Data Analysis Portfolio

DIKSHA NEGI

Professional Background

I am Diksha Negi, a data analyst with a unique journey from zoology to data analytics. My academic background includes a master's degree in Zoology from the University of Delhi and have qualified for the UGC-JRF, reflecting a solid research and analytical foundation. My transition from biological sciences to data analytics has been fuelled by a passion for uncovering actionable insights and supporting strategic decision-making, I transitioned into data analysis to apply my investigative skills to new challenges.

Analytical Expertise and Skills

I have developed hands-on experience with SQL, Excel, Tableau, and Power BI. I enjoy turning raw data into valuable insights that guide decision-making and drive strategic growth. My expertise includes ensuring data quality through thorough cleaning, using statistical analysis to identify key trends, and creating dynamic, interactive dashboards in tools like Tableau and Power BI. These skills allow me to present complex information in an accessible and visually engaging way, making data more meaningful for stakeholders.

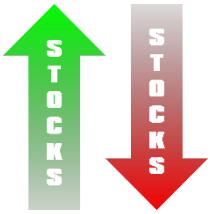
Commitment to Professional Growth

I am dedicated to continuous learning, demonstrated by certifications in Data Analysis Specialization, which reflect my commitment to advancing in this field. Staying current with new tools and methodologies allows me to bring both technical skills and a fresh perspective to every project.

Passion for Impact

With each project, I aim to create analyses that go beyond numbers. I am passionate about turning complex datasets into clear, actionable insights that inspire confidence, inform decisions, and make a real impact. Whether I am supporting a team with metrics or developing a data dashboard, my goal is to provide data solutions that drive meaningful results.

Table of contents



Project 1

Data Analytics- Application in Real Life Scenario Case Study

Description-

This project explores how we unknowingly use data analytics in everyday activities. Specifically, it examines the process of making a purchase, such as going to the market for clothing. By mapping this experience to the data analytics process—Plan, Prepare, Process, Analyze, Share, and Act—this project demonstrates how data-driven decision-making principles apply to common scenarios, highlighting how thoughtful planning and analysis can lead to better outcomes.

Problem statement-

In everyday decisions, such as shopping, people often aim to make choices that maximize value within constraints like budget, trends, and compatibility with existing items. However, without a structured approach, they may struggle to optimize their choices effectively. This project aims to illustrate how applying a data analytics process to such decisions can help in making more informed, satisfying choices, ultimately providing a simple, relatable framework for everyday data-driven decision-making.

Design-

Use of Data analytics in stock selection for long-term investment

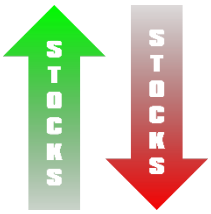
Stocks are most important tool of investment to attain long-term financial goals but the world stock has become so volatile and due to which it is also quite difficult by investing in a particular share or company that will give sustained performance over time. Therefore, taking an informed decision with the investment becomes essential. There is a good amount of data — historical

stock data to financial reports that provide valuable insights, if data are analysed correctly. Investors are able to use data analysis to evaluate potential investments, look for trends or risks associated with them. Through the use of statistical tools and predictive models, investors will take rational judgments on where to put their money, investing in stocks that will give consistent returns over long run and allows to avoid large losses.

The Problem:

Selecting the Right Stocks for Long-Term Investment Selecting the right stocks for long-term investment in this complex, rapidly changing uncertain markets is very challenging. All the stocks performing well in short term does not implies sustained growth over the long term too There are numerous factors to be considered by investors like company's financial health, market trends, competitions, and economic conditions.

The objective is to create a portfolio that balances risk with potential reward, as well as diversifying the portfolio for long term capital gain. Using data-driven approach, like assessing historical stock data, financial ratios, market trends and predictive modelling we can evaluate a range of companies to determine which stock aligns best with long term investment goals.



Six steps for data analysis-

1.Plan

• Objective: Selecting the stocks for the long-term investment that can outperform the market over a long period.

• Different criteria must be established for evaluating stocks such as financial health, growth potential and market position.

• Types of stocks to be considered, large-cap, mid-cap or small cap and the time frame in which you invest – 10/20/30 years.

2.Prepare:

• Data collection: Compile historical data on stock prices, f inancial statements, industry reports and market trends which can provide the foundation for evaluating performance and potential of stock.

• Establish a set of metrics to measure the stock's performance and potential. These could be the price-to-earnings ratio (P/E), return on equity (ROE), earnings per share (EPS) and others that indicate a company's financial health and growth prospects.

3.Process:

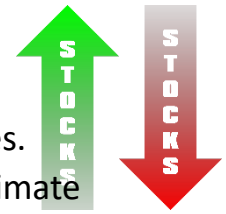
• Data cleaning: The vast amount of data cleaned to remove any inaccuracies, like duplicate entries, missing values, or outliers that could skew the analysis. • Standardize the data to ensure the consistency within it and normalise data formats.

• Create new variables such as growth rates, moving averages that might help in the analysis of data.

4.Analyze:

• Descriptive analysis: Analysis of historical data to understand the past performance of various stocks. This can be done by calculating key metrics like average returns, volatility, dividend yields and various financial ratios.

• Diagnostic analysis: Diagnostic analysis to understand reasons behind past performance. This involves analysing the impact of specific events like economic shifts or market crashes allowing to understand the underlying causes of observed trends.



• Predictive analysis: Predictive models to forecast future performances. Different tools like regression analysis, machine learning models to estimate future returns, risk and growth potential. It provides insights into which stocks are likely to perform well in the future, helping to take decisions.

5.Share:

• Data visualisation: Creating the visualisations using graphs, charts and tables to present the findings of analysis in easily understandable format. This includes performance of various stocks, risk assessments and returns.
• A detailed report which summarizes the analysis insights, including which stocks were selected and why, explaining the methodology, the results and recommendations.
• Presentation: A presentation to share the insights highlighting the key findings.

6.Act:

• Decision can be made based on the insights from analysis on which stocks to invest in. According to the investment strategy developed from the analysis, a portfolio that is diversified to balance potential rewards with risk can be selected.
• Implementation: The funds allocation to selected stocks and setting up the portfolio for long term investment.
• Monitoring the performance of portfolio regularly and make changes where required to stay aligned with long-term investment goals.
• Rebalancing the portfolio and reassessing stocks as new data becomes available.

This approach helps in making informed decisions that are based on comprehensive analysis.

PROJECT 2
INSTAGRAM USER ANALYTICS
SQL Fundamentals

Project description:
As a data analyst on Instagram product team, this project focuses on analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow. Using SQL and MySQL workbench as a tool, the goal is to track how users engage with a digital product, to provide insights derived from this analysis that can help various teams within the business to take informed decisions.

Approach:
• The questions were carefully reviewed to understand the requirements of the provide questions.
• The tables and columns that will be used to answer each question were identified.
• Required information was extracted using MySQL queries using MySQL software.
• The insights derived from the analysis were carefully reviewed to to further take informed decisions.

Tech-Stack Used:
I am using MySQL Workbench 8.0.38-winx64 CE and it is ideal for this project as it provides a user-friendly interface for running SQL queries, managing databases, visualising trends, making it easier to analyze and derive insights efficiently.

ANALYSIS:

SQL Tasks:

A) Marketing Analysis:

1.Loyal User Reward: Identify the five oldest users on Instagram from the provided database.

Run SQL query as:

```
USE ig_clone;

SELECT username, created_at
FROM users
ORDER BY created_at ASC
LIMIT 5;
```

Result:

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

The table shows the five oldest users on Instagram from provided database.

2. Inactive User Engagement: Identify users who have never posted a single photo on Instagram.

Query:

```
USE ig_clone;

select username, t1.id as `user id`
from users t1
left join photos t2
on t1.id = t2.user_id
where t2.user_id is null
```

Result:

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

The table shows 26 users that have never posted a single photo on Instagram. Promotional emails can be sent to these inactive users.

3. Contest Winner Declaration: Determine the winner of the contest and provide their details to the team.

Query:

```
USE ig_clone;

select t3.id, t3.username, photo_id, t2.image_url , count(*) as number_of_likes
from likes t1
inner join photos t2 on t1.photo_id = t2.id
inner join users t3 on t2.user_id = t3.id
group by photo_id
order by number_of_likes DESC
limit 1;
```

Result:

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

The table shows that Zack_kemmer93 has most of the likes in single photo and thus is winner of the contest.

4. Hashtag Research: Identify and suggest the top five most used hashtags on the platform.

Query:

```
use ig_clone;

select tag_name, count(*) as most_used_hashtag
from photo_tags t1
inner join tags t2
on t1.tag_id = t2.id
group by tag_name
order by most_used_hashtag DESC
limit 5;
```

Result:

	tag_name	most_used_hashtag
▶	smile	59
	beach	42
	party	39
	fun	38
	concert	24

The table shows top five most used hashtags on the platform.

5. Ad Campaign Launch: Determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign.

Query:

```
use ig_clone;

select  dayname(created_at) as weeks, count(*) as most_users_registered
from users
group by weeks
order by most_users_registered DESC;
```

Result:

weeks	most_users_registered
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

The table shows most users registered on Thursdays and Sundays on Instagram, so these days are best to schedule an ad campaign.

B) Investor Metrics:

1.User Engagement: 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.

a. Query for average number of posts per user

```
use ig_clone;

select avg(number_of_post) as avg_post_per_user
from
(select user_id, count(*) as number_of_post
from photos
group by user_id ) as average_post;
```

Result:

avg_post_per_user
3.4730

The table shows that average user posts more than 3 photos.

b. Query for total number of photos on Instagram divided by the total number of users.

```
use ig_cone;

select count(id) / (select count(id) as total_users
from users) as `total photos to total users`
from photos ;
```

Result:

	total photos to total users
▶	2.5700

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

Query:

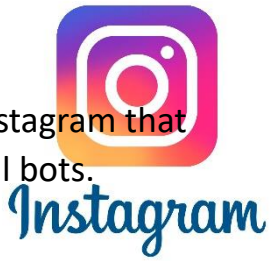
```
use ig_clone;

select username, user_id, count(*) as total_likes
from users
inner join likes
on users.id = likes.user_id
group by user_id
having total_likes = (select count(*) from photos);
```

Result:

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

The table shows the users that have liked every single photo on Instagram that Is typically not normal for very user so these users can be potential bots.



Insights:

The various insights were derived from the analysis of data:

• The most loyal users, i.e., those who have been using the platform for the longest time have been using the platform since 2016

• Some users are inactive on the platform so promotional emails can be scheduled for those users to encourage them to start posting.

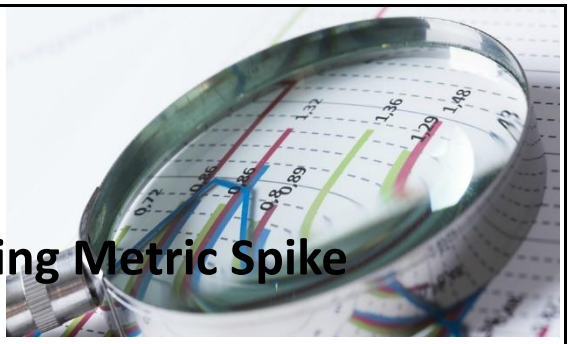
• The most used hashtags on the platform are smile, beach, party, fun and concert.

• Thursdays and Sundays are the days on which most users registered on Instagram so these days can be scheduled to launch an ad campaign.

• Average user posts more than 3 photos on Instagram.

• There are potential bots of fake accounts on Instagram as they liked every post on platform which is not normal for a normal user.

All these insights can be used by various departments to take informed decisions that can help business grow.



PROJECT 3

Operation Analytics and Investigating Metric Spike

Advanced SQL

Project description:

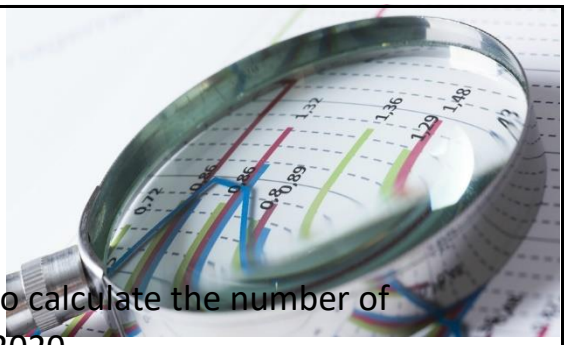
This project focuses on Operational Analytics to enhance the company's efficiency. As a Lead Data Analyst at a major tech company like Microsoft, my responsibility will be to analyze various datasets and tables to derive insights that can answer concerns raised by different departments like operations, support, and marketing.

A key aspect of my role will be investigating sudden changes in key metric like drops in sales or decreases in daily user engagement. By advanced SQL skills, I will analyze causes of sudden changes , enabling the company to make informed decisions.

Approach:

- The database was created using the structured data provided.
- The questions were carefully reviewed to understand the requirements of the provided questions.
- The tables and columns that will be used to answer each question were identified.
- Required information was extracted using MySQL queries using MySQL software.
- The insights derived from the analysis were carefully reviewed to further take informed decisions.

Tech-Stack Used: I am using MySQL Workbench 8.0.38-winx64 CE and it is ideal for this project as it provides a user-friendly interface for running SQL queries, managing databases, visualising trends, making it easier to analyze and derive insights efficiently.



Case Study 1: Job Data Analysis

Tasks:

A. Jobs Reviewed Over Time: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

Query:

```
use project_3;

select ds as review_date,count(job_id) as jobs_per_day, sum(time_spent)/3600 as hours_spent
from job_data
where ds between '2020-11-01'  and  '2020-11-30'
group by ds
order by ds;
```

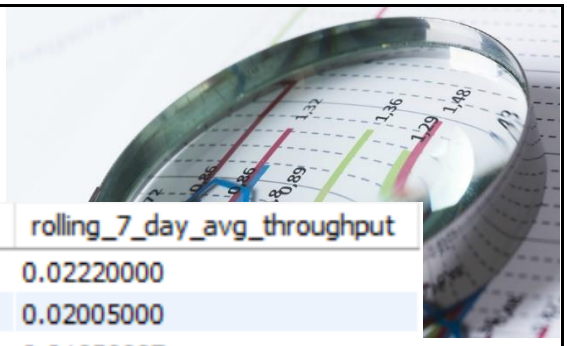
Result:

review_date	jobs_per_day	hours_spent
2020-11-25	1	0.0125
2020-11-26	1	0.0156
2020-11-27	1	0.0289
2020-11-28	2	0.0092
2020-11-29	1	0.0056
2020-11-30	2	0.0111

B. Throughput Analysis: 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.

QUERY:

```
WITH daily_throughput AS (
    SELECT
        ds,
        COUNT(job_id) AS job_count,              -- Total jobs completed per day
        SUM(time_spent) AS total_time_spent      -- Total time spent on jobs per day
    FROM
        job_data
    GROUP BY
        ds
)
SELECT
    ds,
    job_count,
    total_time_spent,
    job_count / NULLIF(total_time_spent, 0) AS jobs_per_time,  -- Throughput (jobs per time)
    AVG(job_count / NULLIF(total_time_spent, 0)) OVER (
        ORDER BY ds
        ROWS BETWEEN 6 PRECEDING AND CURRENT ROW
    ) AS rolling_7_day_avg_throughput              -- 7-day rolling average
FROM
    daily_throughput
ORDER BY
    ds;
```



Result:

ds	job_count	total_time_spent	jobs_per_time	rolling_7_day_avg_throughput
2020-11-25	1	45	0.0222	0.02220000
2020-11-26	1	56	0.0179	0.02005000
2020-11-27	1	104	0.0096	0.01656667
2020-11-28	2	33	0.0606	0.02757500
2020-11-29	1	20	0.0500	0.03206000
2020-11-30	2	40	0.0500	0.03505000

Insights:

Daily Metric vs. 7-Day Rolling Average

1. Daily Metric: The daily metric gives throughput for each specific day. It is useful for identifying day to day fluctuations but daily metric can be volatile, especially if there are irregular patterns in the data.

2. 7-Day Rolling Average: The 7-day rolling average smooths out daily fluctuations by averaging the throughput over a week making it easier to spot long term trends and patterns.

I would prefer the 7-day rolling average for throughput analysis to understand trends over. The rolling average provides a more stable measure by smoothing out fluctuations that helps in making informed decisions based on consistent patterns. But keeping an eye on the daily metric might be necessary as well if immediate responsiveness to changes is important.

C. Language Share Analysis: Write an SQL query to calculate the percentage share of each language over the last 30 days.

Query:

```
use project_3;

Select language, count(language) as total_language,
       ROUND((COUNT(*) * 100.0) / SUM(COUNT(*)) OVER (),2) AS percentage_share
from job_data
group by language
order by percentage_share DESC;
```

Result:

	language	total_language	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

Insights:

The above table shows that Persian language accounts for 37.5% of job reviews so it is the most prevalent language in the dataset.

Other languages each have a 12.5% share, showing a balanced but lesser representation.

D. Duplicate Rows Detection: Write an SQL query to display duplicate rows from the job_data table.

Query:

```
use project_3;

 select *
 from (
      select*, row_number() over(partition by job_id) as row_num
      from job_data) as rowcount
where row_num > 1;
```

Result:

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

Insights:

The above table shows that rows 2 and 3 are duplicates when partitioning the data by job_id.

Case Study 2: Investigating Metric Spike

Tasks:

A. Weekly User Engagement: Write an SQL query to calculate the weekly user engagement.

Query:

```
use project_3;

select extract(week from occurred_at) as weeks , count(distinct user_id) as active_users
from events
where event_type = 'engagement'
group by weeks;
```

Result:

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

Insights:

The table shows weekly user engagement. The user engagement has been increasing from 17th week to 30th week but has been declining after that showing users are not finding quality in product or service.

B. User Growth Analysis: Write an SQL query to calculate the user growth for the product.

Query:

```
use project_3;

SELECT
    DATE_FORMAT(activated_at, '%Y-%m') AS month,
    COUNT(*) AS new_users,
    SUM(COUNT(*)) OVER (ORDER BY DATE_FORMAT(activated_at, '%Y-%m')) AS cumulative_users
FROM
    users
GROUP BY
    month
ORDER BY
    month;
```

Result:

	month	new_users	cumulative_users
▶	2013-01	160	160
	2013-02	160	320
	2013-03	150	470
	2013-04	181	651
	2013-05	214	865
	2013-06	213	1078
	2013-07	284	1362
	2013-08	316	1678
	2013-09	330	2008
	2013-10	390	2398
	2013-11	399	2797
	2013-12	486	3283
	2014-01	552	3835
	2014-02	525	4360
	2014-03	615	4975
	2014-04	726	5701
	2014-05	779	6480
	2014-06	873	7353
	2014-07	997	8350
	2014-08	1031	9381

Insights:

The table shows the growth of users over time for a product. From Jan,2013 to Aug,2014 there are over 9300 active users providing a view of overall user growth.

