DATA ANALYTICS PORTFOLIO

April 2025

Prepared by
Alokk Joshi

Professional Background

Summary

A dynamic professional with 24+ years of leadership experience in pharmaceutical sales, training management, and employability skill development, now strategically pivoting into data analytics to amplify organizational decision-making. Expertise lies in transforming raw data into actionable insights, optimizing training ROI, and driving performance improvements across teams. Adept at merging data visualization, process optimization, and strategic training frameworks to solve complex business challenges.

Why Data Analytics?

My career has been rooted in solving human-centric problems with data:

- As a trainer, I used performance metrics to refine programs.
- As a leader, I relied on sales data to allocate resources.
- Now, I aim to deepen this synergy by mastering analytics to predict trends, automate reporting, and drive scalable solutions for institutions.

Technical Proficiencies

Tools & Applications

MS Excel - Dashboards, Pivot Tables, Data Cleaning, Analysis

Power BI, Tableau - Visualization

Python - Data Analytics (Cleaning, Manipulation, Analysis)

SQL - Data Analysis (MySql)

LMS Platforms - Course Design, Learner Progress Tracking

Canva - Infographics, Training Material Design

Table of Contents (Projects)

Github link to all the project files

Project 1

Project on Data Analytics Fundamentals

Real-Life Scenario: Buying a Gaming Laptop

Step 1: Plan

The first step was to Plan what I really needed. Since I have already used a laptop which was not a gaming one, I suffered a lot in my graphics designing projects. This time, I thought about the key requirements for a gaming laptop (of course as a graphics designer, it should serve all purposes):

- Performance for graphics designing: I needed good RAM and a strong graphics card.
- Speed and storage: A fast processor, SSD storage, and a smooth display.
- Budget: It was important to know how much money I could spend.

In my case, I decided I wanted a laptop that was not only powerful for gaming and video editing but also suitable for heavy graphics work.

Step 2: Prepare

Next, I moved to the Prepare stage. This involved:

- Budgeting: I looked at my savings and decided how much I was willing to pay upfront (the down payment) and what amount I could manage in installments.
- Gathering funds: I checked my finances and planned how to manage the payments over time. I also looked into how much time it will be required to gather the down payment amount and how long should I go for EMIs.

Preparing in this way made sure I knew my spending limit and could avoid overspending and financial burden.

Project 1

Project on Data Analytics Fundamentals

Real-Life Scenario: Buying a Gaming Laptop

Step 3: Process

In the Process stage, I started collecting information. Here is what I did:

- Researching various brands: I collected data on several brands like Dell, HP, Acer, and Lenovo. I searched on Amazon, Flipkart and local electronics stores.
- Identifying specific models: I looked at the models that offered the features I needed. For example, for gaming and graphics work, I needed high-quality graphics and fast processing.
- Comparing specifications: I noted down the key specifications of each model, such as RAM, storage, graphics capability, and display refresh rate. Also the strength of the body was a significant criterion for final selection.

This process helped me to narrow down my options by focusing only on the laptops that met my specific criteria.

Step 4: Analyze

After gathering all the data, I analyzed it to make the best choice:

- Feature comparison: I compared the laptops based on their performance for gaming and graphics work. The ASUS laptop stood out because it met mostly all the essential requirements.
- Budget fit: I checked if the prices were within my budget and how the installment plans would work. I got bigger discount on ASUS than on HP and Lenovo.
- Trends and reviews: I made sure that the chosen laptop was popular and had good reviews in the market. Going through Amazon rating helped me a lot to boil down on this choice.

Analyzing the data allowed me to see that even though there were many options, the ASUS laptop was the best fit for my needs.

Project 1

Project on Data Analytics Fundamentals

Real-Life Scenario: Buying a Gaming Laptop

Step 5: Share

The next step was to Share my findings:

- Discussing with the shopkeeper: I communicated with the store representative about my requirements and the research I had done.
- Seeking advice: I asked for recommendations and confirmed if there were any offers or additional details I might have missed. I asked for additional offers and got further Rs. 1000 discount on the model selected. It taught me the customers who ask for and persist for more offers, usually end up getting benefit of their efforts.

Sharing my analysis helped me get further confirmation that I was making the right decision.

Step 6: Act

Finally, I moved to the Act phase:

- Making the purchase: Based on my analysis and discussions, I went ahead and bought the ASUS gaming laptop.
- Payment method: I made the down payment from my savings and arranged the balance payment in installments.

By following these steps, I ensured that my purchase was well thought out and based on data-driven decision-making.

Project 1

Project on Data Analytics Fundamentals

Real-Life Scenario: Buying a Gaming Laptop

Below are my 5 basic learnings about data analytics that I got from working on this project -

Planning is Essential in any DA project
Before doing any research or collecting data, it's important to plan what you really need. This means setting clear goals and knowing what you're looking for. When you have a plan, you can focus on collecting the right data instead of getting lost in too much information.

Preparation Matters
Getting ready by checking your resources and limits is a key part of data analytics. Whether it's budgeting time or money, preparing helps you know what you have and what you can work with. This step helps in avoiding surprises later on when you start processing the data.

Collecting and Processing Data is to be focused on carefully
One of the big steps is gathering all the relevant data and making sure it's in a useful format. I learned that it's important to look for the details that matter to your goal. Processing the data means organizing it in a way that you can actually compare and understand different options.

Analysis is the Heart and Soul of Decision Making
After collecting the data, analyzing it carefully is what really drives your decision. By comparing the key details and checking if they meet your criteria, you can see which option stands out. This step shows that data analytics is not just about numbers, but about making informed choices.

Clear Communication and Action are Crucial
Once the analysis is done, sharing your findings with others (like discussing with a shopkeeper or team) is important to validate your choice. Finally, taking action based on your research completes the process. This signifies that data analytics isn't complete until the insights are used to make a real decision.

Project 2

Instagram User Analytics

This project aims to analyze user interactions and engagement patterns on Instagram to extract valuable insights. These insights will guide the product team in making data-driven decisions to

- **Improve user experience**
- **Enhance engagement**
- **Support business growth**

Approach to work on the project

Data Extraction: Using MySQL to extract and analyze user data.

User Behavior Analysis: Identifying key engagement metrics, such as likes, comments, shares, and time spent on the platform.

Actionable Insights: Deriving trends and recommendations to assist the product, marketing, and development teams in refining strategies and enhancing user satisfaction.

Decision Support: Presenting findings in a structured format to influence feature enhancements and business initiatives.

Project 2

Instagram User Analytics

Project Tasks

A) Marketing Analysis:

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

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

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

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

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

Project 2

Instagram User Analytics

Project Tasks

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

```
93 •  select * from users order by created_at asc limit 5;
```

Result Grid | Filter Rows:

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

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

```
•  select u.id, u.username, p.id as Photo_id
   from users u
   left join photos p
   on u.id = p.user_id
   where p.id is null
   order by u.id asc;
```

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

```
   select count(u.id) as Never_Posted
   from users u
   left join photos p
   on u.id = p.user_id
   where p.id is null;
```

Result Grid | Filte

Never_Posted
26

Project 2

Instagram User Analytics

Project Tasks

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

For this we will need to connect 3 tables

1. **Photos to users to get the username of the photo id**
2. **Photos to likes to get which photo got the likes on the photos**

The we need to sort the table in descending order after grouping them by photo id and user id

```
select p.id as PhotoID, u.username as Username, count(l.user_id) as TotalLikes
from photos p
join users u on u.id = p.user_id
left join likes l on l.photo_id = p.id
group by p.id, u.username order by TotalLikes desc limit 3;
```

And the output is as below, obviously photo id no 145 and the corresponding name is the winner of the contest

Result Grid | Filter Rows:

PhotoID	Username	TotalLikes
145	Zack_Kemmer93	48
182	Adelle96	43
127	Malinda_Streich	43

Project 2

Instagram User Analytics

Project Tasks

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

Here we need to connect 2 tables together

1. **Tags**
2. **Photo_tags**

using the tag_id

Then we need to find which tag has got maximum on of usages in different photos

```
select h.tag_name as TagName, count(t.photo_id) as UsageCount
from tags h
join photo_tags t on h.id = t.tag_id
group by h.tag_name
order by UsageCount desc limit 5;
```

Tags has been given alias 'h'
Photo_tags has been given alias 't'

Result Grid | Filter Rows:

TagName	UsageCount
smile	59
beach	42
party	39
fun	38
concert	24

Project 2

Instagram User Analytics

Project Tasks

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

Here is what we will need to extract to find the most suitable day for campaign launch

We will user the Users table as has the date to creation as well
We need to find the most no. of Ids created on the particular day
Based on the user registration, we can suggest the days to run ad campaigns on

Below is the query to get the day on which maximum no of users register on Instagram

```
select dayname(created_at) as CreationDay, count(username) as TotalUsers from users
group by CreationDay
order by TotalUsers desc;
```

Result Grid | Filter Rows:

CreationDay	TotalUsers
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Total User count is as below

```
select count(id) as TotalUsers from users;
```

Result Grid

TotalUsers
100

It indicates that on Thursdays and Sundays, maximum no of users register themselves. And these can be the best days to float any ad campaign.

Project 2

Instagram User Analytics

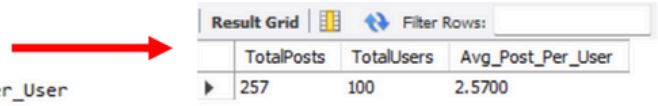
Project Tasks

B) Investor Metrics:

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

For this I need to calculate the total number of users and total number of posts and then find the average of that

```
select
count(p.id) as TotalPosts,
count(distinct u.id) as TotalUsers,
(count(p.id) / count(distinct u.id)) as Avg_Post_Per_User
from users u
left join photos p on p.user_id = u.id;
```

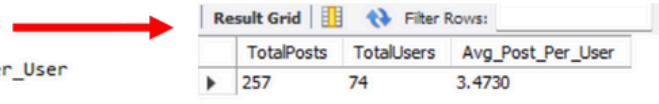


Result Grid | Filter Rows:

TotalPosts	TotalUsers	Avg_Post_Per_User
257	100	2.5700

Left join provides also the users who have not posted anything and this brings our avg down, whereas 'inner join' provide only the matching values and that is more accurate

```
select
count(p.id) as TotalPosts,
count(distinct u.id) as TotalUsers,
(count(p.id) / count(distinct u.id)) as Avg_Post_Per_User
from users u
join photos p on p.user_id = u.id;
```

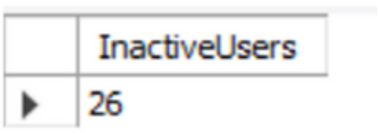


Result Grid | Filter Rows:

TotalPosts	TotalUsers	Avg_Post_Per_User
257	74	3.4730

Inactive users have been confirmed by the below query

```
select
count(u.id) as InactiveUsers
from users u
left join photos p on u.id = p.user_id where p.id is null;
```



InactiveUsers
26

Project 2

Instagram User Analytics

Project Tasks

B) Investor Metrics:

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

Here we need to find the user id from likes table and username from users table linking them with inner join

Using "Having" will make sure that the later condition is met
The later condition is – how many photos(posts) are there and they all have been liked

(the likes count of such users should match the no of photos)

```
select l.user_id, u.username
from likes l
join users u on l.user_id = u.id
group by l.user_id, u.username
having count(l.photo_id) = (select count(*) from photos);
```

Result Grid | Filter Rows:

user_id	username
5	Aniya_Hackett
14	Jadyn81
21	Rocio33
24	Maxwell.Halvorson
36	Ollie_Ledner37
41	Mckenna17

Result Grid | Filter Rows:

user_id	username
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
71	Nia_Haag
75	Leslie67
76	Janelle.Nikolaus81
91	Bethany20

The total count of such fake users = 13

Which is further confirmed with subquery in the next slide

The Subquery for confirming the count of fake users

```
select count(*) as FakeUsers from
(
select l.user_id, u.username
from likes l
left join users u on l.user_id = u.id
group by l.user_id, u.username
having count(l.user_id) = (select count(*) from photos))
As FakeUsers;
```

Result Grid | Filter Rows:

FakeUsers
13

Project 2

Instagram User Analytics

Major learnings from this project

1.Data Can Tell Unexpected Stories

When analyzing user engagement, I expected the oldest users to be the most active, but that wasn't always the case. Some early adopters had stopped posting altogether, while newer users were highly engaged.

2. Identifying Fake Activity Requires a Different Perspective

Finding users who liked every single post seemed straightforward at first, but it made me realize how bots operate differently from normal users. Unlike organic engagement, which is random and varied, bot behavior is systematic and repetitive.

3. Engagement Peaks Can Influence Business Strategy

Discovering the best days for user registrations and activity helped me understand how Instagram could optimize ad campaigns. Instead of just looking at raw numbers, I learned to think strategically and data driven.

4. SQL Can Be More Powerful Than Expected

A single query could reveal behavioral trends, detect anomalies, and drive decision-making, making SQL a crucial skill for any data-driven role. I never thought that it would be that strong and imperative for data analytics.

5. Small Details Can Lead to Big Business Decisions

The hashtag research was interesting. knowing which hashtags are most popular can dramatically impact reach, engagement, and brand partnerships. It reinforced that in data analytics, even seemingly minor insights can have major business implications when used correctly.

Project 3

Operation Analytics and Investigating Metric Spike

Project Description

This project is all about using data to improve how a company runs and spotting sudden changes in key numbers.

As a Lead Data Analyst, my job is to work with different teams—like operations, support, and marketing—to find useful insights in the data they collect.

The project is divided into two main parts:

1.Job Data Analysis – Looking at how jobs are reviewed, how often they happen, which languages are used, and whether there are any duplicate records.

1.Investigating Metric Spikes – Studying user behavior, engagement, and email activity to find patterns, growth trends, and sudden drops or spikes in activity.

Project 3

Operation Analytics and Investigating Metric Spike

Approach to the project

Writing SQL Queries: I'll use SQL to analyze data, track daily job reviews, check language trends, and find duplicate records.

Finding Trends Over Time: Instead of just looking at daily numbers, I'll calculate 7-day averages to get a clearer picture of how things change over time.

Understanding User Activity: I'll track user growth, weekly engagement, and retention to see how well a product is performing.

Detecting Unusual Spikes: If there's a sudden jump or drop in important metrics, I'll investigate what caused it and suggest possible solutions.

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Jobs Reviewed Over Time:

Objective: Calculate the number of jobs reviewed per hour for each day in November 2020.

Throughput Analysis:

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

Language Share Analysis:

Objective: Calculate the percentage share of each language in the last 30 days.

Duplicate Rows Detection:

Objective: Identify duplicate rows in the data.

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Jobs Reviewed Over Time:

Objective: Calculate the number of jobs reviewed per hour for each day in November 2020.

Throughput Analysis:

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

Language Share Analysis:

Objective: Calculate the percentage share of each language in the last 30 days.

Duplicate Rows Detection:

Objective: Identify duplicate rows in the data.

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Jobs Reviewed Over Time:

Objective: Calculate the number of jobs reviewed per hour for each day in November 2020.

```
SELECT
    ds AS DtReviewed,
    count(*) as JobIDs,
    round ((86400 / SUM(time_spent)), 0) AS JobsReviewedPerDay,
    round (((86400 / SUM(time_spent)) / 24), 0) AS JobsReviewedPerHour
FROM job_data
GROUP BY ds
having ds between '2020-11-01' and '2020-11-30'
ORDER BY ds;
```

DtReviewed	JobIDs	JobsReviewedPerDay	JobsReviewedPerHour
2020-11-25	1	1920	80
2020-11-26	1	1543	64
2020-11-27	1	831	35
2020-11-28	2	2618	109
2020-11-29	1	4320	180
2020-11-30	2	2160	90

Consideration

Every day is of 24 hours
Every day = 86400 seconds

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Throughput Analysis:

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

To get the throughput data
We need to divide the number of events by total time spent
Below are the queries for daily and weekly throughput

Query and outcome for

Weekly average throughput

```
select
count(distinct ds) as DateCount,
round(count(event) / sum(time_spent), 2) as WeeklyAvgTP
from job_data;
```

Week = no of distinct dates provided

Result Grid | Filter Rows:

DateCount	WeeklyAvgTP
6	0.03

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Language Share Analysis:

Objective: Calculate the percentage share of each language in the last 30 days.

```
select language,
count(job_id),
round(count(job_id) * 100 /  (select count(*) from job_data), 2) as Percentage
from job_data
group by language;
```

language	count(job_id)	Percentage
English	1	12.50
Arabic	1	12.50
Persian	3	37.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50

Project 3

Operation Analytics and Investigating Metric Spike

Tasks of the project

Duplicate Rows Detection:

Objective: Identify duplicate rows in the data.

```
select ds, Job_id, actor_id, event, language, time_spent, org,
count(*) as duplicate_data
from job_data
group by ds, Job_id, actor_id, event, language, time_spent, org
having count(*) > 1
order by duplicate_data desc;
```

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

ds	Job_id	actor_id	event	language	time_spent	org	duplicate_data

This query does not return any value because there are no duplicate rows in the table, although there are individual duplicate items like job id and actor id but they have different set of row values

Project 4

Hiring Process Analytics

Project Description

In this project, we will analyze the hiring process data of a multinational company. The goal is to uncover key insights related to hiring trends, salary distribution, and departmental composition. By leveraging data analytics techniques in Excel, we will clean, organize, and visualize the data to support better hiring decisions.

