Trainity

Project PORTFOLIO

By Poorva Nimish Nahar

LIST OF CONTENTS

03 PROFESSIONAL BACKGROUND
3

04 MODULE 1 PROJECT
4-9

05 MODULE 2 PROJECT
10-20

06 MODULE 3 PROJECT
21-32

07 MODULE 4 PROJECT
33-41

08 MODULE 5 PROJECT
42-53

09 MODULE 6 PROJECT
54-63

10 MODULE 7 PROJECT
64- 74

11 MODULE 8 PROJECT
75-84

12 LEARNING AND REFLECTIONS
85

Professional BACKGROUND

Poorva Nimish Nahar is a Master's student in Data Analytics and Business Economics at Hong Kong Baptist University, specializing in Big Data and Machine Learning. She is an International Student Ambassador and a recipient of a full tuition waiver scholarship.

With a Bachelor's degree in Economics from Ahmedabad University, Poorva has experience in academic research and tutoring. Her internships include roles as a Student Ambassador, Academic Associate Intern at Ecoholics, and Academic Content Writer at InkCulture, where she developed educational content and supported admissions.

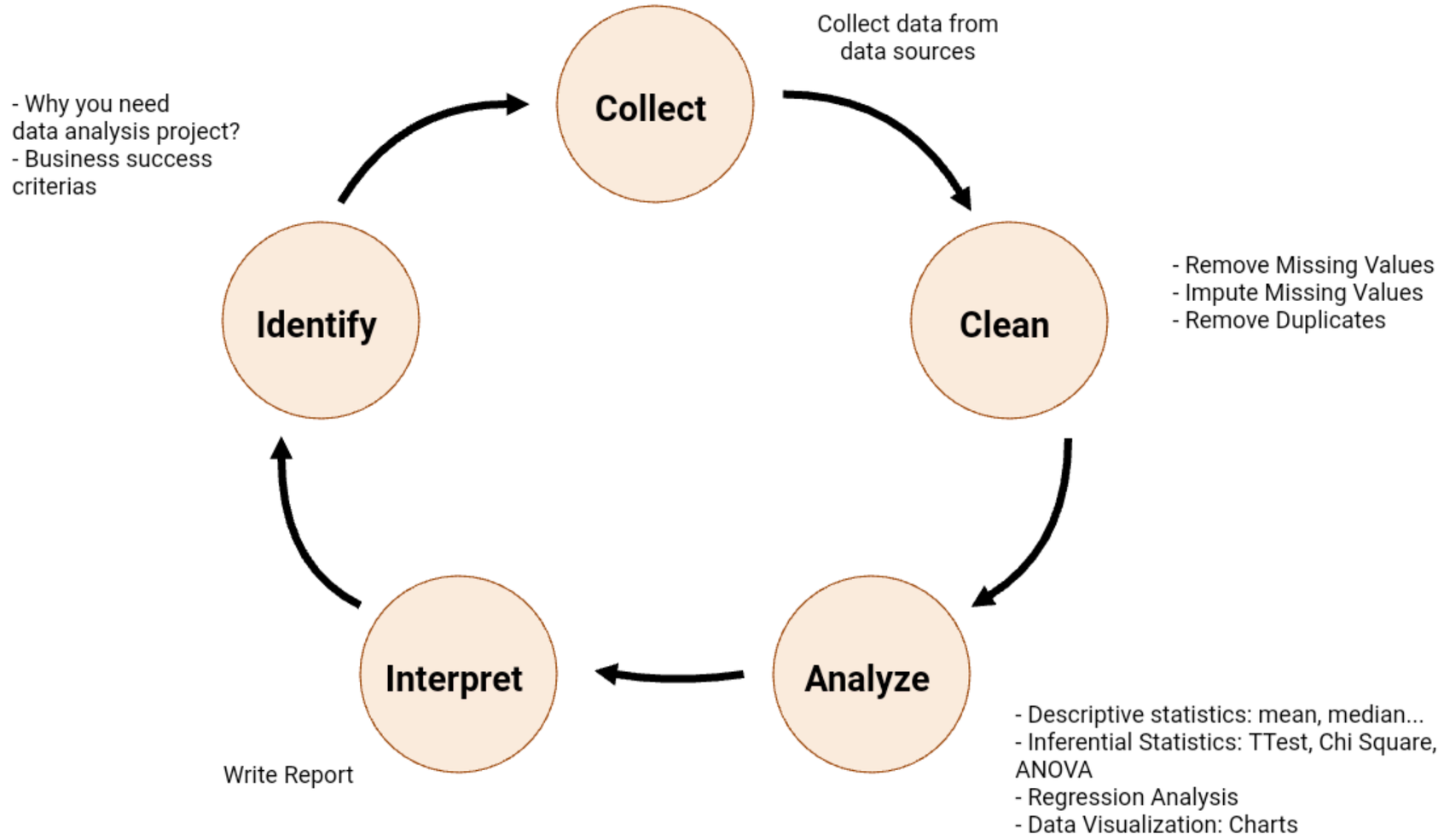
As the founder of The Period Society in Ahmedabad, she leads initiatives for menstrual health education. Poorva is proficient in MS Excel, Python, STATA, SQL, and R Programming, and possesses strong communication and problem-solving skills.

Poorva Nimish Nahar

Aspiring Data Analyst

PROJECT 01

Data Analysis Process



PROJECT 01

MANAGING ULCERATIVE COLITIS : DATA ANALYTICS IN EVERYDAY LIFE

Introduction

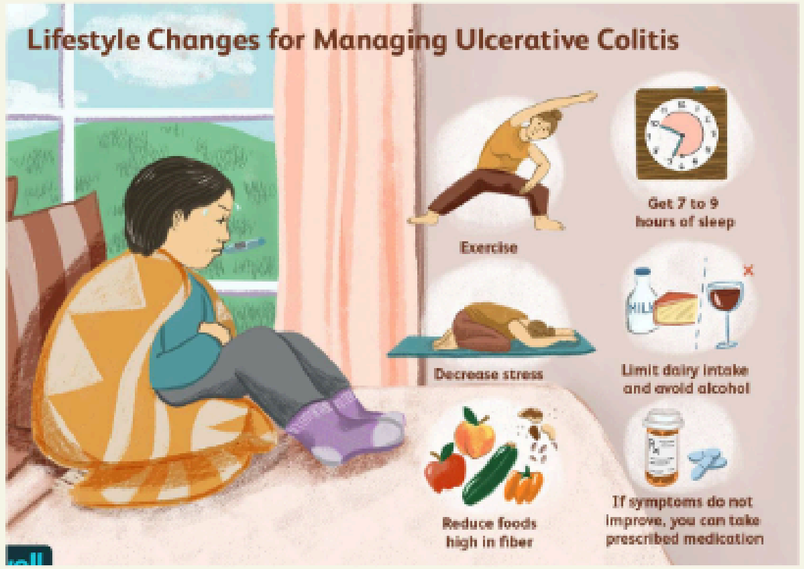
OVERVIEW OF ULCERATIVE COLITIS

A chronic inflammatory bowel disease, **Ulcerative colitis** or often known as UC, is characterised by inflammation in the colon and rectum. There are almost 12 cases of UC out of a population of 100,000.