C. Weekly Retention Analysis: Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.

Query:

```
use project_3;

WITH cohorts AS (
    SELECT
        DATE(activated_at) AS cohort_start_date,
        COUNT(*) AS total_users
    FROM users
    GROUP BY 1
),
weekly_stats AS (
    SELECT
        DATE(u.activated_at) AS cohort_start_date,
        YEARWEEK(e.occurred_at, 0) AS year_week,
        COUNT(DISTINCT e.user_id) AS active_users
    FROM users u
    JOIN events e ON u.user_id = e.user_id
    WHERE e.event_type = 'engagement'
    GROUP BY cohort_start_date, year_week
)
SELECT
    cohorts.cohort_start_date,
    weekly_stats.year_week,
    weekly_stats.active_users,
    cohorts.total_users AS total_users,
    weekly_stats.active_users / cohorts.total_users * 100 AS retention_rate
FROM cohorts
JOIN weekly_stats
    ON cohorts.cohort_start_date = weekly_stats.cohort_start_date
ORDER BY cohort_start_date, year_week;
```

Result:

cohort_start_date	year_week	active_users	total_users	retention_rate
2013-01-01	201417	1	7	14.2857
2013-01-01	201418	1	7	14.2857
2013-01-01	201419	2	7	28.5714
2013-01-01	201420	2	7	28.5714
2013-01-01	201421	1	7	14.2857
2013-01-01	201422	1	7	14.2857
2013-01-01	201423	1	7	14.2857
2013-01-01	201424	2	7	28.5714
2013-01-01	201425	2	7	28.5714
2013-01-01	201426	1	7	14.2857
2013-01-01	201427	1	7	14.2857
2013-01-01	201430	2	7	28.5714
2013-01-01	201431	1	7	14.2857
2013-01-02	201417	1	7	14.2857
2013-01-02	201418	2	7	28.5714
2013-01-02	201419	1	7	14.2857
2013-01-02	201420	1	7	14.2857
2013-01-02	201421	1	7	14.2857
2013-01-02	201422	2	7	28.5714

Insights:

The result shows how many users from a specific signup cohort remain active over subsequent weeks.

D. Weekly Engagement Per Device: Write an SQL query to calculate the weekly engagement per device.

Query:

```
use project_3;

SELECT
    DATE_FORMAT(occurred_at, '%Y-%u') AS week,
    device,
    COUNT(DISTINCT user_id) AS active_users
FROM events
GROUP BY week, device;
```

Result:

week	device	active_users
2014-31	macbook pro	317
2014-32	macbook pro	317
2014-34	macbook pro	308
2014-33	macbook pro	307
2014-28	macbook pro	301
2014-29	macbook pro	295
2014-30	macbook pro	291
2014-35	macbook pro	290
2014-26	macbook pro	276
2014-20	macbook pro	261
2014-24	macbook pro	259
2014-27	macbook pro	259
2014-21	macbook pro	256
2014-23	macbook pro	254
2014-25	macbook pro	251
2014-19	macbook pro	248
2014-22	macbook pro	244
2014-29	lenovo think...	220
2014-30	lenovo think...	209

Insights:

The activeness of users on a weekly basis per device shows that MacBook pro has the most active users.

E. Email Engagement Analysis: Write an SQL query to calculate the email engagement metrics.

Query:

```
SELECT
    DATE_FORMAT(occurred_at, '%Y-%u') AS week,
    COUNT(CASE WHEN action IN ('sent_weekly_digest', 'sent_reengagement_email') THEN 1 ELSE NULL END) AS em
    COUNT(CASE WHEN action = 'email_open' THEN 1 ELSE NULL END) AS emails_opened,
    COUNT(CASE WHEN action = 'email_clickthrough' THEN 1 ELSE NULL END) AS emails_clicked,
    ROUND(
        100.0 * COUNT(CASE WHEN action = 'email_open' THEN 1 ELSE NULL END) /
        NULLIF(COUNT(CASE WHEN action IN ('sent_weekly_digest', 'sent_reengagement_email') THEN 1 ELSE NULL
        2
    ) AS email_open_rate,
    ROUND(
        100.0 * COUNT(CASE WHEN action = 'email_clickthrough' THEN 1 ELSE NULL END) /
        NULLIF(COUNT(CASE WHEN action = 'email_open' THEN 1 ELSE NULL END), 0),
        2
    ) AS email_click_rate
FROM email_events
WHERE action IN ('sent_weekly_digest', 'sent_reengagement_email', 'email_open', 'email_clickthrough')
GROUP BY week
ORDER BY week;
```

Result:

week	emails_sent	emails_opened	emails_clicked	email_open_rate	email_click_rate
2014-18	1006	332	187	33.00	56.33
2014-19	2766	919	434	33.22	47.23
2014-20	2840	971	479	34.19	49.33
2014-21	2912	995	498	34.17	50.05
2014-22	3001	1026	453	34.19	44.15
2014-23	3110	993	492	31.93	49.55
2014-24	3193	1070	533	33.51	49.81
2014-25	3339	1161	563	34.77	48.49
2014-26	3394	1090	524	32.12	48.07
2014-27	3524	1168	559	33.14	47.86
2014-28	3613	1230	622	34.04	50.57
2014-29	3725	1260	607	33.83	48.17
2014-30	3798	1211	584	31.89	48.22
2014-31	3936	1386	633	35.21	45.67
2014-32	3999	1336	432	33.41	32.34
2014-33	4121	1357	430	32.93	31.69
2014-34	4269	1421	487	33.29	34.27
2014-35	4374	1533	493	35.05	32.16

Insights:

The above table shows how users are engaging with the email service. The data is on the weekly basis showing metrics like email open rate and email click rate. This query helps in understanding how users are engaging with email campaigns over time which provides insights into the effectiveness of email content.

Result:

Through this project I was able to manage a structured database for the project. The project helped in gaining understanding of managing tables, applying joins, and performing complex queries to extract relevant information through hands-on experience in performing data analysis using SQL queries, filtering data that provided valuable insights.

From this project, I learned how to identify and analyze unexpected changes in metrics to understand trends, helping improve operational performance and make informed decisions.

Also, applied MySQL in a real-world case, translating theoretical knowledge into practical use which is essential for any data-driven role.

PROJECT-4

Hiring Process Analytics

Statistics

Project description:

As a data analyst, this project involves analyzing the hiring process data of a multinational company, such as Google, to derive valuable insights that can optimize recruitment strategies. The dataset includes records on job types, vacancies, interviews, and rejections.

Using Excel and statistical methods, I will clean, organize, and process the data, identifying key patterns such as rejection rates, interview success rates, and job vacancy trends. Also to provide insights by visualizing data through charts and graphs, that can improve hiring decisions and streamline recruitment efforts. The analysis will help the company make data-driven decisions to enhance the efficiency of the hiring process.

Approach:

- The first step in this analysis will be handling missing data by identifying any gaps in the dataset and deciding the most appropriate strategy to manage them, such as removing incomplete records or imputing missing values.
- Next, I will simplify the dataset by clubbing columns where multiple categories can be combined, which will streamline the analysis and provide clearer insights.
- Outlier detection will be crucial to ensure the data is not skewed by extreme values. Based on the nature and impact of the outliers, I will either remove, replace, or retain them.
- Once the dataset is clean, I will summarize the data by calculating relevant statistical measures such as averages and medians and generating visualizations to uncover trends related to rejections, interviews, job types, and vacancies.

This comprehensive analysis will help in deriving insights that can inform the company's recruitment strategy, leading to a more efficient and data-driven hiring process.

Tech-Stack Used: Microsoft Excel 2019

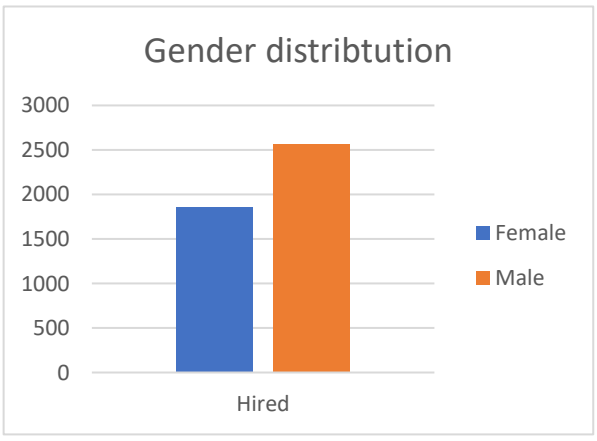
Excel is an excellent tool for handling, analyzing, and visualizing datasets of moderate size. It offers functionalities for data cleaning and statistical analysis.

Excel's built-in charts and pivot tables are useful for creating visual representations like bar charts, pie charts, and histograms, which help in understanding trends in the dataset related to rejections, interviews, job types, and vacancies.

Data Analytics Tasks:

A. Hiring Analysis: Determine the gender distribution of hires. How many males and females have been hired by the company?

Count of new_event_name	Column Labels		
Row Labels	Female	Male	Grand Total
Hired	1854	2562	4416
Grand Total	**1854**	**2562**	**4416**



Insights: The chart shows the gender distribution of hires. 2562 males and 1854 females have been hired by the company.

B. Salary Analysis: What is the average salary offered by this company? Use Excel functions to calculate this.

Function used to calculate the average salary offered by the company

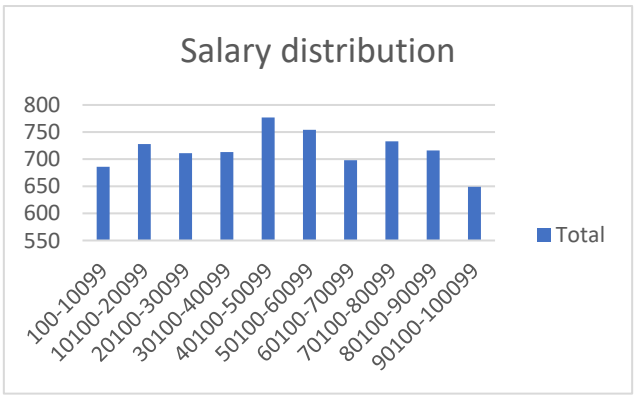
=Average (H2:H7166)

Output: **49878.3464**

Insight: The average salary that is offered by the company is approximately 49878.

C. Salary Distribution: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

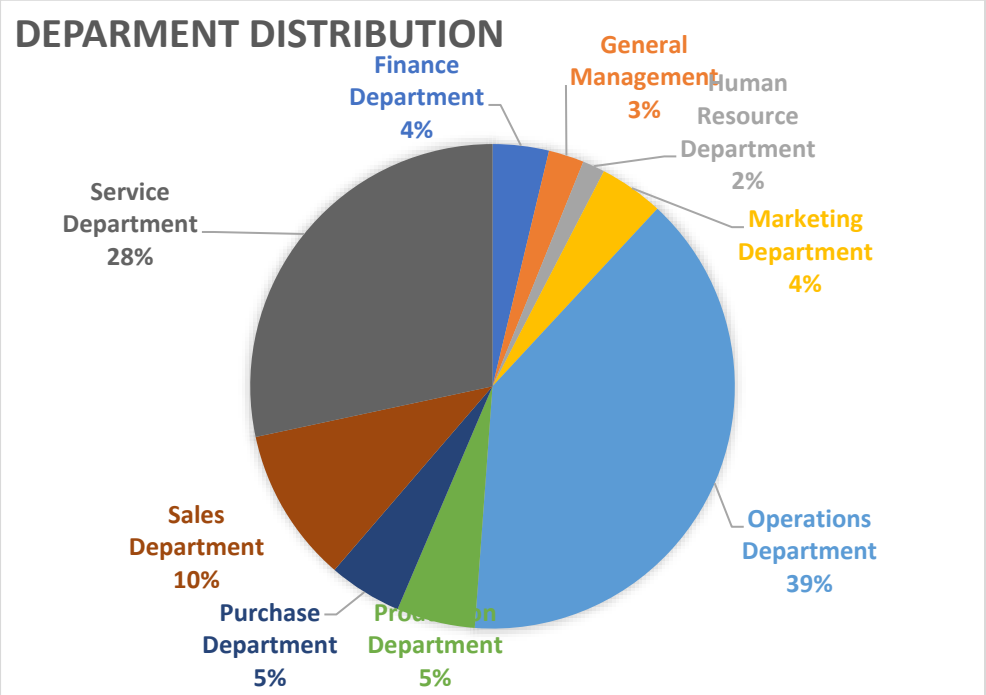
Row Labels	Count of Offered Salary
100-10099	686
10100-20099	728
20100-30099	711
30100-40099	713
40100-50099	777
50100-60099	754
60100-70099	698
70100-80099	733
80100-90099	716
90100-100099	649
Grand Total	**7165**



Insights: The above table shows the salary distribution in company and charts helps us to understand how distribution is visually. A balanced distribution shows a well-compensated workforce across all levels. This analysis helps in assessing pay structure and workforce composition.

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

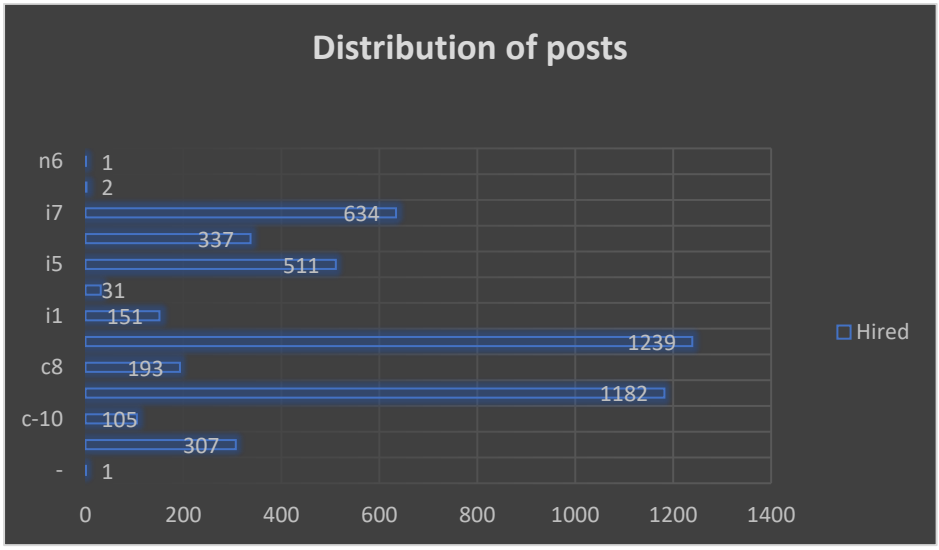
Count of application_id Row Labels	Column Labels Hired
Finance Department	3.75%
General Management	2.36%
Human Resource Department	1.49%
Marketing Department	4.30%
Operations Department	39.26%
Production Department	5.24%
Purchase Department	4.90%
Sales Department	10.33%
Service Department	28.36%
Grand Total	**100.00%**



Insights: The above pie chart shows the proportion of people working in different departments. 39% the highest proportion of people are working in operations department and with 2% the lowest proportion of people are working in human resource department.

E. Position Tier Analysis: 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.

Count of application_id	Column Labels
Row Labels	Hired
-	1
b9	307
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	31
i5	511
i6	337
i7	634
m6	2
n6	1
Grand Total	**4694**



Insights: The above graph shows the distribution of positions across different tiers. The position tier analysis shows how employees are distributed across various levels within the organization.

Result: From this project, gained valuable insights into the company's hiring process, identifying trends such as the number of rejections, interviews, and job vacancies. By handling missing data, detecting and managing outliers, I have more streamlined dataset to work with. Using statistical analysis, I was

able to calculate key metrics such as the average number of interviews per hire or the rejection rate, which will help inform future hiring strategies. The visualizations created provide insights that can improve decision-making in the hiring process, optimize recruitment efforts, and ensure better allocation of resources.

Drive link: For the Excel Sheet file **click this text.**

PROJECT 5

IMDB Movie Analysis

Project description:

This project aims to understand the factors that contribute to a movie's success on IMDB, with success defined by high ratings. This project will dive into a dataset of IMDB movies to explore how different factors like genre, director, budget, release year, and actors impact a movie's rating.

To start, the data will be cleaned and prepped, fixing any missing values, duplicates, or inconsistencies so it is ready for analysis. The main goal is to uncover patterns and relationships between these factors and movie ratings. Using the "Five Whys" method, we will go beyond surface-level insights to dig into the deeper reasons behind the trends we find.

This problem holds great importance for movie producers, directors, and investors who are looking to understand what drives a movie's success. By uncovering the key factors that lead to higher ratings, they can make smarter decisions in their future projects, ensuring better outcomes both creatively and financially.

Approach:

- Import the dataset into Excel and clean it by removing duplicates, handling missing values like adding values like in column duration from searching for it on google and filling these values in place of blanks and deleting rows where value cannot be replaced, dropping columns that are not required, splitting the column and ensuring consistent data types.
- Standardize data using Excel functions.
- Once the dataset is clean, I will summarize the data by calculating relevant statistical measures such as averages and medians and generating visualizations to uncover trends related to rejections, interviews, job types, and vacancies.

Tech-stack used:

Microsoft Excel 2019

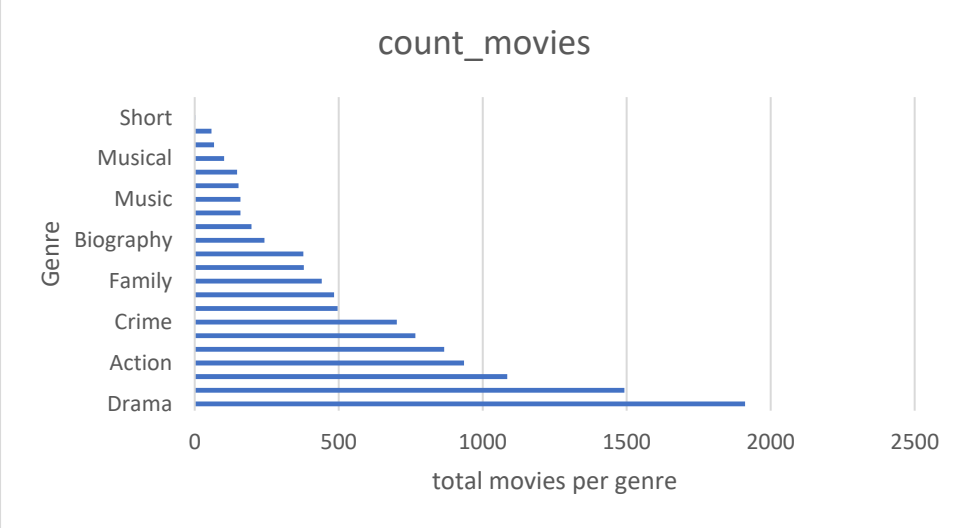
Excel is an excellent tool for handling, analyzing, and visualizing datasets of moderate size. It offers functionalities for data cleaning and statistical analysis. Excel's built-in charts and pivot tables are useful for creating visual representations like bar charts, pie charts, and histograms, which help in understanding trends in the dataset related to rejections, interviews, job types, and vacancies.