Objectives:

- **Understand Hiring Trends: Analyze gender distribution and hiring patterns.**
- **Salary Insights: Determine the average salary and study salary distribution across different levels.**
- **Departmental Overview: Visualize the number of employees across various departments.**
- **Position Tier Analysis: Identify how different job positions are structured within the company.**

Project 4

Hiring Process Analytics

Project Tasks

Cleaning the data

First, we need to clean the data on various parameters

The detailed cleaning steps have been included in project 4 excel file → Cleaning sheet

One of the major steps of cleaning the data was tracing the outliers and managing them

I did it with Z Score method and found 3 of them and decided to delete them as they were quite big on salary figures

Status	event_name	Department	Post Name	Offered Salary	Z score	Column2
Hired	Female	Service Department	b9	200000	5.20	Outl
Hired	Female	General Management	i4	400000	12.13	Outl
Hired	Male	General Management	i7	300000	8.67	Outl

After deleting the outliers, we have total 1765 rows against 1768 of original dataset

Project 4

Hiring Process Analytics

Project Tasks

- **Understand Hiring Trends: Analyze gender distribution and hiring patterns.**

Count has been found with countifs() as well as Pivot table as shown below

Count of Total Males and Females Hired

Males Hired	2562
Females Hiired	1854

<--- with countifs formula

Status	Hired

Event Name	Total Count
Don't want to say	278
Female	1854
Male	2562

<--- with Pivot Table

We can draw additional data on Rejected condition as well

Count of Total Males and Females Rejected

Males Rejected	1522
Females Rejected	819

<--- with countifs formula

Status	Rejected

Event Name	Total Count
Don't want to say	130
Female	819
Male	1522

<--- with Pivot Table

Project 4

Hiring Process Analytics

Project Tasks

- **Understand Hiring Trends: Analyze gender distribution and hiring patterns.**

Status	Hired
Hiring Timeline	Hired Count
⊞ May	1094
⊞ Jun	1078
⊞ Jul	1295
⊞ Aug	1227
Grand Total	4694

Project 4

Hiring Process Analytics

Project Tasks

- **Salary Insights: Determine the average salary and study salary distribution across different levels.**

Department wise average salary offered

Departments	Average Salary Offered
Finance Department	49628.01
General Management	55295.29
Human Resource Department	49002.28
Marketing Department	48489.94
Operations Department	49151.35
Production Department	49448.48
Purchase Department	52564.77
Sales Department	49310.81
Service Department	50557.16

With Pivot Table

Departments	Average Salary Offered
Finance Department	49628.01
General Management	55295.29
Human Resource Department	49002.28
Marketing Department	48489.94
Operations Department	49151.35
Production Department	49448.48
Purchase Department	52564.77
Sales Department	49310.81
Service Department	50557.16

With Averageif() function

Project 4

Hiring Process Analytics

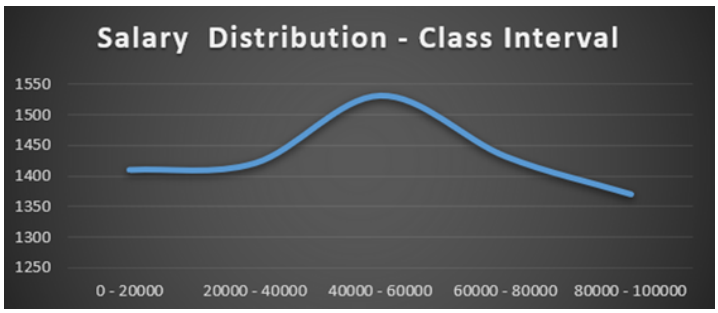
Project Tasks

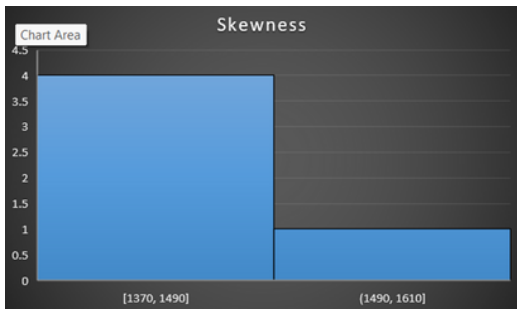
- **Salary Insights: Determine the average salary and study salary distribution across different levels.**

Mean	49878.30
Median	49625
Mode	72843
Min	100
Max	99967

Class Interval	Frequency	C. Frequency
20000	1410	1410
40000	1421	2831
60000	1532	4363
80000	1432	5795
100000	1370	7165

Skew	0.013181923





It shows that maximum salary offered range is between 40000 and 60000

Project 4

Hiring Process Analytics

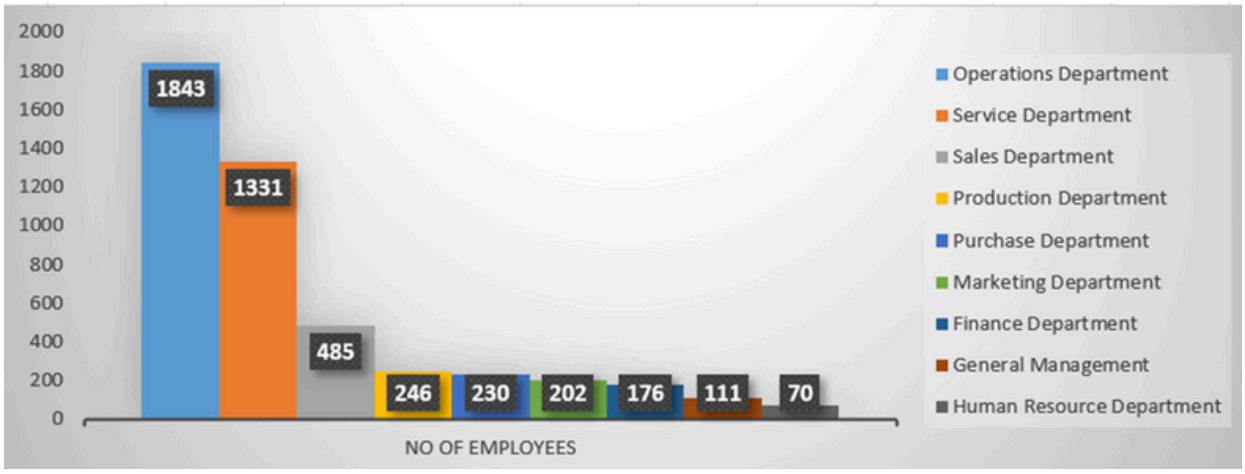
Project Tasks

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

For this we will use pivot table and identify how many employees are working in different departments respectively

We will do it only for the employees who have been hired and working in one or the other department

Department	No of Employees
Operations Department	1843
Service Department	1331
Sales Department	485
Production Department	246
Purchase Department	230
Marketing Department	202
Finance Department	176
General Management	111
Human Resource Department	70



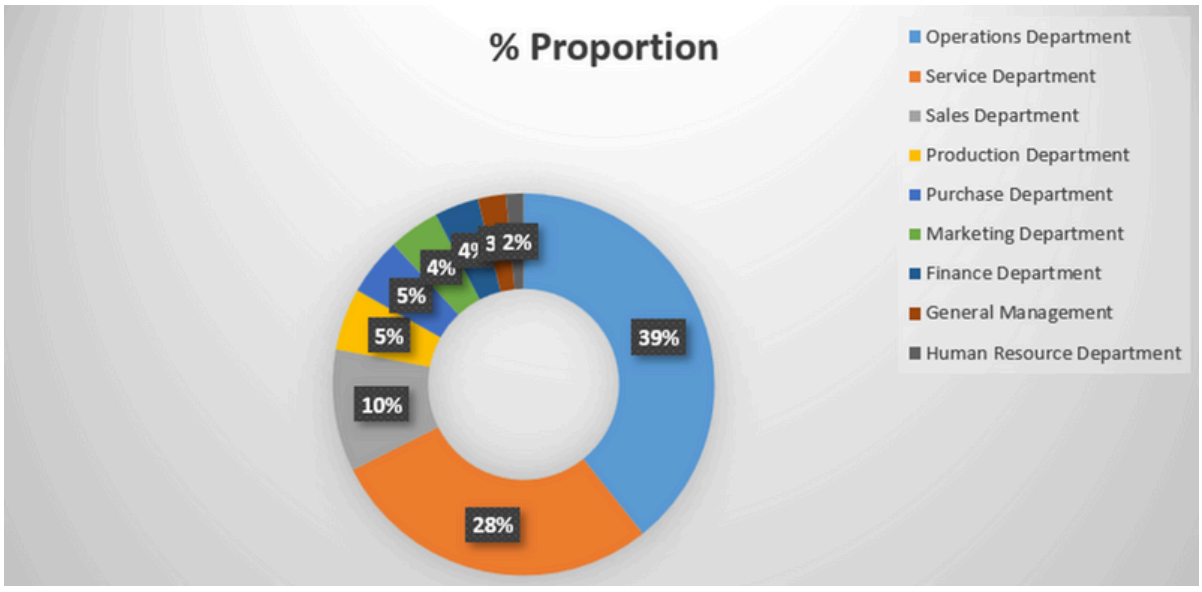
Project 4

Hiring Process Analytics

Project Tasks

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

Department	% Proportion
Operations Department	39.26%
Service Department	28.36%
Sales Department	10.33%
Production Department	5.24%
Purchase Department	4.90%
Marketing Department	4.30%
Finance Department	3.75%
General Management	2.36%
Human Resource Department	1.49%



Project 4

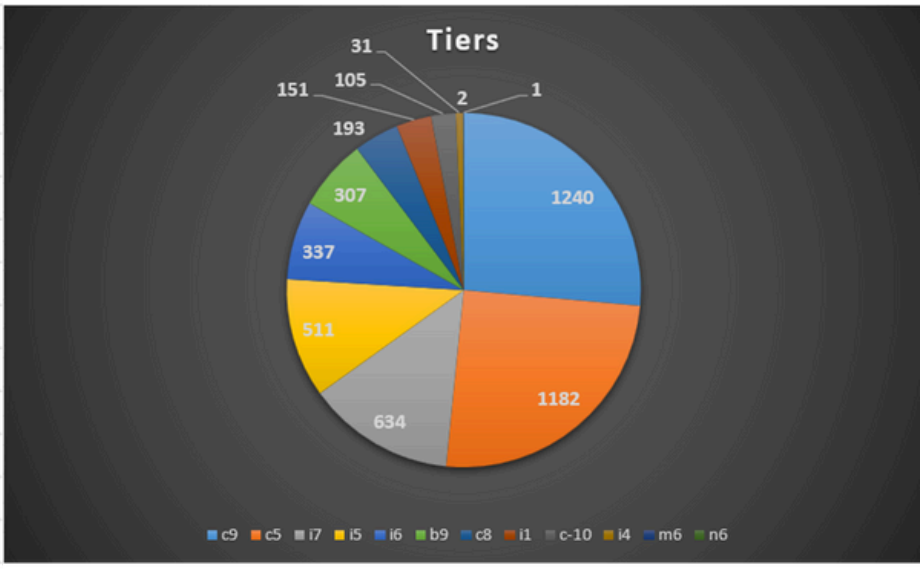
Hiring Process Analytics

Project Tasks

- **Position Tier Analysis: Identify how different job positions are structured within the company.**

Here is the account of all the tiers within the company across departments (for Hired Employees)

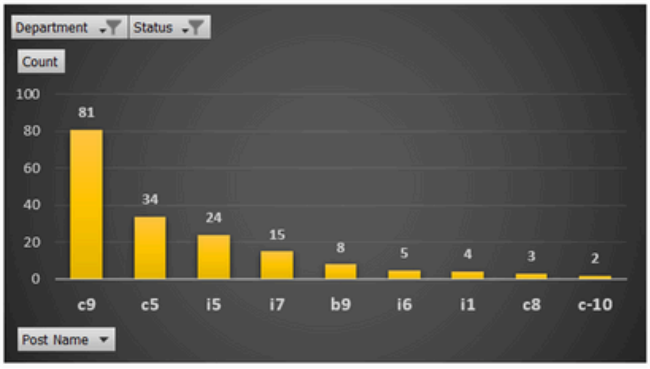
Tiers	Count
c9	1240
c5	1182
i7	634
i5	511
i6	337
b9	307
c8	193
i1	151
c-10	105
i4	31
m6	2
n6	1



We can see the tiers department wise as well

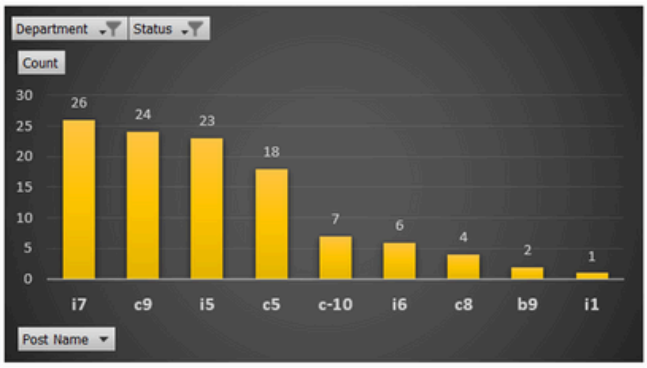
Department	Finance Department
Status	Hired

Tier	Count
c9	81
c5	34
i5	24
i7	15
b9	8
i6	5
i1	4
c8	3
c-10	2
Grand Total	176



Department	General Management
Status	Hired

Tier	Count
i7	26
c9	24
i5	23
c5	18
c-10	7
i6	6
c8	4
b9	2
i1	1
Grand Total	111



Project 4

Hiring Process Analytics

Key Insights from the project

Operations and Service Dominate Hiring:

The Operations Department (39.26%) and Service Department (28.36%) together make up 67.62% of total hires, reflecting a strong emphasis on operational efficiency and customer service.

Moderate Hiring in Sales:

The Sales Department (10.33%) has a balanced workforce, likely indicating controlled turnover and a stable hiring trend.

Specialized Departments Have Fewer Hires:

Departments like Production (5.24%), Purchase (4.90%), and Marketing (4.30%) have lower hiring rates, suggesting these roles require specialized skills and fewer employees.

Minimal Hiring in Support Functions:

Finance (3.75%), General Management (2.36%), and HR (1.49%) have low hiring rates, possibly due to stability in leadership roles and lower workforce requirements.

Hiring Trends Indicate Workforce Priorities:

The focus is on frontline roles (Operations & Service) rather than strategic and support functions, indicating a priority on customer service and business operations.

Project 5

IMDB Movie Analysis

Project Description

The dataset provided is related to IMDB Movies.

A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?"

Here, success can be defined by high IMDB ratings.

The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Project 5

IMDB Movie Analysis

Project Tasks

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

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

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

Director Analysis: Influence of directors on movie ratings.

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

Project 5

IMDB Movie Analysis

Project Approach

Data Cleaning:

- Removed missing or irrelevant data.
- Standardized numerical formats (example - converted figures into millions).

Analysis Techniques Used:

- Correlation Analysis: Measured the strength of relationships between budget and financial success.
- Factor Impact Analysis: Assessed how genre, director, language, budget, and duration influence IMDb scores.
- Profit/Loss Evaluation: Identified the most profitable movies based on budget vs. gross earnings.

Visualization:

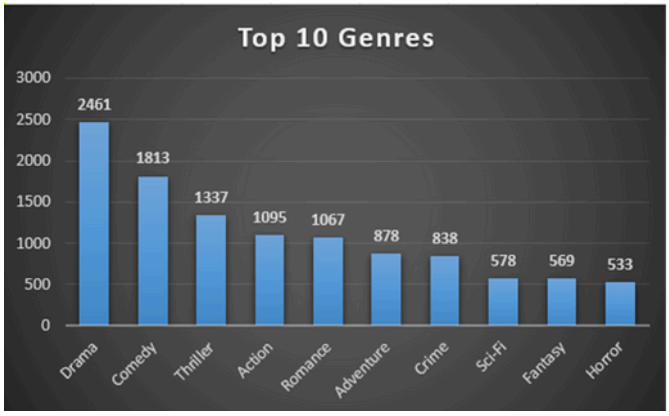
- Created scatter plots, bar charts, and tables for a clear presentation of findings.

Project 5

IMDB Movie Analysis

Project Tasks

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



Unique Genres and their occurance in the data		Descriptive Statistics of all the genres							
Genres	Count	Mean	Median	Mode	Max	Min	Range	VAR	SD
Drama	2461	6.7	6.8	7.2	8.8	3.3	5.5	0.9	0.9
Comedy	1813	6.2	6.3	6.3	9.5	1.9	7.6	1.2	1.1
Thriller	1337	6.3	6.4	6.4	7.5	3.4	4.1	1.1	1.0
Action	1095	6.2	6.3	6.6	9.1	3.7	5.4	1.2	1.1
Romance	1067	6.4	6.5	6.5	6.6	5.1	1.5	1.0	1.0
Adventure	878	6.4	6.6	6.6	7.8	3.3	4.5	1.3	1.1
Crime	838	6.5	6.6	6.6	0.0	0.0	0.0	1.0	1.0
Sci-Fi	578	6.2	6.3	6.7	6.3	2.8	3.5	1.5	1.2
Fantasy	569	6.3	6.4	6.7	6.8	6.4	0.4	1.3	1.2
Horror	533	5.8	5.9	6.2	8.0	2.2	5.8	1.2	1.1
Family	523	6.2	6.3	6.7	8.6	3.2	5.4	1.4	1.20
Mystery	461	6.4	6.5	6.4	0	0	0	1.1	1.06
Music	322	6.4	6.7	7.1	7.2	7.2	0	1.4	1.20
Biography	289	7.1	7.2	7.0	5.7	5.7	0	0.5	0.72
Animation	233	6.5	6.7	6.7	4.8	3.7	1.1	1.3	1.14
War	206	7.1	7.1	7.1	0	0	0	0.8	0.87
History	199	7.1	7.2	7.5	7.5	7.5	0	0.8	0.88
Sport	176	6.6	6.8	7.2	0	0	0	1.2	1.10
Musical	131	6.5	6.7	7.0	3.4	3.4	0	1.5	1.22
Documentary	120	7.2	7.4	7.5	8.7	2.7	6	1.1	1.06
Western	94	6.7	6.8	6.5	8.9	3.8	5.1	1.1	1.05
Film-Noir	6	7.6	7.7	0.0	0	0	0	0.2	0.39
Short	5	6.4	6.5	0.0	0	0	0	0.4	0.67
News	3	7.5	7.4	0.0	0	0	0	0.2	0.42

Unique Genres and their occurance in the data		Descriptive Statistics of all the genres							
Genres	Count	Mean	Median	Mode	Max	Min	Range	VAR	SD
Drama	2461	6.7	6.8	7.2	8.8	3.3	5.5	0.9	0.9
Comedy	1813	6.2	6.3	6.3	9.5	1.9	7.6	1.2	1.1
Thriller	1337	6.3	6.4	6.4	7.5	3.4	4.1	1.1	1.0
Action	1095	6.2	6.3	6.6	9.1	3.7	5.4	1.2	1.1
Romance	1067	6.4	6.5	6.5	6.6	5.1	1.5	1.0	1.0
Adventure	878	6.4	6.6	6.6	7.8	3.3	4.5	1.3	1.1
Crime	838	6.5	6.6	6.6	0.0	0.0	0.0	1.0	1.0
Sci-Fi	578	6.2	6.3	6.7	6.3	2.8	3.5	1.5	1.2
Fantasy	569	6.3	6.4	6.7	6.8	6.4	0.4	1.3	1.2
Horror	533	5.8	5.9	6.2	8.0	2.2	5.8	1.2	1.1

Insights from top 10 genres and their imdb scores

Here are key insights based on the top 10 genres:

- Drama is the most frequent genre (2,461 movies) and has the lowest variance (0.9), indicating stable IMDb ratings.
- Comedy has the widest rating range (7.6), showing strong audience polarization.
- Crime and Romance share the second-lowest variance (1.0), implying consistent audience reception.
- Comedy movies have the highest maximum rating (9.5), suggesting some top-rated blockbuster hits.
- Fantasy and Sci-Fi have the highest standard deviations (1.2), showing varied audience opinions.
- Thriller, Action, and Horror have the lowest median ratings (6.3, 6.3, 5.9), indicating they are more critically mixed.