It is very common in people in 50-55 years of age. It is an autoimmune condition that requires ongoing medical treatments throughout the lifespan of an individual. However, at the age of 16, I was diagnosed with UC, leaving me to miss half a year of my 11th grade. It was when I was introduced to a medical steroid known as "**Prednisolone**" to cure the inflammation in my colon.

PROJECT 01

MANAGING ULCERATIVE COLITIS : DATA ANALYTICS IN EVERYDAY LIFE



Source: Mayo Clinic

Plan

IDENTIFYING NEEDS

As I managed my colitis, I needed to plan my treatment styles and monitor closely my symptoms like abdominal pain, diarrhoea, chronic fatigue, and triggers. All of this were supposed to be done via regular health check up with my doctor, monthly blood tests and colonoscopies once a year. Since I was still a student then, my healthcare providers carefully planned my treatment duration and medications so as to minimise the risks involved and that it aligned with my studies.

I believe this stage was crucial to set goals for a better lifestyle and also long-term health benefits.
this plan also required me to stay informed about my condition and the treatment options that were given to me as a 16 year old. Not only this, the impact of these plans were such that I had to consider taking a break from high school or so called gap year. And thus crucial planning was important in order to think of my educational career too. Now that a few years have passed, I have very well managed to cope up with the triggers of UC and customise my medication treatments so that it well aligns with my goals and lifestyle.

PROJECT 01

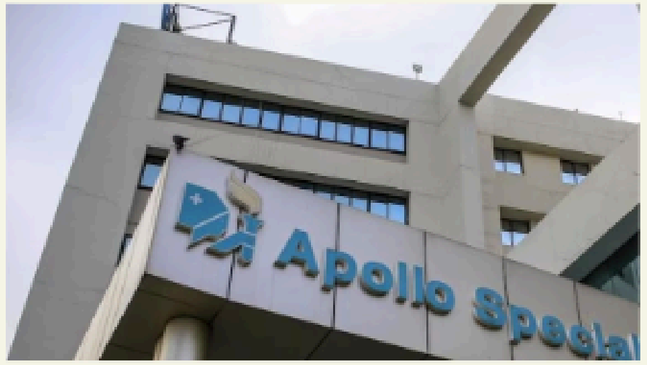
MANAGING ULCERATIVE COLITIS : DATA ANALYTICS IN EVERYDAY LIFE

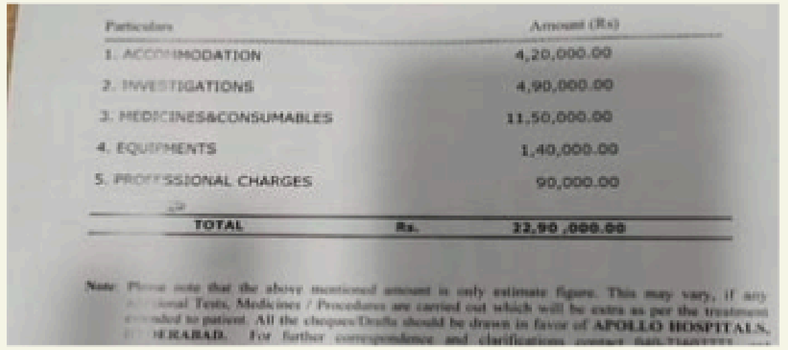
Prepare

TREATMENT BILLS AND MEDICAL COSTS

It requires careful financial planning when it comes to treating a chronic disease.

Thus, from choosing the right doctors, treatments and medications such as Prednisolone can be expensive. In particular, corticosteroids are very expensive to begin treatment. It was crucial for me to consider insurance coverage and the expenses my family would be out of their pockets. Thus, with the help of my father, we made sure to keep a track of all the medical bills and even update our budgets to manage costs effectively.





Particulars		Amount (Rs)
1. ACCOMMODATION		4,20,000.00
2. INVESTIGATIONS		4,90,000.00
3. MEDICINES&CONSUMABLES		11,50,000.00
4. EQUIPMENTS		1,40,000.00
5. PROFESSIONAL CHARGES		90,000.00
TOTAL	Rs.	22,90 ,000.00

Note: Please note that the above mentioned amount is only estimate figure. This may vary, if any additional Tests, Medicines / Procedures are carried out which will be extra as per the treatment provided to patient. All the cheques/Drafts should be drawn in favor of APOLLO HOSPITALS, HYDERABAD. For further correspondence and clarifications contact

Source: Apollo Clinic

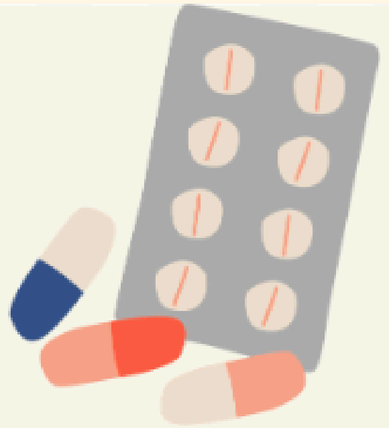
PROJECT 01

MANAGING ULCERATIVE COLITIS : DATA ANALYTICS IN EVERYDAY LIFE

Process

CHOOSING THE RIGHT TREATMENT PLAN

In order for me to recover fast, it was essential to choose the right medication. This involved my healthcare providers at Apollo to analyse my past health records, effectiveness of the previous doses of the steroids, side effects, recovery rates, and costs. Thus, communicating with my doctors seemed important to make informed decisions which would again be in sync with my needs at present and my heath history.



Analyse

TREND ANALYSIS

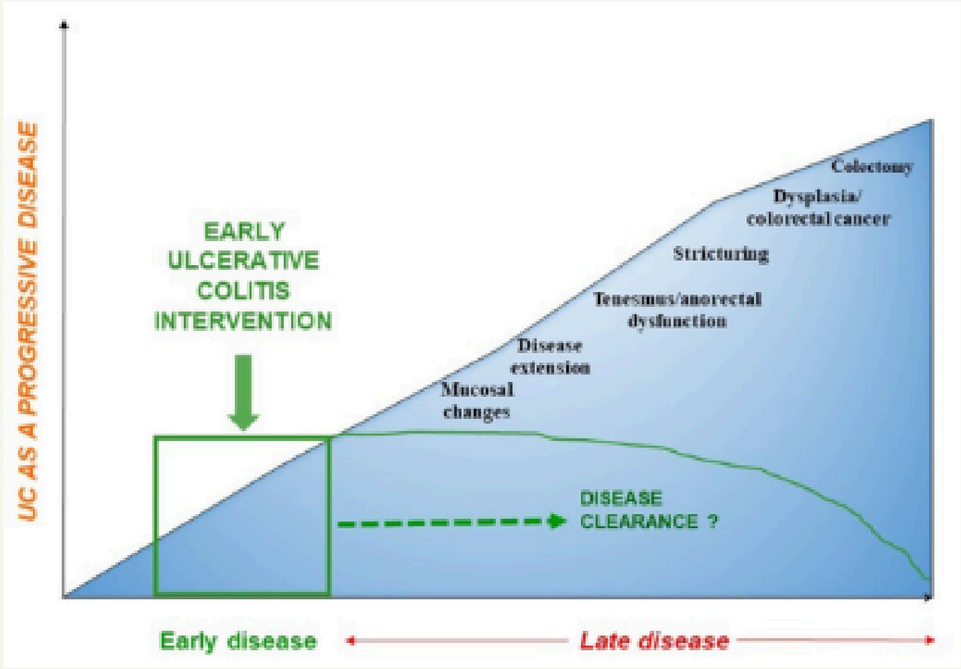
When the process was done, we were able to determine clearly the effectiveness of the treatments. Some medicines worked really well, and some had overpowering side effects, including prednisolone. Thus, until now, I have kept a track of my symptoms and the effects of medication using the Health App in my phone. This way , I was able to identify the trends and patterns such as on which days I experienced the most pain and how does it relate to the dosage I am taking. Or even if my symptoms got worse during menstruation or not. I also used to maintain a journal for me to express more intellectually as sometimes these medicines adversely affects us mentally too.

