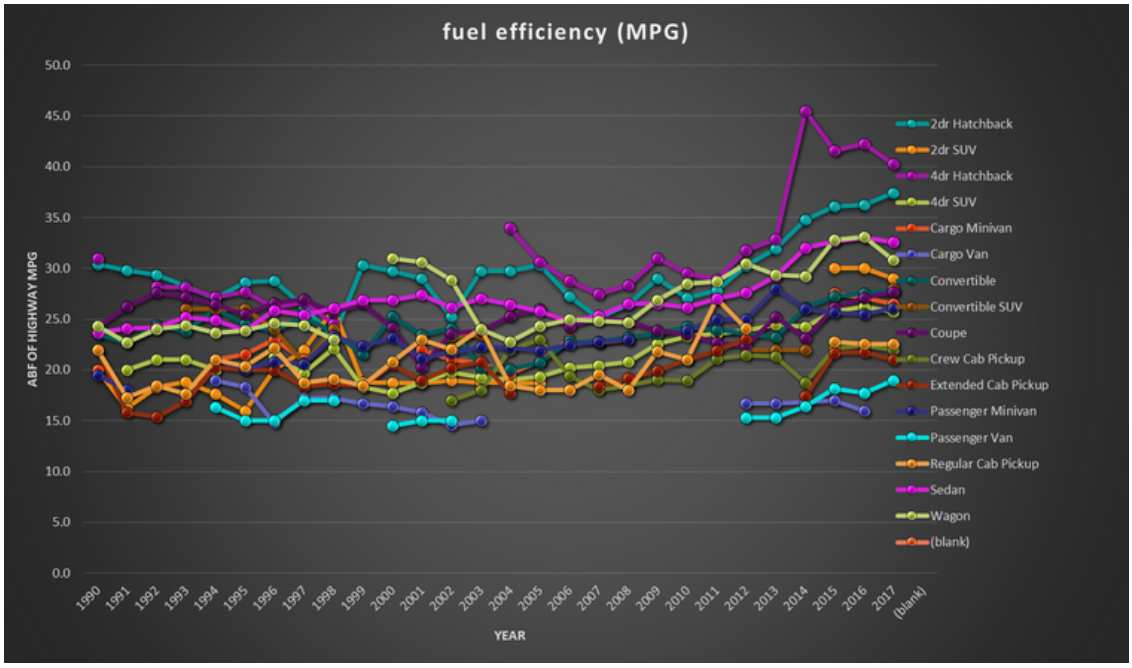


Impact of Car Features

Dashboard Tasks:-

Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

Hints: Line chart to show the trend of fuel efficiency (MPG) over time for each body style. Calculate the average MPG for each combination of body style and model year using AVERAGEIFS or Pivot Tables.

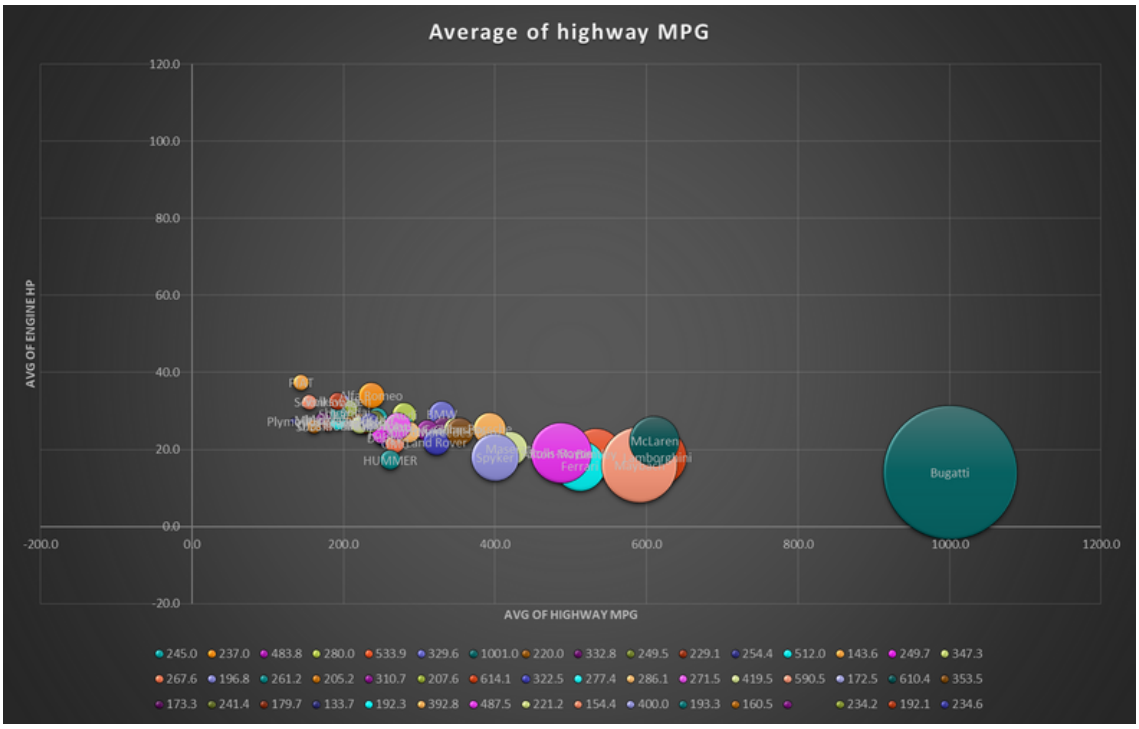


Impact of Car Features

Dashboard Tasks:-

Task 5: How does the car's horsepower, MPG, and price vary across different Brands?

Hints: Bubble chart to visualize the relationship between horsepower, MPG, and price across different car brands. Assign different colors to each brand and label the bubbles with the car model name. Calculate the average horsepower, MPG, and MSRP for each car brand using AVERAGEIFS or Pivot Tables.



Bank Loan Case Study

Description:-

Imagine you're 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 don't 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.

Tech-Stack Used:-

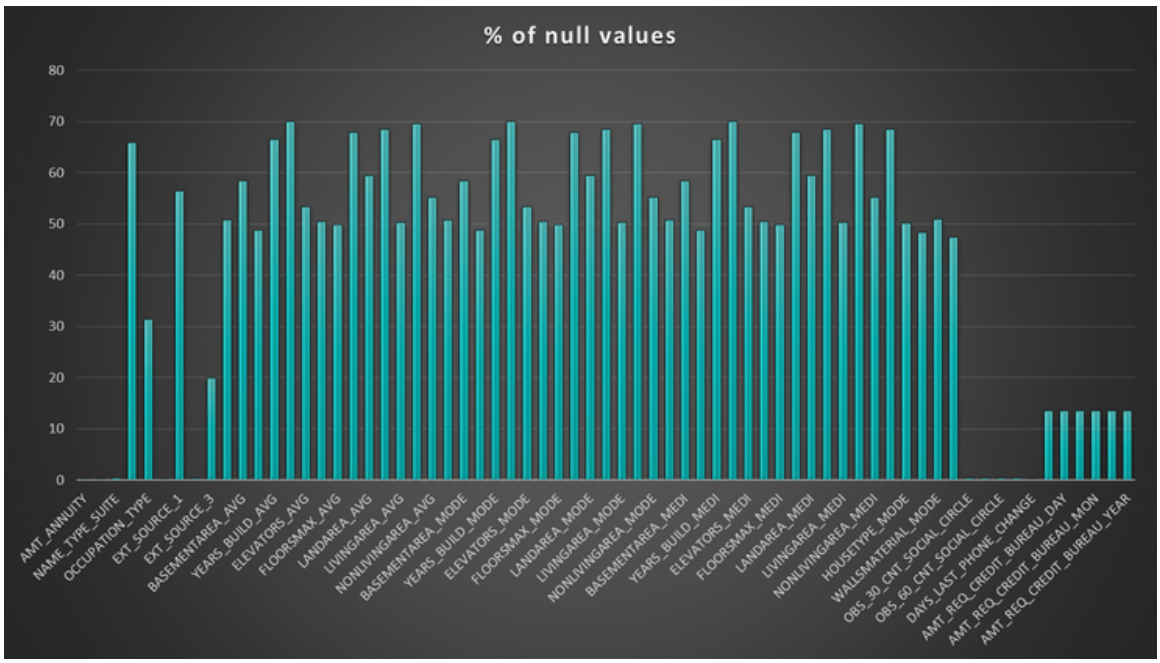
Bank Loan Case Study

Data Analytics Tasks:-
application_data file

A. 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.

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

In this, I have removed all the columns which has null percent more than 35% & replaced missing values with average and median.



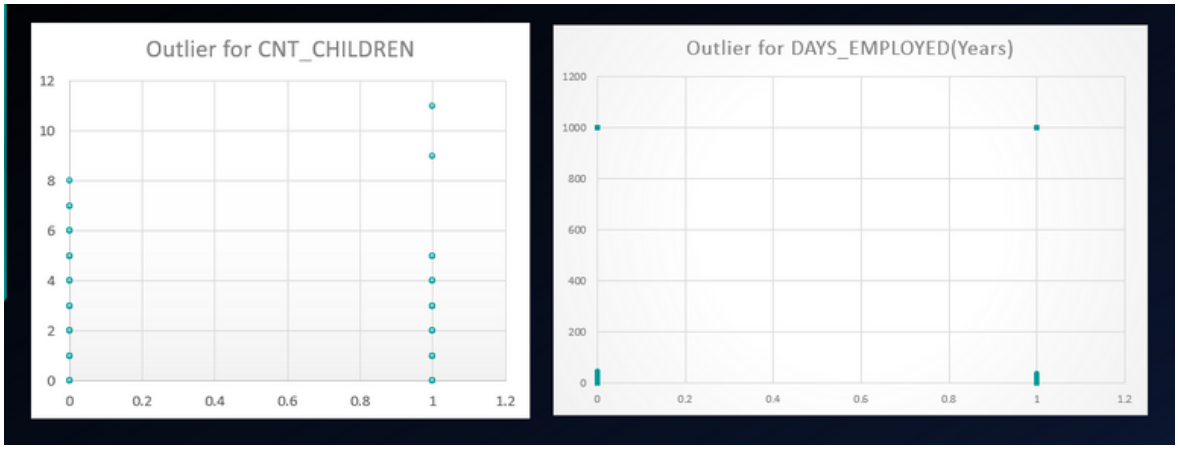
Bank Loan Case Study

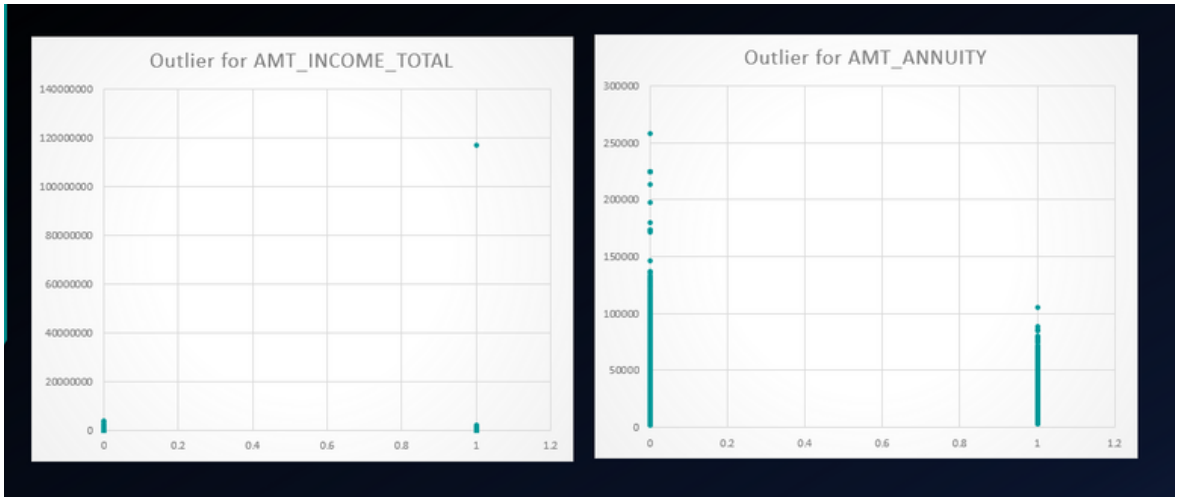
Data Analytics Tasks:-

application_data

B. 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.

Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.