Data Analytics Tasks:

A. Movie Genre Analysis: 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.

1. Movie genre analysis

unique_genre	count_movies	mean	median	mode	max	min	Range	variance	st.Dev
Drama	1911	4.137089981	6.9	6.7	9.3	2.1	7.2	0.793996652	0.891064898
Comedy	1492	5.271996951	6.3	6.3	8.8	1.9	6.9	1.081431552	1.039919012
Thriller	1085	2.234142857	6.4	6.5	9	2.7	6.3	0.938931964	0.968985017
Action	935	6.285989305	6.3	6.6	9	2.1	6.9	1.078186788	1.038357736
Romance	866	2.096107383	6.5	6.5	8.5	2.1	6.4	0.940456888	0.969771565
Adventure	766	4.312827225	6.6	6.6	8.9	2.3	6.6	1.247524378	1.116926308
Crime	702	4.211811966	6.6	6.6	9.3	2.4	6.9	0.968463042	0.984105199
Fantasy	496	2.9923125	6.4	6.7	8.9	2.2	6.7	1.30054464	1.140414241
Sci-Fi	484	2.517368421	6.4	7	8.8	1.9	6.9	1.362318841	1.16718415
Family	441	2.138181818	6.3	5.4	8.6	1.9	6.7	1.367909091	1.169576458
Horror	379	4.092354949	5.9	6.2	8.6	2.3	6.3	0.982127152	0.991023285
Mystery	377	2.669591837	6.5	6.6	8.6	3.1	5.5	1.014838309	1.007391835
Biography	242	6.478361345	7.2	7	8.9	4.5	4.4	0.504237338	0.71009671
Animation	197	3.291309524	6.8	7.3	8.6	2.8	5.8	0.987295659	0.993627525
War	159	2.2	7.1	7.1	8.6	4.3	4.3	0.652386753	0.80770462
Music	159	2.03880597	6.5	6.5	8.5	1.6	6.9	1.473940769	1.214059623
History	152	2.147021277	7.2	7.7	8.9	5.5	3.4	0.451578947	0.671996241
Sport	147	2.08025641	6.8	7.2	8.4	2	6.4	1.09876526	1.048220043
Musical	102	2.4945	6.7	7.1	8.5	2.1	6.4	1.307672297	1.143535
Documentary	67	5.682807018	7.2	6.6	8.5	1.6	6.9	1.439855269	1.199939694
Western	58	3.509090909	6.8	6.8	8.9	4.1	4.8	0.997035693	0.998516746



Insights:

1. The above table shows that “Drama” followed by “comedy” and “thriller” are the most common genres in the dataset. By identifying which genres are most common in the dataset, we can understand the genre distribution and popularity within the dataset.

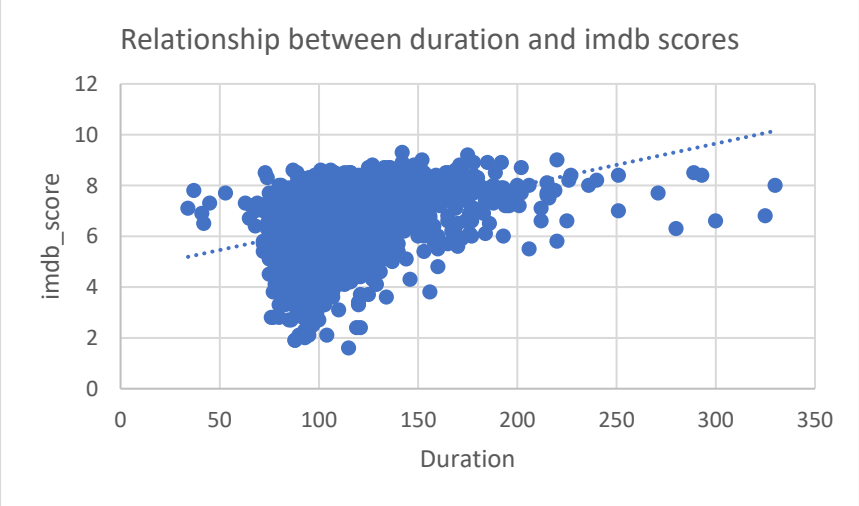
2. The average rating helps identify which genres tend to have higher overall ratings. Here "Action" has a higher average rating compared to "Drama" which suggests that on average, actions are rated more favourably by viewers.
3. The median rating provides a measure of central tendency that is less affected by extreme values. The median rating is significantly different from the average rating, indicating that the ratings for that genre might be skewed by a few very high or very low ratings.
4. Knowing which genres are most common helps understand dataset composition.

B. Movie Duration Analysis: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Distribution of duration

Distribution of movie duration	
Mean	109.8175
Median	105
st.Dev	22.76698

Relationship between movie duration and the imdb scores by scatter plot



Insights:

1. **Distribution:** The descriptive analysis shows the average duration of movies is 109 minutes and the standard deviation of 22.7 reflects the amount of variation or dispersion from that mean.
2. This shows that most values fall within the range of approximately 86.3 to 131.7 as mean ± st.dev.

3. **Relationship:** The above scatter plot shows the relationship between movie duration and imdb is positive correlation as points trend upwards from left to right suggesting that as movie duration increases, imdb scores tend to increase.
4. Also, it shows longer movies may have higher imdb scores but relationship may not be very strong as points are kind of dispersed around the line.

Now if we ask "five whys"

- Why do longer movies receive higher IMDb scores?
  Because they often have more developed plots and characters.

- Why do more developed plots and characters lead to higher scores?
  Because audiences may feel more invested in the story and characters.

- Why do audiences feel more invested in longer films?
  Because they have more time to connect with the characters and plot.

- Why does having more time to connect matter to audiences?
  Because emotional engagement can enhance overall enjoyment of the film.

- Why does emotional engagement enhance enjoyment?
  Because it creates a more immersive experience, making the film memorable.

This helps to dig deeper into the problem.

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

distinct_language	count_movies	mean	median	st.dev
English	3607	6.42107	6.5	1.052583
French	37	7.286486	7.2	0.561329
Spanish	26	7.05	7.15	0.826196
Mandarin	14	7.021429	7.25	0.765786
German	13	7.692308	7.7	0.640913
Japanese	12	7.625	7.8	0.899621
Hindi	10	6.76	7.05	1.111755
Cantonese	8	7.2375	7.3	0.440576
Italian	7	7.185714	7	1.155319
Korean	5	7.7	7.7	0.570088
Portuguese	5	7.76	8	0.978775
Norwegian	4	7.15	7.3	0.574456
Dutch	3	7.566667	7.8	0.404145
Thai	3	6.633333	6.6	0.450925
Danish	3	7.9	8.1	0.52915
Hebrew	3	7.5	7.3	0.43589
Persian	3	8.133333	8.4	0.550757
Aboriginal	2	6.95	6.95	0.777817
Dari	2	7.5	7.5	0.141421
Indonesian	2	7.9	7.9	0.424264

Insights:

The table above shows the distribution of movies based on their language.

The table shows the most common languages for a movie and the average scores for movies in each language. Some languages like Persian, French, German, Dutch have the highest average ratings for score and typical median scores being approximately around 7 for these languages.

Standard deviation shows how spread out scores are for each language and it is low depicting more consistent movie ratings.

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

Output

unique_directors	avg_imbd	percentile
Charles Chaplin	8.6	1
Majid Majidi	8.5	1
Andrew Haigh	7.7	1
Kevin Jordan	7.6	1
Jafar Panahi	7.5	1
Kiyoshi Kurosawa	7.4	1
Shane Carruth	7	1
Neill Dela Llana	6.3	1
Tony Kaye	8.6	0.998
Christopher Nolan	8.425	0.995
Sergio Leone	8.433333333	0.994
Alfred Hitchcock	8.5	0.994
Richard Marquand	8.4	0.992
Damien Chazelle	8.5	0.992
Lee Unkrich	8.3	0.991
Pete Docter	8.233333333	0.991
S.S. Rajamouli	8.4	0.991
Quentin Tarantino	8.2	0.99
Ron Fricke	8.5	0.99
Hayao Miyazaki	8.225	0.988
Lenny Abrahamson	8.3	0.988
Fritz Lang	8.3	0.987
Milos Forman	8.133333333	0.985
Marius A. Markevicius	8.4	0.985
Billy Wilder	8.3	0.983

Insights:

The above table shows the influence of directors on movie ratings.

The directors in 90th percentile or above are the top 10% of directors based on idmb scores. It can be considered these directors to be consistently delivering high quality movies as per the imdb ratings.

Directors around the 50th percentile produce movies that are typically average in ratings.

Directors in lower percentiles may have few poorly received movies dragging d[illegible] average or they might be new.

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

movie_title	budget	gross	Profit
Avatar	237000000	760505847	523505847
Jurassic World	150000000	652177271	502177271
Titanic	200000000	658672302	458672302
Star Wars: Episode IV - A New Hope	11000000	460935665	449935665
E.T. the Extra-Terrestrial	10500000	434949459	424449459
The Avengers	220000000	623279547	403279547
The Lion King	45000000	422783777	377783777
Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677
The Dark Knight	185000000	533316061	348316061
The Hunger Games	78000000	407999255	329999255
Deadpool	58000000	363024263	305024263
The Hunger Games: Catching Fire	130000000	424645577	294645577
Jurassic Park	63000000	356784000	293784000
Despicable Me 2	76000000	368049635	292049635
American Sniper	58800000	350123553	291323553
Finding Nemo	94000000	380838870	286838870
Shrek 2	150000000	436471036	286471036
The Lord of the Rings: The Return of the King	94000000	377019252	283019252
Star Wars: Episode VI - Return of the Jedi	32500000	309125409	276625409
Forrest Gump	55000000	329691196	274691196
Star Wars: Episode V - The Empire Strikes Back	18000000	290158751	272158751
Home Alone	18000000	285761243	267761243
Star Wars: Episode III - Revenge of the Sith	113000000	380262555	267262555
Spider-Man	139000000	403706375	264706375

Output

Correlation coefficient	
0.0965404	
Highest profit margin	
523505847	
Movie with highest profit margin	
1	row number
Avatar	the movie name with highest profit margin

Insights:

A correlation coefficient of 0.09 indicates a very weak positive relationship between movie budgets and gross earnings. The weak correlation shows that spending more on a movie's budget does not relate to significantly higher gross earnings.

The profit margin of some movies is positive showing the movie made more money than costs to produce meaning the higher the margin, the more profitable the movie and some has negative profits showing they were in loss.

The Avatar movie had the maximum profit margin indication the movie was profitable.

Profit margin showed a clear picture of how well a movie performed financially.

Result:

In this project, using Excel's statistical functions for data analysis provided a powerful tool to extract meaningful insights from the data. Functions such as CORREL enable you to assess the strength and direction of relationships between variables, like movie budgets and gross earnings. Calculating AVERAGE and STDEV helps determine central tendencies and variability, revealing overall performance trends and consistency. The PERCENTILE function allows you to understand data distribution and identify key percentiles, such as the top 10% of movie performances. Overall, these Excel functions facilitate a comprehensive statistical analysis, aiding in decision-making and strategic planning by providing clear, data-driven insights.

Link for the excel workbook:
https://docs.google.com/spreadsheets/d/1lwpwG8zr02ogtAmJoCH-P8pgUhJVJ5zV/edit?usp=sharing&ouid=100957567890552950706&rtpof=true&sd=true

Project 6

Bank Loan Case Study

Final Project-2

Project description:

In this project, as an data analyst, I am tasked with analyzing loan application data to help a finance company improve its loan approval process. The company faces a significant risk due to some applicants defaulting on their loans, particularly those with limited or no credit history. The main challenge is to identify patterns in customer and loan attributes that can predict whether an applicant is likely to default.

Using Exploratory Data Analysis (EDA), the goal is to uncover insights into the factors influencing loan defaults. This analysis will help the company make informed decisions, such as approving or rejecting loans, reducing loan amounts, or adjusting interest rates for risky applicants. By identifying key attributes that correlate with payment difficulties, the company can minimize financial loss and ensure that capable applicants are not denied loans.

Approach (Concise and Simple):

1. **Data Exploration**: Understand the dataset by reviewing customer and loan attributes.
2. **Data Cleaning**: Handle missing data, correct any errors, and remove outliers that may impact the analysis.
3. **Univariate Analysis**: Analyze individual features to understand their distribution and key statistics.
4. **Bivariate Analysis**: Explore relationships between customer/loan attributes and loan status (approved, defaulted, rejected).
5. **Correlation Analysis**: Identify which variables are most related to loan defaults.
6. **Insights Generation**: Summarize patterns and insights to recommend better loan approval strategies.

This structured approach helps in understanding key factors that influence loan defaults, providing the company with data-driven strategies to minimize financial risks.

Tech-stack used: Microsoft Excel 2019

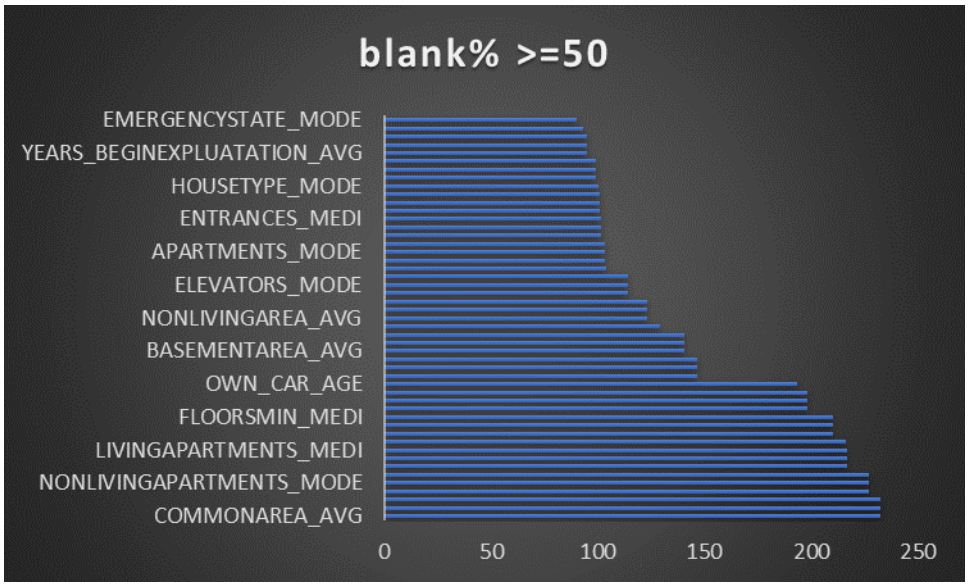
Excel is an excellent tool for handling, analyzing, and visualizing datasets of moderate size. It offers functionalities for data cleaning and statistical analysis. Excel's built-in charts and pivot tables are useful for creating visual representations like bar charts, pie charts, and histograms, which help in understanding trends in the dataset related to rejections, interviews, job types, and vacancies.

Data Analytics Tasks:

A. **Identify Missing Data and Deal with it Appropriately:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Output: PRIMARY FILE

Total columns initially	121	**total column after cleaning**	72
total rows initially	49999	**total rows after cleaning**	49999



Handling of missing data:

1.Categorical data – Using mode

NAME_TYPE_SUITE
OCCUPATION_TYPE

2.Numerical data – Using median for the data with outliers and average for data without the outliers

COLUMN	percentage BLANKS	median	mode
OCCUPATION_TYPE	31.30862617		Unknown
AMT_REQ_CREDIT_BUREAU_HOUR	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_DAY	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_WEEK	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_MON	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_QRT	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_YEAR	13.46826937	1	
NAME_TYPE_SUITE	0.38400768		Unaccompanied
OBS_30_CNT_SOCIAL_CIRCLE	0.33600672	0	
DEF_30_CNT_SOCIAL_CIRCLE	0.33600672	0	
OBS_60_CNT_SOCIAL_CIRCLE	0.33600672	0	
DEF_60_CNT_SOCIAL_CIRCLE	0.33600672	0	
AMT_GOODS_PRICE	0.07600152	450000	
AMT_ANNUITY	0.00200004	24939	
CNT_FAM_MEMBERS	0.00200004	2	
DAYS_LAST_PHONE_CHANGE	0.00200004	0	

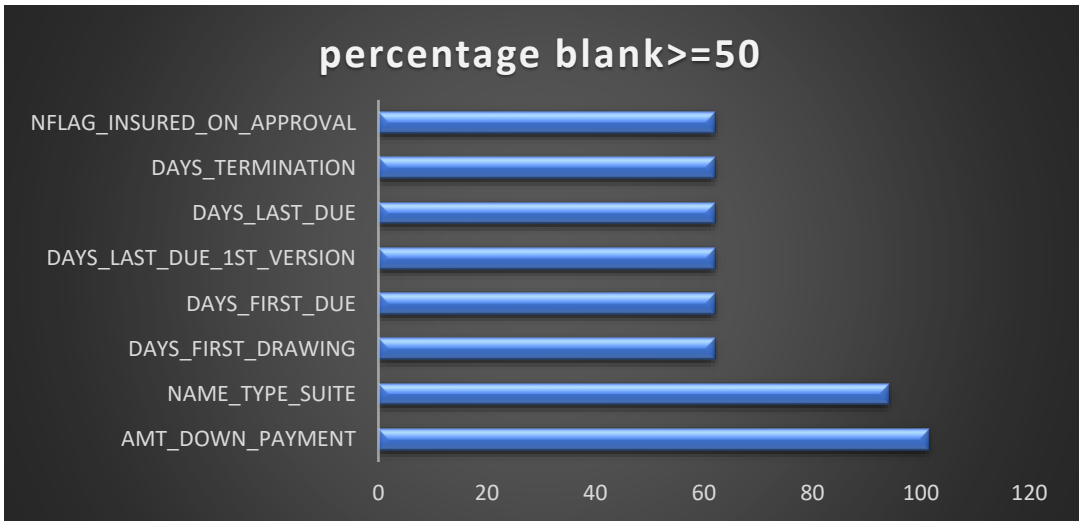
The data above shows there were total 121 column but some column had more than 50% of data missing which can not be imputed so we drop those columns

The missing data was handled by replacing with median where there were outliers present and for the categorical data mode was used to fill the missing values.

After cleaning 72 columns were left.