PROJECT 01

MANAGING ULCERATIVE COLITIS : DATA ANALYTICS IN EVERYDAY LIFE

SHARE AND ACT

On telling my doctors in monthly routine check ups about these side effects, he was able to take actions on the dosage of it post analysing the side effects. If the side effects were manageable he would keep the dosage, but if they were not, if would reduce the dosage. Such insights refined my treatment plan, and my doctors were then able to make necessary adjustments



Source: Journal of Clinical Medicine

PROJECT 02

Instagram User Analytics

PROJECT 02

Instagram User Analytics

Overview:

The second project of this course required me to do a detailed analysis of user interactions and engagements on one of the most famous social platforms, Instagram. As a data analyst working with the product team, I was asked to set a goal so as to utilize SQL and the workbench to get actionable insights from the data the team had gathered. I believe these insights have helped various teams in the organization including areas such as development of products, engagement of users and ultimately optimizing the overall experience of the users on Instagram.

PROJECT 02

Instagram User Analytics

Approach

I approached the project with the following steps in mind:

Step 1: Setting up the database: I first scanned through the database that was provided to us. I also initialized the database using the SQL scripts to ensure that all the data that I would need for the analysis is present.

Step 2: Next, I went ahead and executed the queries to extract data. This was used for analyzing the user engagement, loyalty, the usage of popular hashtags, presence of fake accounts and even the frequency of posting on Instagram by the users.

Step 3: After the data querying, I went ahead to analyze the results to identify trends and patterns. I believe I was able to calculate metrics such as the average number of posts made by the users which would determine if they are active and even identify the inactive users.

Step 4: Furthermore I tried to translate my data findings into actionable insights that could help the departments into taking strategic decisions.

PROJECT 02

Instagram User Analytics

Tech-Stack Used - MySQL Workbench : I used it to accept SQL queries as it is robust and also provides a comprehensive feature set for database management.

PROJECT 02

Instagram User Analytics

Insights

Summarize the insights and knowledge you gained while working on the project. Explain the inferences you made from the data, highlighting any significant findings or patterns. Keep the insights concise and relevant to the project.

1. I was able to identify the longest-standing users, providing a target for the reward of loyalty. I was even able to identify a large segment of users who were inactive, which indicated that there was potential in improving the engagement.

2. I was able to detect the unusually high activity from accounts that could possibly be fake accounts or bots. This affects the integrity of the data and even user experience.
3. I was able to identify the most used hashtags which could maximize the reach for the partner brands and could help in influencing the content strategy further.
4. I was also able to find the day on which most of the registrations were made. I believe it is useful in order to know the engagement levels to maximize the efficiency of ads campaigns to launch features.

PROJECT 02

Instagram User Analytics

Result

Here are some of the achievements I have tried to accomplish through the project and the benefits I have experienced and learned about.

1. Strategic Decision Making. I believe the insights have a direct influence on the marketing strategies, development of new features, and even tactics of user engagement.
2. I learned that the recommendations that were based on the user activity and even registrations data lead to managing targeted content and updates of the features.
3. The major learning came from the high potential of bot activity. As I believe that due to the rise in Artificial Intelligence , it has become much easier to make fake accounts, and bots to get maximum recognition on a platform like Instagram. This query was important to understand the security concerns that arise and take necessary actions.

PROJECT 02

Instagram User Analytics

A) Marketing Analysis:

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

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

Code:

```
SELECT * FROM users
ORDER BY created_at
LIMIT 5;
```

Results:

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

Insights: The above 5 users should be rewarded for being the most loyal users. This would again encourage them to keep supporting the platform. This would also make them feel appreciated and seen.

PROJECT 02

Instagram User Analytics

2. **Inactive User Engagement:** The team wants to encourage inactive users to start posting by sending them promotional emails.
   Your Task: Identify users who have never posted a single photo on Instagram.

Code:

```
SELECT username
 FROM users
 LEFT JOIN photos
 ON  users.id= photos.user_id
 WHERE photos.id IS NULL;
```

Results: Usernames who have not posted any pictures are as follows:

1. Tierra.Trantow
2. Rocio33
3. Pearl7
4. Ollie_Ledner37
5. Nia_Haag
6. Morgan.Kassulke
7. Mike.Auer39
8. Mckenna17
9. Maxwell.Halvorson
10. Linnea59
11. Leslie67
12. Kasandra_Homenick
13. Julien_Schmidt
14. Jessyca_West
15. Janelle.Nikolaus81
16. Jaclyn81
17. Hulda.Macejkovic
18. Franco_Keebler64
19. Esther.Zulauf61
20. Esmeralda.Mraz57
21. Duane60
22. David.Osinski47
23. Darby_Herzog
24. Bethany20
25. Bartholome.Bernhard
26. Aniya_Hackett

Insights: I believe the above user can make use of the promotional emails and start posting more pictures with good quality content. The promotions would motivate them to gain an internet presence and even help small-scaled brands grow.

PROJECT 02

Instagram User Analytics

Insights:

As seen above, the winner of the contest is Zack Kemmer. His instagram id is Zack_Kemmer 95 and the image with the most likes is https://jarret.name . This image has received a total of 48 likes.

---

4. Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.
   Your Task: Identify and suggest the top five most commonly used hashtags on the platform.

Code:

```
SELECT tags.tag_name, COUNT(*) AS total
FROM photo_tags
JOIN tags
ON photo_tags.tag_id= tags.id
GROUP BY tags.id
ORDER BY total DESC
LIMIT 5;
```

Results:

tag_name	total
smile	59
beach	42
party	39
fun	38
concert	24

Insights:

As listed above, these hashtags can be utilized by the partner brand in order to reach more people through their posts. This will further increase engagement with the brand products and eventually increase demand and sales.

PROJECT 02

Instagram User Analytics

5. Ad Campaign Launch: The team wants to know the best day of the week to launch ads.
   Your Task: Determine the day of the week when most users register on Instagram.
   Provide insights on when to schedule an ad campaign.