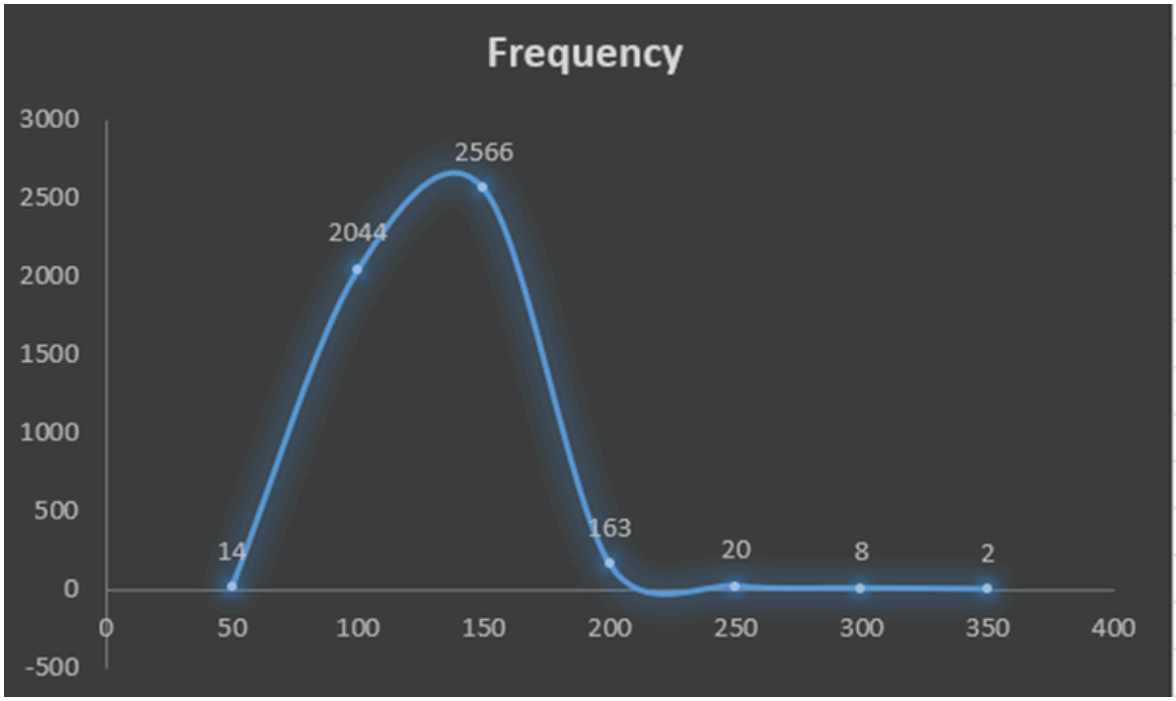
Project 5

IMDB Movie Analysis

Project Tasks

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

Class	Frequency
1-50	14
50-100	2044
100-150	2566
150-200	163
200-250	20
250-300	8
300-350	2



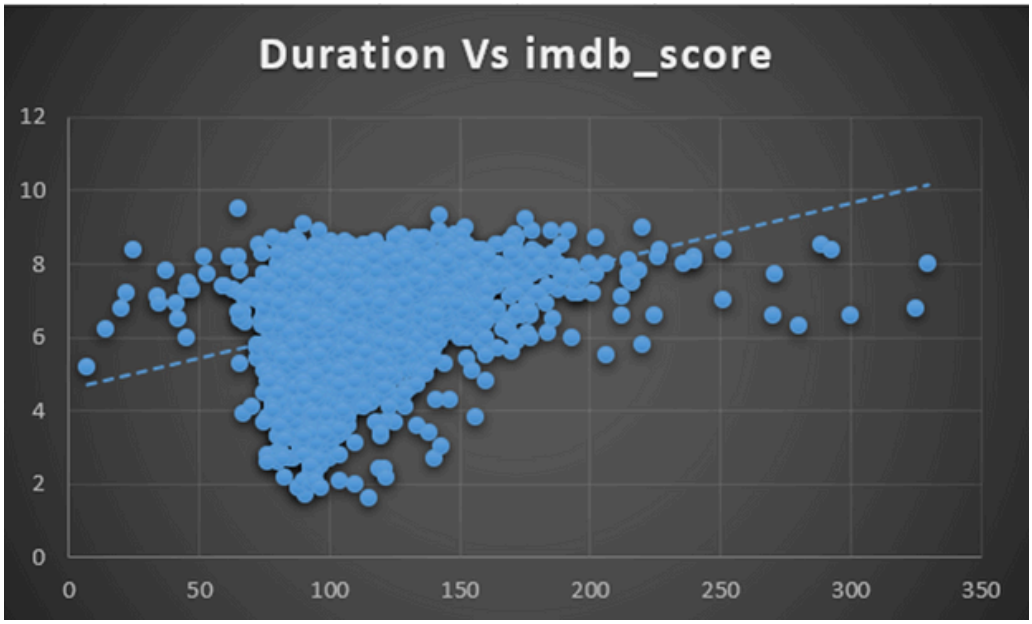
Maximum movies fall between 50 and 150 minutes of duration

Project 5

IMDB Movie Analysis

Project Tasks

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



Key insights

THERE IS A WEAK POSITIVE CORRELATION BETWEEN MOVIE DURATION AND IMDB SCORE, MEANING LONGER MOVIES TEND TO HAVE SLIGHTLY HIGHER RATINGS.

MOST MOVIES FALL BETWEEN 80-150 MINUTES, FORMING A DENSE CLUSTER, INDICATING A STANDARD DURATION RANGE FOR FILMS.

MOVIES SHORTER THAN 60 MINUTES MIXED RATINGS BETWEEN 5.5 AND 8.2, WHEREAS MANY MOVIES WITH DURATION BETWEEN 80 AND 130 ALSO RECEIVE LOWER RATINGS.

A FEW OUTLIERS ABOVE 250 MINUTES EXIST, BUT THEIR RATINGS BETWEEN 6 AND 8, VARY WIDELY, SHOWING NO STRONG TREND AMONG EXTREMELY LONG FILMS.

THE TRENDLINE SUGGESTS A GRADUAL INCREASE IN IMDB SCORE WITH DURATION, BUT THE SPREAD OF DATA POINTS IMPLIES OTHER FACTORS ALSO INFLUENCE RATINGS.

Project 5

IMDB Movie Analysis

Project Tasks

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

Descriptive Statistics

Language	Movie Count	Mean	Median	Mode	Max	Min	Range	VAR	SD
English	4500	6.37	6.50	6.70	9.50	1.60	7.90	1.24	1.12
French	72	7.02	7.20	7.20	8.40	4.90	3.50	0.51	0.71
Spanish	40	6.94	7.15	7.20	8.20	4.40	3.80	0.71	0.84
Hindi	28	6.63	6.95	7.80	8.50	2.80	5.70	1.89	1.37
Mandarin	24	6.79	7.05	7.60	7.90	3.20	4.70	1.03	1.02
German	19	7.34	7.60	7.40	8.50	4.90	3.60	0.86	0.93
Japanese	16	7.37	7.60	8.20	8.70	5.60	3.10	0.99	1.00
Cantonese	11	6.95	7.20	6.50	7.80	5.30	2.50	0.45	0.67
Russian	11	6.36	6.50	5.30	8.10	4.10	4.00	1.74	1.32
Italian	10	7.08	7.15	0.00	8.90	5.10	3.80	1.31	1.14

Filtering of language data in different category

Movie count-wise top 10 languages

Language	No of Movies	Average of imdb_score
English	4500	6.4
French	72	7.0
Spanish	40	6.9
Hindi	28	6.6
Mandarin	24	6.8
German	19	7.3
Japanese	16	7.4
Russian	11	6.4
Cantonese	11	7.0
Italian	10	7.1

Highest imdb rating wise top 10 languages

Language	No of Movies	Average of imdb_score
Telugu	1	8.4
None	2	8.0
Indonesian	2	7.9
Maya	1	7.8
Hebrew	5	7.6
Persian	4	7.6
Danish	5	7.5
Dari	2	7.5
Dzongkha	1	7.5
Portuguese	8	7.5

Project 5

IMDB Movie Analysis

Key insights

English Dominates Quantity but Not Quality –
English is the most common language with 4500 movies, but its average IMDB score (6.4) is lower than several other languages.

Japanese and German Have High Ratings –
Among the top 10 most common languages, Japanese (7.37) and German (7.34) have the highest average IMDB scores.

French and Italian Perform Well –
French (7.02) and Italian (7.08) maintain strong average IMDB scores, suggesting that movies in these languages are generally well-received.

Higher Scores in Less Common Languages –
Less frequently used languages like Telugu (8.4), Indonesian (7.9), and Hebrew (7.6) tend to have higher IMDB scores, indicating that movies in these languages may be more selective or higher quality. Due to a smaller number of movies, they cannot be considered in drawing big conclusions.

Wide Score Range in English and Hindi –
English movies have the widest IMDB score range (7.9), followed by Hindi (5.7), showing that quality varies significantly within these languages.

Lowest Variance in Cantonese Movies –
Cantonese has the lowest variance (0.45) and standard deviation (0.67), meaning its IMDB scores are more consistent compared to other languages.

Project 5

IMDB Movie Analysis

Project Tasks

Director Analysis: Influence of directors on movie ratings.

Here are the analytics

Movie Count-wise top directors			
Directors	**Movies**	**Avg IMDB Score**	**Percentile Rank**
Steven Spielberg	26	7.5	89%
Woody Allen	22	7.0	75%
Martin Scorsese	20	7.7	93%
Clint Eastwood	20	7.2	83%
Spike Lee	16	6.6	56%
Ridley Scott	16	7.1	79%
Steven Soderbergh	15	6.7	61%
Renny Harlin	15	5.7	28%
Tim Burton	14	7.1	76%
Oliver Stone	14	7.0	72%
Ron Howard	13	6.9	72%
Robert Zemeckis	13	7.3	86%
Robert Rodriguez	13	5.7	26%
Joel Schumacher	13	6.4	51%
Barry Levinson	13	6.6	57%

Total Unique Directors 2398

IMDB Score-wise Directors			
Directors	**Movies**	**Avg IMDB Score**	**Percentile Rank**
John Blanchard	1	9.5	100%
Sadyk Sher-Niyaz	1	8.7	100%
Mitchell Altieri	1	8.7	100%
Cary Bell	1	8.7	100%
Mike Mayhall	1	8.6	100%
Charles Chaplin	1	8.6	100%
Ron Fricke	1	8.5	100%
Raja Menon	1	8.5	100%
Majid Majidi	1	8.5	100%
Damien Chazelle	1	8.5	100%
Sergio Leone	4	8.5	100%

Project 5

IMDB Movie Analysis

Key insights

Martin Scorsese has the best score among top directors –
He has directed 20 movies with an average IMDB score of 7.7, placing him in the 93rd percentile, meaning his movies are highly rated.

Steven Spielberg is the most consistent –
With 26 movies and an average IMDB score of 7.5 (89th percentile), he has maintained strong ratings across many films.

Clint Eastwood and Woody Allen have good ratings –
Both have directed over 20 movies. Clint Eastwood has an average IMDB score of 7.2 (83rd percentile), while Woody Allen has 7.0 (75th percentile).

Some directors have made many movies but have lower ratings – Renny Harlin (5.7 IMDB score) and Steven Soderbergh (6.7 IMDB score) have directed several movies, but their ratings are not as high.

Directors with the highest IMDB scores have made only one movie – John Blanchard (9.5 IMDB score) and Sadyk Sher-Niyaz (9.0 IMDB score) have the best scores, but they have only directed a single film.

Tim Burton and Robert Zemeckis are consistently good –
Tim Burton (14 movies, 7.1 score) and Robert Zemeckis (13 movies, 7.3 score) have made multiple movies that are well-rated by audiences.

Project 5

IMDB Movie Analysis

Project Tasks

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

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

Correlation = 0.0993131

Top 10 most profit making movies using Large() function					
Max Profit	Budget	Director	Movie_title	% Profit over Budget	IMDB Score
523505847	237000000	James Cameron	Avatar	221%	7.9
502177271	150000000	Colin Trevorrow	Jurassic World	335%	7
458672302	200000000	James Cameron	Titanic	229%	7.7
449935665	11000000	George Lucas	Star Wars: Episode IV - A New Hope	4090%	8.7
424449459	10500000	Steven Spielberg	E.T. the Extra-Terrestrial	4042%	7.9
403279547	220000000	Joss Whedon	The Avengers	183%	8.1
377783777	45000000	Roger Allers	The Lion King	840%	8.5
359544677	115000000	George Lucas	Star Wars: Episode I - The Phantom Menace	313%	6.5
348316061	185000000	Christopher Nolan	The Dark Knight	188%	9
329999255	78000000	Gary Ross	The Hunger Games	423%	7.3

Project 5

IMDB Movie Analysis

Key insights

A high-profit movie doesn't always have the best rating –
The Dark Knight has the highest IMDb rating (9.0) but made a lower profit ($348M, 188% profit) than Avatar ($523M, 221%) and Jurassic World ($502M, 335%), which have lower ratings (7.9 and 7.0).

Small-budget movies can still make huge profits –
Star Wars: Episode IV - A New Hope had a budget of only $11M but made a massive profit of $449M, achieving an outstanding 4090% profit margin with a strong IMDb rating of 8.7. Similarly, E.T. the Extra-Terrestrial turned a $10.5M budget into $424M profit (4042%). This shows that a good story and a strong fan base matter more than just big budgets.

Superhero and animated movies are both profitable and well-rated –
The Avengers (IMDb 8.1, $403M profit, 183%) and The Lion King (IMDb 8.5, $377M profit, 840%) show that popular franchises can make big money and also get high ratings.

A low rating doesn't always mean low profit –
Star Wars: Episode I - The Phantom Menace has the lowest IMDb score (6.5) but still made $359M profit (313%), likely because Star Wars has a huge fan following.

Christopher Nolan's The Dark Knight is the best-rated movie on the list –
It has the highest IMDb rating (9.0) and still made a strong profit of $348M (188%), proving that a movie can be both critically loved and financially successful.

Project 5

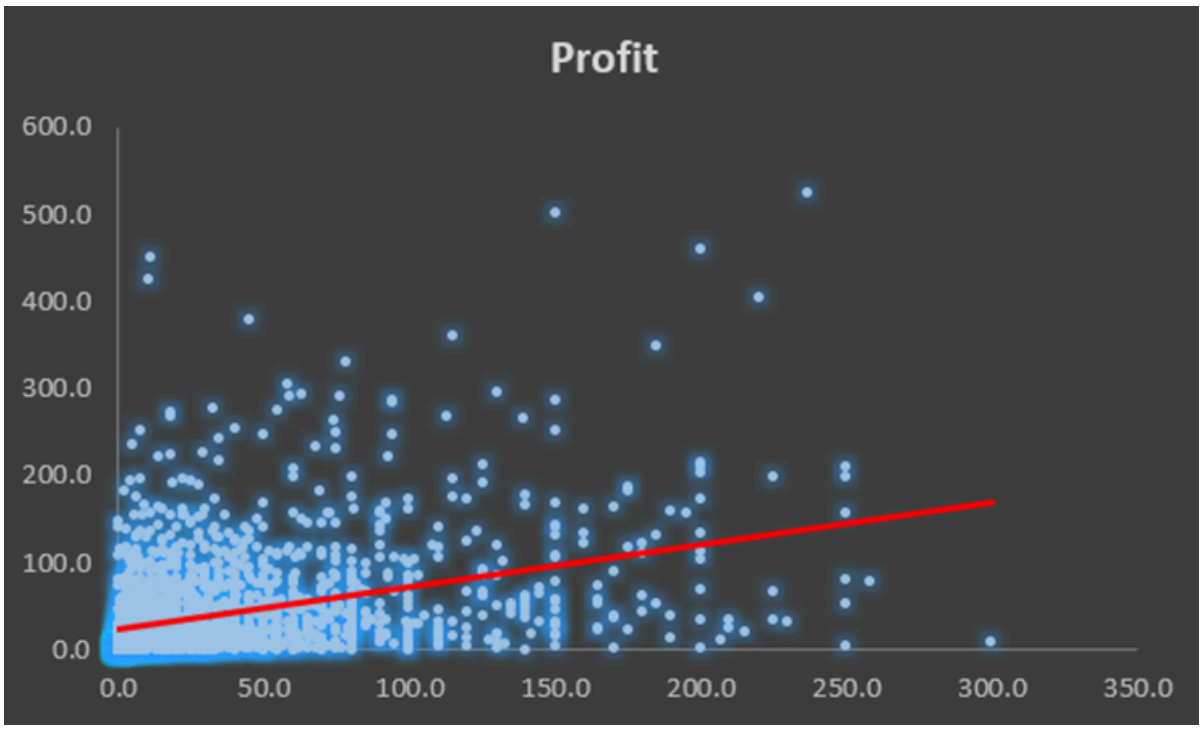
IMDB Movie Analysis

Project Tasks

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

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

Scatter Plot



Project 5

IMDB Movie Analysis

5 Why approach of analysis - 1

Why do some low-budget movies still achieve high profits?
They have strong storytelling and audience engagement.