Bank Loan Case Study

Data Analytics Tasks:-

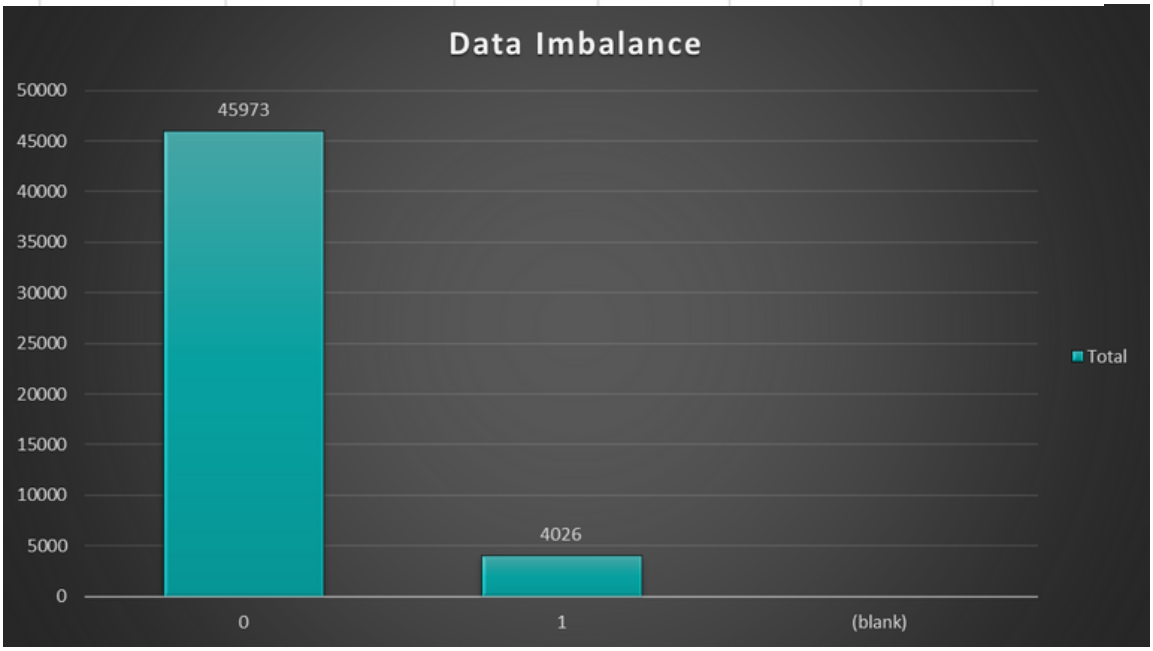
application_data

C. 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.

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Target	Count of TARGET	Percentage
0	45973	91.94784
1	4026	8.052161
(blank)		
Grand Total	49999	100

Ratio	11.41903



Bank Loan Case Study

Data Analytics Tasks:-

application_data

D. 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.

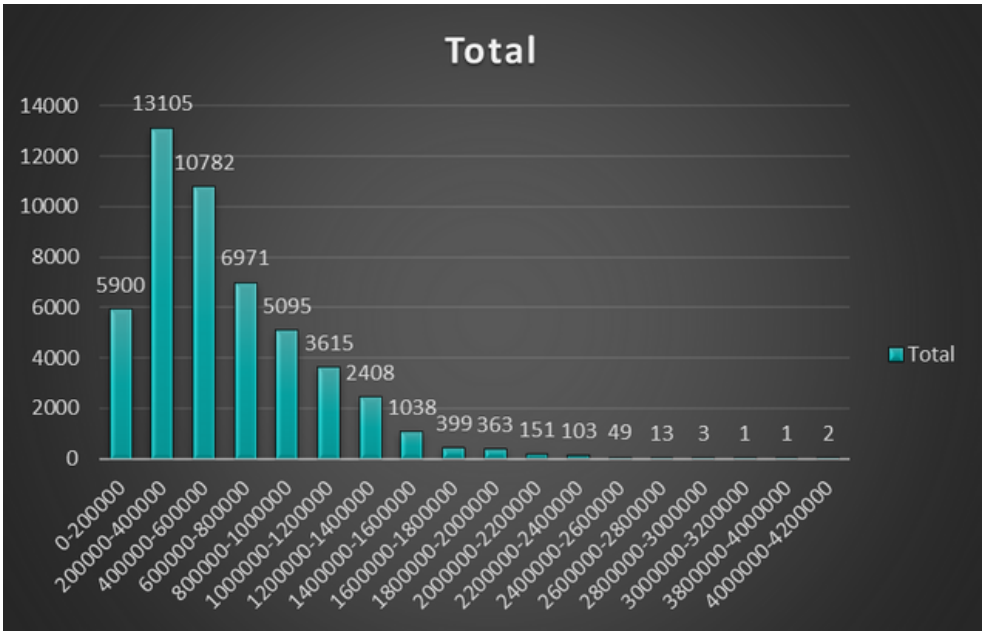
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