Output: SECONDARY FILE

The columns with more than 50% blanks



drop columns	
AMT_DOWN_PAYMENT	RATE_INTEREST_PRIMARY
NAME_TYPE_SUITE	RATE_INTERSET_PRIVILAGED
DAYS_FIRST_DRAWING	
DAYS_FIRST_DUE	
DAYS_LAST_DUE_1ST_VERSION	
DAYS_LAST_DUE	
DAYS_TERMINATION	
NFLAG_INSURED_ON_APPROVAL	

The handling of the missing data:

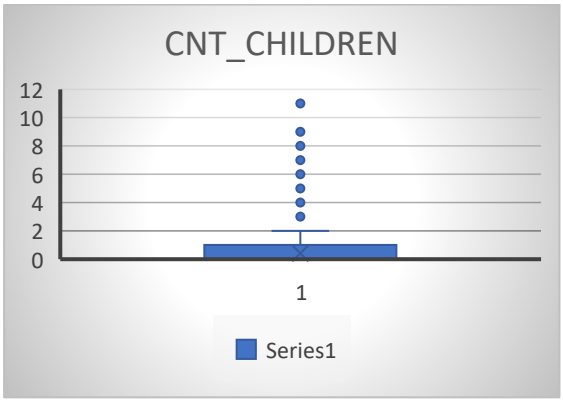
column	percentblank	median	replace
AMT_GOODS_PRICE	27.36976181	104017.5	
AMT_ANNUITY	26.87847337	10879.92	
CNT_PAYMENT	26.87847337		0
PRODUCT_COMBINATION	0.016002881		unknown

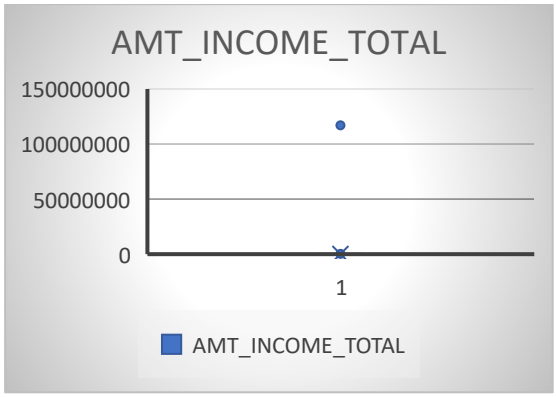
Insights:

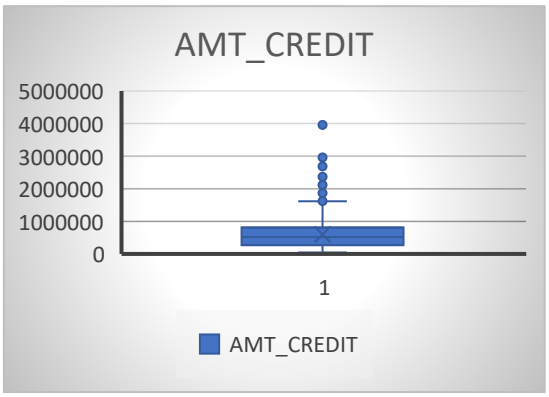
- There were about 36 columns in total initially and after dropping the columns with more than 50% of missing data the total column left was 26.
- The missing data was imputed using median as outliers were present and for categorical data in product_combination we put unknown in blanks as we cannot fil it with mode and in cnt_payment with filled the blanks with 0.

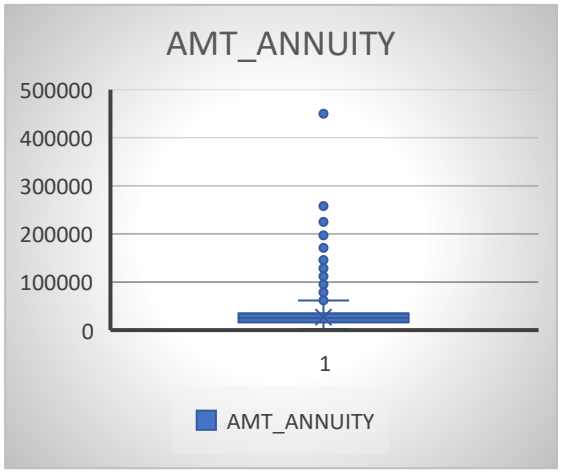
B. Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Output: PRIMARY FILE









Insights:

- Based on the box plot analysis of the variables CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE, the presence of outliers in each column indicates variability in the dataset.
- For CNT_CHILDREN, outliers may suggest atypical family structures that could influence repayment behavior or it might be an error during data collection.
- In AMT_INCOME_TOTAL, outliers point to extreme income levels which highlight a potential risk factor for default if high-income clients are over-leveraged.
- The outliers in AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE suggest that certain clients are taking on unusually large loans or purchasing high-value goods, which may require closer scrutiny to mitigate risk.

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

OUTPUT: PRIMARY FILE

TARGET = 1: Clients who had **payment difficulties**, meaning they were **late by more than a certain number of days** on at least one of the early instalments of their loan.

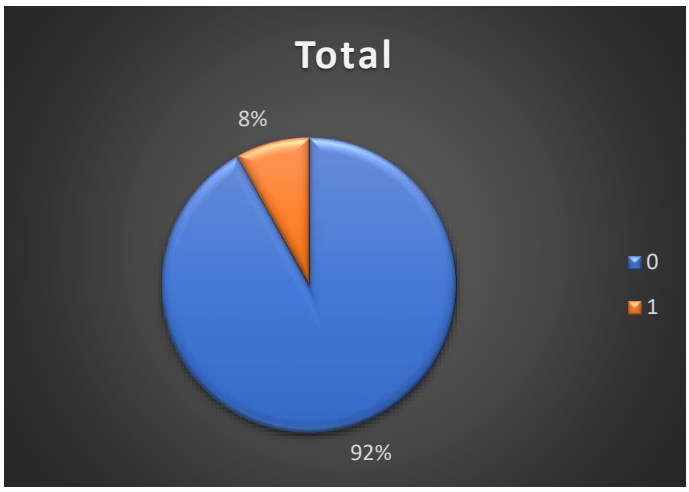
TARGET = 0: Clients who had **no payment difficulties**, meaning they made all their payments on time.

This means

Defaulted clients (TARGET = 1): These clients struggled to meet their loan repayment obligations. They missed or were late in making payments by more than a specified threshold on at least one instalment.

Non-defaulted clients (TARGET = 0): These clients managed their loan repayments well and made their payments on time without any issues.

Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**



Insights:

- Analysis shows that the number of non-defaulters (clients with TARGET = 0) is greater than the number of defaulters (clients with TARGET = 1). The data imbalance, with more non-defaulters than defaulters, reflects a generally good lending portfolio.

- The higher number of non-defaulters suggests that the lending institution's risk assessment and management strategies are effective in identifying and approving clients who are likely to repay their loans.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: 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.

OUTPUT: PRIMARY FILE

Univariate analysis/ segmented univariate analysis

1. For numerical data

univariate analysis of AMT_CREDIT	
AMT_CREDIT	
Mean	599700.5815
Standard Error	1799.674528
Median	514777.5
Mode	450000
Standard Deviation	402415.4339
Sample Variance	1.61938E+11
Kurtosis	1.917459058
Skewness	1.223668739
Range	4005000
Minimum	45000
Maximum	4050000
Sum	29984429376
Count	49999

univariate analysis of AMT_INCOME_TOT	
AMT_INCOME_TOTAL	
Mean	170767.5905
Standard Error	2378.391081
Median	145800
Mode	135000
Standard Deviation	531819.0951
Sample Variance	2.82832E+11
Kurtosis	46582.52582
Skewness	212.0777967
Range	116974350
Minimum	25650
Maximum	117000000
Sum	8538208758
Count	49999

Insights:

1.Income Distribution:

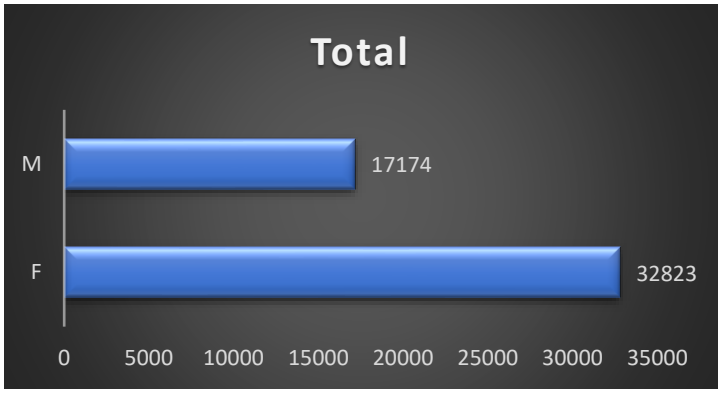
- **Mean and Median Income**: The mean income is higher than the median which indicate a right-skewed distribution where a small number of clients earn much more than the majority. This suggests a concentration of wealth and potential risk if the high earners are over-leveraged.
- **Income Range**: A wide range between the minimum and maximum income values indicates significant variability in client financial backgrounds.

2.Credit Amount Analysis:

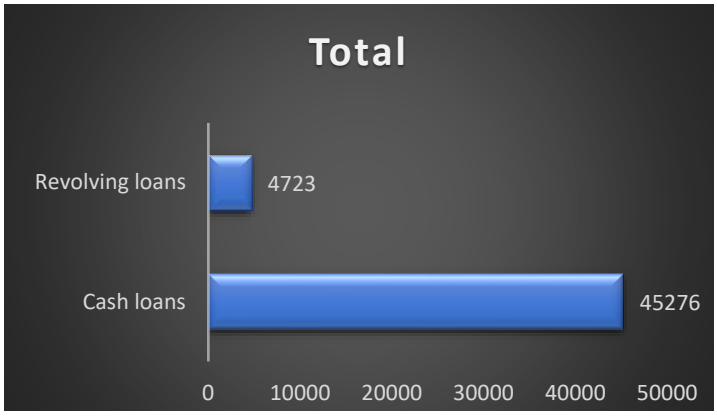
- **Average and Median Loan Amounts**: The average loan amount is considerably higher than the median, this suggests that a few clients are taking out disproportionately large loans. Understanding the reasons behind high loan amounts can help identify potential risks.

2.FOR CATEGORICAL DATA:

Univariate analysis for code_gender	
Row Labels	**Count of CODE_GENDER**
F	32823
M	17174
Grand Total	**49997**



Univariate ananlysis for name_contract	
Row Labels	**Count of NAME_CONTRACT_TYPE**
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**



Insights:

1.Gender Distribution:

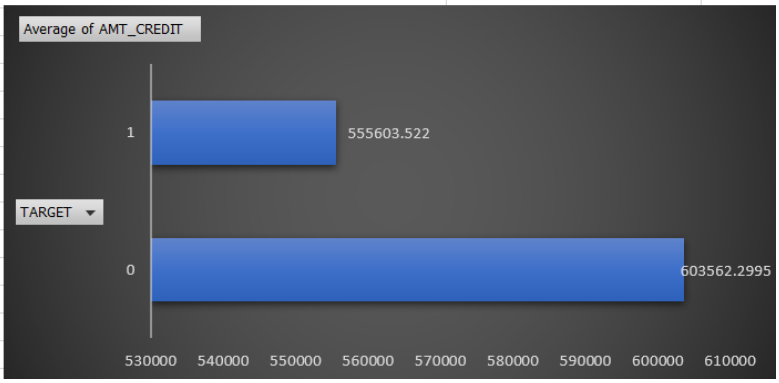
- **Higher Female Clientele**: The predominance of female clients in the dataset suggests that the bank may have successfully targeted or appealed to female borrowers.
- **Implications for Marketing and Product Development**: Understanding the needs of female borrowers could help the institution tailor products or services that meet their specific financial goals. This might include offerings like family loans, education loans or flexible repayment options.

2.Prevalence of Cash Loans:

- **Dominance of Cash Loans**: The higher number of cash loans indicates a strong demand for immediate liquidity among borrowers.
- **Understanding Loan Utilization**: Analyzing the reasons behind taking cash loans can provide valuable insights into client behavior. For example, if cash loans are primarily used for emergencies, lenders might consider offering more tailored financial products that provide better support in urgent situations.

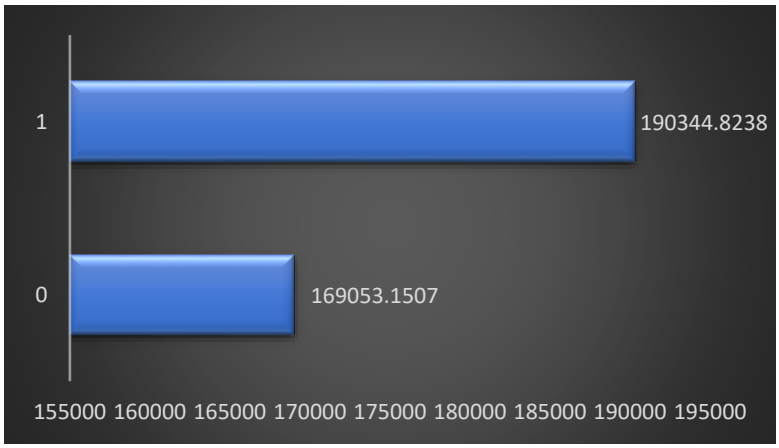
1.

Segmenting loan ammount by loan default status	
Row Labels	**Average of AMT_CREDIT**
0	603562.2995
1	555603.522
Grand Total	**599700.5815**



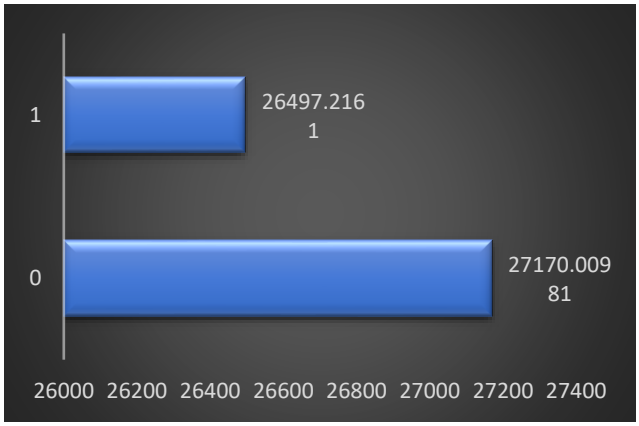
2.

segmenting income by loan default status	
Row Labels	**Average of AMT_INCOME_TOTAL**
0	169053.1507
1	190344.8238
Grand Total	**170767.5905**



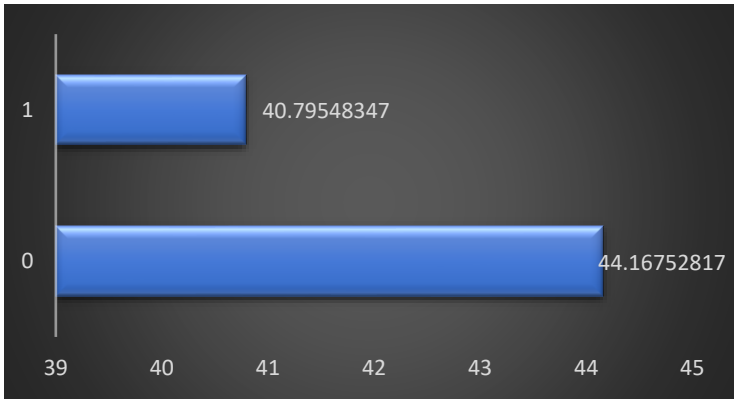
3.

segmenting loan annuity by default status	
Row Labels	**Average of AMT_ANNUITY**
0	27170.00981
1	26497.2161
Grand Total	**27115.83538**



4.

segmenting clients age by default status	
Row Labels	**Average of client's age**
0	44.16752817
1	40.79548347
Grand Total	**43.8960057**



Insights:

1.Non-Defaulters Have Higher Credit, Annuity, and Age:

- **Higher Credit Amounts:** The finding that non-defaulters have greater credit amounts suggests that these clients may be more financially stable and capable of managing higher loans. This could indicate a more favorable risk profile among non-defaulters.
- **Loan Annuity Insights:** A higher average loan annuity among non-defaulters may suggest that they are able to comfortably meet their repayment obligations.
- **Client Age Factor:** The greater age of non-defaulters might imply that older clients tend to have more established financial habits and stability. This could point to the importance of age as a potential risk factor in loan approvals.

2.Defaulters Have Higher Income:

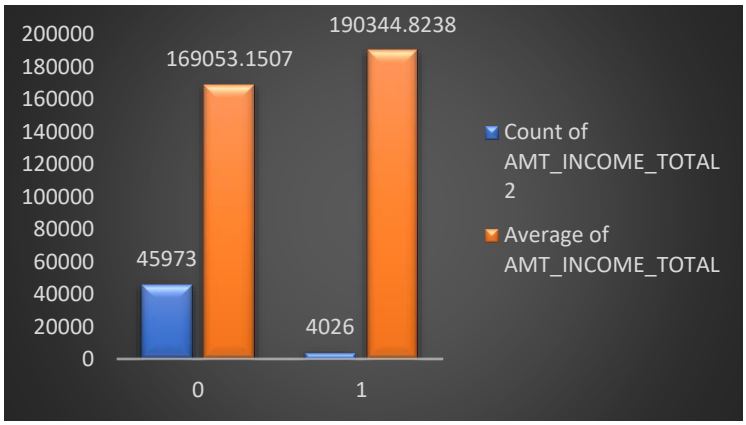
- **Income Disparity in Defaulters**: The clients in the default group have greater income which indicate that higher income alone does not guarantee repayment ability. This might suggest that income is not the only factor influencing loan default risk and that other elements, such as financial management skills or unexpected financial burdens, come into play.

Bivariate analysis:

1.

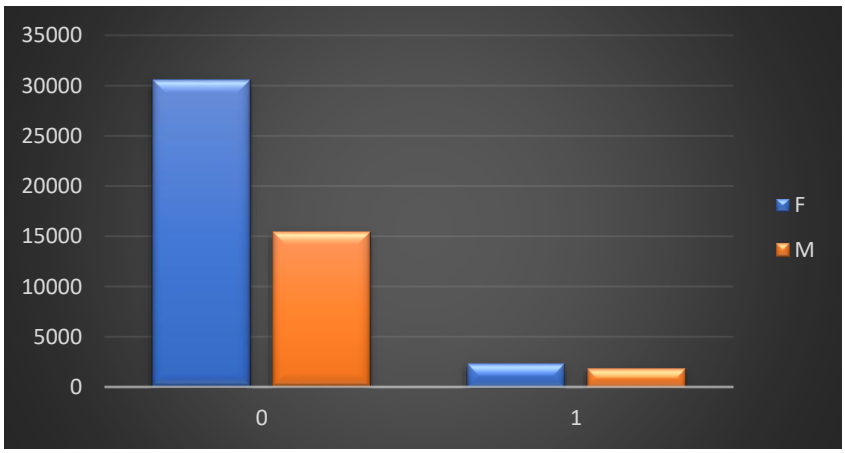
relationship analysis

Row Labels	Count of AMT_INCOME_TOTAL2	Average of AMT_INCOME_TOTAL
0	45973	169053.1507
1	4026	190344.8238
Grand Total	**49999**	**170767.5905**



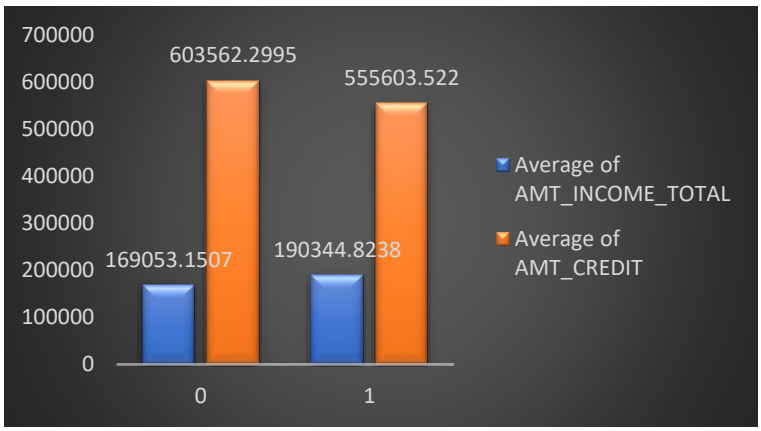
2.