Code:

```
SELECT
 DAYNAME(created_at) AS day, COUNT(*) AS total
 FROM users
 GROUP BY day
 ORDER BY total DESC
 LIMIT 1;
```

Insights: As we have set the limit to 1, we can see that most of the users have registered on a Thursday. This means that the ad campaign should run around this day. When we set the limit to 2 , we get Sunday as another day when most of the registrations are made. Thus, this could also potentially mean that the team can run their campaigns starting from Thursday till Sundays to ensure more engagement.

Results:

Day	Total
Thursday	16

PROJECT 02

Instagram User Analytics

B) Investor Metrics:

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

Code:

```
SELECT
(SELECT COUNT(*) FROM  photos) /   (SELECT COUNT(*) FROM users) AS AVG;
```

Results: 2.5700

Insights: The results showed that at an average the number of posts per user on Instagram is 2.57. Comparatively, this shows that the users are posting on average few posts.

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

Code:

```
SELECT user_id, COUNT(*) as num_likes
FROM likes
GROUP BY user_id
HAVING num_likes = (SELECT COUNT(*) FROM photos);
SELECT u.username, COUNT(*) as num_likes
FROM users u
JOIN likes l ON u.id = l. user_id
GROUP BY u.id
HAVING num_likes = (SELECT COUNT(*) FROM photos);
```

Results:

username	num_likes
Aniya_Hackett	257
Jaclyn81	257
Rocio33	257
Maxwell.Halvorson	257
Ollie_Ledner37	257
Mckenna17	257
Duane60	257
Julien_Schmidt	257
Mike.Auer39	257
Nia_Haag	257
Leslie67	257
Janelle.Nikolaus81	257
Bethany20	257

Insights: As we can see that the above accounts are potential bots and fake accounts. The number of likes from these accounts is 257 likes, which is not possible for an individual to do.

PROJECT 03

Operation Analytics and Investigating Metric Spike

PROJECT 03

Operation Analytics and Investigating Metric Spike

Overview:

The third project of this course required me to do operational analytics and also to invetigate metric spikes. As a data analyst working with the product team, I was asked to set a goal so as to utilize SQL and the workbench to get actionable insights from the data the team had gathered. I believe these insights have helped various teams in the organization including areas such as engagement of users , mailing strategies, language and use of the devices by the users , that could further help in modifying users' experiences. Here, we used the company's end-to-end operational data to identify key improvements within the conernced company.

PROJECT 03

Operation Analytics and Investigating Metric Spike

Approach

I approached the project with the following steps in mind:

Step 1: Setting up the database: I first scanned through the database that was provided to us. I also initialized the database using the SQL scripts to ensure that all the data that I would need for the analysis is present.

Step 2: Next, I went ahead and executed the queries to extract data. This was used for investigating metric spikes, such as dip in daily user engagement or even a drop in sales of a particular product.

Step 3: After the data querying, I went ahead to analyze the results to identify trends and patterns. I believe I was able to calculate metrics such as job reviews over the time, language share analysis, duplication of rows, weekly user engagement, device usage, retention analysis and even email engagement analysis.

Step 4: Furthermore I tried to translate my data findings into actionable insights that could help the departments into taking strategic decisions.

PROJECT 03

Operation Analytics and Investigating Metric Spike

Tech-Stack Used - MySQL Workbench : I used it to accept SQL queries as it is robust and also provides a comprehensive feature set for database management. I also used Excel for visualising and summarising the results.

PROJECT 03

Operation Analytics and Investigating Metric Spike

Insights

1. From case study 1 , I identified that the number of jobs reviewed per house varied daily with most days averaging around 0.0417 jobs/hours. Also, I understood that the 7-day rolling average of the throughput is much better compared to daily throughput as it's more comprehensive. I learned how to detect duplicates in the table as well.
2. Another interesting insight I was able to identify was from case study 2, where the results from the email engagement metrics were rather shocking. Thus, it suggested for iptiomising email campaigns. I was also able to identify that laptops are mostly preferred by the users.
3. I was able to detect the retention and engagement of users per week. The growth of active users was rather significant with a short term retention of 81%.

PROJECT 03

Operation Analytics and Investigating Metric Spike

Result

Here are some of the achievements I have tried to accomplish through the project and the benefits I have experienced and learned about: I believe the insights have a direct influence on the marketing strategies, job reviews, and even tactics of user engagement. The major learning came from the email engagement metrics. As I believe that due to the rise in Artificial Intelligence , it has become much easier to send automated mails to everyone. However, this also reduces the reader's attention.

PROJECT 03

Operation Analytics and Investigating Metric Spike

SQL Tasks:

Case Study 1: Job Data Analysis
You will be working with a table named job_data with the following columns:

- **job_id:** Unique identifier of jobs
- **actor_id:** Unique identifier of actor
- **event:** The type of event (decision/skip/transfer).
- **language:** The Language of the content
- **time_spent:** Time spent to review the job in seconds.
- **org:** The Organization of the actor
- **ds:** The date in the format yyyy/mm/dd (stored as text).

Tasks:

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

Code:

```
SELECT
  ds,
  COUNT(*) / 24.0 AS jobs_per_hour
FROM
  job_data
WHERE
  ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY
  ds
ORDER BY
  ds ;
```

Results:

ds	jobs_per_hour
25-11-2020	0.0417
26-11-2020	0.0417
27-11-2020	0.0417
28-11-2020	0.0833
29-11-2020	0.0417
30-11-2020	0.0833

Insights: The number of jobs revived per hour on each day in November 2020 differed for each working day. For most days, the jobs reviewed per hour was around 0.0417 and the other days, 0.0833. Strategies need to be made in order to ensure the maximum jobs reviwed, while considering manpower and overworking.

PROJECT 03

Operation Analytics and Investigating Metric Spike

B. Throughput Analysis:

- Objective: Calculate the 7-day rolling average of throughput (number of events per second).
- Your Task: Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the

7-day rolling average for throughput, and why.

Code:

```
SELECT
  ds, COUNT(*) / SUM(time_spent) AS daily_throughput
FROM
  job_data
GROUP BY ds;
WITH daily_throughput AS (
  SELECT
    ds,
    COUNT(*) / SUM(time_spent) AS throughput
  FROM
    job_data
  GROUP BY
    ds

SELECT
  ds,
  throughput,
  AVG(throughput) OVER (
    ORDER BY ds
    ROWS BETWEEN 6 PRECEDING AND CURRENT ROW
  ) AS rolling_avg_throughput

FROM
  daily_throughput
ORDER BY
```

Results:

ds	throughput	rolling_avg_throughput
25-11-2020	0.0222	0.0222
26-11-2020	0.0179	0.02005
27-11-2020	0.0096	0.01656667
28-11-2020	0.0606	0.027575
29-11-2020	0.05	0.03206
30-11-2020	0.05	0.03505

Insights:

In Order to calculate the throughput I believe we should use the 7 day rolling because it gives us the average for all days beginning from day 1 to day 7. On the otherhand, when we use the daily metrics, it would only give us the avergae for that particular day. In order to calculate the 7 day rolling daily metrics average of throughput, I have used the count of job_id and have then ordered them according to the date of interviews or the ds. I have then used the row function and considered the rows between 6 and preceding rows and the current row. Afterwhich, I proceeded to take the average of the job_reviewed.