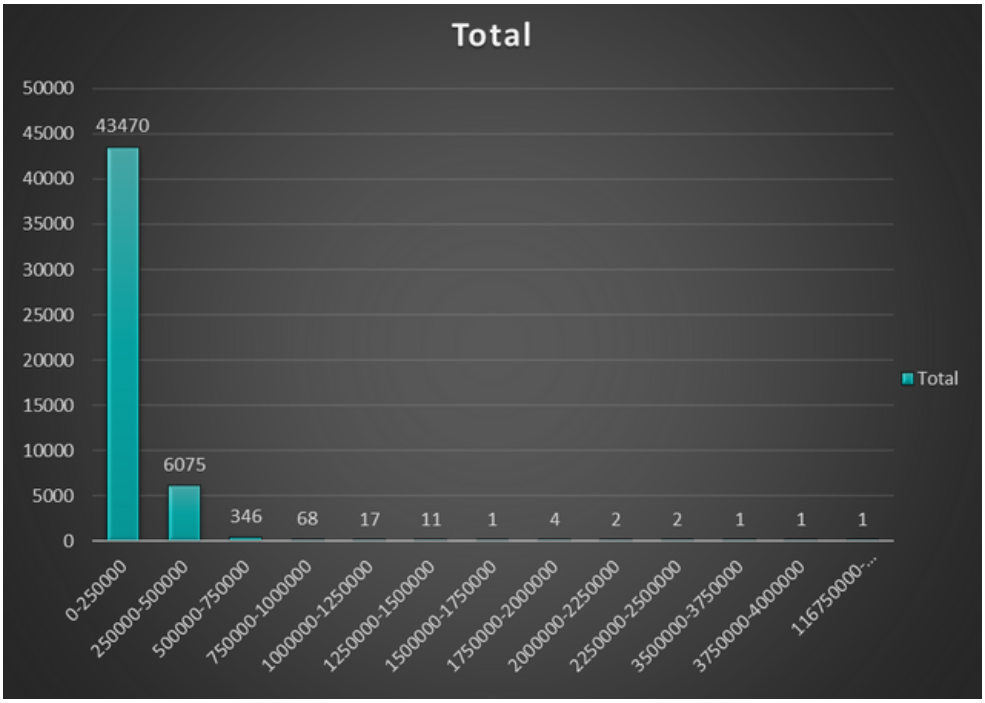
Bank Loan Case Study

Data Analytics Tasks:-

application_data

-Univariate analysis



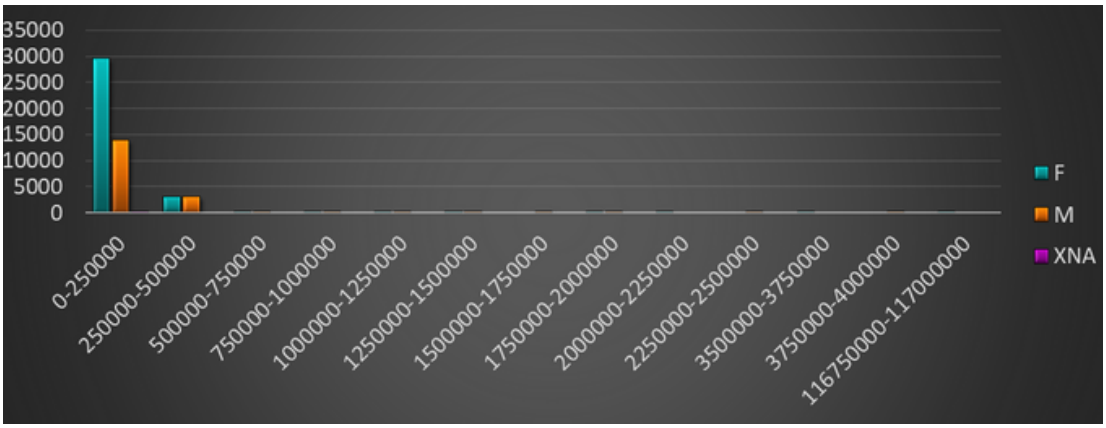


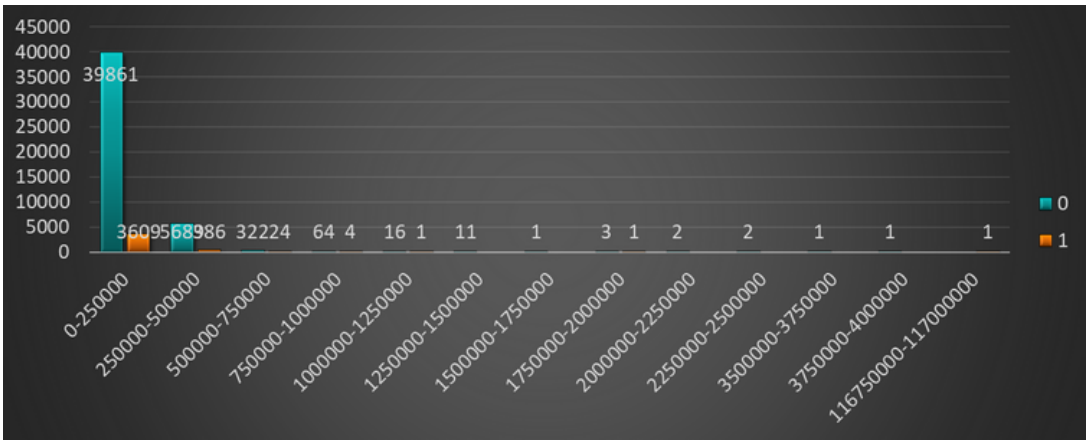
Bank Loan Case Study

Data Analytics Tasks:-

application_data

-Segmented Univariate analysis



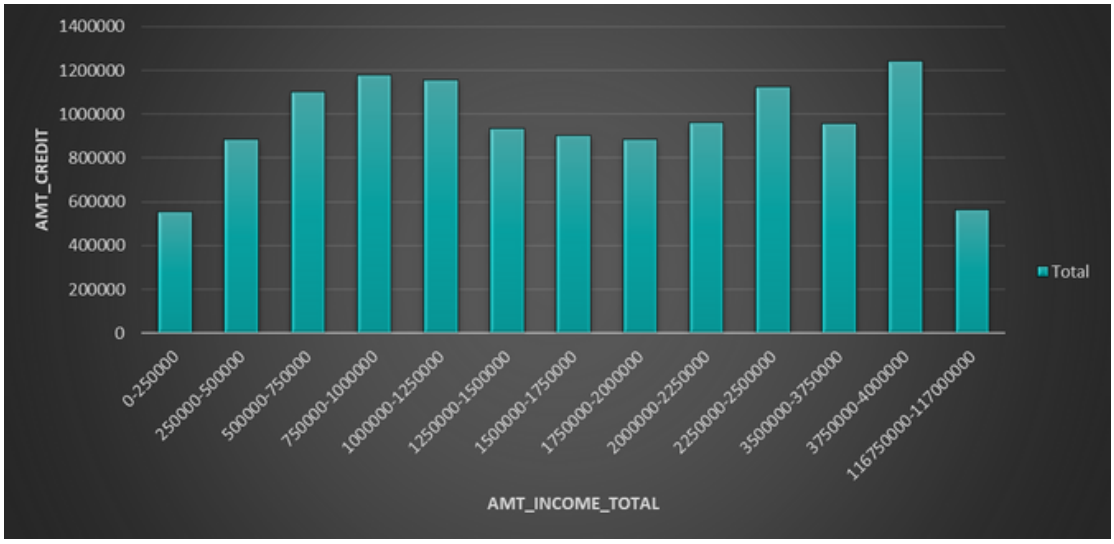


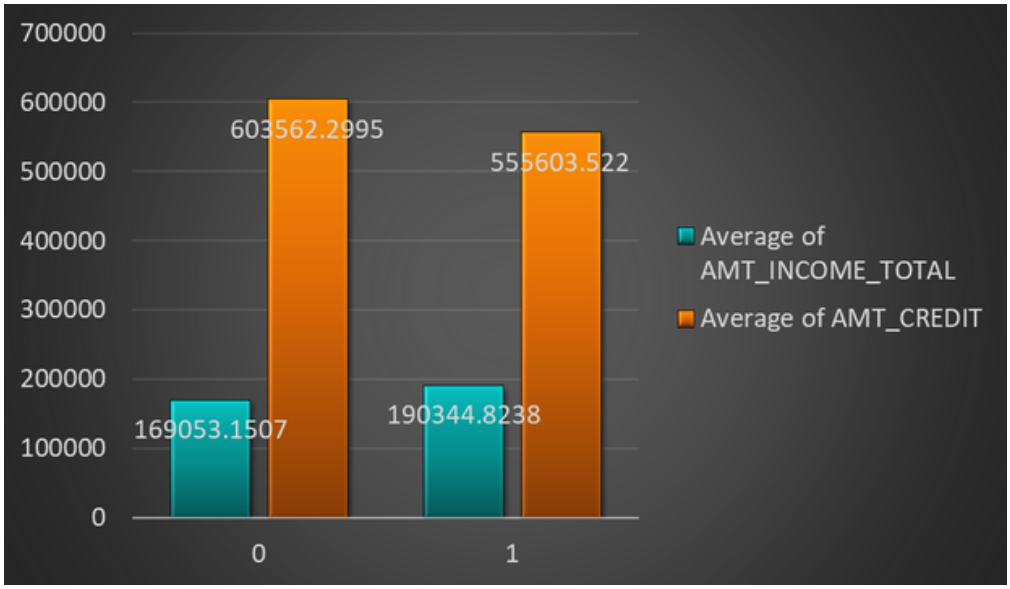
Bank Loan Case Study

Data Analytics Tasks:-

application_data

-Bivariate analysis





Bank Loan Case Study

Data Analytics Tasks:-

application_data

E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: 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.

	CNT_CHILDREN	IT_INCOME_TOT	AMT_CREDIT	AMT_ANNUIT	YS_BIRTH(Year	EMPLOYED(Y	OPULATION	_RATING_CLIENT
CNT_CHILDREN	1							
AMT_INCOME_TOTAL	0.009588558	1						
AMT_CREDIT	0.00497156	0.069315897	1					
AMT_ANNUITY	0.025458299	0.08201331	0.759902085	1				
DAYS_BIRTH(Years)	-0.329263754	-0.016002774	0.059342658	-0.008471227	1			
DAYS_EMPLOYED(Years)	-0.241539565	-0.03151033	-0.06773941	-0.107713382	0.62172831	1		
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	0.113996867	0.032513748	-0.00415834	1	
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.124346744	-0.016779196	0.03455866	-0.532667	1

Bank Loan Case Study

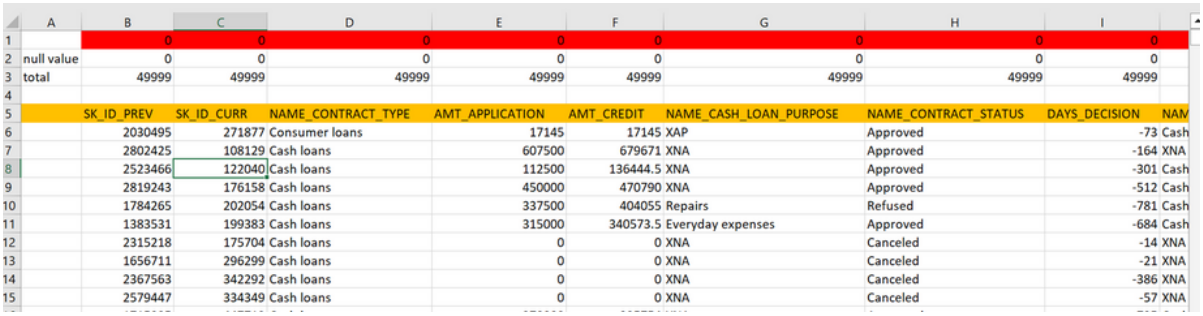
Data Analytics Tasks:-

previous_application file

A. 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.

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

In this, I have removed all the columns which has null percent more than 20%. So there is no missing value left



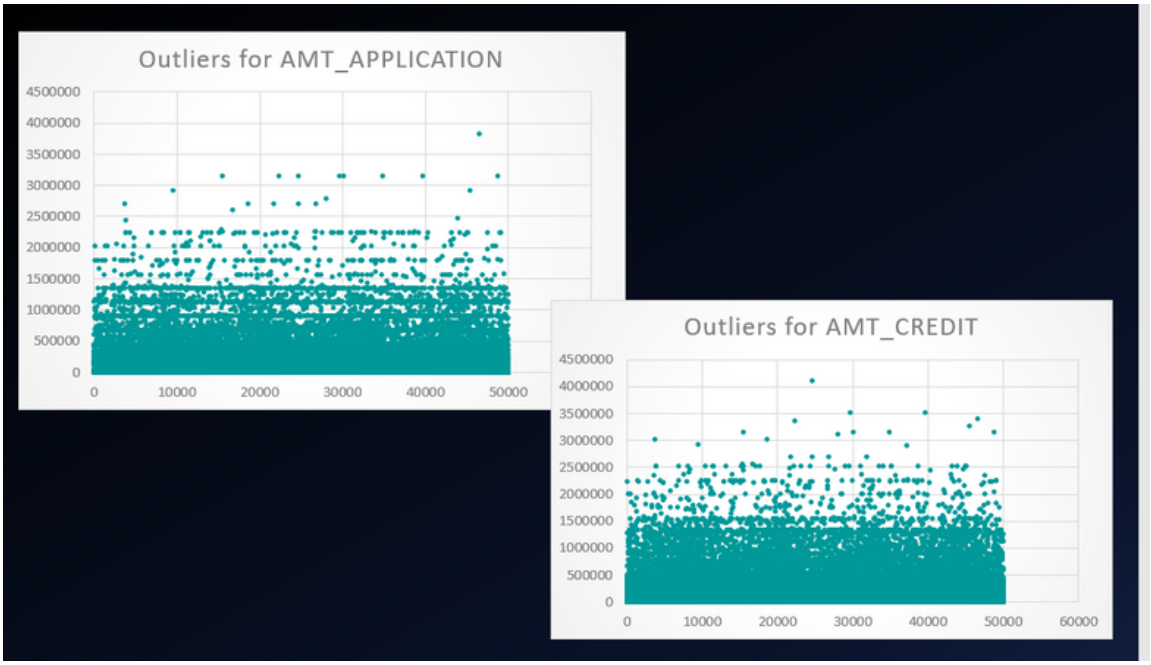
	A	B	C	D	E	F	G	H	I	
1		0	0	0	0	0	0	0	0	
2	null value	0	0	0	0	0	0	0	0	
3	total	49999	49999	49999	49999	49999	49999	49999	49999	
4										
5		SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_APPLICATION	AMT_CREDIT	NAME_CASH_LOAN_PURPOSE	NAME_CONTRACT_STATUS	DAYS_DECISION	NAM
6		2030495	271877	Consumer loans	17145	17145	XAP	Approved	-73	Cash
7		2802425	108129	Cash loans	607500	679671	XNA	Approved	-164	XNA
8		2523466	122040	Cash loans	112500	136444.5	XNA	Approved	-301	Cash
9		2819243	176158	Cash loans	450000	470790	XNA	Approved	-512	Cash
10		1784265	202054	Cash loans	337500	404055	Repairs	Refused	-781	Cash
11		1383531	199383	Cash loans	315000	340573.5	Everyday expenses	Approved	-684	Cash
12		2315218	175704	Cash loans	0	0	XNA	Canceled	-14	XNA
13		1656711	296299	Cash loans	0	0	XNA	Canceled	-21	XNA
14		2367563	342292	Cash loans	0	0	XNA	Canceled	-386	XNA
15		2579447	334349	Cash loans	0	0	XNA	Canceled	-57	XNA

Bank Loan Case Study

Data Analytics Tasks:-
previous_application

B. 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.

Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.