Count of CODE_ Column Labels			
Row Labels	F	M	Grand Total
0	30559	15412	45971
1	2264	1762	4026
Grand Total	**32823**	**17174**	**49997**



3.

Row Labels	Average of AMT_INCOME_TOTAL	Average of AMT_CREDIT
0	169053.1507	603562.2995
1	190344.8238	555603.522
Grand Total	**170767.5905**	**599700.5815**



Insights:

1.Higher Count of Non-Defaulters (TARGET = 0):

The larger count of clients classified as non-defaulters indicates that a significant portion of the client base is effectively managing their loans and obligations. This may suggest that the lending institution's risk assessment processes are working well for the majority of borrowers.

Since there are more non-defaulters, this demographic could be a focus for future marketing and lending strategies, as they represent a stable client base. Understanding their demographics can help in tailoring products that fit their needs.

2. Higher Average Income Among Defaulters (TARGET = 1):

Income Discrepancy: The finding that the average income is higher among defaulters raises important questions about the nature of financial risk. It suggests that having a higher income does not automatically correlate with better repayment behavior.

3. The finding that non-defaulters have a higher average credit amount implies that these clients may be more confident in their repayment capabilities, resulting in them taking larger loans. This could indicate that they have a more stable financial foundation, potentially backed by good financial management practices.

OUTPUT: SECONDARY FILE

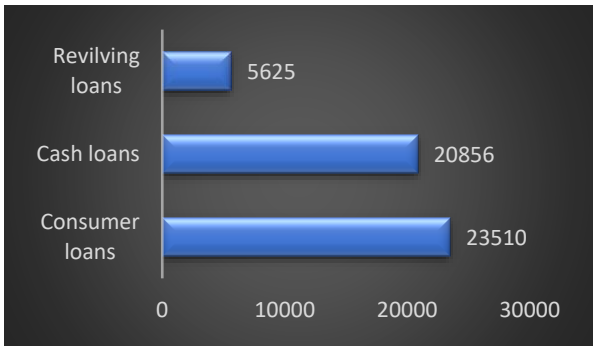
Univariate/segmented univariate analysis:

univariate analysis	
AMT_ANNUITY	
Mean	14507.54628
Standard Error	58.30233413
Median	10879.92
Mode	10879.92
Standard Deviation	13036.66787
Sample Variance	169954709.1
Kurtosis	17.66913934
Skewness	3.164915849
Range	234478.395
Minimum	0
Maximum	234478.395
Sum	725362806.6
Count	49999

AMT_CREDIT	
Mean	188542.8855
Standard Error	1379.549679
Median	78907.5
Mode	0
Standard Deviation	308473.6014
Sample Variance	95155962744
Kurtosis	14.88061385
Skewness	3.344679263
Range	4104351
Minimum	0
Maximum	4104351
Sum	9426955730
Count	49999

NAME_CONTRACT_TYPE

type	total
Consumer loans	23510
Cash loans	20856
Revilving loans	5625



Insights:

1.Annuity Distribution:

- **Average Loan Annuity**: The mean or average annuity amount indicates the typical loan payment clients are making on a periodic basis.
- **Comparison of Mean and Median**: The mean is higher than the median, it suggests that there are a few clients with very high annuities, which are pulling the average up.
- **Loan Annuity Variability**: A high standard deviation or wide range indicates significant variability in loan annuities across the client base, meaning that some clients have very large payments while others have much smaller ones.

2.Credit Amount Analysis:

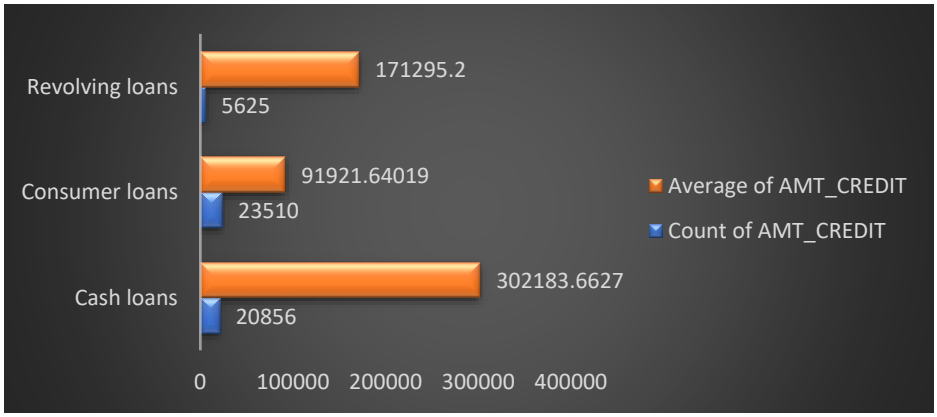
- **Average and Median Loan Amounts**: The average loan amount is considerably higher than the median, this suggests that a few clients are taking out disproportionately large loans. Understanding the reasons behind high loan amounts can help identify potential risks.

3.Contract type analysis:

- **Revolving Loans vs. Cash Loans**: The revolving loans are typically smaller than cash loans is consistent with the nature of revolving credit. Revolving loans, such as credit cards, generally offer smaller more flexible borrowing limits that can be reused once paid off. This reflects a lower average loan size and suggests that clients using revolving loans may be seeking short-term, smaller financing solutions. Cash loans, often used for one-time, larger tend to have higher amounts
- **Prevalence of Consumer Loans**: If consumer loans re the most frequent type, it suggests that the majority of clients are borrowing for everyday expenses or personal consumption. This indicates that the lender's target market is primarily focused on consumers seeking financing for personal or family needs rather than business or investment purposes.

Segmented univariate analysis:

Row Labels	Count of AMT_CREDIT	Average of AMT_CREDIT
Cash loans	20856	302183.6627
Consumer loans	23510	91921.64019
Revolving loans	5625	171295.2
Grand Total	**49991**	**188573.0578**



Insights:

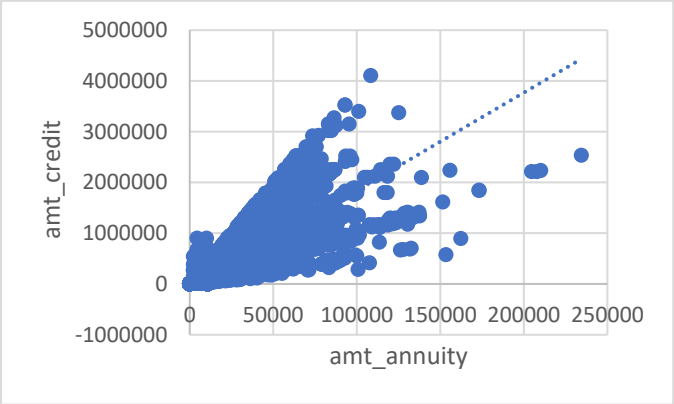
1. Higher Average Credit for Cash Loans:

- Cash loans typically have higher average amounts as they are often used for long-term purchases such as homes, vehicles, or other large expenses.

- Revolving loans, such as credit cards or lines of credit, tend to offer smaller, flexible borrowing limits compared to cash loans. Clients may be using these loans to cover short-term expenses or maintain liquidity, which explains the lower average credit amount.
- The consumer loans have the highest count which indicates strong demand for this type of loan. Consumer loans are often used for a wide range of personal expenses (e.g., home improvement, education, large purchases). Banks can leverage this insight by focusing on expanding and marketing consumer loans, as they clearly cater to a large segment of the borrower population.

Bivariate analysis:

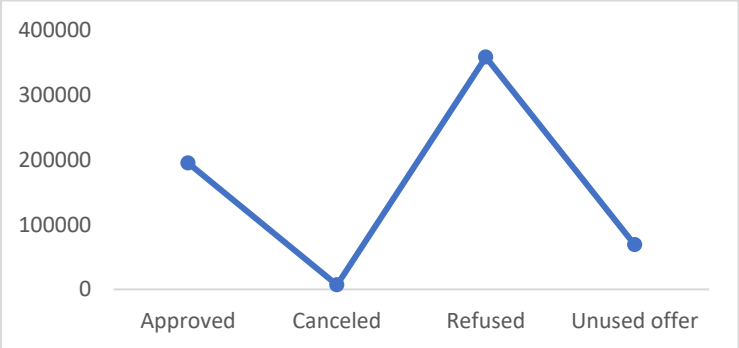
1. AMT_CREDIT vs. AMT_ANNUITY



2.

AMT_CREDIT vs. NAME_CONTRACT_STATUS

Row Labels	Average of AMT_CREDIT
Approved	194728.2553
Canceled	6867.494764
Refused	357962.9779
Unused offer	68754.13737
Grand Total	**188542.8855**



Insights:

1. The linear trendline suggests a positive correlation between the loan amount (AMT_CREDIT) and the loan annuity (AMT_ANNUITY). This means that as the loan amount increases, the annuity payments also tend to increase proportionally. This is logical because larger loans typically require higher regular payments.

There are some outliers they could indicate unusual cases where the annuity is disproportionately high or low relative to the loan amount. These could be due to special loan terms, financial difficulties, or miscalculated risk profiles.

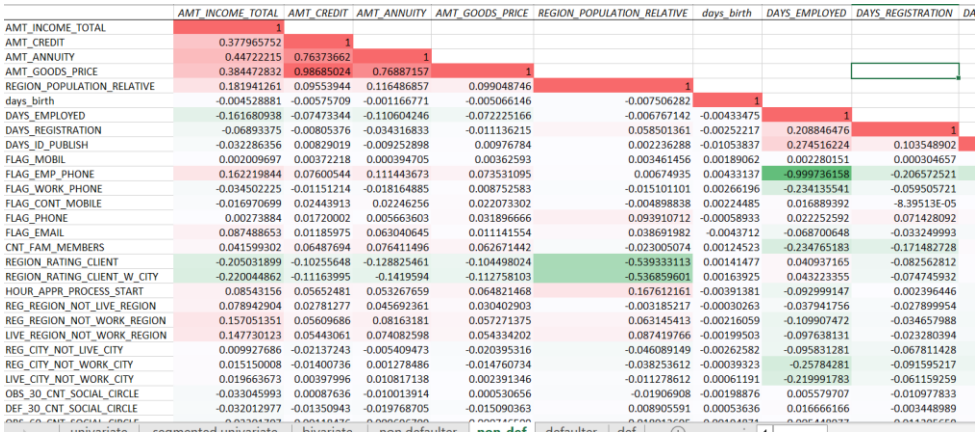
2. Larger loans come with higher risk, both in terms of repayment capability and potential default. Banksmay be refusing these larger loans as part of their risk management strategy.

The cancelled loans have the least average AMT_CREDIT suggests that clients who initially apply for smaller loan amounts may be more likely to cancel their applications. This could indicate that clients seeking smaller loans may have less financial commitment leading them to cancel.

E. **Identify Top Correlations for Different Scenarios:** 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**.**

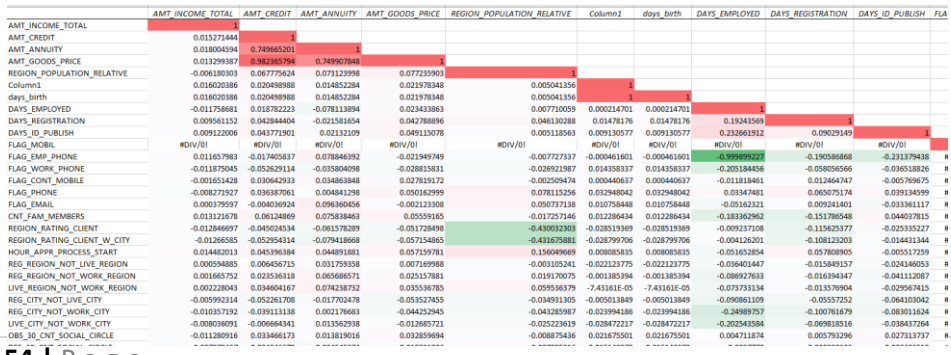
OUTPUT: Primary file

NON-DEFAULTERS:



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	days_birth	DAYS_EMPLOYED	DAYS_REGISTRATION
AMT_INCOME_TOTAL	1							
AMT_CREDIT	0.377965752	1						
AMT_ANNUITY	0.44722215	0.76373662	1					
AMT_GOODS_PRICE	0.384472832	0.98685024	0.76887157	1				
REGION_POPULATION_RELATIVE	0.181941261	0.09553944	0.116486857	0.099048746	1			
days_birth	-0.004528881	-0.00575709	-0.001166771	-0.005066146	-0.007506282	1		
DAYS_EMPLOYED	-0.161680938	-0.07473344	-0.110604246	-0.072225166	-0.006767142	-0.00433475	1	
DAYS_REGISTRATION	-0.06893375	-0.00805376	-0.034316833	-0.011136215	0.058501361	-0.00252217	0.208846476	1
DAYS_ID_PUBLISH	-0.032286356	0.00829019	-0.009252898	0.00976784	0.002236288	-0.01053837	0.274516224	0.103548902
FLAG_MOBIL	0.002009697	0.00372218	0.000394705	0.00362593	0.003461456	0.00189062	0.002280151	0.000304657
FLAG_EMP_PHONE	0.162219844	0.07600544	0.111443673	0.073531095	0.00674935	0.00433137	-0.999736158	-0.206572521
FLAG_WORK_PHONE	-0.034502225	-0.01151214	-0.018164885	0.008752583	-0.015101101	0.00266196	-0.234135541	-0.059505721
FLAG_CONT_MOBILE	-0.016970699	0.02443913	0.02246256	0.022073302	-0.004898838	0.00224485	0.016889392	-8.39513E-05
FLAG_PHONE	0.00273884	0.01720002	0.005663603	0.031896666	0.093910712	-0.00058933	0.022252592	0.071428092
FLAG_EMAIL	0.087488653	0.01185975	0.063040645	0.011141554	0.038691982	-0.0043712	-0.068700648	-0.033249993
CNT_FAM_MEMBERS	0.041599302	0.06487694	0.076411496	0.062671442	-0.023005074	0.00124523	-0.234765183	-0.171482728
REGION_RATING_CLIENT	-0.205031899	-0.10255648	-0.128825461	-0.104498024	-0.539333113	0.00141477	0.040937165	-0.082562812
REGION_RATING_CLIENT_W_CITY	-0.220044862	-0.11163995	-0.1419594	-0.112758103	-0.536859601	0.00163925	0.043223355	-0.074745932
HOUR_APPR_PROCESS_START	0.08543156	0.05652481	0.053267659	0.064821468	0.167612161	-0.00391381	-0.092999147	0.002396446
REG_REGION_NOT_LIVE_REGION	0.078942904	0.02781277	0.045692361	0.030402903	-0.003185217	-0.00030263	-0.037941756	-0.027899954
REG_REGION_NOT_WORK_REGION	0.157051351	0.05609686	0.08163181	0.057271375	0.063145413	-0.00216059	-0.109907472	-0.034657988
LIVE_REGION_NOT_WORK_REGION	0.147730123	0.05443061	0.074082598	0.054334202	0.087419766	-0.00199503	-0.097638131	-0.023280394
REG_CITY_NOT_LIVE_CITY	0.009927686	-0.02137243	-0.005409473	-0.020395316	-0.046089149	-0.00262582	-0.095831281	-0.067811428
REG_CITY_NOT_WORK_CITY	0.015150008	-0.01400736	0.001278486	-0.014760734	-0.038253612	-0.00039323	-0.25784281	-0.091595217
LIVE_CITY_NOT_WORK_CITY	0.019663673	0.00397996	0.010817138	0.002391346	-0.011278612	0.00061191	-0.219991783	-0.061159259
OBS_30_CNT_SOCIAL_CIRCLE	-0.033045993	0.00087636	-0.010013914	0.000530656	-0.01906908	-0.00198876	0.005579707	-0.010977833
DEF_30_CNT_SOCIAL_CIRCLE	-0.032012977	-0.01350943	-0.019768705	-0.015090363	0.008905591	0.00053636	0.016666166	-0.003448989

Defaulter:



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	Column1	days_birth	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
AMT_INCOME_TOTAL	1									
AMT_CREDIT	0.015271444	1								
AMT_ANNUITY	0.018004594	0.749665201	1							
AMT_GOODS_PRICE	0.013299387	0.982365794	0.749907848	1						
REGION_POPULATION_RELATIVE	-0.006180303	0.067775624	0.073123998	0.077235903	1					
Column1	0.016020386	0.020498988	0.014852284	0.021978348	0.005041356	1				
days_birth	0.016020386	0.020498988	0.014852284	0.021978348	0.005041356	1	1			
DAYS_EMPLOYED	-0.011758681	0.018782223	-0.078113894	0.023433863	0.007710059	0.000214701	0.000214701	1		
DAYS_REGISTRATION	0.009561152	0.042844404	-0.021581654	0.042788896	0.046130288	0.01478176	0.01478176	0.19241569	1	
DAYS_ID_PUBLISH	0.009122006	0.043771901	0.02132109	0.049115078	0.005118563	0.009130577	0.009130577	0.232661912	0.09029149	1
FLAG_MOBIL	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_EMP_PHONE	0.011657983	-0.017405837	0.078846392	-0.021949749	-0.007727337	-0.000461601	-0.000461601	-0.999899227	-0.190586868	-0.231379438
FLAG_WORK_PHONE	-0.011875045	-0.052629114	-0.035804098	-0.028815831	-0.026921987	0.014358337	0.014358337	-0.205184456	-0.058056566	-0.036518826
FLAG_CONT_MOBILE	-0.001651428	0.030642933	0.034863848	0.027819172	-0.002509474	0.000440637	0.000440637	-0.011818461	0.012464747	-0.005769675
FLAG_PHONE	-0.008271927	0.036387061	0.004841298	0.050162999	0.078115256	0.032948042	0.032948042	0.03347481	0.065075174	0.039134599
FLAG_EMAIL	0.000179597	-0.004036924	0.096360456	-0.002123308	0.050737138	0.010758448	0.010758448	-0.05162321	0.009241401	-0.033361117
CNT_FAM_MEMBERS	0.013121678	0.06124869	0.075838463	0.05559165	-0.017257146	0.012286434	0.012286434	-0.183362962	-0.151786548	0.044037815
REGION_RATING_CLIENT	-0.012846697	-0.045024534	-0.061578289	-0.051728498	-0.430032303	-0.028519369	-0.028519369	-0.009237108	-0.115625377	-0.025335227
REGION_RATING_CLIENT_W_CITY	-0.01266585	-0.052954314	-0.079418668	-0.057154865	-0.431675881	-0.028799706	-0.028799706	-0.004126201	-0.108123203	-0.014431344
HOUR_APPR_PROCESS_START	0.014482013	0.045396384	0.044891881	0.057159781	0.156049669	0.008085835	0.008085835	-0.051652854	0.057808905	-0.005517259
REG_REGION_NOT_LIVE_REGION	0.000594885	0.006456715	0.031759358	0.007169988	-0.003105241	-0.022123775	-0.022123775	-0.036401447	-0.015849157	-0.024146053
REG_REGION_NOT_WORK_REGION	0.001665752	0.023536318	0.065686571	0.025157881	0.019170075	-0.001385394	-0.001385394	-0.086927633	-0.016394347	-0.041112087
LIVE_REGION_NOT_WORK_REGION	0.002228043	0.034604167	0.074238732	0.035536785	0.059536379	-7.431616-05	-7.431616-05	-0.073733134	-0.013576904	-0.029567415
REG_CITY_NOT_LIVE_CITY	-0.005992314	-0.052261708	-0.017702478	-0.035527455	-0.034931305	-0.005013849	-0.005013849	-0.090861109	-0.05557252	-0.064103042
REG_CITY_NOT_WORK_CITY	-0.010357192	-0.039113138	0.002176683	-0.044252945	-0.043285987	-0.023994186	-0.023994186	-0.24989757	-0.100761679	-0.083011624
LIVE_CITY_NOT_WORK_CITY	-0.008036091	-0.006664341	0.013562938	-0.012685721	-0.025223619	-0.028472217	-0.028472217	-0.202543584	-0.069818516	-0.038437264
OBS_30_CNT_SOCIAL_CIRCLE	-0.011280916	0.033466173	0.013819016	0.032859694	-0.008875436	0.021675501	0.021675501	0.004711874	0.005793296	0.027313737

Insights:

Positive correlation

If certain variables (e.g. AMT_CREDIT, AMT_ANNUITY) show a strong positive correlation with the other variable, it suggests that as these amounts increase, the likelihood of defaulting also increases. The values in red shows the positive correlation.