Why does strong storytelling lead to high profits?
Engaging plots attract word-of-mouth marketing, leading to more viewers.

Why does word-of-mouth marketing increase viewership?
People trust recommendations from friends and family more than advertisements.

Why do personal recommendations matter more than advertisements?
They feel more authentic and unbiased.

Why does authenticity matter in movie success?
It builds credibility and long-term fan loyalty.

Project 5

IMDB Movie Analysis

5 Why approach of analysis - 2

Why do some movies with lower IMDb ratings still make high profits?
They belong to well-established franchises with loyal fan bases.

Why do established franchises attract large audiences despite lower ratings?
Fans are emotionally invested in the characters and storylines.

Why does emotional investment lead to box office success?
Fans watch the movies regardless of critical reception.

Why do fans continue watching despite poor reviews?
They want to stay updated with the franchise's story.

Why does staying updated matter?
It keeps them connected to the larger fan community and cultural trends.

Project 5

IMDB Movie Analysis

5 Why approach of analysis - 3

Why do movies directed by well-known directors tend to perform better?
Famous directors have built trust and a loyal audience.

Why does a director's reputation impact movie success?
Audiences associate their name with quality filmmaking.

Why do people trust well-known directors?
Their past successful movies have created a strong track record.

Why does a strong track record influence audience behavior?
It reduces the perceived risk of watching a new movie.

Why does reducing risk make people more likely to watch a movie?
Viewers feel confident they will get a good experience, leading to strong box office performance.

Project 5

IMDB Movie Analysis

My Key Learnings from IMDb Movie Analysis

Clean data is crucial – Before analyzing, it's important to remove errors, standardize formats, and organize data properly.

Numbers need context – Looking at percentages (like profit % over budget) alongside absolute values gives a clearer picture.

Data reveals patterns – Genre, director, and budget influence a movie's success, just like key factors drive outcomes in any industry.

Statistics support decisions – Using metrics like correlation, outliers, and averages helps in making informed business choices.

Presentation makes insights powerful – Presenting data in a clear and structured way is as important as analyzing it.

Project 5

IMDB Movie Analysis

Key insights

Scatter Plot (Budget vs. Profit)

Positive Correlation: The red trend line indicates a positive relationship between budget and profit—higher budgets generally lead to higher profits.

High Variability: While some high-budget movies earn significant profits, many lower-budget movies also perform well, showing that budget alone doesn't guarantee success.

Outliers Exist: A few movies with extremely high profits stand out, suggesting that exceptional films can break trends and outperform expectations.

Cluster at Lower Budgets: Most movies have lower budgets and profits, implying that smaller films dominate the industry, with only a few high-budget blockbusters.

Budget is a Factor, Not a Guarantee: While there is a general upward trend, the wide spread of points shows that other factors play a crucial role in profitability.

Project 6

Bank Loan Case Study

Project Description

As a data analyst at a finance company specializing in urban loans, my role is to analyze customer loan applications to minimize financial risk.

The company faces two challenges simultaneously: rejecting creditworthy applicants leads to lost business, while approving high-risk applicants results in defaults and financial losses.

Using Exploratory Data Analysis (EDA), this project aims to identify patterns in customer attributes and loan characteristics that influence loan repayment behavior.

The goal is to help the company make informed decisions on loan approvals, reducing risks while ensuring business growth.

Project 6

Bank Loan Case Study

Project Tasks

Identify and Handle Missing Data – Detect missing values and decide on the best approach (imputation or removal) to maintain data integrity.

Identify Outliers – Find extreme values in numerical variables using statistical techniques to prevent data distortion.

Analyze Data Imbalance – Examine the distribution of loan default cases to assess if the dataset is skewed towards a particular class.

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Identify Top Correlations – Determine the strongest relationships between customer attributes and loan default risk for better decision-making.

Project Files

Project has 2 database files

Project 6

Bank Loan Case Study

Project Tasks

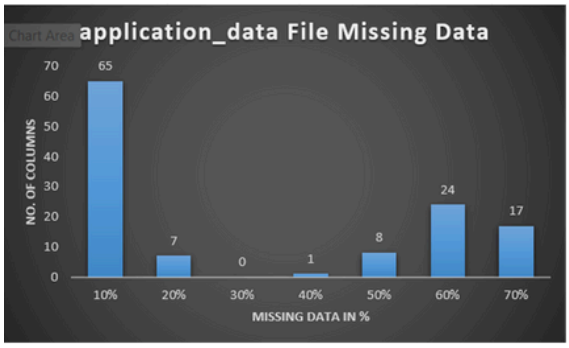
Identify and Handle Missing Data – Detect missing values and decide on the best approach (imputation or removal) to maintain data integrity.

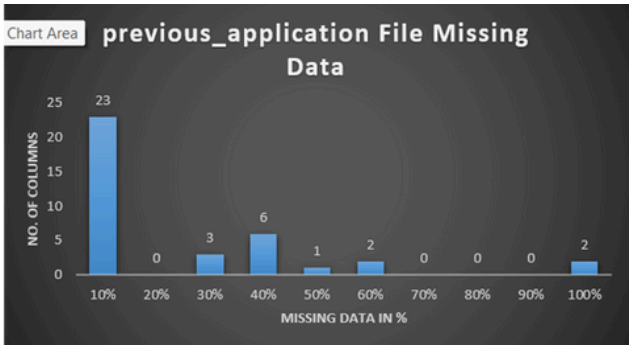
Application_data file

Total Columns	122
Total Rows	49999
Total Blank Cells	1488212
Columns with > 50% Blank Cells	41
% of cells having >50% missing data	34%

Previous_application file

Total Columns	37
Total Rows	49999
Total Blank Cells	321203
Columns with > 50% Blank Cells	4
% of cells having >50% missing data	11%





Column Name	Missing Values	% of Missing Values
COMMONAREA_AVG	34960	70%
COMMONAREA_MODE	34960	70%
COMMONAREA_MEDI	34960	70%
NONLIVINGAPARTMENTS_AVG	34714	69%
NONLIVINGAPARTMENTS_MODE	34714	69%
NONLIVINGAPARTMENTS_MEDI	34714	69%
LIVINGAPARTMENTS_AVG	34226	68%
LIVINGAPARTMENTS_MODE	34226	68%
LIVINGAPARTMENTS_MEDI	34226	68%
FONDKAPREMONT_MODE	34191	68%
FLOORSMIN_AVG	33894	68%
FLOORSMIN_MODE	33894	68%
FLOORSMIN_MEDI	33894	68%
YEARS_BUILD_AVG	33239	66%
YEARS_BUILD_MODE	33239	66%
YEARS_BUILD_MEDI	33239	66%
OWN_CAR_AGE	32950	66%
LANDAREA_AVG	29721	59%
LANDAREA_MODE	29721	59%
LANDAREA_MEDI	29721	59%
BASEMENTAREA_AVG	29199	58%
BASEMENTAREA_MODE	29199	58%
BASEMENTAREA_MEDI	29199	58%
EXT_SOURCE_1	28172	56%
NONLIVINGAREA_AVG	27572	55%
NONLIVINGAREA_MODE	27572	55%
NONLIVINGAREA_MEDI	27572	55%
ELEVATORS_AVG	26651	53%
ELEVATORS_MODE	26651	53%
ELEVATORS_MEDI	26651	53%
WALLSMATERIAL_MODE	25459	51%
APARTMENTS_AVG	25385	51%
APARTMENTS_MODE	25385	51%
APARTMENTS_MEDI	25385	51%
ENTRANCES_AVG	25195	50%
ENTRANCES_MODE	25195	50%
ENTRANCES_MEDI	25195	50%
LIVINGAREA_AVG	25137	50%
LIVINGAREA_MODE	25137	50%
LIVINGAREA_MEDI	25137	50%
HOUSETYPE_MODE	25075	50%

Column Name	Missing Values	% of Missing Values
RATE_INTEREST_PRIMARY	49834	99.7%
RATE_INTEREST_PRIVILEGED	49834	99.7%
AMT_DOWN_PAYMENT	25198	50.4%
RATE_DOWN_PAYMENT	25198	50.4%

Application_data file

Previous_application file

We will choose to delete these columns as they lack more than 50% data inside them, and it will prevent us from getting meaningful insights

Project 6

Bank Loan Case Study

Project Tasks

Identify and Handle Missing Data – Detect missing values and decide on the best approach (imputation or removal) to maintain data integrity.

Application_data file missing data replacement strategy

AMT_ANNUITY		CNT_FAM_MEMBERS		DAYS_LAST_PHONE_CHANGE		AMT_GOODS_PRICE	
Mean	27107.37736	Mean	2.158946	Mean	-964.296	Mean	539060.0361
Standard Error	65.12877001	Standard Error	0.004076	Standard Error	3.709646	Standard Error	1654.67948
Median	24939	Median	2	Median	-755	Median	450000
Mode	9000	Mode	2	Mode	0	Mode	450000
Standard Deviation	14562.94444	Standard Deviation	0.911332	Standard Deviation	829.4856	Standard Deviation	369853.2527
Sample Variance	212079350.6	Sample Variance	0.830527	Sample Variance	688046.3	Sample Variance	1.36791E+11
Kurtosis	9.412028546	Kurtosis	1.715436	Kurtosis	-0.32418	Kurtosis	2.486844961
Skewness	1.688525905	Skewness	0.949618	Skewness	-0.71092	Skewness	1.347815809
Range	255973.5	Range	12	Range	4002	Range	4005000
Minimum	2052	Minimum	1	Minimum	-4002	Minimum	45000
Maximum	258025.5	Maximum	13	Maximum	0	Maximum	4050000
Sum	1355314653	Sum	107943	Sum	-4.8E+07	Sum	26931978465
Count	49998	Count	49998	Count	49998	Count	49961

Will replace the null values with **Median as skewness is either positive or negative**

DS	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
Mean	0.007095805	0.00751185	0.03238183	0.27028776	0.26097307	1.88103548
Mode	0	0	0	0	0	0
Median	0	0	0	0	0	1
Min	0	0	0	0	0	0
Max	3	6	6	24	8	25
SD	0.087708647	0.10799223	0.19408035	0.92856012	0.60699573	1.8650543
VAR	0.007692807	0.01166232	0.03766718	0.8622239	0.36844381	3.47842755
Skewness	13.56285992	22.2738602	7.92758702	7.97367358	2.70297074	1.29628222

Will replace the null values with **Median as skewness are highly positive**

DS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
Mean	1.420782244	0.141819349	1.403664386	0.098332363
Mode	0	0	0	0
Median	0	0	0	0
Min	0	0	0	0
Max	28	6	28	5
SD	2.302085879	0.440539565	2.281781763	0.357263762
VAR	5.299599396	0.194075109	5.206528013	0.127637396
Skewness	2.525749911	3.865176954	2.530120243	4.460229402

Will replace the null values with **Median as skewness are highly positive**

Project 6

Bank Loan Case Study

Project Tasks

Identify and Handle Missing Data – Detect missing values and decide on the best approach (imputation or removal) to maintain data integrity.

Row Labels	Count of OCCUPATION_TYPE
	15654
Laborers	8952
Sales staff	5160
Core staff	4434
Managers	3489
Drivers	3044
High skill tech staff	1852
Accountants	1621
Medicine staff	1403
Security staff	1140
Cooking staff	963
Cleaning staff	739
Private service staff	447
Low-skill Laborers	357
Waiters/barmen staff	228
Secretaries	212
Realty agents	123
HR staff	101
IT staff	80
(blank)	
Grand Total	**49999**

We will choose **MODE** imputation as they are text columns

Imputation values

Laborers
Unaccompanied
No

Row Labels	Count of EMERGENCYSTATE_MODE
No	25944
	23698
Yes	357
(blank)	
Grand Total	**49999**

Row Labels	Count of NAME_TYPE_SUITE
Unaccompanied	40435
Family	6549
Spouse, partner	1849
Children	542
Other_B	259
	192
Other_A	137
Group of people	36
(blank)	
Grand Total	**49999**

DS	FLOORSMIN_AVG	FLOORSMAX_MODE	FLOORSMAX_MEDI	YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	TOTALAREA_MODE
Mean	0.23165	0.221489	0.225081	0.978036	0.977404	0.978031	0.102690027
Mode	0.2083	0.1667	0.1667	0.9871	0.9871	0.9871	0
Median	0.2083	0.1667	0.1667	0.9816	0.9816	0.9816	0.0685
Min	0	0	0	0	0	0	0
Max	1	1	1	1	1	1	1
SD	0.161545	0.144289	0.145574	0.056486	0.061657	0.057363	0.107950724
VAR	0.026097	0.020819	0.021192	0.003191	0.003802	0.00329	0.011653359
Skewness	0.964059	1.273558	1.265787	-16.21588	-15.42783	-16.23244	2.776809393

Will replace the null values with **Median as skewness are either positive or negative**

Project 6

Bank Loan Case Study

Project Tasks

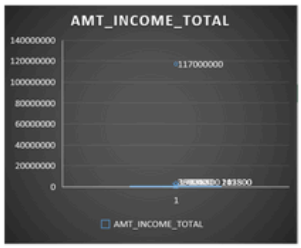
Identify Outliers – Find extreme values in numerical variables using statistical techniques to prevent data distortion.

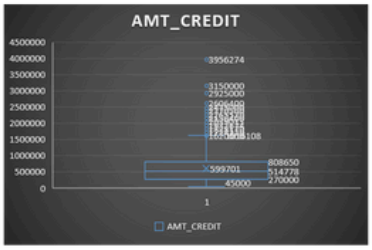
Setting the layout for outlier detection

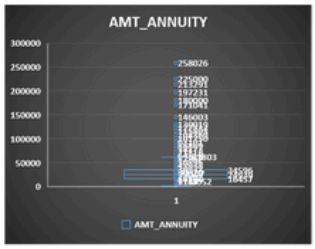
Application_data file

		AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_CHILDREN	YEARS_EMPLOYED
Quartile 1 = QUARTILE(B2:B50000,1)	Q1	112500	270000	16457	238500	0	3
Quartile 3 = QUARTILE(B2:B50000,3)	Q3	202500	808650	34596	679500	1	16
IQR = Quartile 3 -Quartile 1	IQR	90000	538650	18140	441000	1	13
Upper Limit = Quartile 3 + (1.5*IQR)	Upper Limit	337500	1616625	61805	1341000	3	36
Lower Limit = Quartile 1 – (1.5*IQR)	Lower Limit	-22500	-537975	-10753	-423000	-2	-17