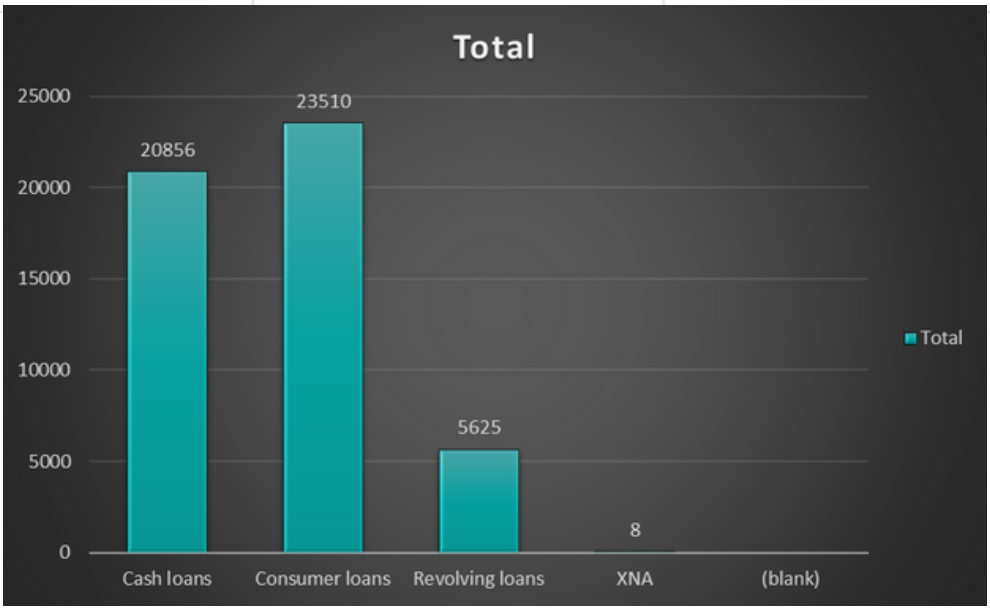
Bank Loan Case Study

Data Analytics Tasks:-
previous_application

C. 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.

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

NAME_CONTRACT_T	Count of NAME_CONTRACT_TYPE	Percentage
Cash loans	20856	41.71283426
Consumer loans	23510	47.02094042
Revolving loans	5625	11.250225
XNA	8	0.01600032
(blank)		
Grand Total	**49999**	**100**



Bank Loan Case Study

Data Analytics Tasks:-
previous_application

D. 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.

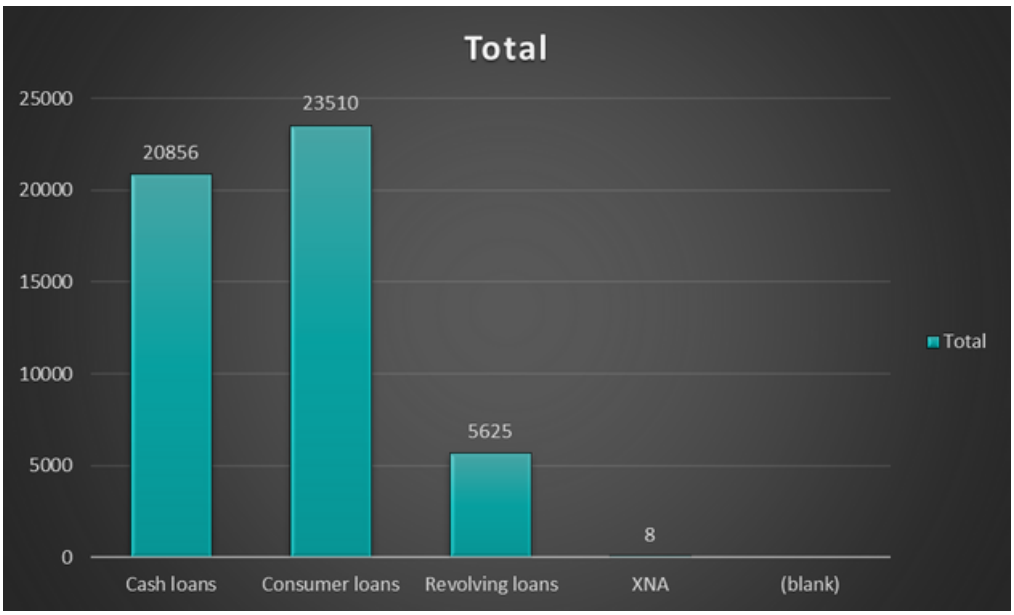
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

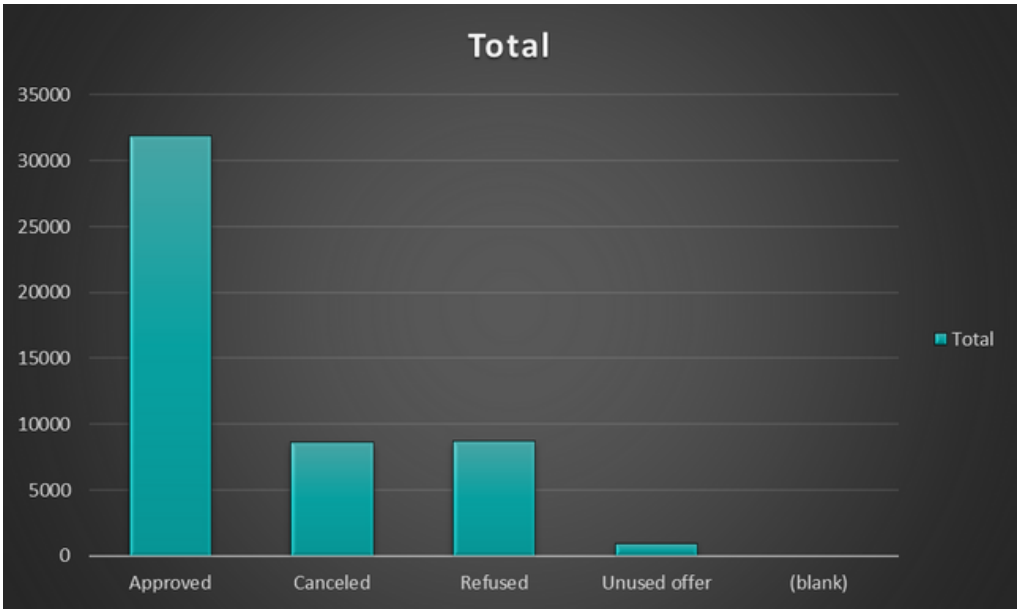
Bank Loan Case Study

Data Analytics Tasks:-

previous_application

-Univariate analysis



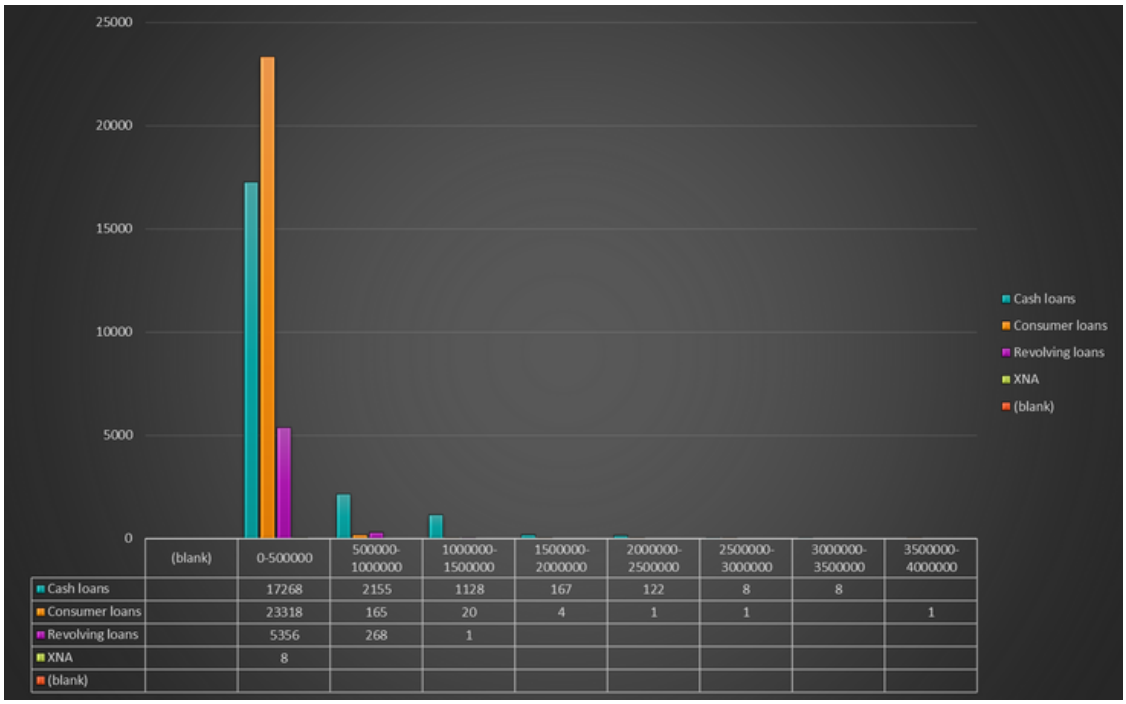


Bank Loan Case Study

Data Analytics Tasks:-

previous_application

-Segmented Univariate analysis



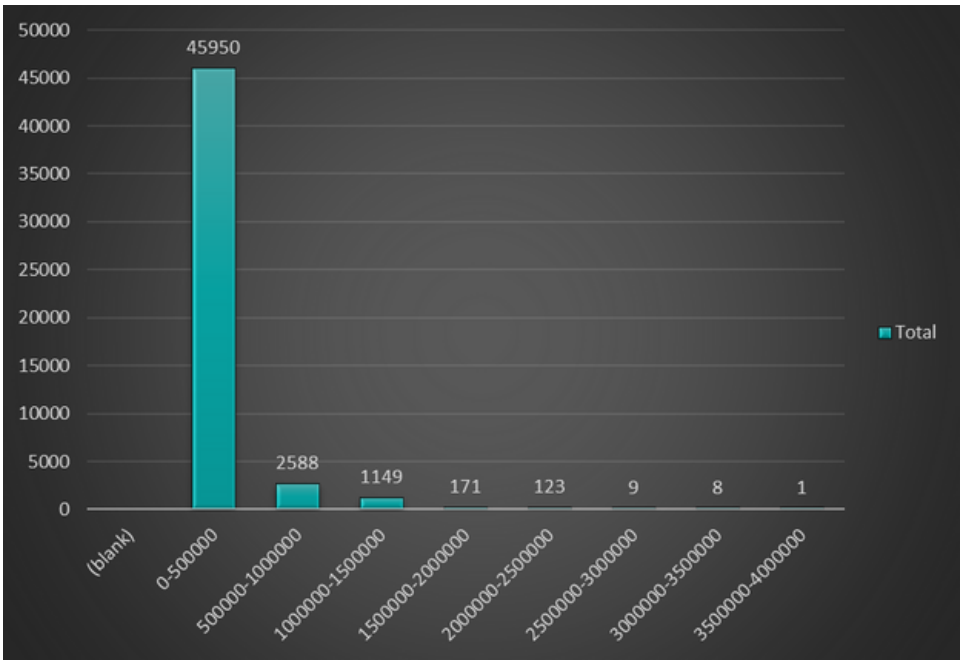
	(blank)	0-500000	500000-1000000	1000000-1500000	1500000-2000000	2000000-2500000	2500000-3000000	3000000-3500000	3500000-4000000
Cash loans		17268	2155	1128	167	122	8	8	
Consumer loans		23318	165	20	4	1	1		1
Revolving loans		5356	268	1					
XNA		8							
(blank)									