Negative Correlation

When one variable increases, the other variable tends to decrease. Conversely, when one variable decreases, the other variable tends to increase. The green ones shows the negative correlation between variables.

In a loan default analysis, negative correlations can be significant predictors of risk.

RESULT:

In this project, I conducted a comprehensive analysis of loan applications to understand the factors influencing loan defaults. By performing univariate analysis to explore the distribution of individual variables, such as income, credit amount, and loan types. I conducted bivariate analysis to investigate relationships between variables, particularly focusing on how attributes like average income and credit amount correlate with the likelihood of defaulting. I calculated correlation coefficients to identify strong predictors, discovering that higher income often correlates with a lower likelihood of default.I also assessed data imbalance within the target variable, noting that non-defaulters significantly outnumbered defaulters, which highlighted potential challenges in predictive modeling. Additionally, I examined outliers using box plots, providing insights into how extreme values can influence the analysis.

Overall, this project enhanced my analytical skills and deepened my understanding of credit risk assessment. I learned how to utilize Excel tools for data analysis effectively, interpret statistical relationships, and derive actionable insights to inform lending strategies. This experience has equipped me with valuable knowledge that I can apply in real-world data-driven decision-making scenarios.

Drive link:

This my excel workbook 1 and 2

Project - 7

Analyzing the Impact of Car Features on Price and Profitability

Final Project-3

Project description:

This project focuses on helping a car manufacturer optimize pricing and product development strategies to maximize profitability while meeting evolving consumer demands. With the increasing popularity of electric and hybrid vehicles, along with ongoing demand for traditional gasoline-powered cars, it has become essential to understand which car features and market categories drive consumer preferences and profitability.

The primary business question we aim to address is: *How can the manufacturer balance pricing and product development to ensure profitability while satisfying consumer demand?*

For this analysis, we are using a dataset that includes information on car features, market categories, fuel types, and pricing. Data cleaning and preprocessing involved handling missing values, standardizing categories, and ensuring data consistency across variables. Assumptions made during the project include that the dataset reflects the current market landscape and that factors like car features and market category significantly influence pricing decisions.

By applying data analysis techniques such as regression and market segmentation, this project will identify trends that can guide future pricing strategies and product development, ultimately enhancing the manufacturer's competitiveness and profitability.

Approach:

- Cleaned and preprocessed the dataset using Excel functions to handle missing data and standardize variables.
- Applied descriptive statistics, pivot tables, scatter plots, regression analysis, and interactive dashboards to extract insights on pricing, features, and market categories for optimizing profitability.
- Descriptive statistics and visualizations provide a straightforward way to explore the data and reveal initial insights. Excel's built-in functions and charting tools make it easy to summarize and interpret key patterns.

Tech-Stack Used:

- For this project, **Microsoft Excel 2019** was the primary tool used for data analysis, visualization, and modeling. Excel was chosen for its user-friendly interface, built-in functions, and versatility in handling various types of data analysis tasks.

Tasks: Analysis:

Task 1: Popularity of Car Models Across 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.

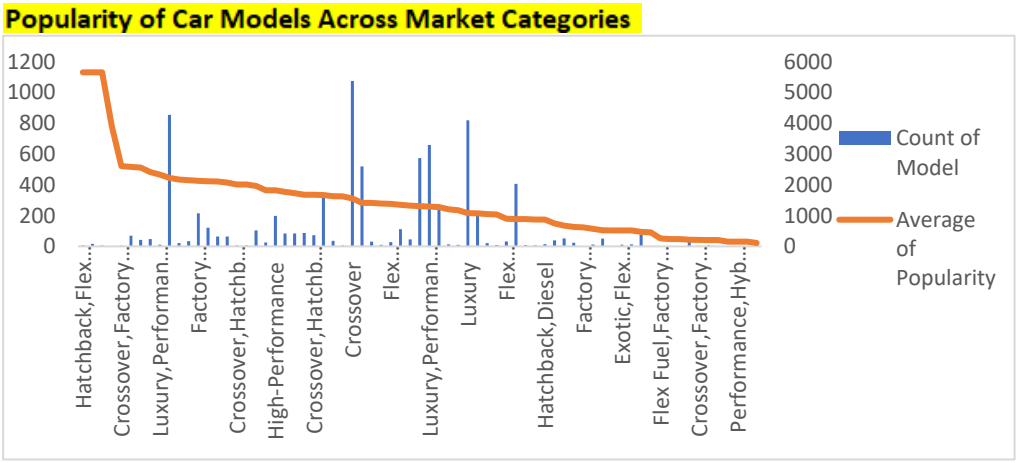
Output:

Factory Tuner,Luxury,Performance	31	1413.419355
Exotic,Performance	10	1391
Flex Fuel,Luxury,Performance	28	1380.071429
Crossover,Luxury,Performance	112	1349.089286
Hatchback,Luxury	45	1323.133333
Hatchback	574	1308.65331
Luxury,Performance	659	1293.062215
Exotic,High-Performance	254	1280.047244
Hatchback,Factory Tuner,High-Performance	13	1205.153846
Crossover,Flex Fuel,Luxury	10	1173.2
Luxury	819	1079.214896
Hatchback,Performance	198	1073.661616
Exotic,Factory Tuner,High-Performance	21	1046.380952
Crossover,Luxury,High-Performance	9	1037.222222
Flex Fuel,Luxury,High-Performance	32	898.3125
Crossover,Luxury	406	889.2142857
Hatchback,Factory Tuner,Luxury,Performance	9	886.8888889
Crossover,Diesel	7	873
Hatchback,Diesel	14	873
Flex Fuel,Luxury	39	746.5384615
Luxury,Hybrid	52	673.6346154
Crossover,Luxury,Hybrid	24	630.9166667
Factory Tuner,Luxury	2	617
Luxury,High-Performance,Hybrid	12	568.8333333
Exotic,Factory Tuner,Luxury,High-Performance	51	523.0196078
Exotic,Factory Tuner,Luxury,Performance	3	520

Row Labels	Count of Model	Average of Popularity
Hatchback,Flex Fuel	7	5657
Flex Fuel,Diesel	16	5657
Crossover,Flex Fuel,Performance	6	5657
Crossover,Luxury,Performance,Hybrid	2	3916
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Performance	69	2585.956522
Crossover,Hybrid	42	2563.380952
Diesel,Luxury	47	2416.106383
Luxury,Performance,Hybrid	11	2333.181818
Flex Fuel	855	2225.71345
Hatchback,Factory Tuner,Performance	21	2173.714286
Crossover,Luxury,Diesel	34	2149.411765
Factory Tuner,Luxury,High-Performance	215	2133.367442
Hybrid	121	2116.586777
Hatchback,Hybrid	64	2111.15625
Crossover,Flex Fuel	64	2073.75
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Performance	6	2009
Factory Tuner,High-Performance	104	1966.442308
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.461538
High-Performance	198	1823.378788
Factory Tuner,Performance	84	1774.047619
Diesel	84	1730.904762
Flex Fuel,Performance	87	1680.471264
Crossover,Hatchback	72	1675.694444
Luxury,High-Performance	334	1668.017964
Hatchback,Luxury,Performance	36	1632.25
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover	1075	1556.168372
Performance	520	1415.209615

The pivot table shows the count of models in each category and their average popularity.

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



Insights:

1. Consumer Preference by Category:

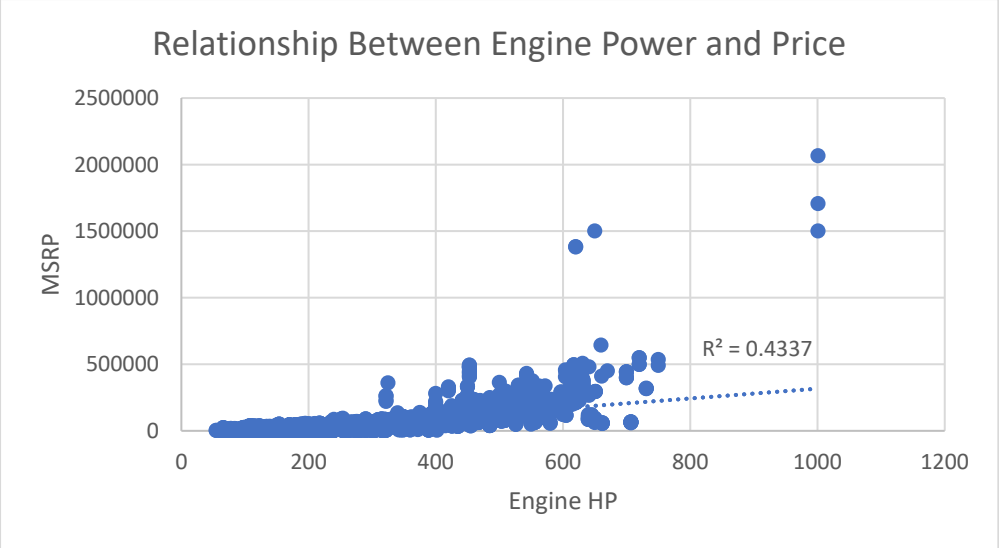
- Higher average popularity in certain market categories like hatchback,flexfuel or flexfuel,diesel indicate that consumers tend to prefer car models within these segments, reflecting strong demand.
- Lower average popularity in other categories flexfuel,hybrid or exotic,luxury suggest lower consumer interest.

2. Market Segments with High Potential: Categories with a consistently high average popularity across models could be identified as high-potential segments for manufacturers to focus on for future product development and marketing.

Task 2: Relationship Between Engine Power and Price

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.

Output:



Insights:

An R^2 value of 0.4337 suggests a moderate positive correlation between engine power and price.

This means that about 43.37% of the variation in car prices can be explained by differences in engine power.

While there is a relationship but not very strong, indicating that other factors also significantly influence the price.

Task 3: Car Features Impacting Price

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.

Output:

For this we take numerical columns and some categorical columns like fuel type, transmission type but they need to be assessed as they are also important car features so we assign them numbers to make them numerical column for further analysis.

Engine Fuel Type	
assign	**as**
electric	0
premium unleaded(required)	1
premium unleaded (recommended)	2
regular unleaded	3
flex-fuel (unleaded/natural gas)	4
flex-fuel (unleaded/E85)	5
flex-fuel (premium unleaded required/E85)	6
flex-fuel (premium unleaded recommended/E85)	7
diesel	8
natural gas	9

transmission type	
assign	**as**
AUTOMATED_MANUAL	0
AUTOMATIC	1
DIRECT_DRIVE	2
MANUAL	3
UNKNOWN	4

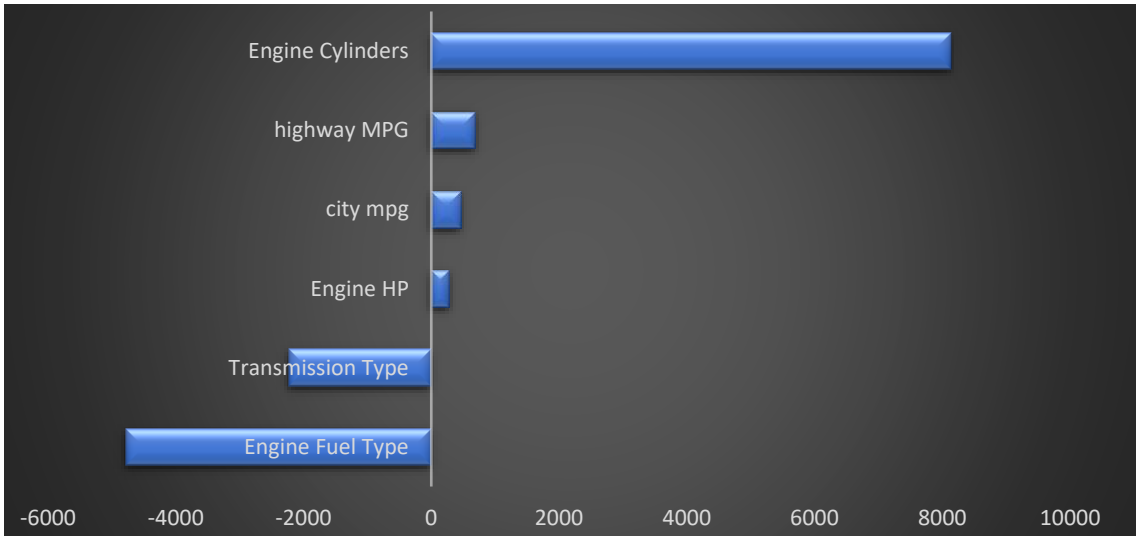
Regression analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.679358454
R Square	0.461527909
Adjusted R Square	0.461239236
Standard Error	45166.92479
Observations	11199

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-85955.67802	3494.163642	-24.5998	3.5048E-130	-92804.85362	-79106.50242	-92804.85362	-79106.50242
Engine Fuel Type	-4780.786894	375.2990523	-12.7386	6.5068E-37	-5516.439078	-4045.134711	-5516.439078	-4045.134711
Engine HP	279.532205	6.814786898	41.01848	0	266.1740235	292.8903865	266.1740235	292.8903865
Engine Cylinders	8139.33074	457.1028186	17.80635	5.79805E-70	7243.32878	9035.3327	7243.32878	9035.3327
Transmission Type	-2230.969668	488.1143522	-4.57059	4.91566E-06	-3187.759691	-1274.179645	-3187.759691	-1274.179645
highway MPG	689.6090577	106.9158254	6.450019	1.16447E-10	480.0352261	899.1828893	480.0352261	899.1828893
city mpg	461.5416294	101.5428298	4.54529	5.54347E-06	262.4998146	660.5834441	262.4998146	660.5834441

Coefficient vs variables



Insights:

The positive coefficient of variables like engine cylinders, highway MPG, City MPG, engine HP shows that these features increase the price while the negative ones tend to decrease the price.

Bar chart also shows the Variables with larger coefficients (both positive and negative) have a greater impact on the price.

Positive coefficients indicate that the variable increases the price, while negative coefficients indicate a decrease.

Task 4: Average Price by Manufacturer

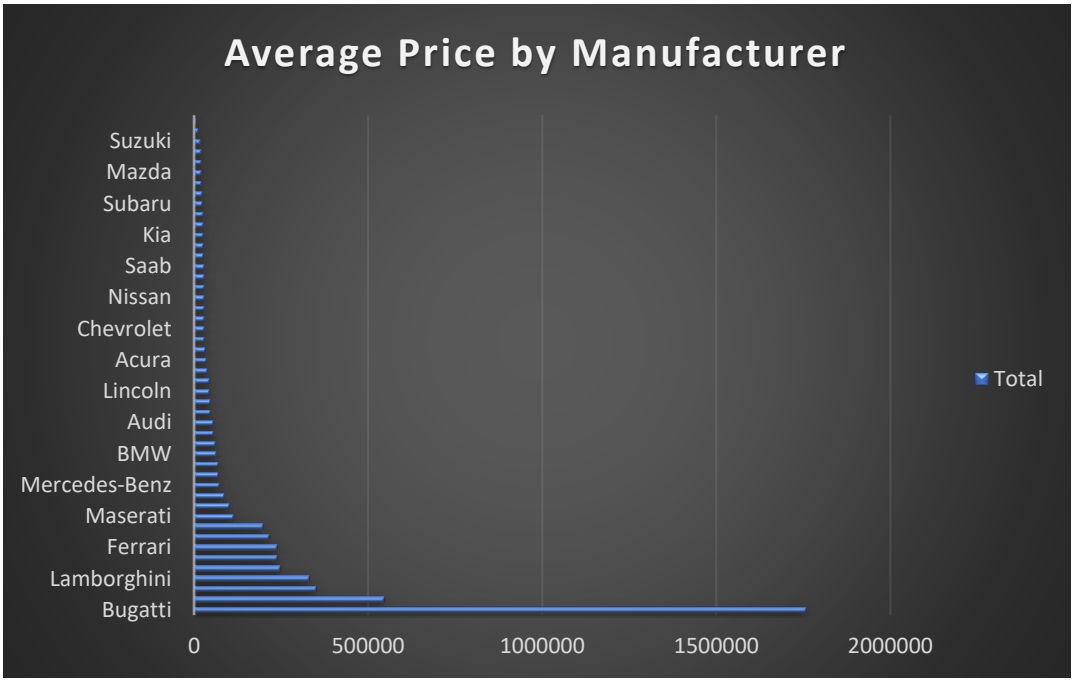
- **Task 4.A:** Create a pivot table that shows the average price of cars for each manufacturer.

Output:

Row Labels	Average of MSRP
Bugatti	1757223.667
Maybach	546221.875
Rolls-Royce	351130.6452
Lamborghini	331567.3077
Bentley	247169.3243
McLaren	239805
Ferrari	238218.8406
Spyker	214990
Aston Martin	198123.4615
Maserati	113684.4909
Porsche	101622.3971
Tesla	85255.55556
Mercedes-Benz	72069.52786
Lotus	68377.14286
Ford	28511.30788
Saab	27879.80734
Chrysler	26722.96257
Honda	26655.14781
Kia	25513.75546
Hyundai	24926.26255
Dodge	24857.04537
Subaru	24240.67364
FIAT	22670.24194
Mitsubishi	21340.5625
Mazda	20416.62379
Scion	19932.5
Pontiac	19800.0442
Suzuki	18026.4152
Oldsmobile	12843.79545
Plymouth	3296.873239
Grand Total	**41925.92714**

The table shows the highest and lowest average prices based on the manufacturer.