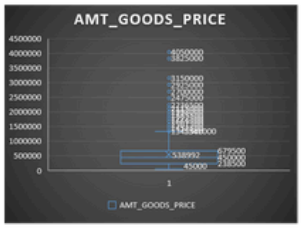
Outlier Detection

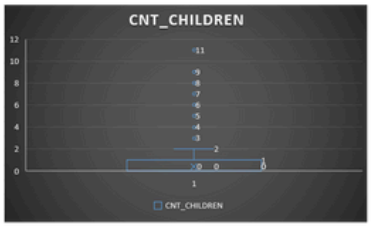


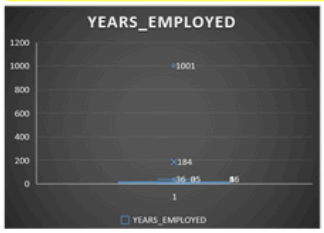




This column was additionally inserted








```
=COUNT(FILTER($B$2:$B$50000,($B$2:$B$50000>K8)+($B$2:$B$50000<K9)))
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_CHILDREN	YEARS_EMPLOYED
Outlier Count	2295	1063	1188	2387	723	9076

Project 6

Bank Loan Case Study

Project Tasks

Identify Outliers – Find extreme values in numerical variables using statistical techniques to prevent data distortion.

Outlier-handling Strategy

Column YEARS_EMPLOYED, we can see people being employed for 1000 yrs which is beyond human capacity. We will need to correct it.

Column CNT_CHILDREN shows people are having 11 children which is not impractical but rare in normal situations. This will require data validation again.

AMT_INCOME_TOTAL one of the extreme outlier is 117000000 but we will not remove it because income of people may have different figures and it could be a real case too.

AMT_CREDIT and AMT_INCOME_TOTAL where amount is higher than the entire usual trend. We need to verify it again.

We will not remove outlier from AMT_CREDIT too as it may be one of the actual cases.

AMT_GOODS_PRICE also shows outliers but we will first understand it from the clients and then decide on it.

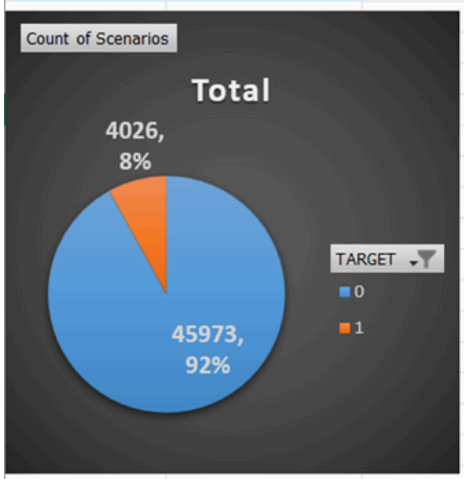
Project 6

Bank Loan Case Study

Project Tasks

Analyze Data Imbalance – Examine the distribution of loan default cases to assess if the dataset is skewed towards a particular class.

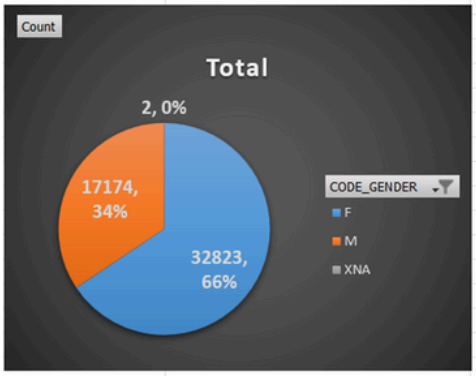
Target	Count of Scenarios
0	45973
1	4026
Grand Total	**49999**



Interpretation

Almost 92% don't fall under defaulters
8% are defaulters

Gender	Count
F	32823
M	17174
XNA	2
Grand Total	**49999**



Interpretation

Females are the major debtors up to 66%
(Maybe the schemes are more liked by females)
Males account for 34%
A negligible count is of XNA

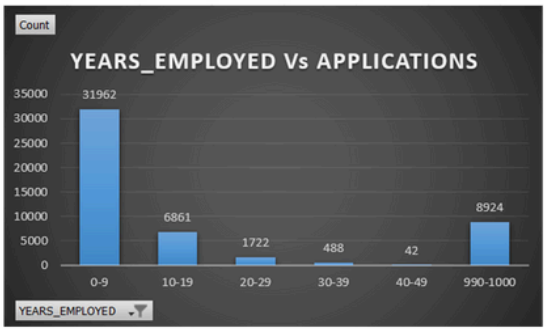
Project 6

Bank Loan Case Study

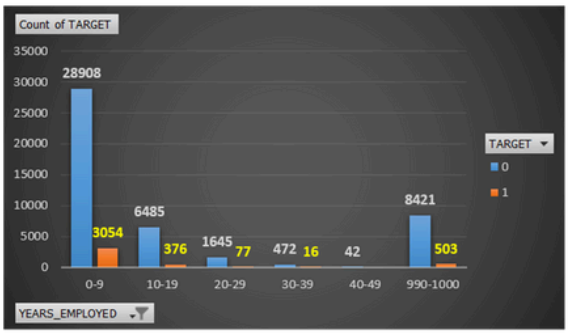
Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Duration of Employment of applicants Vs no. of applicants

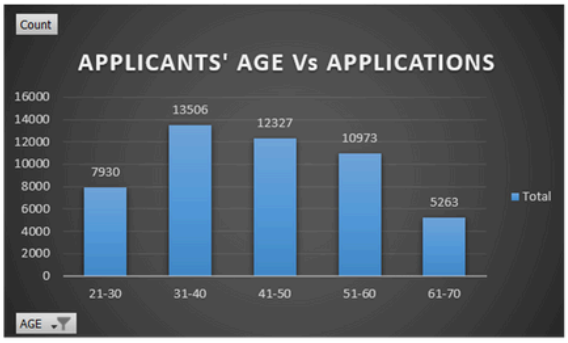


Duration of Employment of applicants Vs no. of applicants & Repayment trend

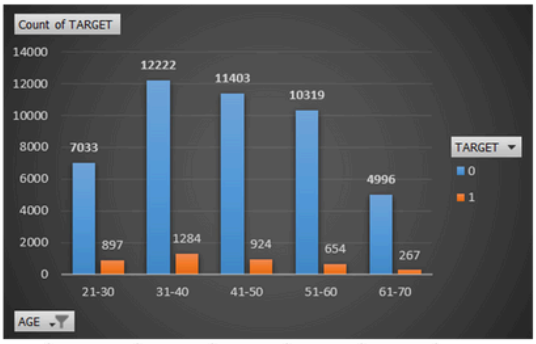


- Maximum applicants are from 0-9 years of employment range (brush off outliers here)
- As the employment age progresses, applicants have lower rate of defaulting on repayment

Applicant's age data Vs count



Duration of Employment of applicants Vs no. of applicants & Repayment trend



Majority of applicant's age is between 31-40
Applications decrease with age after 40
Repayment difficulty decreases with age

AGE Vs Repayment Trend				
AGE	0	1	0	1
21-30	7033	897	89%	11%
31-40	12222	1284	90%	10%
41-50	11403	924	93%	7%
51-60	10319	654	94%	6%
61-70	4996	267	95%	5%

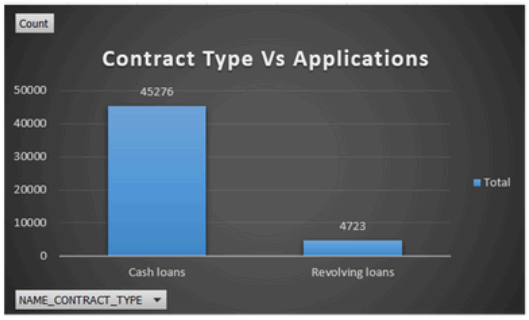
Project 6

Bank Loan Case Study

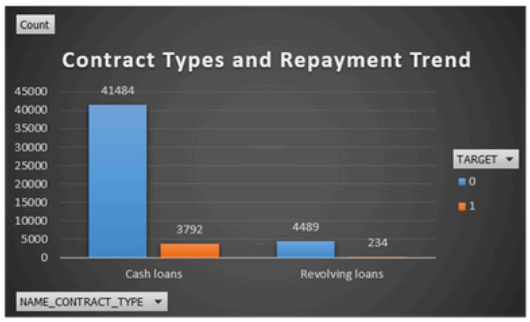
Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Types of loans and their frequency



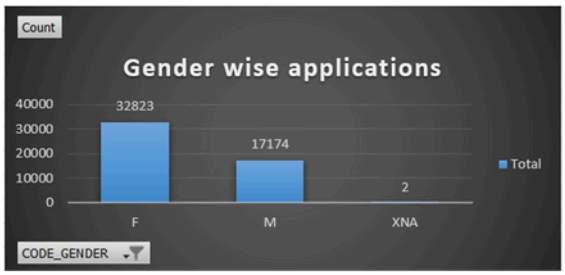
Types of loans and repayment trend



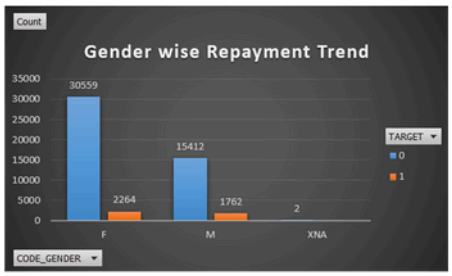
Contract Type	0	1	0	1
Cash loans	41484	3792	92%	8%
Revolving loans	4489	234	95%	5%

Cash loans are more in trend
Revolving loans have lesser default tendency, although the difference is huge between both

Gender wise application data



Gender wise application and repayment trend



Females are major applicants
Females have slightly better repayment record than males

Repayment Trend				
Gender	0	1	0	1
F	30559	2264	93%	7%
M	15412	1762	90%	10%
XNA	2			

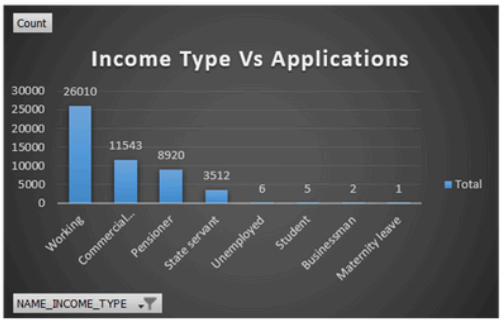
Project 6

Bank Loan Case Study

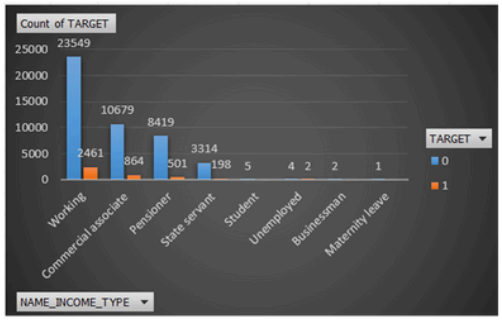
Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Income Type Vs Applications



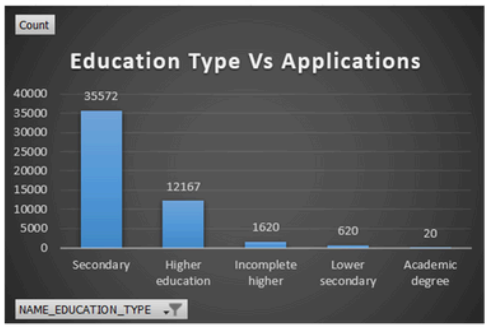
Income Type application and repayment trend



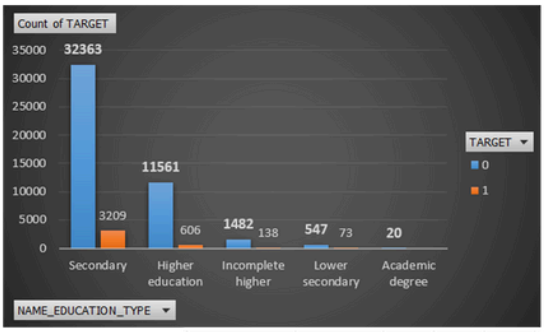
- The highest number of loan applications come from the working category followed by commercial associate and pensioner
- Students, businessmen, and maternity leave applicants have a 100% repayment rate
- Higher defaults among working applicants

Repayment Trend				
Income Type	0	1	0	1
Working	23549	2461	91%	9%
Commercial associate	10679	864	93%	7%
Pensioner	8419	501	94%	6%
State servant	3314	198	94%	6%
Student	5		100%	0%
Unemployed	4	2	67%	33%
Businessman	2		100%	0%
Maternity leave	1		100%	0%

Education Type Vs Applications



Education Type application and repayment trend



•Secondary education dominates loan applications
•Higher education borrowers show better repayment behavior, with a 95% repayment rate
•Those with an academic degree have a 100% repayment rate, but their numbers are extremely low (only 20 applicants), making it less impactful overall.

Repayment Trend				
Education Type	0	1	0	1
Secondary	32363	3209	266%	26%
Higher education	11561	606	95%	5%
Incomplete higher	1482	138	91%	9%
Lower secondary	547	73	88%	12%
Academic degree	20		100%	0%

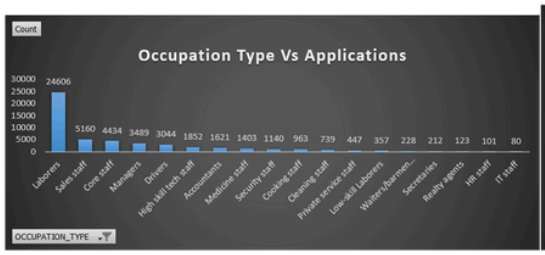
Project 6

Bank Loan Case Study

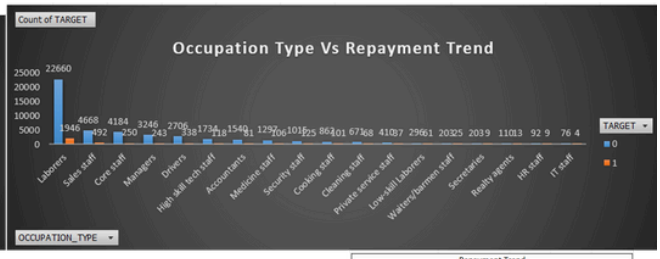
Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Occupation Type Vs Applications



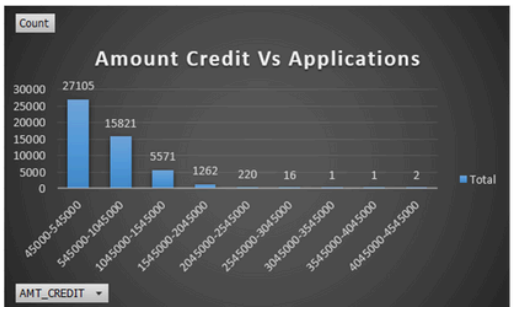
Occupation Type application and repayment trend



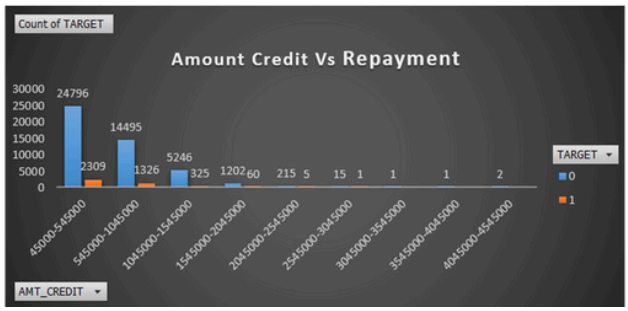
- Laborers form the largest group of borrowers (22,660) but have a higher default rate (8%), indicating financial instability in this segment.
- IT staff and accountants show the best repayment behavior, with 95% and 91% repayment rates, respectively, suggesting stable and well-paying jobs.
- Waiters/barmen and low-skill laborers have the highest default rates (11% and 17%), making them riskier borrower groups compared to others.

Repayment Trend				
Occupation Type	0	1	0	1
Laborers	22660	1946	92%	8%
Sales staff	4668	492	90%	10%
Core staff	4184	250	94%	6%
Managers	3246	243	93%	7%
Drivers	2706	338	89%	11%
High skill tech staff	1734	118	94%	6%
Accountants	1540	81	95%	5%
Medicine staff	1297	106	92%	8%
Security staff	1015	125	89%	11%
Cooking staff	862	101	90%	10%
Cleaning staff	671	68	91%	9%
Private service staff	410	37	92%	8%
Low-skill Laborers	296	61	83%	17%
Waiters/barmen staff	203	25	89%	11%
Secretaries	203	9	96%	4%
Realty agents	110	13	89%	11%
HR staff	92	9	91%	9%
IT staff	76	4	95%	5%

Amount Credit Vs Applications



Amount Credit Vs Applications



Project 6

Bank Loan Case Study

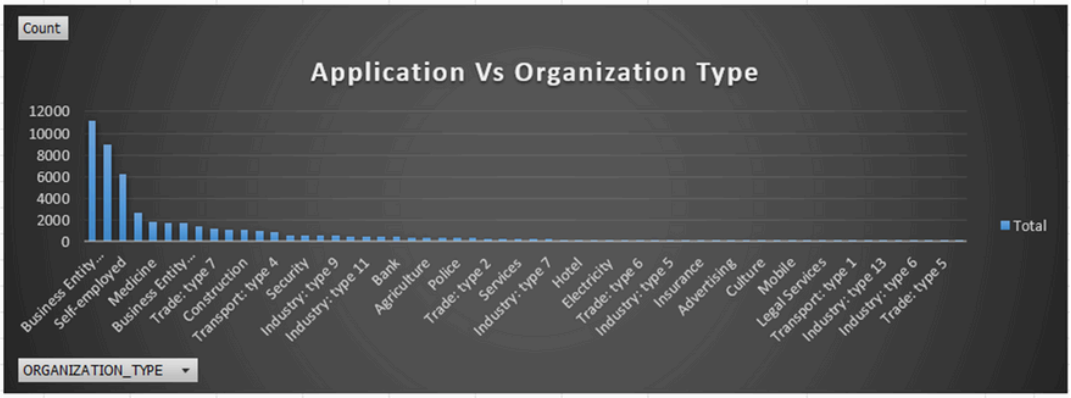
Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

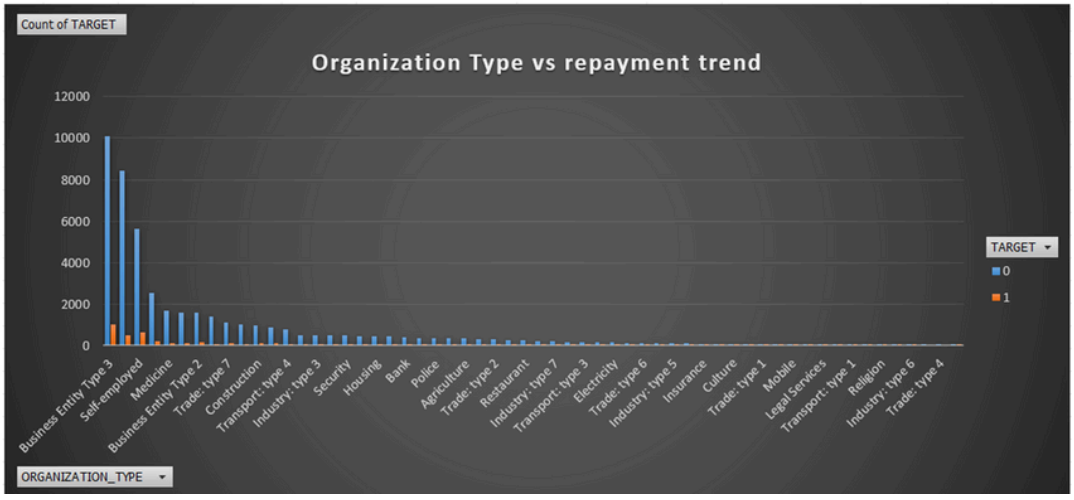
- Lower credit amounts (45,000 - 5,45,000) have the highest number of borrowers but also show a higher default rate (9%), indicating financial strain on lower-credit borrowers.
- As the credit amount increases, the repayment rate improves, with loans between 10,45,000 - 20,45,000 showing a default rate of only 5-6%, suggesting that higher-credit borrowers are more financially stable.
- Loans above 30,45,000 have a 100% repayment rate, showing that high-credit borrowers are highly reliable in loan repayments.

	Repayment			
Amount Credit	0	1	0	1
45000-545000	24796	2309	91%	9%
545000-1045000	14495	1326	92%	8%
1045000-1545000	5246	325	94%	6%
1545000-2045000	1202	60	95%	5%
2045000-2545000	215	5	98%	2%
2545000-3045000	15	1	94%	6%
3045000-3545000	1		100%	0%
3545000-4045000	1		100%	0%
4045000-4545000	2		100%	0%

Organization Type Vs Applications



Organization Type Vs Repayment Trend



Project 6

Bank Loan Case Study

Key insights

Government and stable institutions (like Schools, Police, Universities) show the highest repayment rates (95%+), indicating that employees in these sectors have more financial stability.

Self-employed individuals, construction workers, and trade professionals show lower repayment rates (around 90%) with higher default risks (10%+), highlighting financial unpredictability in these sectors.

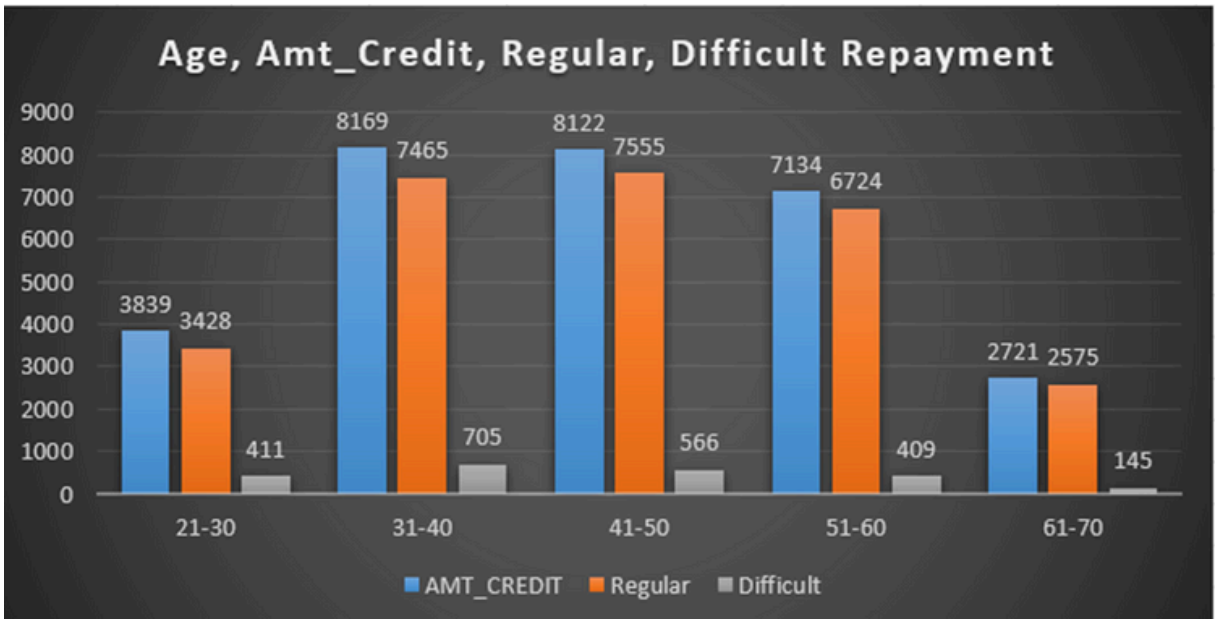
Certain industries like Agriculture (87%), Transport (87%), and Industry type 13 (73%) have the highest default rates, signaling financial instability and potential risk for lenders.

Project 6

Bank Loan Case Study

Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.



Key Insights

Credit amounts peak in the 31-50 age range, aligning with prime working years.

Regular repayment trends follow a similar pattern, meaning most loans are repaid on time.

Difficult repayment cases are higher among younger(31-50) borrowers, possibly due to unstable income sources.

Older borrowers (61-70) take lower credit and have fewer repayment issues, likely because they are more financially stable or take conservative loans.

Project 6

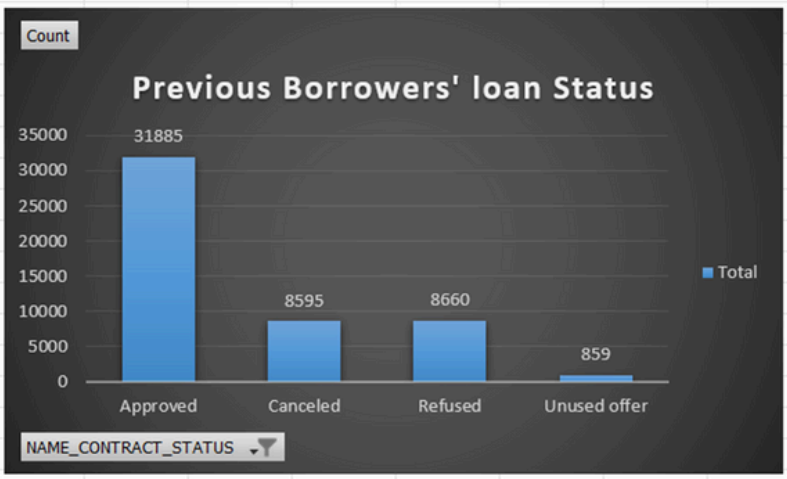
Bank Loan Case Study

Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Previous Applications' loan status

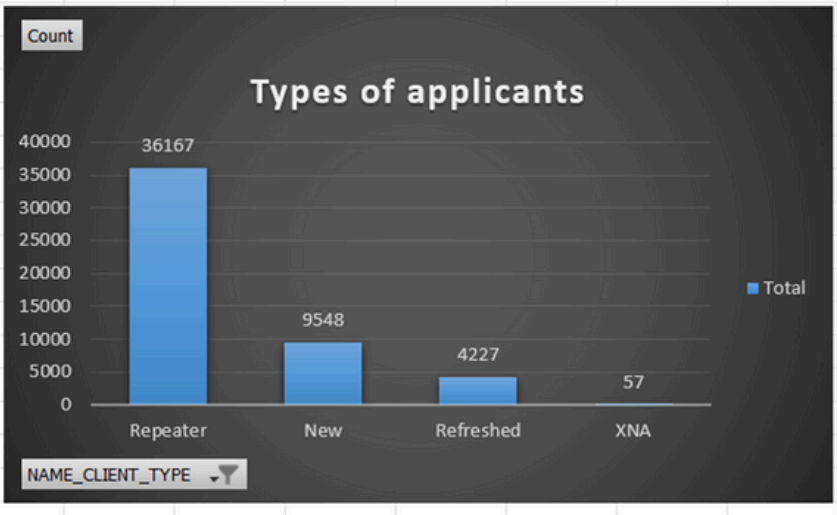
Previous Borrowers	Count
Approved	31885
Canceled	8595
Refused	8660
Unused offer	859
Grand Total	**49999**



Higher number of previous applicants' loans have been approved
859 have not used their loans
8660 have refused to borrow

Previous Applications' loan status

Client Type	Count
Repeater	36167
New	9548
Refreshed	4227
XNA	57
Grand Total	**49999**



Majority of the borrowers are repeaters. Focusing on old borrowers should be focused in future business.

Project 6

Bank Loan Case Study

Project Tasks

Perform Univariate, Segmented Univariate, and Bivariate Analysis – Explore how individual and combined attributes impact loan repayment behavior.

Here is how the application amounts were processed in previous applications

Applied Amount - Credit Amount	Count
Approved more than applied	19546
Approved less than applied	11418
Refused	9600
Not Applied	9435
Total	49999

It shows that majority the borrowers got more credit than they applied for.

Project 6

Bank Loan Case Study

Project Tasks

Identify Top Correlations – Determine the strongest relationships between customer attributes and loan default risk for better decision-making.

Regular Repayment Correlation - 0		
AMT_CREDIT	AMT_ANNUITY	0.770772818
AMT_CREDIT	AMT_GOODS_PRICE	0.986999774
NAME_INCOME_TYPE	YEARS_EMPLOYED	0.797293628
CNT_CHILDREN	CNT_FAM_MEMBERS	0.879238049
AMT_ANNUITY	AMT_GOODS_PRICE	0.775835204
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950468157
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861374946
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825358079
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	0.973531781
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.994674497
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	0.966738323
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.98958882
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998357563
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.850995792

EMPLOYMENT STABILITY MATTERS – "YEARS EMPLOYED" HAS A STRONG CORRELATION (0.7973) WITH REPAYMENT, INDICATING THAT STABLE EMPLOYMENT LEADS TO BETTER REPAYMENT BEHAVIOR.

HIGHER CREDIT AMOUNTS ALIGN WITH HIGHER ANNUITIES – "AMT_CREDIT" AND "AMT_ANNUITY" HAVE A HIGH CORRELATION (0.7708), SUGGESTING THAT LARGER LOAN AMOUNTS CORRESPOND WITH HIGHER INSTALLMENT PAYMENTS.

SOCIAL CIRCLE AND DEFAULT MONITORING ARE STRONGLY LINKED – "OBS_30_CNT_SOCIAL_CIRCLE" AND "OBS_60_CNT_SOCIAL_CIRCLE" SHOW EXTREMELY HIGH CORRELATION (0.9984), MEANING PEOPLE WHO DEFAULT IN ONE SHORT-TERM PERIOD ARE HIGHLY LIKELY TO DEFAULT IN ANOTHER.

PROPERTY AGE AND LOAN REPAYMENT CONNECTION – THE STRONG CORRELATION OF "YEARS_BEGINEXPLUATATION_AVG" (0.9947) WITH OTHER EXPLOITATION-RELATED VARIABLES SUGGESTS THAT PROPERTY AGE OR OWNERSHIP HISTORY SIGNIFICANTLY IMPACTS LOAN REPAYMENT BEHAVIOR.

Project 6

Bank Loan Case Study

Project Tasks

Identify Top Correlations – Determine the strongest relationships between customer attributes and loan default risk for better decision-making.

Difficulty Repayment Correlation - 1		
AMT_CREDIT	AMT_ANNUITY	0.749665201
AMT_CREDIT	AMT_GOODS_PRICE	0.982267963
CNT_CHILDREN	CNT_FAM_MEMBERS	0.892521875
AMT_ANNUITY	AMT_GOODS_PRICE	0.74950403
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950768899
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.806743886
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.783754676
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	0.969745206
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.983626828
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	0.979592562
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.989772825
FLOORSMAX_AVG	FLOORSMAX_MODE2	0.987677718
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998065853
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.89051161

Loan Amount and Goods Price Are Highly Correlated – "AMT_CREDIT" and "AMT_GOODS_PRICE" have an extremely high correlation (0.9823), suggesting that borrowers struggling with repayment often take loans close to the exact value of their purchased goods, leaving little financial flexibility.