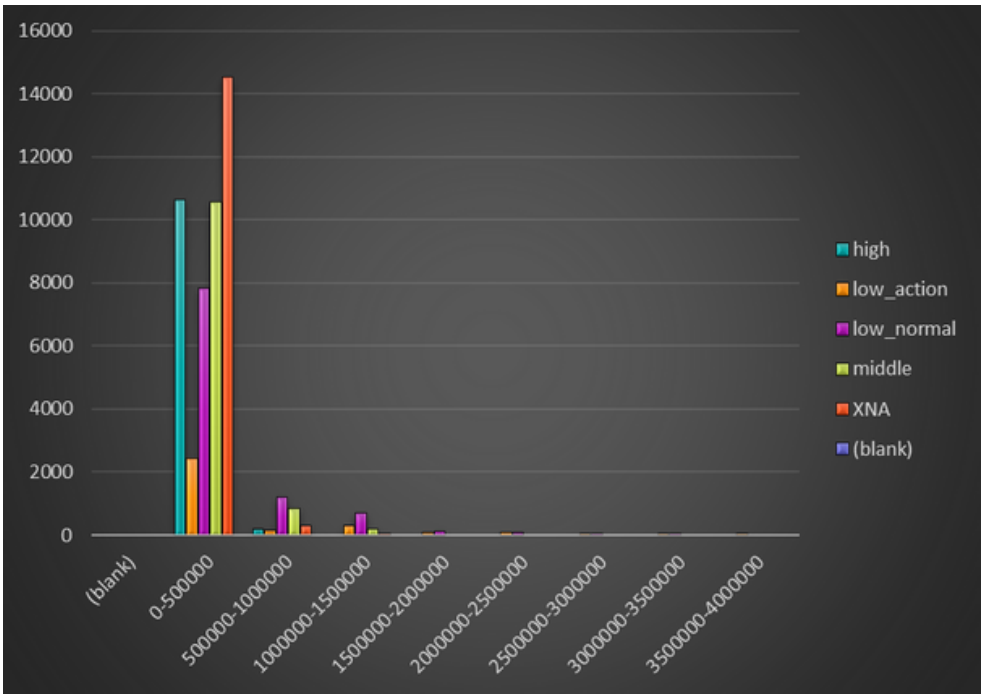
Bank Loan Case Study

Data Analytics Tasks:-

previous_application

-Bivariate analysis





Bank Loan Case Study

Data Analytics Tasks:-
previous_application

E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: 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.

	SK_ID_PREV	SK_ID_CURR	AMT_APPLICATION	AMT_CREDIT	DAYS_DECISION	SELLERPLACE_AREA
SK_ID_PREV	1					
SK_ID_CURR	0.001990413	1				
AMT_APPLICATION	-0.002219254	0.001996305	1			
AMT_CREDIT	0.000130108	0.000964228	0.975771049	1		
DAYS_DECISION	0.01551255	-0.005152718	0.13399106	0.137431211	1	
SELLERPLACE_AREA	0.007010428	0.000240656	-0.003965725	-0.004949463	-0.008624778	1

IMDB Movie Analysis

Description:-

Problem Statement: 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.

Tech-Stack Used:-

IMDB Movie Analysis

Approach:-

- We have given a dataset of movies.
- Firstly, I have cleaned the whole dataset to get valuable insights.
- Data Cleaning involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering. So,

1. In total, there are 5044 rows (including headers)
2. There are 45 duplicate values.
3. And after cleaning and removing errors and blank cells in final clean dataset:
4. There are 3724 rows (including headers)

IMDB Movie Analysis

Data Analytics Tasks:-

A. Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.

Task: 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.

Genre	Count	Mean of imdb
Action	951	6.289905363
Adventure	773	6.452393273
Animation	196	6.70255102
Biography	238	7.157563025
Comedy	1455	6.188109966
Crime	704	6.541903409
Documentary	45	6.988888889
Drama	1876	6.792537313
Family	440	6.216363636
Fantasy	503	6.286309524
Film-Noir	1	7.7
History	147	7.157823129
Horror	386	5.92253886
Music	149	6.454545455
Musical	94	6.596875
Mystery	377	6.480687831
Romance	845	6.435840188
Sport	146	6.589115646
Sci-Fi	485	6.325813008
Thriller	1101	6.376651584
War	149	7.062666667
Western	57	6.812280702

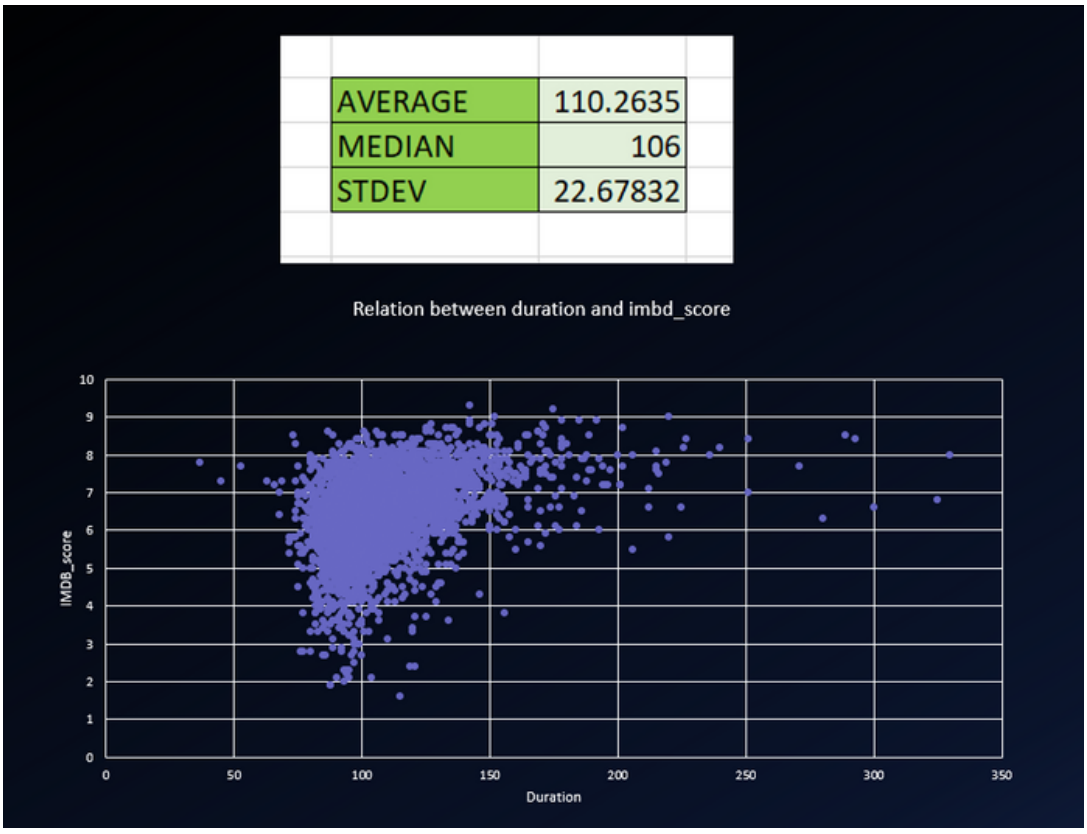
AVERAGE	505.3636364
MEDIAN	381.5
MODE	149
MIN	1
MAX	1876
VARIANCE	243380.8139
STDEV	493.3364104

IMDB Movie Analysis

Data Analytics Tasks:-

B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

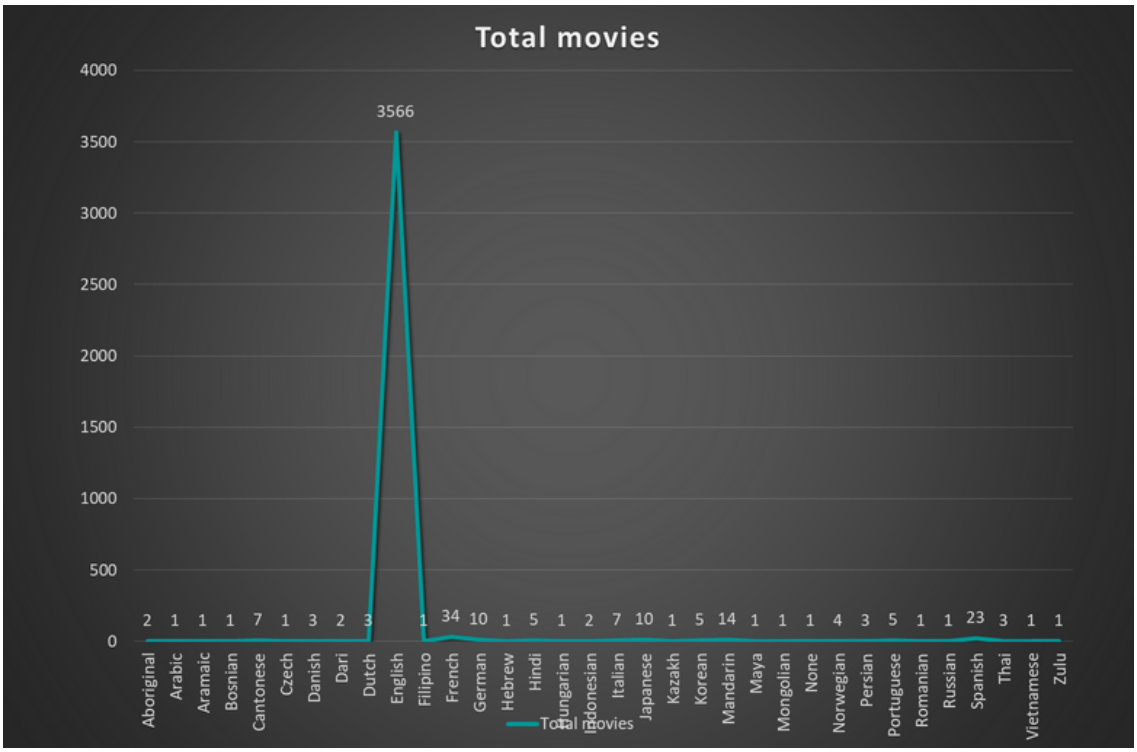


IMDB Movie Analysis

Data Analytics Tasks:-

C. Language Analysis: Situation: Examine the distribution of movies based on their language.

Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.



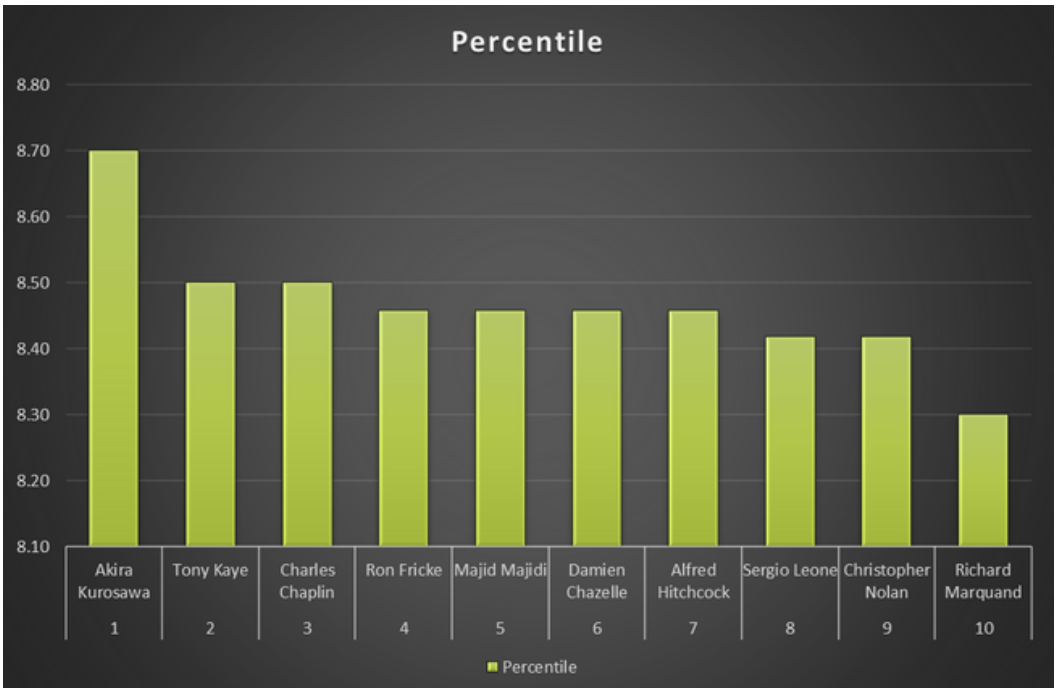
IMDB Movie Analysis

Data Analytics Tasks:-

D. Director Analysis: Influence of directors on movie ratings.

Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

S.no.	Top 10	Total movies	Avg of IMDB	Percent rank	Percentile
1	Akira Kurosawa	1	8.7	1	8.70
2	Tony Kaye	1	8.6	0.998	8.50
3	Charles Chaplin	1	8.6	0.998	8.50
4	Ron Fricke	1	8.5	0.996	8.46
5	Majid Majidi	1	8.5	0.996	8.46
6	Damien Chazelle	1	8.5	0.996	8.46
7	Alfred Hitchcock	1	8.5	0.996	8.46
8	Sergio Leone	3	8.43333333	0.995	8.42
9	Christopher Nolan	8	8.425	0.995	8.42
10	Richard Marquand	1	8.4	0.993	8.30