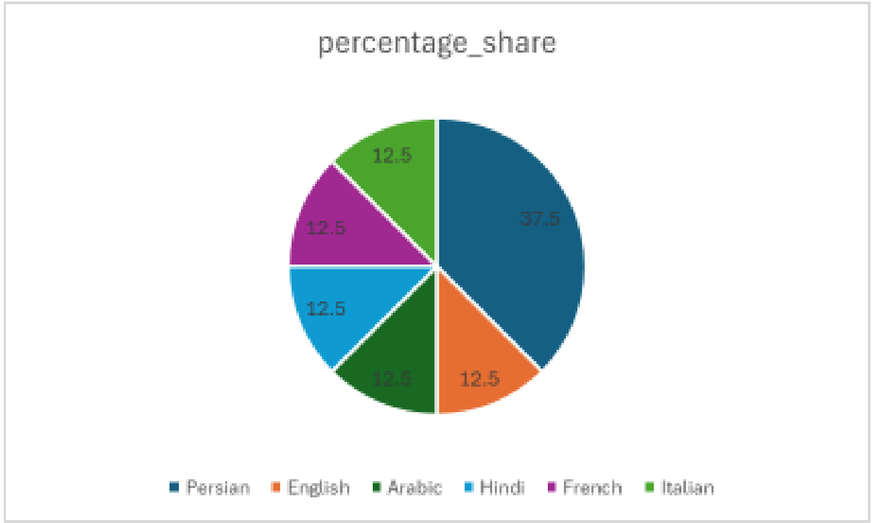
PROJECT 03

Operation Analytics and Investigating Metric Spike

C. Language Share Analysis:

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

```
;WITH last_30_days AS (
    SELECT *
    FROM job_data
    WHERE ds >= DATE_SUB((SELECT MAX(ds) FROM job_data), INTERVAL 30 DAY)
)
SELECT
    language,
    COUNT(*) * 100.0 / (SELECT COUNT(*) FROM last_30_days) AS percentage_share
FROM
    last_30_days
GROUP BY
    language
ORDER BY
    percentage_share DESC;
```



Insights: As from the above chart, we can see that Persian had the highest percentage among all the languages used. Meanwhile, the other languages shared an equal percentage. This means that a majority of the content was delivered in the Persian Language. And it also ensured diversity as the other languages were also used to deliver the content.

PROJECT 03

Operation Analytics and Investigating Metric Spike

D. Duplicate Rows Detection:

- Objective: Identify duplicate rows in the data.
- Your Task: Write an SQL query to display duplicate rows from the job_data table.

Code:

```
Select * from

(

select *,

row_number () over (partition by job_id) as row_num

from job_data

) a

where row_num > 1;
```

Results:

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

Insights:

In order to know which rows were duplicated, or had the same values, I had to firstly decide the column or th eparemeter where we had the duplicated row values, and then used the **row_number** function to further find the row numbers which had the same values. The column that was to be considered was job_id would be portioned using the previous function. Then, the ufnction **where** was used tofind the row _num which have the values greater than 1, that is, row num> 1. I believe, it's quite an interesting feature of MySql to detect the duplications as it saves time and reduces errors.

PROJECT 03

Operation Analytics and Investigating Metric Spike

Case Study 2: Investigating Metric Spike

Tables:

- **users**: Contains one row per user, with descriptive information about that user's account.
- **events**: Contains one row per event, where an event is an action that a user has taken (e.g login, messaging, search).
- **email_events**: Contains events specific to the sending of emails.

Tasks:

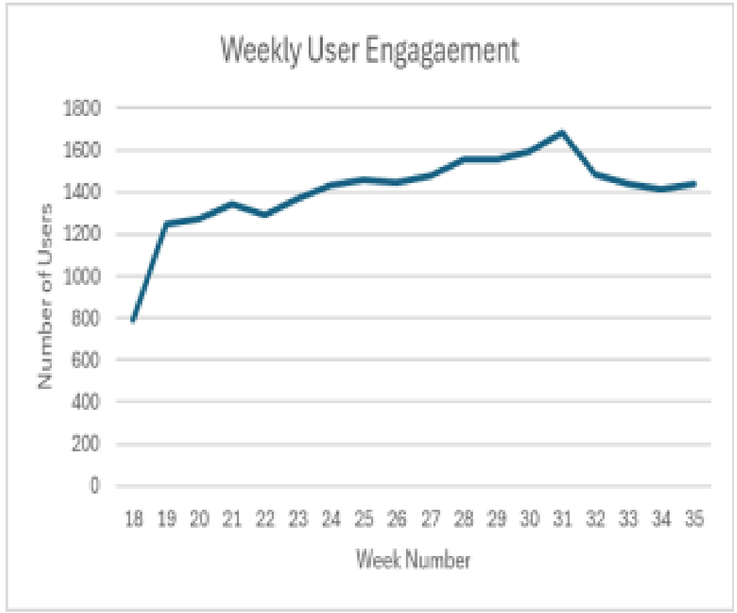
A. **Weekly User Engagement:**
   - Objective: Measure the activeness of users on a weekly basis.
   - Your Task: Write an SQL query to calculate the weekly user engagement.

Code:

```
SELECT
    EXTRACT(WEEK FROM occurred_at) AS week_num,
    COUNT(DISTINCT user_id)
FROM
    events
GROUP BY
    week_num;
```

Results:

week number	number of users
18	791
19	1244
20	1270
21	1341
22	1293
23	1366
24	1434
25	1462
26	1443
27	1477
28	1556
29	1556
30	1593
31	1685
32	1483
33	1438
34	1412
35	1442



Insights:

The above results show that theuser engagement increased from the weeks 18-35. The highest enegamenet was in week 16. The users engagement showed a steep increase during the first 10 weeks, reaching its peak at week 16 and then declining post that.

This can be interpreted as a significant event or even a festival during week 16 , which the concerned department can leverage it. To increase the enegagement further, the team can also conduct events to experiment weekly to understand the activness of the users.

PROJECT 03

Operation Analytics and Investigating Metric Spike

B. User Growth Analysis:

- Objective: Analyze the growth of users over time for a product.
- Your Task: Write an SQL query to calculate the user growth for the product

```
year_num,
week_num,
num_active_users,
SUM(num_active_users)OVER(ORDER BY year_num, week_num ROWS BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW) AS cum_active_users
from
(
select
extract(year from a.activated_at) as year_num,
extract(week from a.activated_at) as week_num,
count(distinct user_id) as num_active_users
from
users
WHERE
state = 'active'
group by year_num,week_num
order by year_num,week_num
) a;

SELECT
COUNT(*)
```

year_num	week_num	num_active_users	cum_active_users
2013	1	67	67
2013	2	29	96
2013	3	47	143
2013	4	36	179
2013	5	30	209
2013	6	48	257
2013	7	41	298
2013	8	39	337
2013	9	33	370
2013	10	43	413
2013	11	33	446
2013	12	32	478
2013	13	33	511
2013	14	40	551
2013	15	35	586
2013	16	42	628
2013	17	48	676
2013	18	48	724
2013	19	45	769
2013	20	55	824
2013	21	41	865
2013	22	49	914
2013	23	51	965
2013	24	51	1016
2013	25	46	1062
2013	26	57	1119
2013	27	57	1176
2013	28	52	1228
2013	29	71	1299
2013	30	66	1365
2013	31	69	1434
2013	32	66	1500
2013	33	73	1573
2013	34	70	1643
2013	35	80	1723
2013	36	65	1788
2013	37	71	1859
2013	38	84	1943
2013	39	92	2035
2013	40	81	2116
2013	41	88	2204
2013	42	74	2278
2013	43	97	2375
2013	44	92	2467
2013	45	97	2564
2013	46	94	2658
2013	47	82	2740

Insights:

The above results show us the cumulative active users along with the weeks . As we can see that the number of the cumulative active users grew from 67 at the start of the year 2013, to 9381 by the 35th week of the year 2014. This in turn represents an immense growth of 13,900% over the course of 87 weeks, which is commendable . Furthermore, we can see that the average number of weekly active users in 2013 was about 56, and in 2014 it was about 173, showing a significant increase in the user engagement per week.