- **Task 4.B:** Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



Insights:

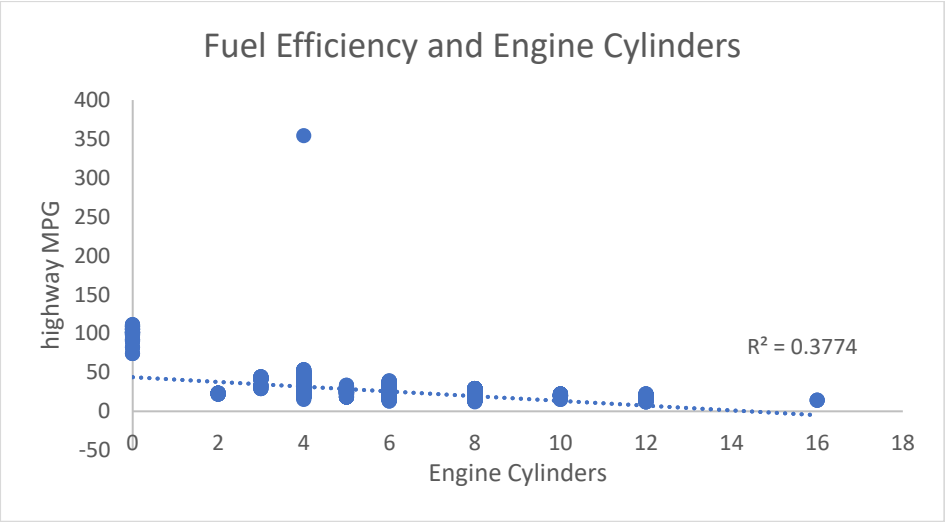
Variation in Car Prices by Manufacturer:

- we can observe how different manufacturers position their cars in the market based on average price.
- Luxury brands like Bugatti, Lamborghini, Rolls-Royce may have significantly higher average prices compared to economy brands like Suzuki, Honda, Ford.

Task 5: Fuel Efficiency and Engine Cylinders

- **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.

Output:



- **Task 5.B:** Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

Output:

correlation coefficient	-0.614333

Insights:

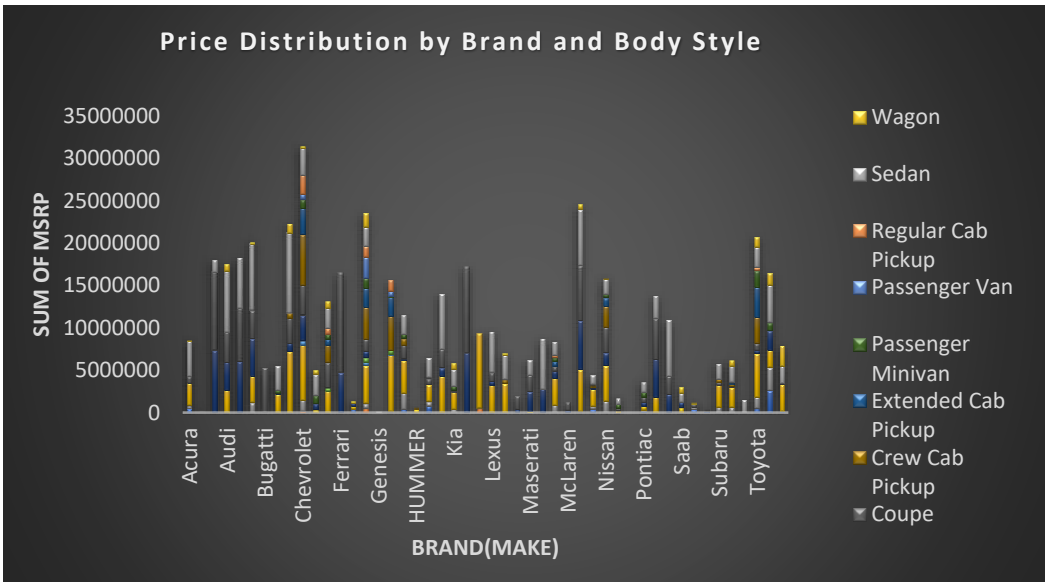
A negative correlation suggests that as the number of cylinders increases, the highway MPG decreases. This is a common finding, as vehicles with more cylinders tend to consume more fuel and therefore have lower fuel efficiency.

Building the Dashboard:

Task 1: Price Distribution 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.

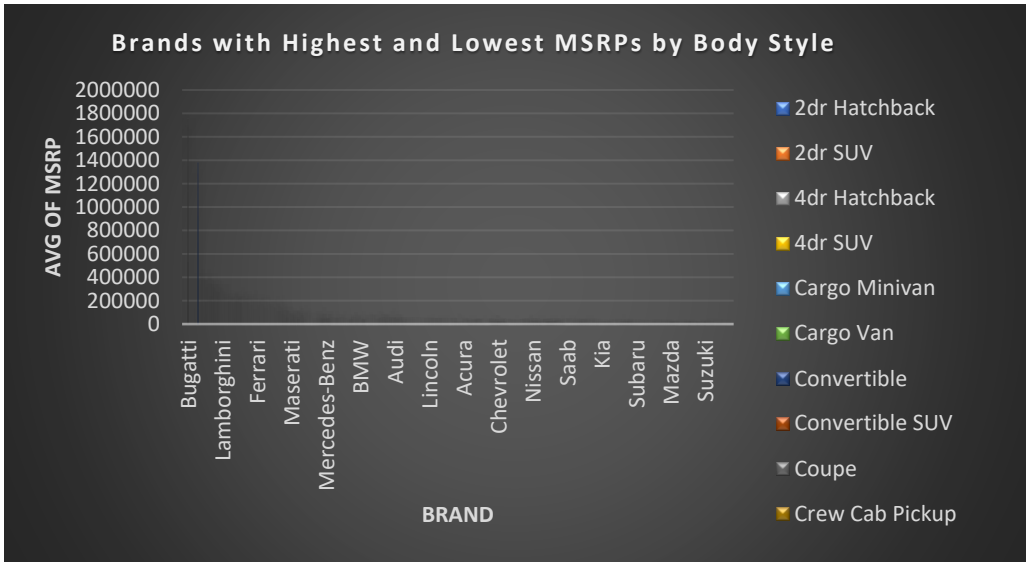
Output:



Task 2: Brands with Highest and Lowest MSRPs 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.

Output:



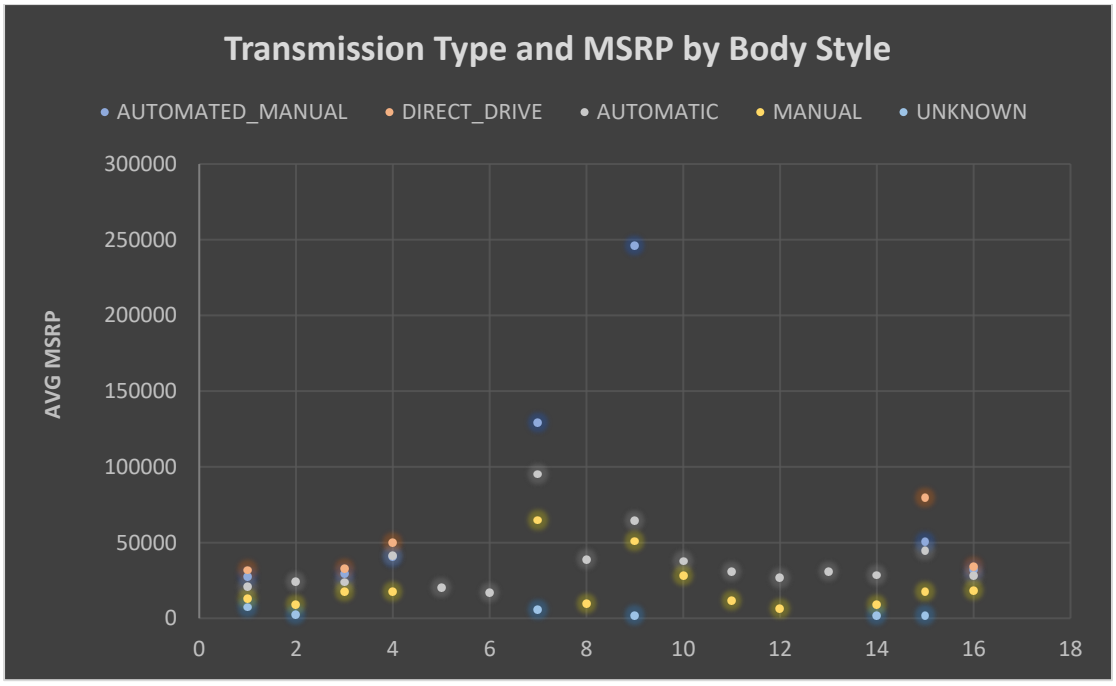
Insights:

- The brands like Bugatti, Maybach have the highest average MSRP and Bugatti having only coupe body style cars at highest average. The brands like Plymouth, Oldsmobile have the lowest average MSRP with having body styles like hatchback, sedan, wagon, and coupe.
- This analysis helps understand how different brands position themselves in the market, whether they focus on budget-friendly models or luxury vehicles.
- We can observe certain body styles coupe tend to have higher average prices compared to others like hatchbacks.

Task 3: Transmission Type and MSRP 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.

Output:



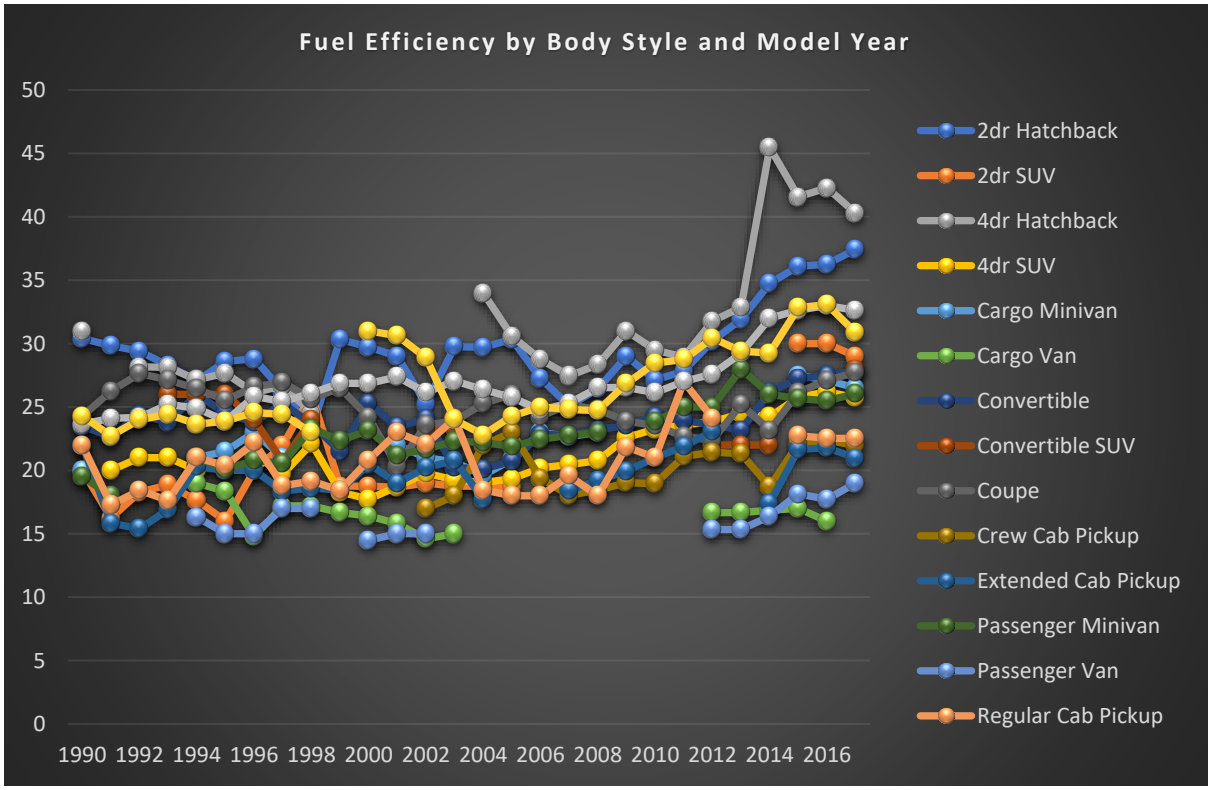
Insights:

You might find that certain body styles like convertible and coupe have higher average MSRPs with automatic transmissions compared to manual ones, indicating consumer preference.

Task 4: Fuel Efficiency by Body Style and Model Year

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.

Output:



Insights:

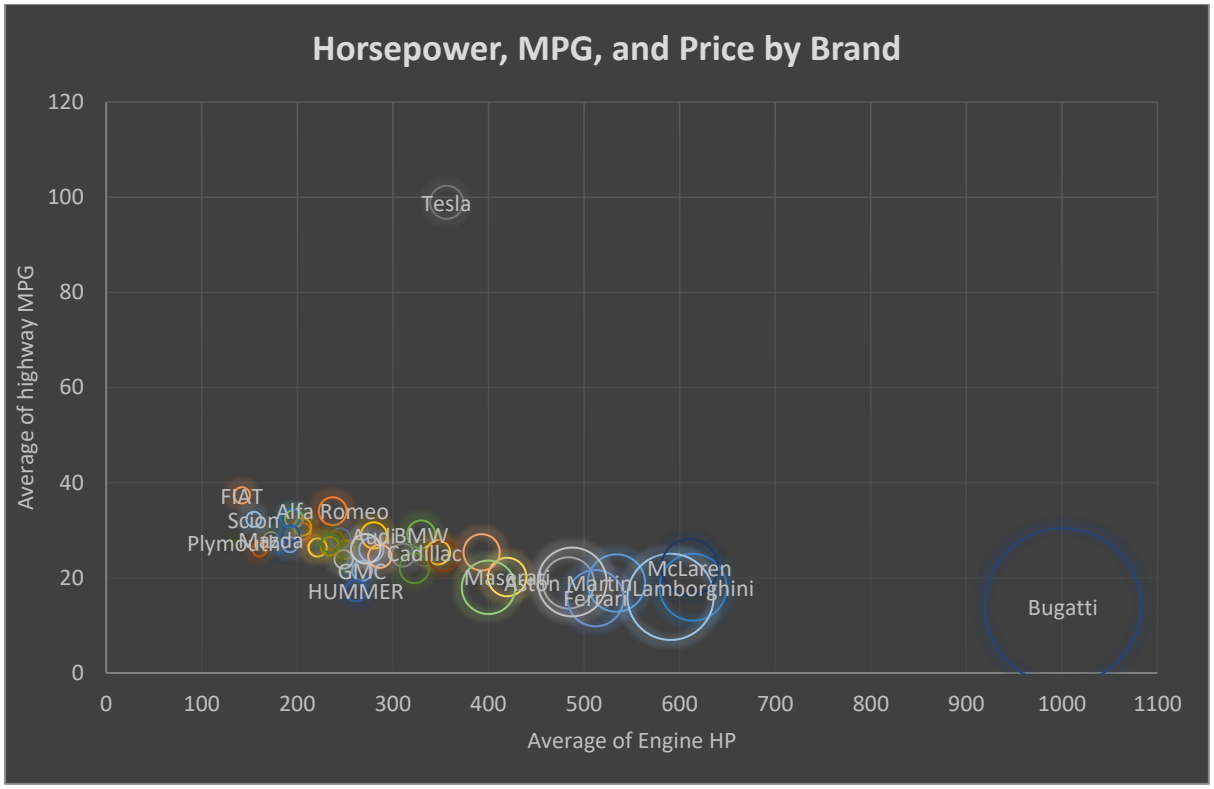
The fuel efficiency of 4dr Hatchback has increased over the years but cargo van, coupe, passenger van and few more fuel efficiency has been stagnant over the years.

Body styles like hatchback, sedan generally have better fuel efficiency than larger models like SUV, cargo van.

Task 5: Horsepower, MPG, and Price by Brand

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.

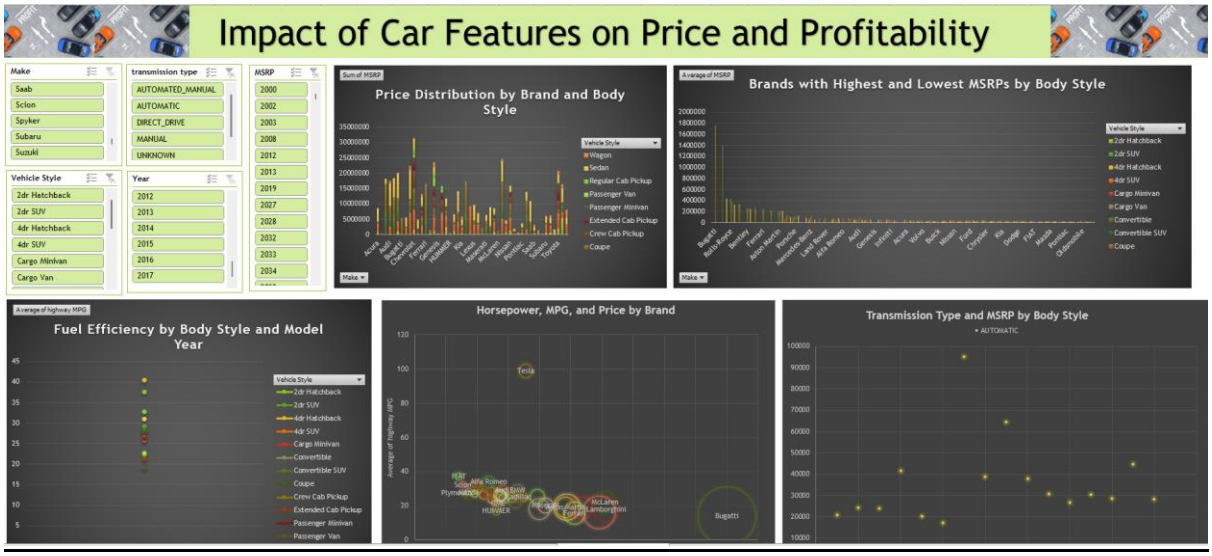
Output:



Insights:

- Cars with higher horsepower like Bugatti have higher prices.
- The chart can reveal higher horsepower comes at the expense of fuel efficiency lower MPG for certain brands.
- Brands cluster in certain areas of the chart. For example, Fiat and Honda have smaller bubbles (lower price), higher MPG, and moderate horsepower, while some brands like Ferrari, Bentley have higher horsepower but lower MPG and larger bubbles (higher price).