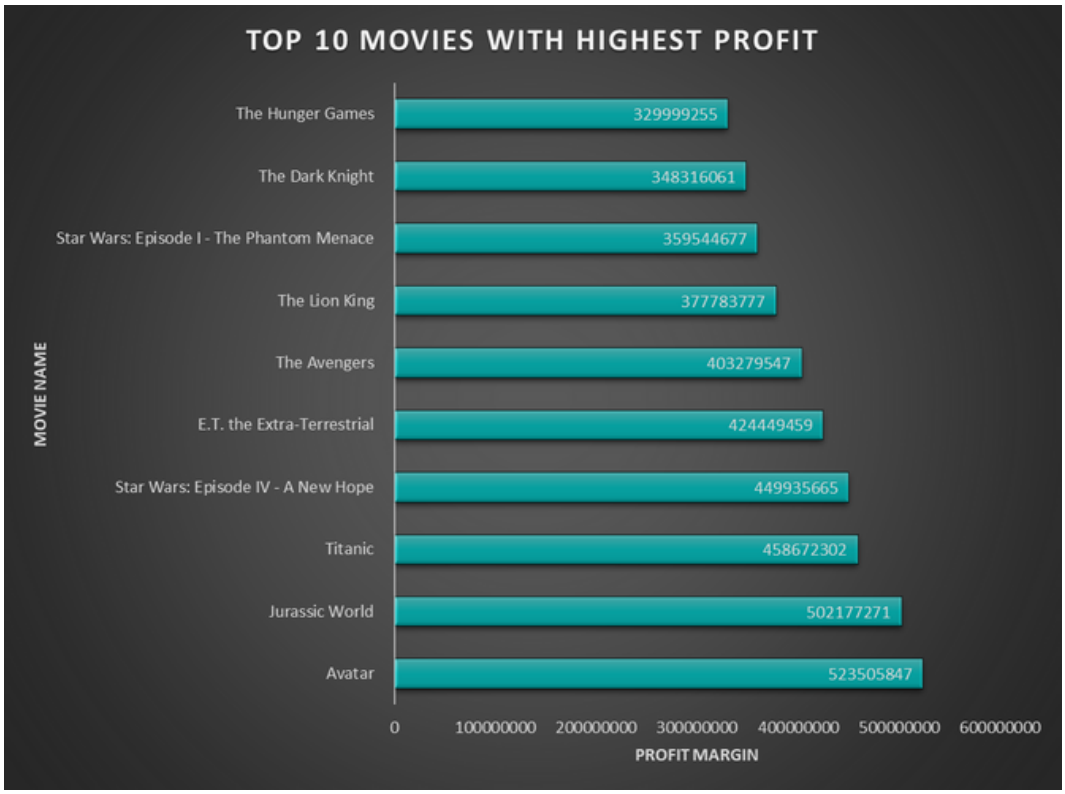


IMDB Movie Analysis

Data Analytics Tasks:-

E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.



Correlation coefficient	0.098318102

Movie with highest profit margin	Profit
Avatar	523505847

IMDB Movie Analysis

Results & Insights :-

1. The top most common genres are Drama, Comedy and Thriller, respectively with average IMDB of more than 6.0.
2. If movie duration is more than 75 then there is a chance of getting more imdb_score, people often prefer to watch longer movie because longer the movie, more drama will be there.
3. So, movies with more than 100 duration will have more ratings.
4. The most common language used in movies is English, second most is French and then Spanish.
5. The director with highest average imdb_score is Akira Kurosawa for the movie Seven Samurai.
6. The director who have made movies on popular genres has more average imdb_score.
7. Avatar is the movie with highest profit margin of 523505847 .

Hiring Process Analytics

Description:-

Our task is to analyze the company's hiring process data and draw meaningful insights from it. The hiring process is a crucial function of any company, and understanding trends such as the number of rejections, interviews, job types, and vacancies can provide valuable insights for the hiring department.

As a data analyst, you'll be given a dataset containing records of previous hires.

Our job is to analyze this data and answer certain questions that can help the company improve its hiring process.

Tech-Stack Used:-

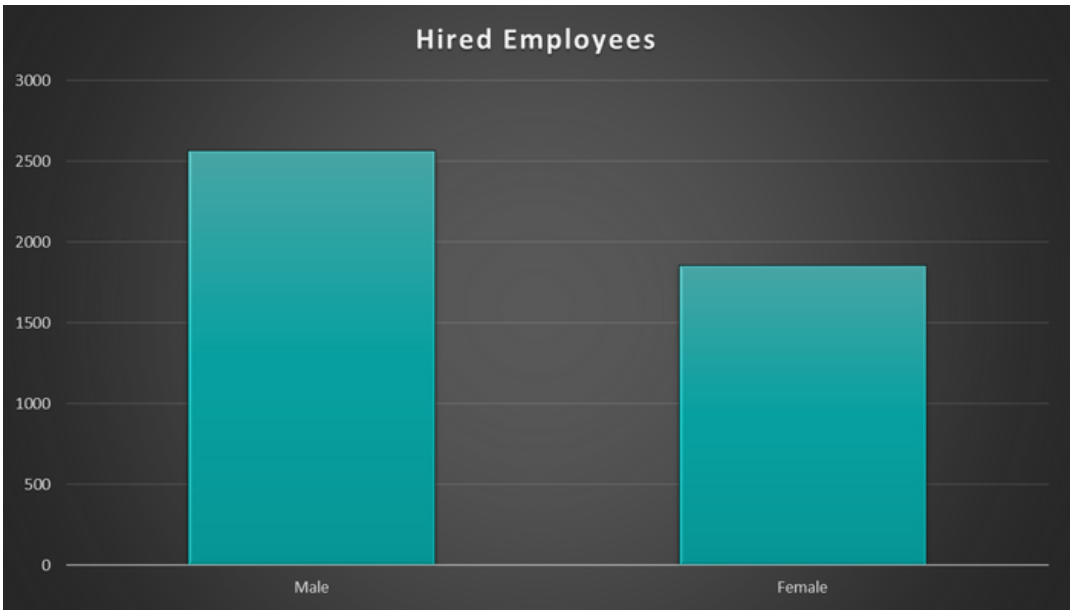
Hiring Process Analytics

Data Analytics Tasks:-

1. Hiring Analysis: The hiring process involves bringing new individuals into the organization for various roles.

Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company?

Hired	Total
Male	2563
Female	1856



Hiring Process Analytics

Data Analytics Tasks:-

2. 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.

Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.

Average_salary	₹49983.03

Hiring Process Analytics

Data Analytics Tasks:-

3. 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.

Your Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

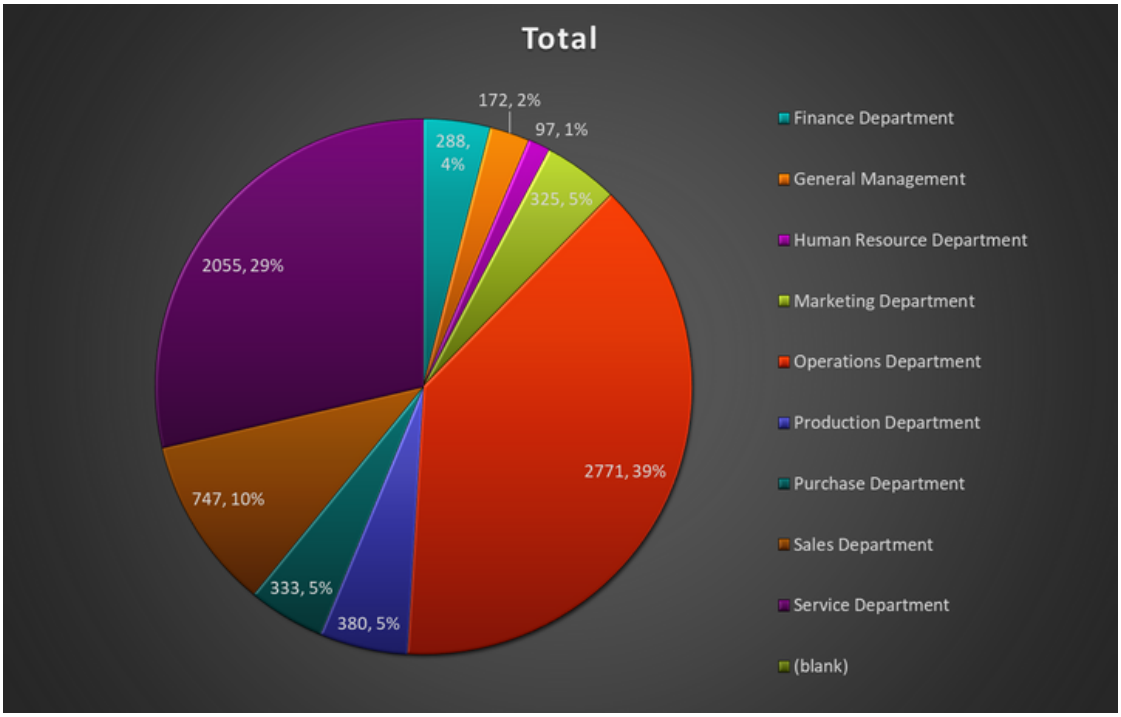
Salary Range	No. of employees
<0 or (blank)	1
0-24999	1757
25000-49999	1854
50000-74999	1797
75000-99999	1756
200000-224999	1
300000-324999	1
375000-400000	1
Grand Total	7168

Hiring Process Analytics

Data Analytics Tasks:-

4. Departmental Analysis: Visualizing data through charts and plots is a crucial part of data analysis.

Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.



Hiring Process Analytics

Data Analytics Tasks:-

5. Position Tier Analysis: Different positions within a company often have different tiers or levels.

Your Task: 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.

Hiring Process Analytics

Insights & Results :-

- Total no. of hired males are 2563 and females are 1865.
- Average salary is ₹49983.03.
- Only 3 Employees has salary more than Rs. 2,00,000.
- Operations Department have maximum employees.
- Human Resource Department have minimum employees.
- C9 has most employees.