12

Next, we can see that there is an upward trend in the user activity towards the end of each year. Also, we can see that weeks 50-52 saw a higher activity and also during the beginning of 2014. This suggests that possibly due to the New Years and Chistrmas celebrations, the engagement increased. We can also notice a consistency in the weekly growth in the year 2014, with more lower weeks in 2013. The lowest as we can see is the 91 users in week 1 in 2014, but is still higher than most weeks in 2013. This suggests that our platform is maturing as well as finding the correct product-market fit. So, some of the factors could definitely be the following:

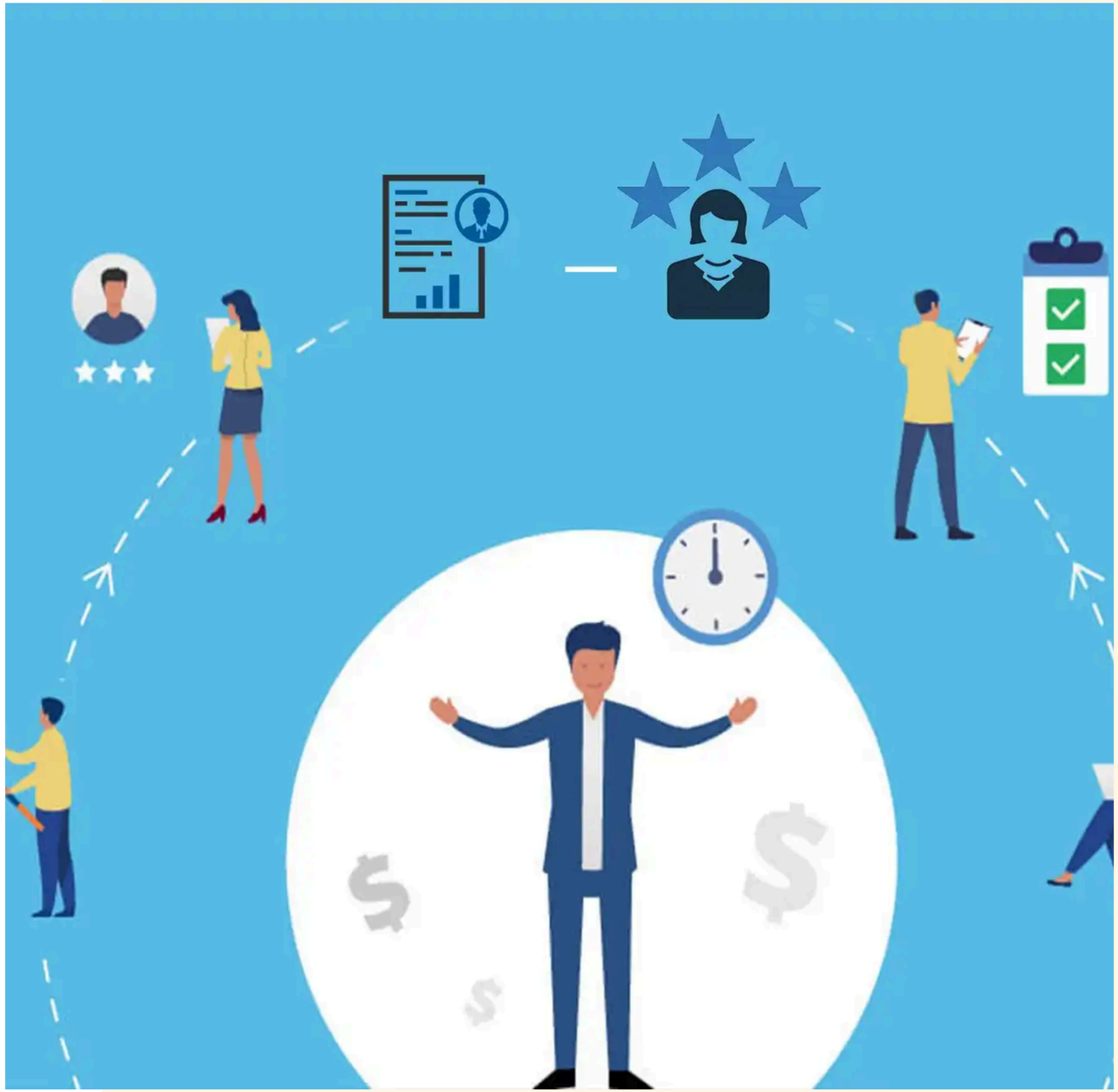
a) Word of mouth referrals
b) Seasonal factors
c) Marketing campaigns
d) Improvement in the features of our products

Again, we if compare the results year-toyear, let say , the first 35 weeks, we can see that

2013 > total new users amounted to 1723 and in 2014 , the total new users accounted for 6007. Thus, year-to-year growth is about 248% for the same period. I believe that these specific insights can definitely help the concerned departments to provide a good overview of the growth trajectory of the platform over the course of two years.Again, it would have been beneficial to have more einformation on marketing efforts, updates on the product features or even any external factors such as elections or calamities, or even important communities functions that could have influcnced the user engagement.

PROJECT 04

Hiring Process Analysis



PROJECT 04

Hiring Process Analysis

Project Description

The fourth project of this course required me to do hiring process analytics . As a data analyst working with the hiring team at Google , I was asked to anlayse the company's goals and also draw insights from it. As the hiring process is crucial, I believe it is important to understand trends such as the job types, interviews conducted, number of rejects and even vacancies. I was expected to handle the missing data, club columns, detcet outliers and remove them. I Believe with the following insights , the hiring team can strategize better.

Approach

I approached the project with the following steps in mind: I checked for any missing datas and dealt with outliers in the data provided. Further which, I proceeded to perform the tasks designated to me. I made sure that for each task, I have created visualisations wherever necessary to better understand the results and insights.

PROJECT 04

Hiring Process Analysis

Tech-Stack Used - Excel

Insights

1. The lack of gender diversity could be improved in the long run while hiring. This can be done through campaigns running specifically for the encouragement of the female conditions
2. The average salaries varied across departments, which indicated a discrepancy and needs to be understood in the context of responsibilities, positions and skills.
3. The salary distribution also helps us to understand and design better compensation packages and career progression plans.
4. The departmental and position tier analysis could also help us understand the career growth and promotions of teams at Google.