Larger Families Struggle More – "CNT_CHILDREN" and "CNT_FAM_MEMBERS" show a strong correlation (0.8925), indicating that borrowers with larger families might face financial strain, making repayment more difficult.

Living and Working in Different Regions Affects Repayment – The correlation between "REG_REGION_NOT_WORK_REGION" and "LIVE_REGION_NOT_WORK_REGION" (0.8067) suggests that people who live far from their workplace may have increased financial burdens, impacting their ability to repay loans.

Property Characteristics Strongly Influence Repayment – The correlation of "YEARS_BEGINEXPLUATATION_AVG" with different property condition measures (ranging from 0.9697 to 0.9836) indicates that property age and condition significantly impact repayment behavior, possibly because older or lower-value properties are linked to financial instability.

Social Default Risk Is Highly Predictable – "OBS_30_CNT_SOCIAL_CIRCLE" and "OBS_60_CNT_SOCIAL_CIRCLE" have an almost perfect correlation (0.9980), meaning that borrowers who struggle with repayment in the short term are almost certain to continue struggling in the longer term. Additionally, "DEF_30_CNT_SOCIAL_CIRCLE" and "DEF_60_CNT_SOCIAL_CIRCLE" (0.8905) reinforce this pattern, showing that social circles with defaulters are a strong indicator of financial distress.

Project 6

Bank Loan Case Study

Key insights of the project

- The majority of clients take cash loans. Most clients are loan repayers rather than defaulters.
- The bank lends more to women than men, but women have a lower default rate compared to men.
- Clients with higher education levels are less likely to default compared to those with lower education, such as those with only secondary special education.
- The bank should prioritize lending to clients with higher educational qualifications.
- As age and experience increase, the likelihood of default decreases.
- Older clients tend to borrow larger loan amounts, but their default rate is lower, making them less risky and more profitable for the bank.
- As the number of children increases, the number of clients taking loans decreases.
- The bank should exercise extra caution when lending to unemployed clients, as they have the highest default rate and take larger loan amounts.

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Description

This project explores how various car features influence pricing and profitability in the automotive industry. With evolving consumer preferences and increasing competition, car manufacturers need data-driven strategies to optimize pricing and product development.

Using data analysis techniques such as regression analysis and market segmentation, the project examines the relationship between car features (like horsepower, fuel type, body style) and their market value. The goal is to identify the most popular and profitable car attributes, enabling manufacturers to balance consumer demand with profitability.

By leveraging insights from the analysis, manufacturers can make informed pricing decisions, enhance product development, and maintain a competitive edge in the market.

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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

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

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

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

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

Building the Dashboard:

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

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

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

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

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

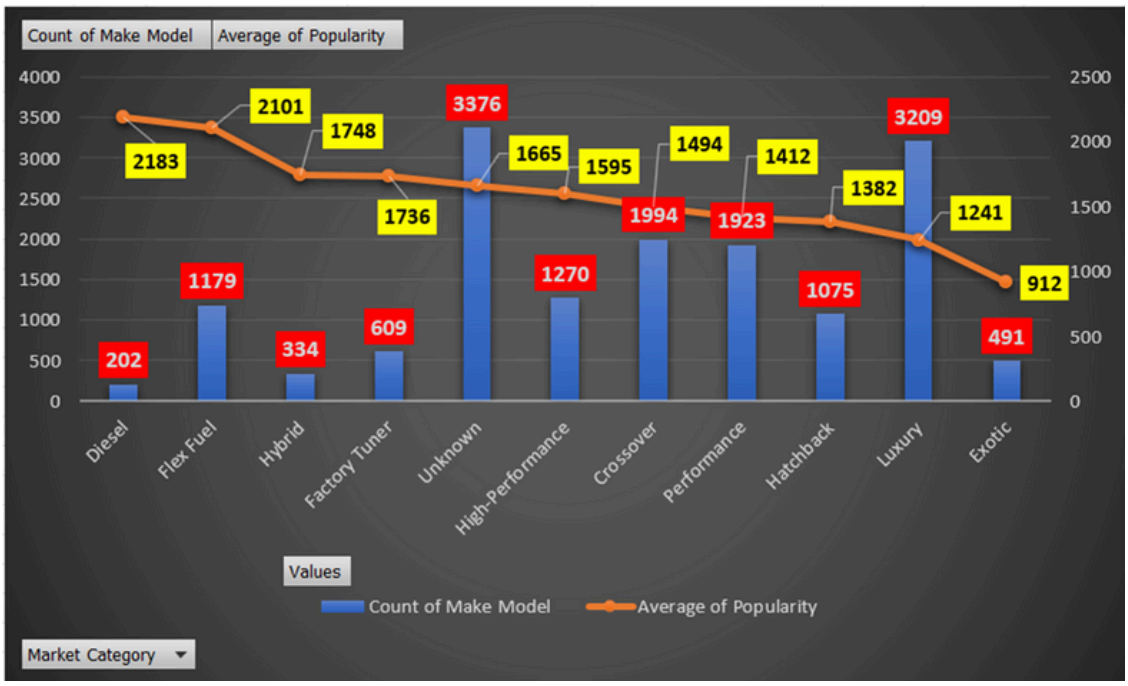
Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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

Market Category	Count of Make Model	Average of Popularity
Diesel	202	2183
Flex Fuel	1179	2101
Hybrid	334	1748
Factory Tuner	609	1736
Unknown	3376	1665
High-Performance	1270	1595
Crossover	1994	1494
Performance	1923	1412
Hatchback	1075	1382
Luxury	3209	1241
Exotic	491	912
Grand Total	**15662**	**1521**



Project 7

Analyzing the Impact of Car Features on Price and Profitability

Key insights

- "Unknown" category has the highest count (3376) but only moderate popularity (1665).
- Luxury cars are widely available (3209) but have a declining popularity score (1382).
- Exotic cars are rare (491) and have the lowest popularity (912).
- High-Performance cars (1270) are more popular (1494) than Performance cars (1923, 1412).
- Hybrid cars (334) are fewer but more popular (2101) than Flex Fuel cars (1179, 1748).

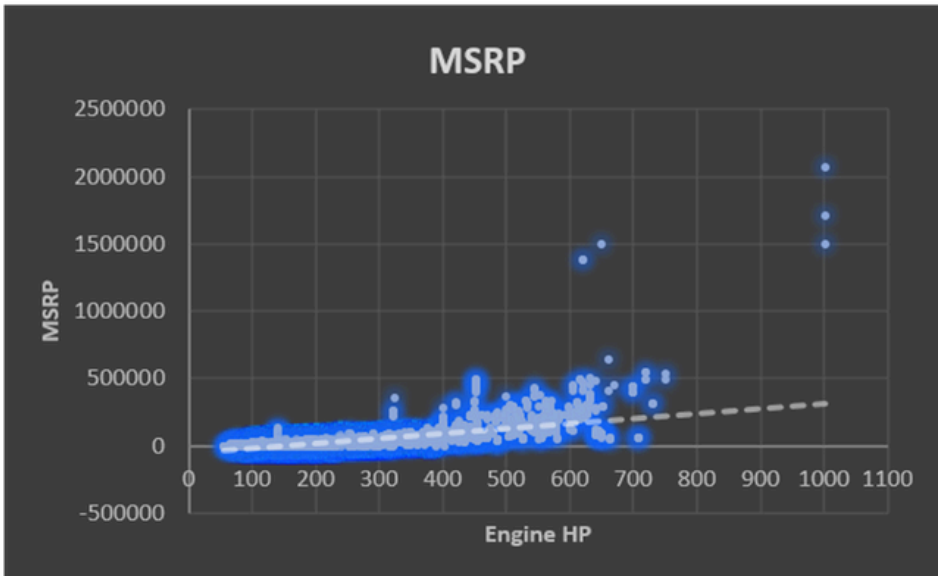
Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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

Scatter chart
- **Engine power on the x-axis**
- **MSRP on the y-axis**



Key insights

Positive Correlation – Higher engine horsepower (HP) generally leads to a higher MSRP, as indicated by the upward trend in the dotted trendline.

Luxury & Supercars Outliers – A few vehicles with very high HP (>800) have significantly higher MSRPs, likely representing exotic or luxury cars.

Majority Cluster in Lower Range – Most cars are concentrated below 500 HP and under $100,000, suggesting that mainstream vehicles dominate the dataset.

Non-Linear Pricing – While HP increases gradually, MSRP shows sharp jumps, indicating that luxury branding, exclusivity, and features also impact pricing.

Some Affordable High-HP Cars – There are instances where cars with relatively high HP (~500-600) still have moderate MSRPs, possibly due to performance-focused mainstream models.

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

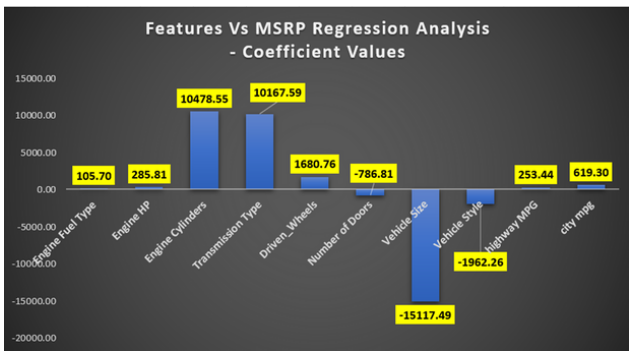
Which car features are most important in determining a car's price?

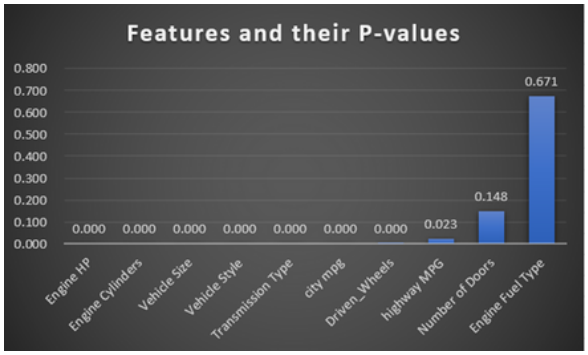
Regression Statistics	
Multiple R	0.704965493
R Square	0.496976346
Adjusted R Square	0.496526736
Standard Error	43662.71619
Observations	11199

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	2.10728E+13	2.10728E+12	1105.34988	0
Residual	11188	2.13292E+13	1906432785		
Total	11198	4.24019E+13			

Factors	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-90283.6972	4881.954398	-18.49335119	3.07217E-75	-99853.18726	-80714.20714	-99853.18726	-80714.20714
Engine Fuel Type	105.7029477	249.2023898	0.424165064	0.671453593	-382.777607	594.1835023	-382.777607	594.1835023
Engine HP	285.80656	6.664692006	42.88368612	0	272.7425904	298.8705297	272.7425904	298.8705297
Engine Cylinders	10478.55251	457.5844774	22.89971147	1.8535E-113	9581.606375	11375.49864	9581.606375	11375.49864
Transmission Type	10167.58652	878.4914764	11.57391596	8.36638E-31	8445.588572	11889.58447	8445.588572	11889.58447
Driven_Wheels	1680.76274	449.317091	3.740705113	0.000184421	800.0221417	2561.503338	800.0221417	2561.503338
Number of Doors	-786.8117167	543.2036171	-1.448465533	0.147514923	-1851.586434	277.9630008	-1851.586434	277.9630008
Vehicle Size	-15117.49134	663.5696129	-22.78207297	2.427E-112	-16418.2046	-13816.77809	-16418.2046	-13816.77809
Vehicle Style	-1962.256302	142.03293	-13.81550251	4.6353E-43	-2240.665849	-1683.846755	-2240.665849	-1683.846755
highway MPG	253.4435608	111.1624762	2.279938065	0.022630095	35.54553796	471.3415836	35.54553796	471.3415836
city mpg	619.2986168	101.7405952	6.087035522	1.1879E-09	419.8691394	818.7280942	419.8691394	818.7280942





Project 7

Analyzing the Impact of Car Features on Price and Profitability

Key insights

A high coefficient does not always mean a feature is important unless it has a low p-value.

Look for both high coefficients and low p-values when identifying the most important features.

Most impactful and statistically significant features:
Engine HP (Coefficient: 285.81, P-value: 0.000) → Strong, significant impact.

Engine Cylinders (Coefficient: 10,478.55, P-value: 1.85E-113) → Very significant.

Transmission Type (Coefficient: 10,167.59, P-value: 8.36E-31) → Strong impact and highly significant.

Vehicle Size (Coefficient: -15,117.49, P-value: 2.43E-112) → Strong negative impact and highly significant.

Top 4 most important features:

Vehicle Size
Engine Cylinders
Transmission Type
Engine HP.

We need to avoid using Number of Doors & Engine Fuel Type as key predictors since their impact is not statistically significant.

City & Highway MPG affect MSRP but not as much as engine and transmission-related features.

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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

Manufacturer	Avg MSRP
Bugatti	1757224
Maybach	546222
Rolls-Royce	351131
Lamborghini	331567
Bentley	247169
McLaren	239805
Ferrari	238219
Spyker	214990
Aston Martin	198123
Maserati	113684
Porsche	101622
Tesla	85256
Mercedes-Benz	72070
Lotus	68377
Land Rover	68067
BMW	62163
Alfa Romeo	61600
Cadillac	56368
Audi	54574
Lexus	47549

Manufacturer	Avg MSRP
Genesis	46617
Lincoln	43861
Infiniti	42640
HUMMER	36464
Acura	35087
GMC	32444
Volvo	29725
Chevrolet	29075
Buick	29034
Volkswagen	28979
Nissan	28921
Toyota	28847
Ford	28511
Saab	27880
Chrysler	26723
Honda	26655
Kia	25514
Hyundai	24926
Dodge	24857
Subaru	24241

Manufacturer	Avg MSRP
FIAT	22670
Mitsubishi	21341
Mazda	20417
Scion	19933
Pontiac	19800
Suzuki	18026
Oldsmobile	12844
Plymouth	3297

MRSP wise Top 10 Manufacturers

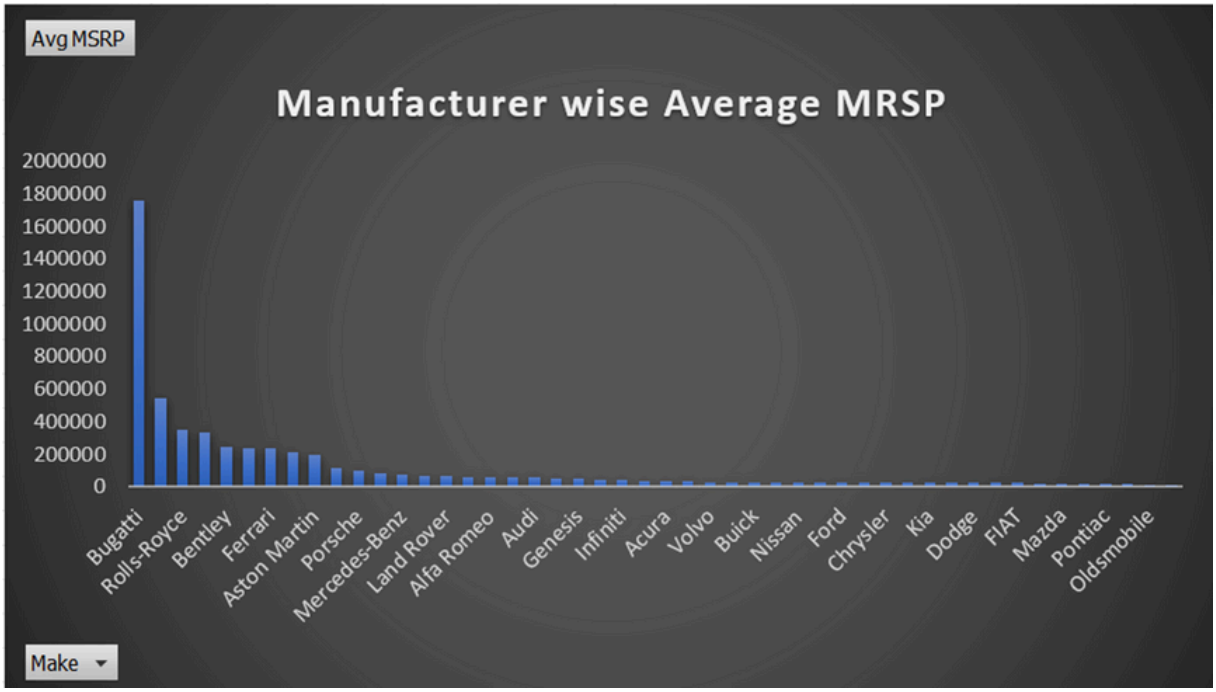
Manufacturer	Avg MSRP	Rank
Bugatti	1757224	1
Maybach	546222	2
Rolls-Royce	351131	3
Lamborghini	331567	4
Bentley	247169	5
McLaren	239805	6
Ferrari	238219	7
Spyker	214990	8
Aston Martin	198123	9
Maserati	113684	10
Porsche	101622	11
Tesla	85256	12
Mercedes-Benz	72070	13
Lotus	68377	14
Land Rover	68067	15

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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

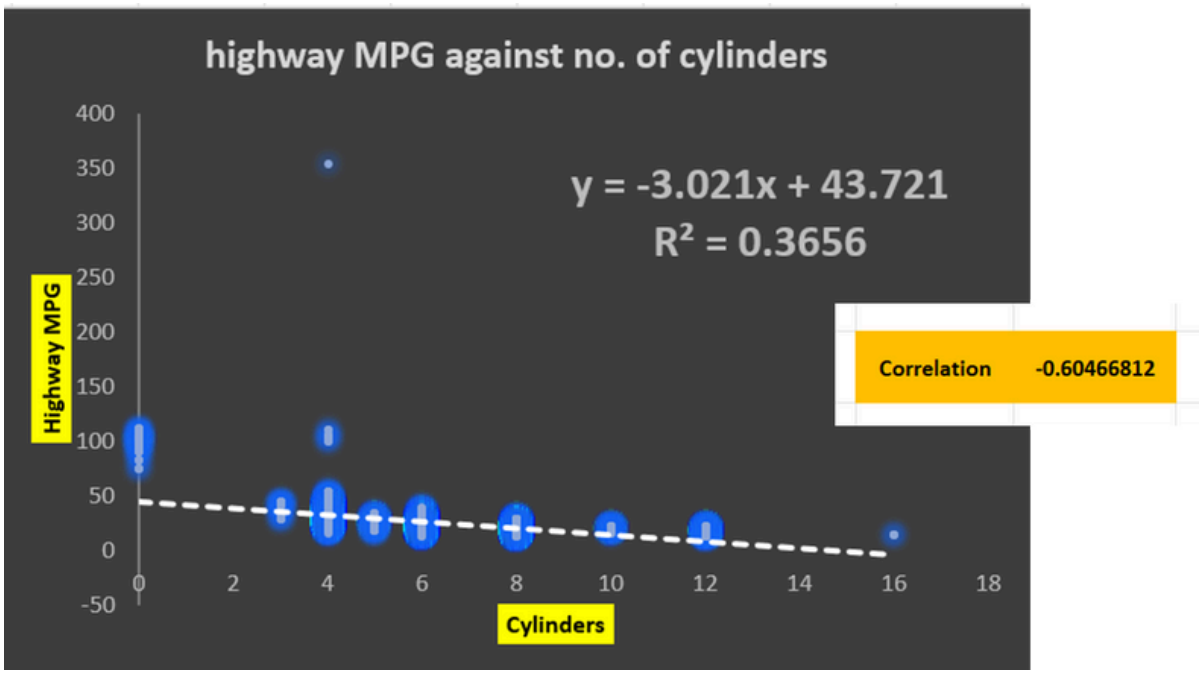


Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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



Insights from the Scatter Plot

Negative Slope (-3.021) - The equation y = -3.021x + 43.721 suggests that for each additional cylinder, the highway MPG decreases by approximately 3.021 units, reinforcing that higher-cylinder engines are generally less fuel-efficient.

Intercept (43.721) - When the number of cylinders is zero (hypothetically), the predicted highway MPG would be 43.721, though this is not realistic for real-world vehicles. (Although 0-cylinder vehicles are Electric Vehicles)

R^2 Value (0.3656) - Weak Correlation - The R^2 = 0.3656 indicates that only 36.56% of the variation in highway MPG is explained by the number of cylinders. This means other factors (like vehicle weight, aerodynamics, fuel type) may also significantly influence MPG, beyond just the number of cylinders.

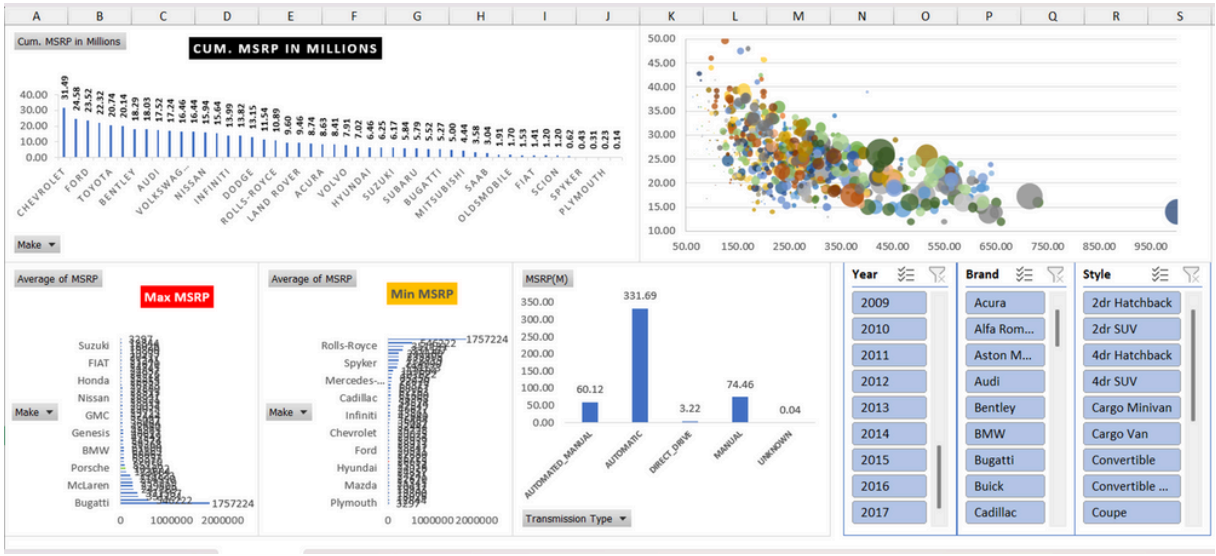
Correlation value of -0.604 indicates that as the cylinders increase in number the milage takes a backseat.

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Project Tasks

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



Building the Dashboard:

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

Stacked Bar - Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

Bar Chart - How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

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

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

Project 7

Analyzing the Impact of Car Features on Price and Profitability

Key insights of the project

- Higher engine HP strongly correlates with higher MSRP, but luxury branding and exclusivity also impact pricing.
- Vehicle Size, Engine Cylinders, Transmission Type, and Engine HP are the top four factors influencing MSRP.
- Most cars are below 500 HP and $100,000, while luxury and exotic cars create sharp price jumps.
- Each additional engine cylinder decreases highway MPG by ~3.021 units, confirming lower fuel efficiency.
- Cylinder count weakly explains MPG variation (R^2 = 0.3656), suggesting other factors like weight and aerodynamics matter.

Project 8

ABC Call Volume Trend Analysis

Project Description

- This project focuses on Customer Experience (CX) analytics, specifically analyzing inbound call volume trends in an insurance company's call center.
- The dataset provided spans 23 days and contains details like agent ID, queue time, call time, call duration, and call status (answered, abandoned, or transferred).
- The objective is to analyze call trends, agent workload, and manpower planning to optimize customer service and reduce abandoned calls.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Average Call Duration Analysis

Calculate the average call duration for each time bucket.

Call Volume Analysis

Create charts/graphs to visualize the total number of calls received per time bucket.

Manpower Planning (Day Shift: 9 AM - 9 PM)

Determine the minimum number of agents required per time bucket to reduce the abandoned call rate from 30% to 10%.

Night Shift Manpower Planning (9 PM - 9 AM)

Since 30% of the day's call volume occurs at night, propose a manpower plan to ensure a maximum 10% abandon rate at night.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Average Call Duration Analysis