Operation & Metric Analytics

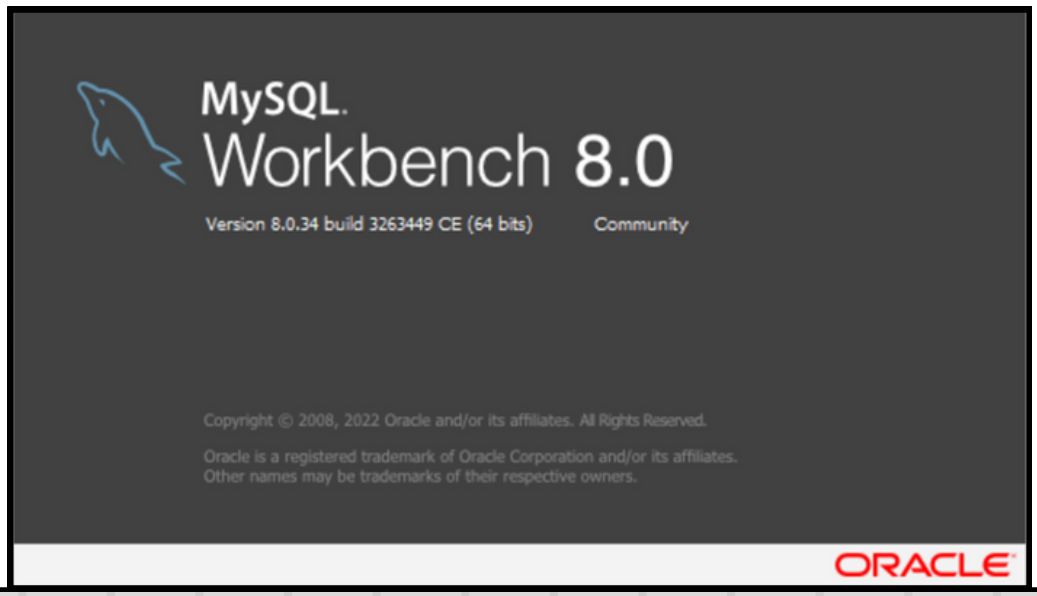
Description:-

Operational Analytics is a crucial process that involves analyzing a company's end-to-end operations. This analysis helps identify areas for improvement within the company.

As a Data Analyst, you'll work closely with various teams, such as operations, support, and marketing, helping them derive valuable insights from the data they collect. One of the key aspects of Operational Analytics is investigating metric spikes.

This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales.

Tech-Stack Used:-



Operation & Metric Analytics

Approach :-

- In this project, as we are working as the role of a Lead Data Analyst at a company like Microsoft
- We have few tables containing data relevant to the platform.
- We have,
- 1 table:- job_data for case study 1
- 3 tables: users, events, email_events for case study 2
- By these tables we can give provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics.

Operation & Metric Analytics

Case Study -1 :- Job Data Analysis

1. Jobs Reviewed Over Time : Calculate the number of jobs reviewed per hour for each day in November 2020.

Your Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

```
110
111     #Jobs Reviewed Over Time
112 •   SELECT
113         ds AS 'date',
114         SUM(time_spent / (60 * 60)) AS 'time_in_hours',
115         COUNT(job_id) AS no_of_jobs
116     FROM
117         job_data
118     WHERE
119         ds >= '2020-11-01'
120             AND ds <= '2020-11-30'
121     GROUP BY ds
122     ;
123
124
```

Result Grid | Filter Rows:

date	time_in_hours	no_of_jobs
2020-11-30	0.0111	2
2020-11-29	0.0056	1
2020-11-28	0.0092	2
2020-11-27	0.0289	1
2020-11-26	0.0156	1
2020-11-25	0.0125	1

Operation & Metric Analytics

Case Study -1 :- Job Data Analysis

2. Throughput Analysis : Calculate the 7-day rolling average of throughput (number of events per second).

Your Task: Write an SQL query to calculate the 7-day rolling average of throughput.

```
#Throughput Analysis
with throughput as(
    SELECT ds as day, count(job_id) / sum(time_spent) AS throughput
    FROM job_data
    GROUP BY ds)

SELECT *,
avg(throughput.throughput) OVER ( ORDER BY day ) AS avg_of_throughput
from throughput
group by day
;
```

Result Grid | Filter Rows: | Expo

day	throughput	avg_of_throughput
2020-11-25	0.0222	0.02220000
2020-11-26	0.0179	0.02005000
2020-11-27	0.0096	0.01656667
2020-11-28	0.0606	0.02757500
2020-11-29	0.0500	0.03206000
2020-11-30	0.0500	0.03505000

Operation & Metric Analytics

Case Study -1 :- Job Data Analysis

3. Language Share Analysis: Calculate the percentage share of each language in the last 30 days.

Your Task: Write an SQL query to calculate the percentage share of each language over the last 30 days.

```
140
141     #Language Share Analysis
142 •   SELECT language,
143     count(job_id) as no_of_jobs,
144     round(count(job_id)*100 / sum(count(*)) OVER(),2) as percentage_share
145     FROM job_data
146     WHERE ds between '2020-11-01' and '2020-12-01'
147     GROUP by language;
148
149
150
```

Result Grid | Filter Rows:

language	no_of_jobs	percentage_share
English	1	12.50
Arabic	1	12.50
Persian	3	37.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50

Operation & Metric Analytics

Case Study -1 :- Job Data Analysis

4. Duplicate Rows Detection : Identify duplicate rows in the data.

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

```
156
157     #Duplicate Rows Detection
158 •   with dup as(
159         SELECT *,
160         row_number() OVER (partition by ds, job_id, actor_id, event, language, time_spent, org)
161         as no_of_rows
162         FROM job_data )
163     select * ,
164     case
165     when no_of_rows=1
166     then 'no'
167     else'yes'
168     end as Duplicate
169     from dup;
170
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_id	actor_id	event	language	time_spent	org	ds	no_of_rows	Duplicate
20	1003	transfer	Italian	45	C	2020-11-25	1	no
23	1004	skip	Persian	56	A	2020-11-26	1	no
11	1007	decision	French	104	D	2020-11-27	1	no
23	1005	transfer	Persian	22	D	2020-11-28	1	no
25	1002	decision	Hindi	11	B	2020-11-28	1	no
23	1003	decision	Persian	20	C	2020-11-29	1	no
21	1001	skip	English	15	A	2020-11-30	1	no
22	1006	transfer	Arabic	25	B	2020-11-30	1	no

Operation & Metric Analytics

Case Study -2 :- Investigating Metric Spike

1. Weekly User Engagement: Measure the activeness of users on a weekly basis.

Your Task: Write an SQL query to calculate the weekly user engagement.

```
177
178     #Weekly User Engagement
179 •   SELECT
180         WEEK(occurred_at) AS week,
181         COUNT(DISTINCT user_id) AS no_of_user
182     FROM
183         events
184     GROUP BY WEEK(occurred_at)
185     ORDER BY WEEK(occurred_at);
186
187
188
```

Result Grid | Filter Rows:

week	no_of_user
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299

Result 17

Operation & Metric Analytics

Case Study -2 :- Investigating Metric Spike

2. User Growth Analysis: Analyze the growth of users over time for a product.

Your Task: Write an SQL query to calculate the user growth for the product.

```sql
with user1 as
(select week(activated_at) as week,
    year(activated_at) as year,
    count(distinct user_id) as no_of_user
    from users
    group by week(activated_at),year(activated_at)
    order by week(activated_at),year(activated_at)
)
select week,
year,
no_of_user,sum(no_of_user)
over(order by year,week rows between unbounded preceding and current row) as growth
from user1
;
```

Result Grid | Filter Rows:

week	year	no_of_user	growth
0	2013	23	23
1	2013	30	53
2	2013	48	101
3	2013	36	137
4	2013	30	167
5	2013	48	215
6	2013	38	253
7	2013	42	295
8	2013	34	329
9	2013	43	372
10	2013	32	404
11	2013	31	435

Result 17

Operation & Metric Analytics

Case Study -2 :- Investigating Metric Spike

3. Weekly Retention Analysis: Analyze the retention of users on a weekly basis after signing up for a product.

Your Task: Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.

```
206
207 • with cte1 as (
208   select distinct user_id,
209   week(occurred_at) as login_week
210   from events
211   where event_type='signup_flow'
212   and event_name='complete_signup'
213   and week(occurred_at)=18),
214   cte2 as
215   (select distinct user_id,
216   week(occurred_at) as engage_week
217   from events
218   where event_type='engagement')
219   select count(user_id) as total_user,
220       sum(case  when r_week>0 then 1 else 0 end ) as r_user
221       from (select a.user_id,a.login_week,b.engage_week,b.engage_week - a.login_week as r_week
222       from cte1 a
223       left join cte2 b
224       on a.user_id=b.user_id
225       order by a.user_id) c;
```

Result Grid | Filter Rows:

total_user	r_user
615	452

Operation & Metric Analytics

Case Study -2 :- Investigating Metric Spike

4. Weekly Engagement Per Device: Measure the activeness of users on a weekly basis per device.

Your Task: Write an SQL query to calculate the weekly engagement per device.

```
226    #Weekly Engagement Per Device
227    SELECT
228        WEEK(occurred_at) AS week,
229        COUNT(DISTINCT user_id) AS no_of_user,device
230    FROM
231        events
232    GROUP BY WEEK(occurred_at),device
233    ORDER BY WEEK(occurred_at);
234
235
236
237
```

Result Grid | Filter Rows:

week	no_of_user	device
17	9	acer aspire desktop
17	20	acer aspire notebook
17	4	amazon fire phone
17	21	asus chromebook
17	18	dell inspiron desktop
17	46	dell inspiron notebook
17	14	hp pavilion desktop
17	16	htc one
17	27	ipad air
17	19	ipad mini
17	21	iphone 4s
17	65	iphone 5
17	42	iphone 5s
17	6	kindle fire
17	86	lenovo thinkpad

Result 23

Operation & Metric Analytics

Case Study -2 :- Investigating Metric Spike

5. Email Engagement Analysis: Analyze how users are engaging with the email service.

Your Task: Write an SQL query to calculate the email engagement metrics.