Result

Here are some of the achievements I have tried to accomplish through the project and the benefits I have experienced and learned about:I believe the insights have a direct influencehiring process, interviews, positions , salaries offered and even the gender diversity in the company. The major learning came from Task 1 , where it was found that the male candidates were much higher than the females. I believe that our company is widely known and not having gender diversirty might create a problem in the long run.

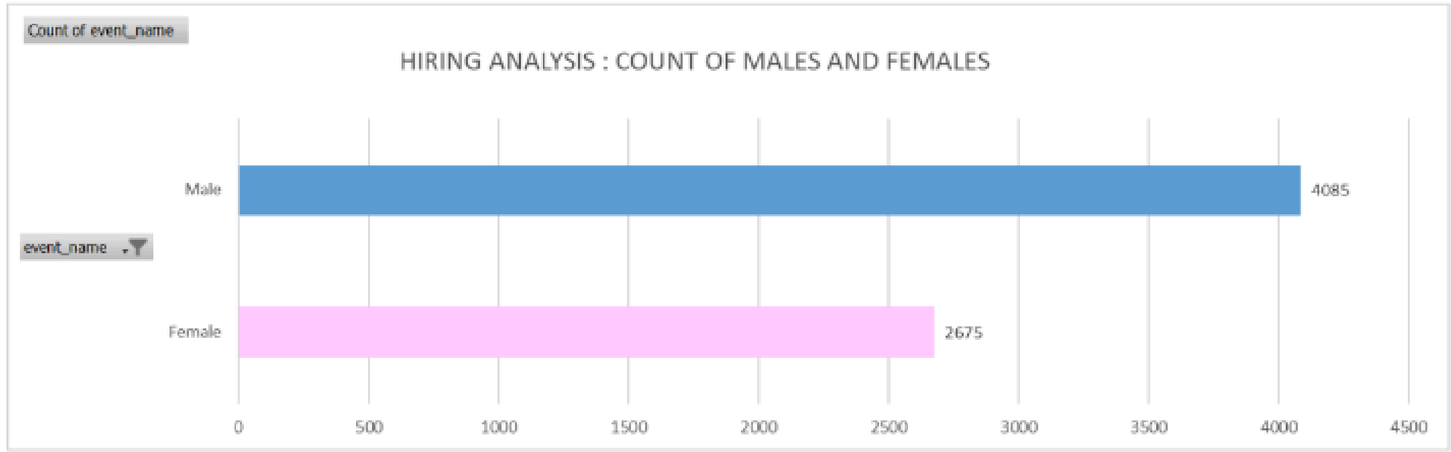
PROJECT 04

Hiring Process Analysis

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

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

Results:



Insights: As observed above, we can see that the count of males hired by the company is much more than the females. Only 39% are females. Thus, it shows a noticeable gender disaprity. The cause of this could be the applicant pool, that is, there may be more applications by make than females. It could also be the case of industry norms or biases in hiring. Thus, it is recommended to promote gender diverity in the workplace. We can impliment initiatives like targeted recruitment campaigns, partnerships with organisations supporting women in the industry in order to attract more famles candidates. We can even promote career development programs specififaclly for women to help them excel within our firm.

Process used: Pivot tables to filter other columns and count the number of males and females. A bar graph was further used to visualise the findings from the pivot table.

PROJECT 04

Hiring Process Analysis

B. Salary Analysis: The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.

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

Sum of Salaries	Number of employees	Average salary (Sum of salaries / Number of employees
358228369	7168	49976.05594

Thus, the average salary provided to an employee is 499976.03 for this company

Insights: As we can see from the above picture, the average salary rounds up to Rs. 50,000 in the company. However, we also need to keep in mind that average salaries vary across different departments. This could also cause a discrepancy in the average salary.

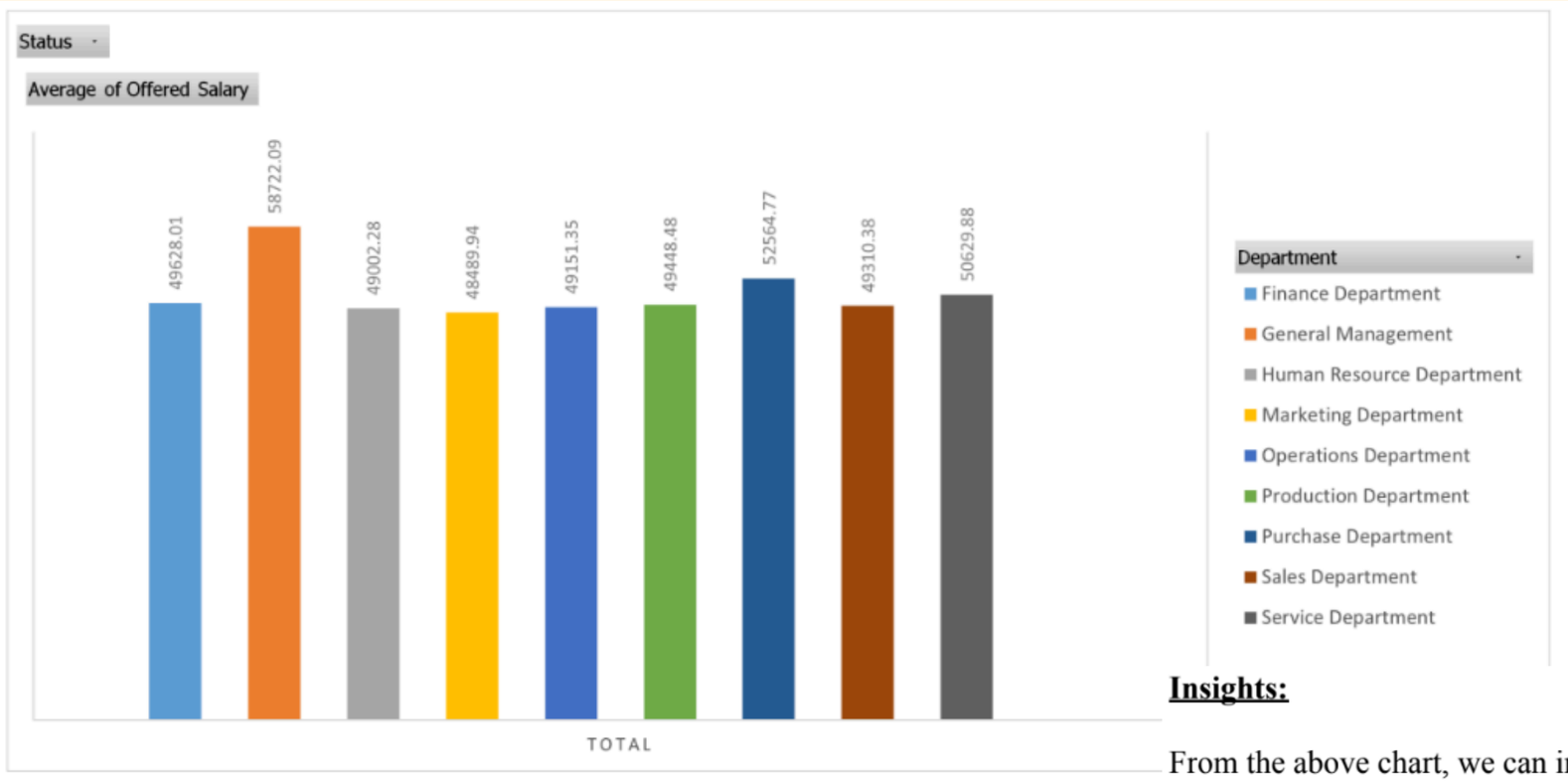
Process Used: To perform this task I used the concept of dividing the sum of salaries by the number of employees in Excel.