FINAL DASHBOARD



Conclusion:

- Higher average popularity in certain market categories like hatchback,flexfuel or flexfuel,diesel indicate that consumers tend to prefer car models within these segments, reflecting strong demand.
- Lower average popularity in other categories flexfuel,hybrid or exotic,luxury suggest lower consumer interest.
- Categories with a consistently high average popularity across models could be identified as high-potential segments for manufacturers to focus on for future product development and marketing.
- An R^2 value of 0.4337 suggests a moderate positive correlation between engine power and price. This means that about 43.37% of the variation in car prices can be explained by differences in engine power.
- While there is a relationship but not very strong, indicating that other factors also significantly influence the price.
- The positive coefficient of variables like engine cylinders, highway MPG, City MPG, engine HP shows that these features increase the price while the negative ones tend to decrease the price.
- Bar chart also shows the Variables with larger coefficients (both positive and negative) have a greater impact on the price.Positive coefficients indicate that the variable increases the price, while negative coefficients indicate a decrease.
- we can observe how different manufacturers position their cars in the market based on average price.Luxury brands like Bugatti, Lamborghini, Rolls-Royce may have significantly higher average prices compared to economy brands like Suzuki, Honda, Ford.
- A negative correlation suggests that as the number of cylinders increases, the highway MPG decreases. This is a common finding, as vehicles with more cylinders tend to consume more fuel and therefore have lower fuel efficiency.

- You might find that certain body styles like convertible and coupe have higher average MSRPs with automatic transmissions compared to manual ones, indicating consumer preference.

Result:

From working on this project, I gained valuable insights into the relationship between a car's features and its market performance, as well as how these factors influence pricing decisions.

Through tasks such as analyzing the popularity of car models across market categories and examining the relationship between engine power and price, I developed a deeper understanding of how consumer preferences and technical specifications play a role in shaping the automotive market.

This project improved my skills in data analysis using Excel, particularly in areas like pivot tables, regression analysis, and visualizations through scatter plots and trendlines. Additionally, working with real-world automotive data helped me appreciate the importance of normalizing comparisons and understanding how various factors contribute to overall trends. This experience strengthened my ability to extract actionable insights and make data-driven recommendations, which will be essential for future projects in the field of data analysis.

The excel sheet: linked here

PROJECT – 8

ABC Call Volume Trend Analysis

Final Project-4

Project description:

This project focuses on analyzing the inbound calling team's performance as part of the company's Customer Experience (CX) strategy. I will be working with a 23-day dataset that includes details such as agent IDs, customer queue time, call durations, and call statuses.

The goal is to understand trends in call volume, evaluate how efficiently the team is handling inbound support, and uncover any issues affecting the customer experience. By diving into the data, I will look for patterns like peak call times, long wait periods, or high abandonment rates, and provide insights to optimize the customer journey.

The ultimate objective is to help the company enhance customer satisfaction by making data-driven decisions to improve the support process, streamline operations, and ensure a more seamless experience for customers when they reach out. Through this analysis, we aim to support the company in turning first impressions into long-term loyalty.

Approach:

- Data cleaning is performed to identify and address inconsistencies, missing values, or outliers to enhance overall data quality.
- Exploratory data analysis (EDA) is conducted to uncover patterns, trends, and relationships within the data, providing insights into overall behaviour and performance.
- Statistical methods are applied to interpret the data, test hypotheses, and validate findings.
- Visualizations like creating graphs, charts, and dashboards to effectively communicate the findings and make insights easily digestible.

Finally, actionable insights are generated from the analysis to inform decision-making and strategy development based on observed data trends.

Tech-stack used:

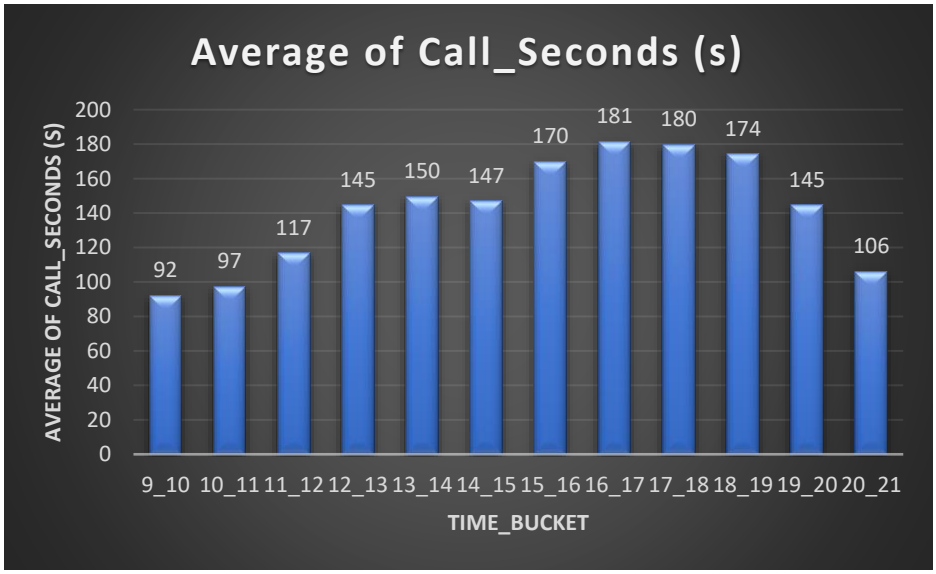
For this project, Microsoft Excel 2019 was the primary tool used for data analysis and visualization. Excel was chosen for its user-friendly interface, built-in functions, and versatility in handling various types of data analysis tasks.

Data Analytics Tasks:

1. Average Call Duration: What is the average duration of calls for each time bucket?

Output:

Time_bucket	Average of Call_Seconds (s)
9_10	92
10_11	97
11_12	117
12_13	145
13_14	150
14_15	147
15_16	170
16_17	181
17_18	180
18_19	174
19_20	145
20_21	106
Grand Total	140



Insights:

The table shows the average call duration for each time bucket.

The average call duration is longer from 3 to 7pm in evening and shorter in mornings.

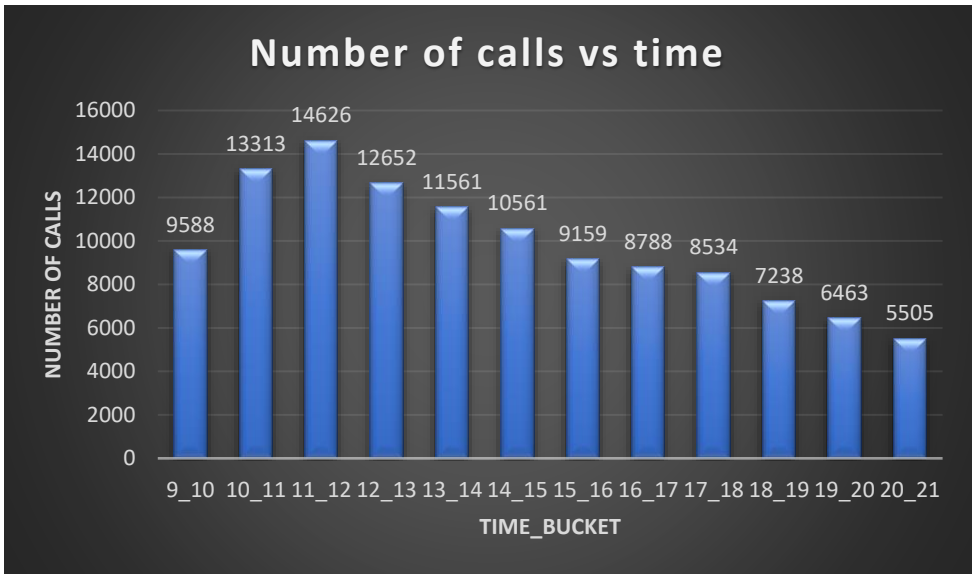
The average duration of all incoming calls received by agents is 140 seconds.

This information can be used to understand which time periods have longer or shorter average call durations.

2. Call Volume Analysis: Can you create a chart or graph that shows the number of calls received in each time bucket?

Output:

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



Insights:

Time buckets from 10-3pm experience the highest call volumes. This helps in understanding when customers are most likely to reach out, allowing for better staffing and resource allocation.

After 3pm call volumes are lower. So, company can reduce staffing during these off-peak hours, potentially leading to cost savings.

By understanding peak and off-peak times, we can make data-driven decisions about staffing levels. For example, 10- 3pm there is significantly higher call volume, so can schedule more agents during that time.

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

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%.

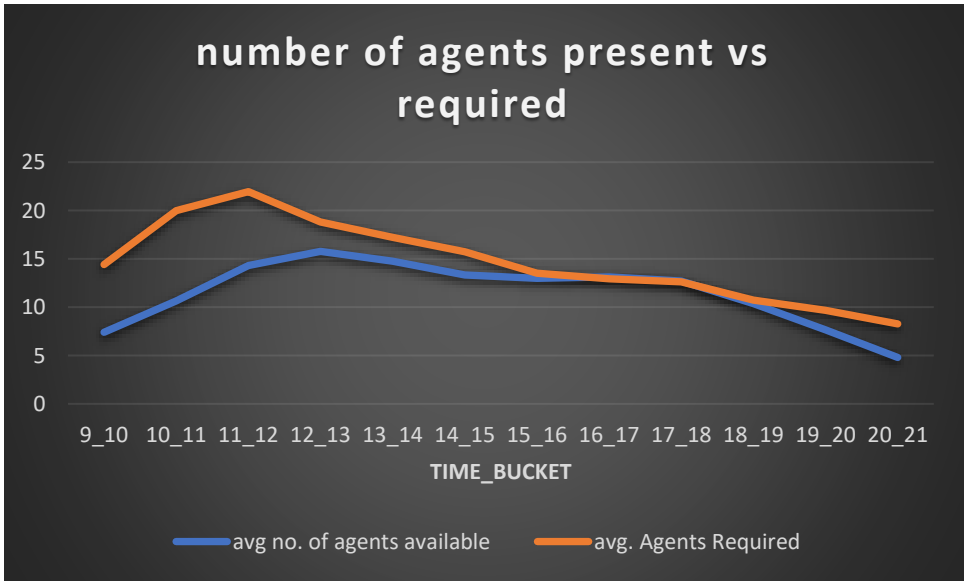
Parameters:

- **Current Abandon Rate**: 30% (70% of calls are answered)
- **Target Abandon Rate**: 10% (90% of calls should be answered)

Output:

Working Hours	9
Break	1.5
Actual Working Hours	7.5
calls taken	60%
Total work seconds	16200
Average Call Time/Agent	140
Calls by Agent/day	116
Calls by an Agent/Hour	26

Time_bucket	avg no. of agents availabl	avg. Agents Require
9_10	7	14
10_11	11	20
11_12	14	22
12_13	16	19
13_14	15	17
14_15	13	16
15_16	13	14
16_17	13	13
17_18	13	13
18_19	10	11
19_20	8	10
20_21	5	8



Insights:

To reduce the abandon rate to 10% we need to increase the number of agents to maintain this.

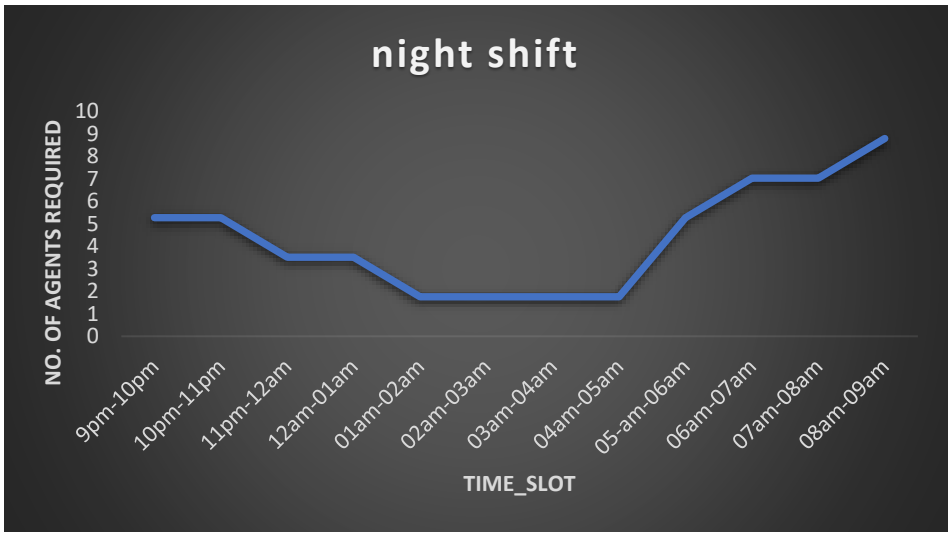
The manpower needs to be increased for morning and afternoon as significant number of calls are being abandoned during these times.

4.Night Shift Manpower Planning: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

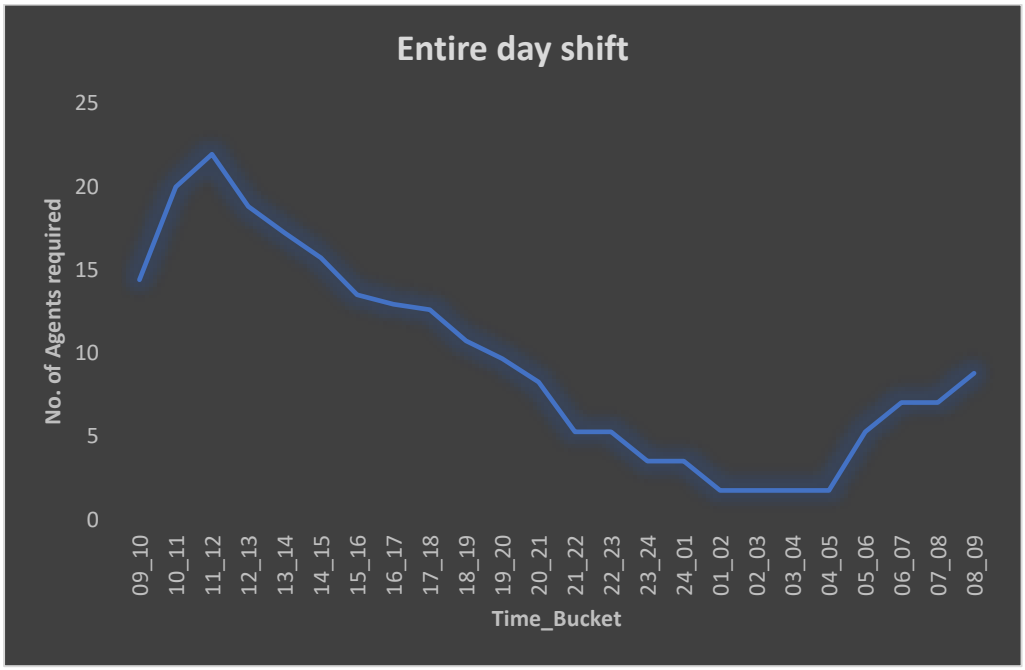
Output:

Manpower required for night shift

Time_Slot	no. of agents required
9pm-10pm	5
10pm-11pm	5
11pm-12am	4
12am-01am	4
01am-02am	2
02am-03am	2
03am-04am	2
04am-05am	2
05-am-06am	5
06am-07am	7
07am-08am	7
08am-09am	9



Manpower for the entire day and night shift



INSIGHTS:

- The average call duration is longer from 3 to 7pm in evening and shorter in mornings. The average duration of all incoming calls received by agents is 140 seconds. This information can be used to understand which time periods have longer or shorter average call durations.
- Time buckets from 10-3pm experience the highest call volumes. This helps in understanding when customers are most likely to reach out, allowing for better staffing and resource allocation. After 3pm call volumes are lower. So, company can reduce staffing during these off-peak hours, potentially leading to cost savings. By understanding peak and off-peak times, we can make data-driven decisions about staffing levels. For example, 10- 3pm there is significantly higher call volume, so can schedule more agents during that time.
- To reduce the abandon rate to 10% we need to increase the number of agents to maintain this.
- The manpower needs to be increased for morning and afternoon as significant number of calls are being abandoned during these times.

Conclusion:

This project focused on customer experience analytics provided valuable insights into analyzing call data for better service delivery. It highlighted how understanding customer behaviour can lead to improved operations. Identifying call volume trends was crucial for resource, ensuring adequate staffing during peak times.

Calculating metrics planning like average call duration and abandonment rates shed light on areas needing improvement. Using data visualization tools made it easier to interpret complex information and derive actionable insights. Collaboration with different teams emphasized the role of data analytics in addressing operational challenges. Overall, this experience underscored how data-driven approaches can enhance customer service and support business success.

Drive link : here

Key Learnings

Through my diverse project experiences, I have gained a comprehensive skill set in data analytics, spanning from SQL database management to data visualization and analysis in Excel. Each project has enriched my abilities in unique ways, as detailed below:

SQL-Based Projects

- Instagram User Analytics
  This project taught me how to manage large datasets using MySQL Workbench to track user engagement and generate insights for strategy. I learned to optimize queries for efficiency, manage relational data, and draw actionable conclusions that supported growth objectives.
- Operation Analytics and Investigating Metric Spike
  Through this analysis, I refined my skills in troubleshooting and anomaly detection within a data-driven operation. SQL was essential for identifying data spikes, conducting time-based analysis, and isolating variables impacting performance. I learned how SQL can streamline complex operations and enhance metric tracking for quick response.

Excel-Based Projects

- Hiring Process Analytics
  This project improved my Excel data-cleaning skills, allowing me to efficiently prepare and analyze data. Using formulas and pivot tables, I provided a clear breakdown of hiring metrics and trends, helping stakeholders understand hiring patterns and make data-driven decisions.
- IMDB Movie Analysis
  In this movie analysis project, I used Excel to explore audience preferences and trends. This involved calculating descriptive statistics and building visualizations to uncover patterns.
- Bank Loan Case Study
  This case study enhanced my understanding of customer analytics by examining loan default risk. Through exploratory data analysis (EDA) in Excel, I identified patterns and variables influencing loan outcomes. This project improved my ability to handle imbalanced datasets and draw insights that support risk assessment.

- Analyzing the Impact of Car Features on Price and Profitability
  Here, I explored profitability factors by analyzing car features. Excel's regression analysis tools helped me determine which specifications impacted price the most, while pivot tables allowed for clear visualizations. This project reinforced my ability to communicate complex analyses through interactive dashboards.
- ABC Call Volume Trend Analysis
  I applied time-series analysis in Excel to study customer support trends, helping optimize manpower planning. I learned to structure and visualize call volume data, calculate averages and trends, and provide recommendations for improving customer experience.

General Learnings

Across all projects, I have strengthened my data analytics skills, including:

- Data Cleaning and Preparation: Ensuring data quality by handling missing values, removing duplicates, and maintaining consistency.
- Descriptive and Predictive Analysis: Using statistical techniques and trend analysis to uncover insights and inform future strategies.
- Visualization and Communication: Building intuitive, impactful visualizations that make complex insights accessible to stakeholders, using both SQL and Excel.

These experiences have given me a well-rounded skill set, allowing me to tackle real-world problems with structured analysis, whether using SQL or Excel.