```
avg_call_duration = df.groupby('time_bucket')['call_seconds_s'].mean()
avg_call_duration = avg_call_duration.astype(int).sort_values(ascending = True)
avg_call_duration.sort_values(ascending = True)
```

time_bucket	call_seconds_s
9_10	92
10_11	97
20_21	105
11_12	116
12_13	144
19_20	144
14_15	146
13_14	149
15_16	169
18_19	174
17_18	179
16_17	181

Project 8

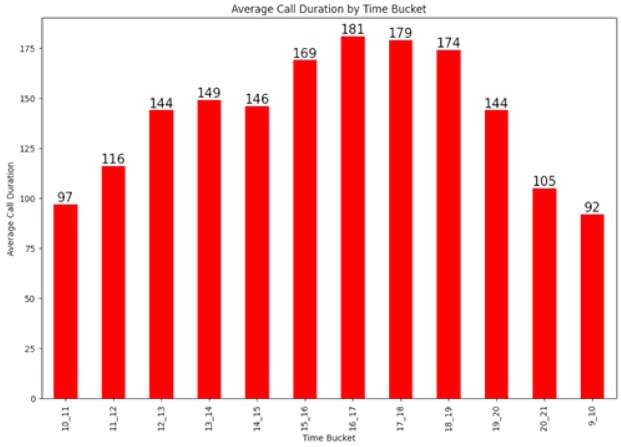
ABC Call Volume Trend Analysis

Project Tasks

Average Call Duration Analysis

Creating the bar chart for the above data

```
plt.figure(figsize = (12,8))
x = avg_call_duration.plot(kind='bar', color='Red')
for container in x.containers:
  x.bar_label(container, fontsize = 10)
plt.xlabel('Time Bucket')
plt.ylabel('Average Call Duration')
plt.title('Average Call Duration by Time Bucket')
plt.show()
```



Insights from Task Call Duration Analysis

Peak Call Duration Timing: The highest average call duration is observed between 4:00 PM and 5:00 PM, indicating that callers prefer detailed discussions during this slot.

Afternoon & Evening Preference: Time slots 3-4 PM, 4-5 PM, 5-6 PM, and 6-7 PM show maximum average call duration, suggesting that more staff is required during these hours.

Lower Call Duration in the Morning: The 9-10 AM time slot has the lowest average call duration, meaning calls during this period might be for quick queries rather than detailed discussions.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

1Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time.

Time should be represented in buckets.