```
#Email Engagement Analysis

SELECT week(occurred_at) as Week,
count( DISTINCT ( CASE WHEN action = "sent_weekly_digest"
THEN user_id end )) as weekly_digest,
count( distinct ( CASE WHEN action = "sent_reengagement_email"
THEN user_id end )) as reengagement_mail,
count( distinct ( CASE WHEN action = "email_open"
THEN user_id end )) as email_open,
count( distinct ( CASE WHEN action = "email_clickthrough"
THEN user_id end )) as email_click
FROM email_events
GROUP BY week(occurred_at)
ORDER BY week(occurred_at);

```

Result Grid | Filter Rows: | Export: | Wrap Ce

Week	weekly_digest	reengagement_mail	email_open	email_click
17	908	73	310	166
18	2602	157	900	425
19	2665	173	961	476
20	2733	191	989	501
21	2822	164	996	436
22	2911	192	965	478
23	3003	197	1057	529
24	3105	226	1136	549
25	3207	196	1084	524
26	3302	219	1149	550
27	3399	213	1207	613
28	3499	213	1228	594

Result 30

Operation & Metric Analytics

Results & Insights :-

- No. of jobs reviewed per day per hour is very less.
- 7-day rolling average is best for throughput.
- The rows which have ‘yes’ value in duplicate columns are duplicate rows else if ‘no’ is there then there are no duplicates.
- Weekly user engagement is highest in 30th week.
- Persian language is most popular.
- Most of the users are using macbook pro.
- By growth insights we can see that users increasing rate has almost 30-40 add-ons each week.

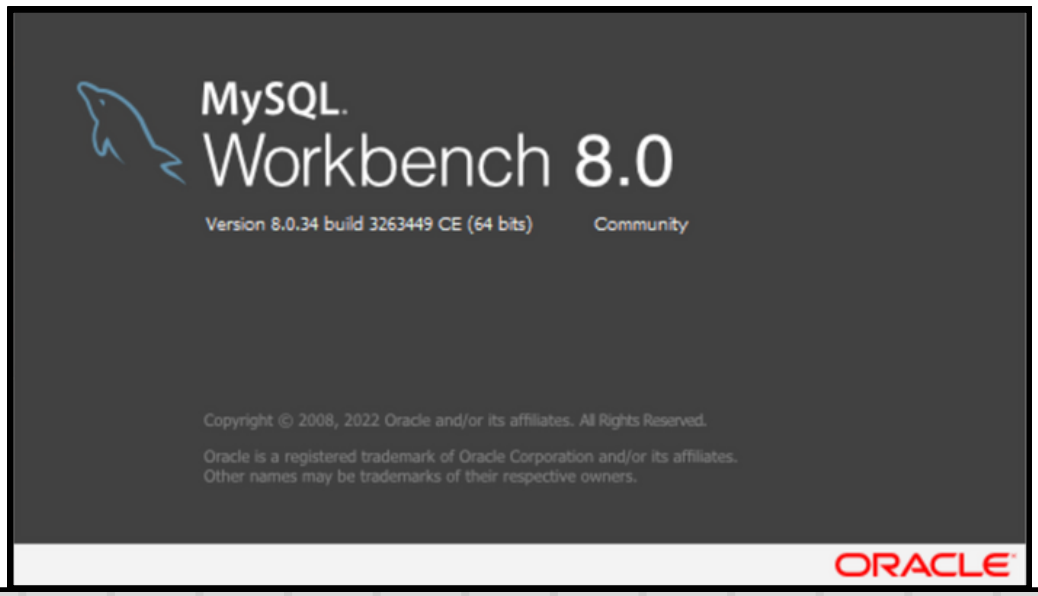
Instagram User Analytics

Description:-

Imagine you're a data analyst working with the product team at Instagram. Your role involves analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow.

User analysis involves tracking how users engage with a digital product, such as a software application or a mobile app. The insights derived from this analysis can be used by various teams within the business. Your insights will help the product manager and the rest of the team make informed decisions about the future direction of the Instagram app.

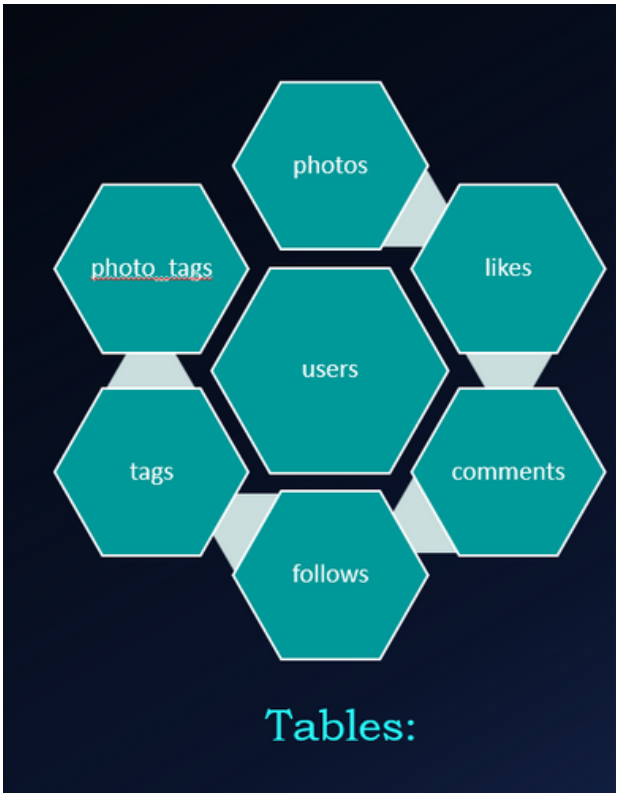
Tech-Stack Used:-



Instagram User Analytics

Approach :-

- In this project, as we are working with product team at Instagram.
- We have few tables containing data relevant to the platform.
- By these tables we can give insights to the marketing and product team to improve overall user experience.



Instagram User Analytics

Marketing Analysis :-

1. Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time.

Your Task: Identify the five oldest users on Instagram from the provided database.

```
80
81
82      #loyalty user reward
83  •   select id,username, min(created_at) as oldest_user
84      from users
85      group by id
86      order by oldest_user
87      limit 5
88      ;
89
90
```

Result Grid | Filter Rows:

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

Instagram User Analytics

Marketing Analysis :-

2. Inactive User Engagement: The team wants to encourage inactive users to start posting by sending them promotional emails.

Your Task: Identify users who have never posted a single photo on Instagram.

```
94
95      #Inactive User Engagement
96  •   select distinct users.id,username
97      from users where id not in
98      (select users.id
99      from users
100     inner join photos
101     on users.id=photos.user_id )
102     ;
103
104
105
```

Result Grid | Filter Rows:

id	username
5	Aniya_Hackett
7	Kasandra_Homenick
14	Jadyn81
21	Rocio33
24	Maxwell.Halvorson
25	Tierra.Trantow
34	Pearl7
36	Ollie_Ledner37

users 5

Instagram User Analytics

Marketing Analysis :-

3. Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins.

Your Task: Determine the winner of the contest and provide their details to the team.

```
157
158     #Contest Winner Declaration
159 •   with likes_no as(
160     select photo_id,photos.user_id, count(likes.user_id) as no_of_users_l
161     from likes
162     inner join users
163     on likes.user_id=users.id
164     inner join photos
165     on likes.photo_id=photos.id
166     group by photo_id
167     )
168     select photo_id,user_id,no_of_users_liked,username from likes_no
169     inner join (select username,id from users) as username_info
170     on likes_no.user_id=username_info.id
171     order by no_of_users_liked desc
172     limit 1
173     ;
```

Result Grid | Filter Rows: | Export:

photo_id	user_id	no_of_users_liked	username
145	52	48	Zack_Kemmer93

Instagram User Analytics

Marketing Analysis :-

4. Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.

Your Task: Identify and suggest the top five most commonly used hashtags on the platform.

```
176
177     #Question 4 hashtag research
178 •   select tag_id,tags.tag_name,count(photo_id) as tag_used_per_photo
179     from photo_tags
180     inner join tags
181     on photo_tags.tag_id=tags.id
182     group by tag_id
183     order by tag_used_per_photo desc
184     limit 5;
185
```

Result Grid | Filter Rows:

tag_id	tag_name	tag_used_per_photo
21	smile	59
20	beach	42
17	party	39
13	fun	38
18	concert	24

Result 7

Instagram User Analytics

Marketing Analysis :-

5. Ad Campaign Launch: The team wants to know the best day of the week to launch ads.

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

```sql
#Question 5 Ad Campaign Launch
with weekdays as (
select id,username,weekday(created_at) as week_day
from users)
select count(id) as no_of_users,week_day,
    case
        when week_day = 0
            then "Monday"
        when week_day = 1
            then "Tuesday"
            when week_day = 2
            then "Wednesday"
            when week_day = 3
            then "Thursday"
            when week_day = 4
            then "Friday"
            when week_day = 5
            then "Saturday"
            when week_day = 6
            then "Sunday"
    end as days
from weekdays
```

Result Grid | Filter Rows:

no_of_users	week_day	days
16	3	Thursday
16	6	Sunday
15	4	Friday
14	1	Tuesday
14	0	Monday
13	2	Wednesday
12	5	Saturday

```sql
    end as days
from weekdays
group by week_day
order by no_of_users desc;
```

Instagram User Analytics

Investor Metrics :-

1. User Engagement: Investors want to know if users are still active and posting on Instagram or if they are making fewer posts.

Your Task: 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.

```
221
222  #average post per user
223  select users.id,count(photos.id)/users.id as numPost
224  from users
225  left join photos
226  on users.id=photos.user_id
227  group by users.id;
```

Result Grid | Filte

id	numPost
1	5.0000
2	2.0000
3	1.3333
4	0.7500
5	0.0000
6	0.8333
7	0.0000
8	0.5000
9	0.4444
10	0.3000
11	0.4545
12	0.3333
13	0.3846
14	0.0000

Result 12 ×

```
#Total posts/total user
SELECT
count(distinct NumPosts) / COUNT(distinct NumUsers) as 'Total_posts/Total_users'
FROM (select users.id as NumUsers,photos.id as NumPosts
from users
left join photos
on users.id=photos.user_id
) as num_of_posts;
```

Result Grid | Filter Rows:

Total_posts/Total_users
2.5700

Instagram User Analytics

Investor Metrics :-

2. Bots & Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts.

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

```
240
241   #counting no. of posts
242 • select count(photos.id) as total_photos
243   from photos;
244
245   #bots and fake accounts
246 • set @total_posts=257;
247 • select user_id,users.username,count(photo_id) as no_of_posts_liked
248   from likes
249   inner join users
250   on likes.user_id=users.id
251   group by user_id
252   having no_of_posts_liked=@total_posts;
253
```

Result Grid | Filter Rows: | Exp

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

Conclusion

In conclusion, my journey as a data analyst has been marked by a relentless pursuit of insights and a commitment to transforming raw data into actionable strategies. Throughout my portfolio, I have showcased proficiency in data analysis, leveraging programming languages, and employing advanced visualization tools to drive informed decision-making. The ability to dissect complex datasets, identify patterns, and communicate findings effectively has been at the core of my contributions.

As I continue to evolve in this dynamic field, I am excited about the prospect of tackling new challenges, embracing emerging technologies, and contributing to data-driven innovations. Thank you for exploring my portfolio, and I look forward to the opportunity to bring my analytical skills and passion for data to new and exciting projects.

Appendix

--->> Links to the file

ABC Call Volume Trend

Impact of Car Features

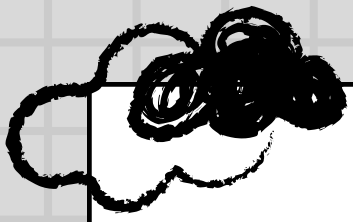
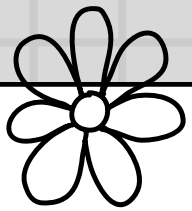
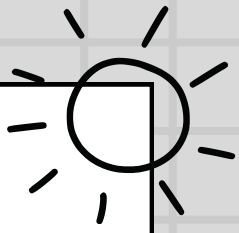
Bank Loan Case Study

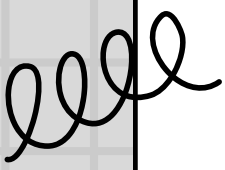
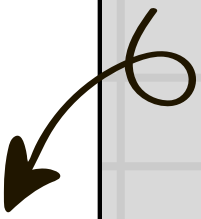
IMDB Movie Analysis

Hiring Process Analytics

Operation & Metric Analytics

Instagram User Analytics

Thank You