```
no_of_calls = df.groupby('time_bucket')['call_seconds_s'].count().sort_values(ascending = True)
no_of_calls
```

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

dtype: int64

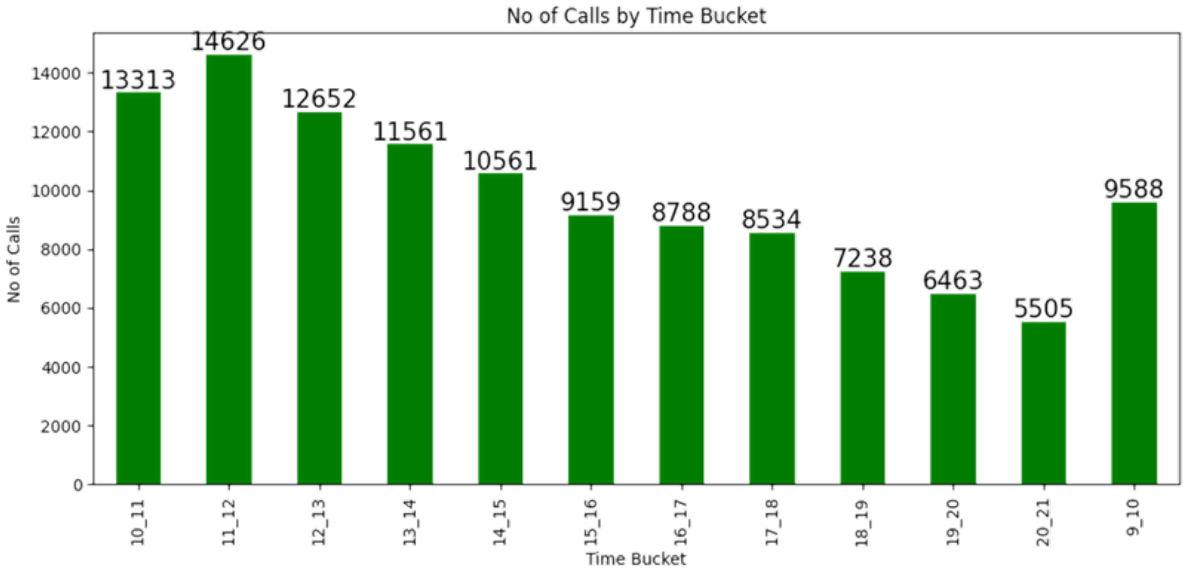
Project 8

ABC Call Volume Trend Analysis

Project Tasks

Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time.

Time should be represented in buckets.



Insights from Call Volume Analysis

- Highest Call Volume Timing: The 11 AM to 12 PM slot receives the most calls, but the average call duration is relatively lower, indicating more quick queries.
- Early Morning Call Trends: The 9-10 AM slot has significant call volume but with a lower duration, suggesting it's a secondary peak for quick customer inquiries.
- Lowest Call Volume: The 8-9 PM slot receives the least number of calls, possibly due to customer preference for resolving issues earlier in the day.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Manpower Planning:

The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%.

In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.

Assumptions: An agent works for 6 days a week; On average, each agent takes 4 unplanned leaves per month; An agent's total working hours are 9 hours, out of which 1.5 hours are spent on lunch and snacks in the office.

On average, an agent spends 60% of their total actual working hours (i.e., 60% of 7.5 hours) on calls with customers/users. The total number of days in a month is 30.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Manpower Planning

```
[120] manpower = round(df.groupby('time_bucket')['call_seconds_s'].sum()/3600/4.5, 0)
      print(manpower)

      time_bucket
      10_11     80.0
      11_12    105.0
      12_13    113.0
      13_14    107.0
      14_15     96.0
      15_16     96.0
      16_17     98.0
      17_18     95.0
      18_19     78.0
      19_20     58.0
      20_21     36.0
      9_10      54.0
      Name: call_seconds_s, dtype: float64
      1016.0
```

time bucket wise manpower for 23 days and futher calculations

```
[121] daily_manpower = round(manpower / 23, 0) # for a single day at 30% call abandon rate
      print(daily_manpower)
      req_manpower = round((daily_manpower * 90)/70, 0) # requirement of manpower for 10% call abandon rate
      print(req_manpower)
      total_req_manpower = req_manpower.sum()
      print(total_req_manpower)
```

```
time_bucket
10_11    3.0
11_12    5.0
12_13    5.0
13_14    5.0
14_15    4.0
15_16    4.0
16_17    4.0
17_18    4.0
18_19    3.0
19_20    3.0
20_21    2.0
9_10     2.0
Name: call_seconds_s, dtype: float64
44.0
```

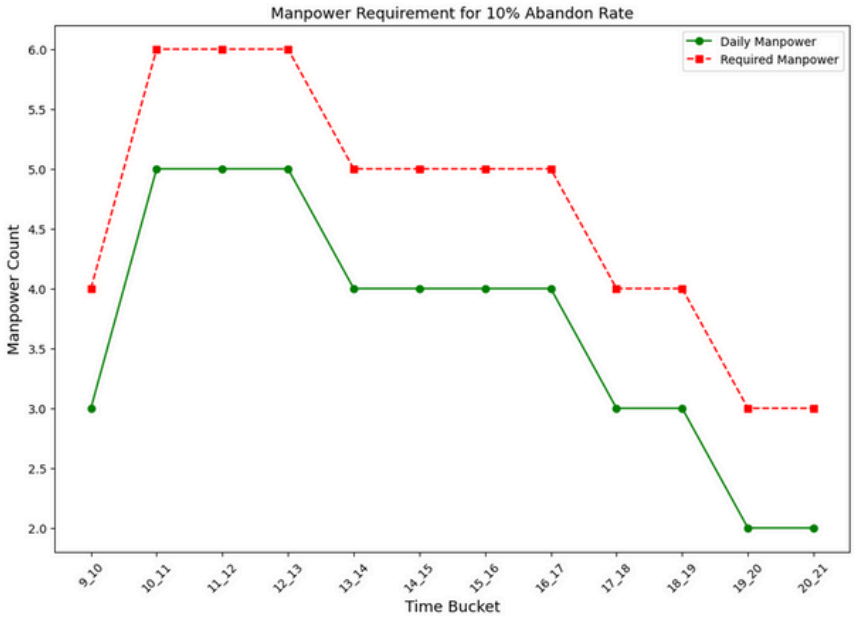
```
time_bucket
10_11    4.0
11_12    6.0
12_13    6.0
13_14    6.0
14_15    5.0
15_16    5.0
16_17    5.0
17_18    5.0
18_19    4.0
19_20    4.0
20_21    3.0
9_10     3.0
Name: call_seconds_s, dtype: float64
56.0
```

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Manpower Planning



Current manpower with 30% call abandon rate = **44**

Manpower required to reduce it to 10% = **56**

Insights from Manpower Planning

- Current vs. Required Manpower: The existing workforce of 44 agents leads to a 30% call abandonment rate, while 56 agents are needed to reduce the abandonment rate to 10%.
- Agent Efficiency Considerations: Each agent effectively works 4.5 hours daily on calls, accounting for breaks and other activities.
- Staffing Adjustments Needed: Increasing the workforce by 12 agents would ensure that at least 90 out of every 100 calls are answered, improving customer satisfaction.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Night Shift Manpower Planning

Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience.

Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am.

The distribution of these 30 calls is as follows:

Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

Creating a dataframe with nightshift time_bucket and call distributions ratio

Creating a dataframe for calculation of night shift

```
nightshift = { 'shift' : ['21_22', '22_23', '23_24', '0_1', '1_2', '2_3', '3_4', '4_5', '5_6', '6_7', '7_8', '8_9'],
               'call_distribution' : [3, 3, 2, 2, 1, 1, 1, 1, 3, 4, 4, 5]
             }
nshiftdf = pd.DataFrame(nightshift)
nshiftdf
```

	shift	call_distribution
0	21_22	3
1	22_23	3
2	23_24	2
3	0_1	2
4	1_2	1
5	2_3	1
6	3_4	1
7	4_5	1
8	5_6	3
9	6_7	4
10	7_8	4
11	8_9	5

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Night Shift Manpower Planning

Calculating the night calls as per the 30% of day calls quantum

```
[86] dailycallseconds = (df['call_seconds_s'].sum() / 23)
     nightcallseconds = (dailycallseconds * 0.30).astype(int)
     nightcallseconds
```

```
np.int64(214736)
```

Calculating the night call_seconds, creating a new columnn and calculating the regualar manpower

```
[87] nshiftdf['nightcallseconds'] = (nshiftdf['call_distribution'] * nightcallseconds) / 30
```

```
[137] nshiftdf['nightcallseconds'] = nshiftdf['nightcallseconds'].astype(int)
```

Calculating the manpower for the calculated callseconds

```
nshiftdf['nmanpower'] = round(((nshiftdf['nightcallseconds'] / 3600) / 4.5), 0)
nshiftdf
```

Replacing less than 1 with 1 as there are calls in the time_buckets and they cannot unresponded

```
nshiftdf['nmanpower'] = nshiftdf['nmanpower'].transform(lambda x: 1 if x < 1 else x)
nshiftdf
```

	shift	call_distribution	nightcallseconds	nmanpower
0	21_22	3	21473	1.0
1	22_23	3	21473	1.0
2	23_24	2	14315	1.0
3	0_1	2	14315	1.0
4	1_2	1	7157	0.0
5	2_3	1	7157	0.0
6	3_4	1	7157	0.0
7	4_5	1	7157	0.0
8	5_6	3	21473	1.0
9	6_7	4	28631	2.0
10	7_8	4	28631	2.0

Before = 9

After = 13

	shift	call_distribution	nightcallseconds	nmanpower
0	21_22	3	21473	1.0
1	22_23	3	21473	1.0
2	23_24	2	14315	1.0
3	0_1	2	14315	1.0
4	1_2	1	7157	1.0
5	2_3	1	7157	1.0
6	3_4	1	7157	1.0
7	4_5	1	7157	1.0
8	5_6	3	21473	1.0
9	6_7	4	28631	2.0
10	7_8	4	28631	2.0

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Night Shift Manpower Planning

Calculating the required night shift manpower for 10% abandon rate

```
rmanpower = round((nshiftdf['nmanpower'] * 90 / 70), 0)
rmanpower.sum()
```

```
np.float64(18.0)
```

Inserting the required manpower column to the dataframe

```
nshiftdf['req_nightshift_manpower'] = rmanpower
nshiftdf
```

Inserting the required manpower column to the dataframe

```
nshiftdf['req_nightshift_manpower'] = rmanpower
nshiftdf['req_nightshift_manpower'] = nshiftdf['req_nightshift_manpower'].astype(int)
nshiftdf['nmanpower'] = nshiftdf['nmanpower'].astype(int)
print(nshiftdf.to_string())
```

	shift	call_distribution	nightcallseconds	nmanpower	nightcallseconds90	req_nightshift_manpower
0	21_22	3	21473	1	27608	1
1	22_23	3	21473	1	27608	1
2	23_24	2	14315	1	18405	1
3	0_1	2	14315	1	18405	1
4	1_2	1	7157	1	9201	1
5	2_3	1	7157	1	9201	1
6	3_4	1	7157	1	9201	1
7	4_5	1	7157	1	9201	1
8	5_6	3	21473	1	27608	1
9	6_7	4	28631	2	36811	3
10	7_8	4	28631	2	36811	3
11	8_9	5	35789	2	46014	3

Project 8

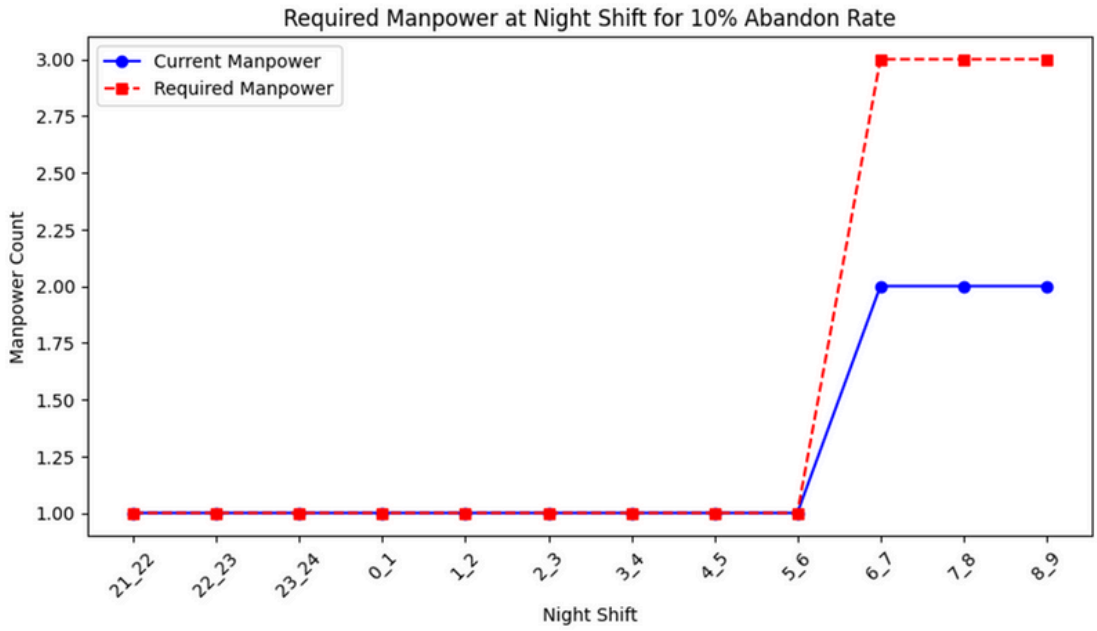
ABC Call Volume Trend Analysis

Project Tasks

Night Shift Manpower Planning

Creating the chart with the above table

```
[216] plt.figure(figsize=(10, 5))
      # linechart
      plt.plot(nshiftdf['shift'], nshiftdf['nmanpower'], marker='o', linestyle='-', label='Current Manpower', color='b')
      plt.plot(nshiftdf['shift'], nshiftdf['req_nightshift_manpower'], marker='s', linestyle='--', label='Required Manpower', color='r')
      # Titles
      plt.xlabel('Night Shift')
      plt.ylabel('Manpower Count')
      plt.title('Required Manpower at Night Shift for 10% Abandon Rate')
      plt.xticks(rotation=45)
      # Adding legend
      plt.legend()
      plt.show()
```



Insights from Task 4 (Night Shift Manpower Planning)

Night Call Volume Distribution: For every 100 calls made during the day (9 AM - 9 PM), an additional 30 calls are made at night (9 PM - 9 AM), requiring staffing adjustments.

Manpower Increase Required: Initially, 9 agents were planned for night shifts, but after adjustments, 13 agents were assigned to meet demand while keeping the call abandonment rate under 10%.

Staffing Redistribution for Coverage: Executives were assigned even in time slots where call volume was low to ensure there were no periods of zero availability.

Project 8

ABC Call Volume Trend Analysis

Project Tasks

Key insights of the project

Peak Call Handling Strategy:

Most calls occur between 11 AM - 12 PM, while the longest call durations happen after 3 PM, requiring different staffing strategies for handling volume vs. duration.

Manpower Efficiency is Critical:

The current workforce is insufficient, and increasing it from 44 to 56 agents can significantly reduce call abandonment.

Night Call Handling Needs Attention:

30% of daytime calls also happen at night, requiring a dedicated night shift to improve customer experience.

Staffing Optimization Reduces Customer Frustration:

Without proper staffing, high abandonment rates (30%) lead to poor customer service. Proper workforce planning can cut it down to 10%.

Data-Driven Decision-Making Works:

Analysis of call volume, duration, and efficiency metrics led to a structured manpower planning approach, demonstrating how data insights optimize operations.

Key Learnings from the Projects

Data Handling & Cleaning – Learned how to clean, transform, and prepare raw data for analysis using Excel, Python, and Power BI, ensuring accuracy and consistency.

Excel for Data Analysis – Gained proficiency in advanced Excel functions like PivotTables, VLOOKUP, INDEX-MATCH, and data visualization techniques for insightful reporting.

Power BI & Tableau for Visualization – Developed skills in creating interactive dashboards and reports, making data-driven insights more accessible and actionable.

Python for Data Analytics – Understood how to use Python libraries like Pandas, NumPy, and Matplotlib to manipulate, analyze, and visualize data efficiently.

Statistical Analysis & Data Interpretation – Built a strong foundation in statistics, enabling me to draw meaningful conclusions from data using measures like mean, median, correlation, and hypothesis testing.

Problem-Solving with Data – Learned how to approach real-world problems using data-driven methodologies, breaking down complex scenarios into actionable insights.

Logical Thinking & Decision Making – Enhanced my ability to analyze patterns, detect anomalies, and make informed decisions based on data trends.

Storytelling with Data – Understood the importance of presenting data in a structured, compelling way to communicate insights effectively to stakeholders.

Project Management & Time Efficiency – Developed skills in managing multiple projects efficiently, meeting deadlines, and handling data tasks systematically.

Industry-Ready Mindset – Gained a practical, hands-on experience that bridges the gap between theoretical knowledge and real-world data analytics applications.

Feedback on Trainity Data Analytics Internship Program

trainity

Comprehensive Learning Experience – The course provided a well-structured and detailed learning path, making it easier to understand the fundamentals of data analytics.

Beginner-Friendly Approach – Despite having a moderate background in analytics, the program made complex topics accessible through step-by-step explanations and practical demonstrations.

Hands-on Project Work – The inclusion of real-world projects enhanced my practical understanding, allowing me to apply theoretical concepts in a meaningful way.

Expert-Guided Learning – Each project was reviewed by experts, ensuring valuable feedback and helping me improve my analytical approach and problem-solving skills.

Diverse Toolset Training – The program covered essential tools like MS Excel, Power BI, Tableau, and Python for data analytics, helping me to be proficient in multiple platforms used in the industry.

Engaging and Immersive Content – The video lectures were clear, concise, and engaging, making learning both interactive and enjoyable.

Career-Oriented Methodology – The program is thoughtfully designed for professionals transitioning into data analytics, providing industry-relevant knowledge and skills.

Structured and Systematic Learning Path – The curriculum followed a logical sequence, ensuring a smooth transition from basic concepts to advanced applications.

Self-Paced Learning – The balance between flexibility and quality made it easy to complete the course efficiently while receiving necessary support.

Gratitude to Trainity – I am grateful to Trainity for offering such an insightful and well-organized program. It has significantly contributed to my learning journey, and I now feel more confident in my data analytics skills.