PROJECT 04

Hiring Process Analysis

C. Salary Distribution: Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.

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



Insights:

From the above chart, we can infer that the salaries in the class intervals differ from department to department. The lowest accounts to the Marketing Department, whereas the General Management Department has the highest. This also shows how different skills are paid differently at every company. In addition to this, the salary also changes with the positions in individual departments.

Process Used:

The process used here was of pivot tables. Where the salary was averged according to the concerned departments.

PROJECT 04

Hiring Process Analysis

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

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

Results: Please refer to Task E results

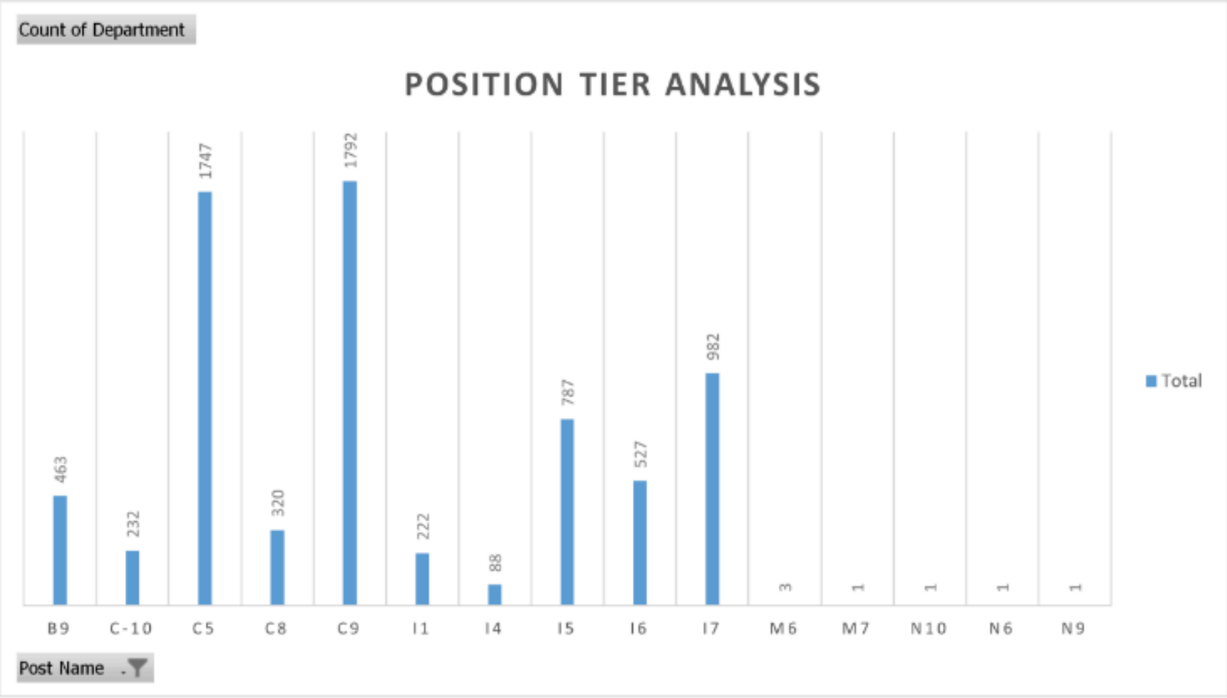
PROJECT 04

Hiring Process Analysis

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

Your Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.

Results:



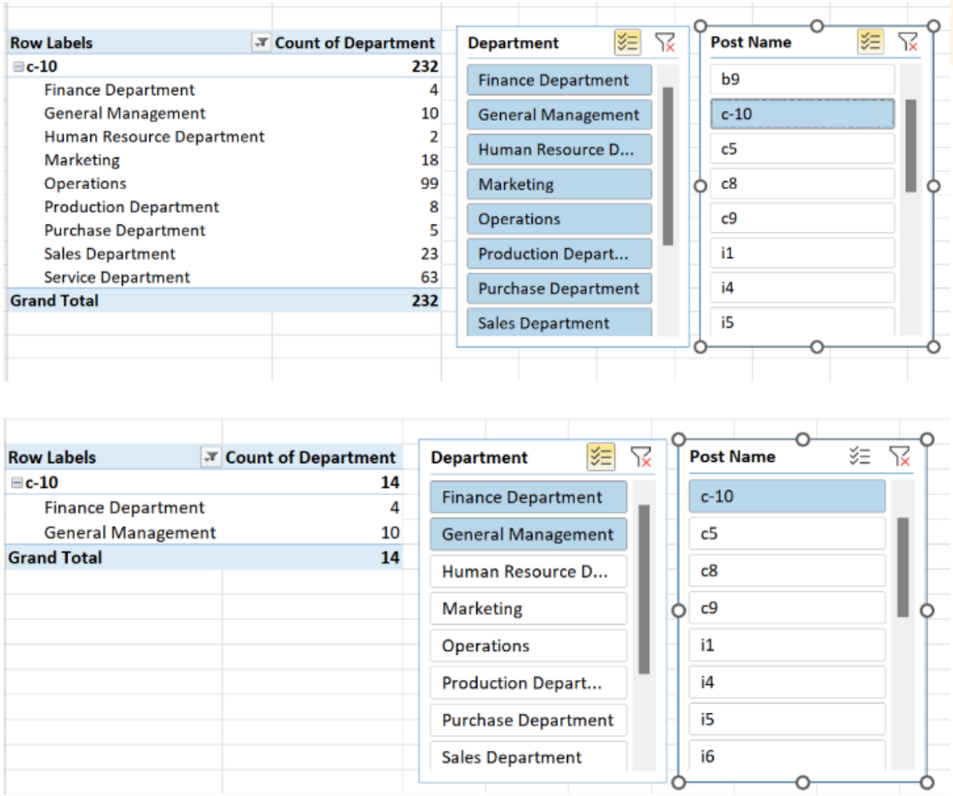
Insights:

So for Task D and E, I tried to come up with a solution that would show us both the departmental analysis as well as position tier analysis. The above chart is a representation of the number of members at each position in each department. The highest number of members hold the position of C9 and the lowest belong to M7, N10, N6 and even N9. This again represents how every company works internally. The members that hold positions can also be promoted to the other up-ranking positions at every stage of growth.

PROJECT 04

Hiring Process Analysis

Process Used:



Row Labels	Count of Department
c-10	232
Finance Department	4
General Management	10
Human Resource Department	2
Marketing	18
Operations	99
Production Department	8
Purchase Department	5
Sales Department	23
Service Department	63
Grand Total	232

Row Labels	Count of Department
c-10	14
Finance Department	4
General Management	10
Grand Total	14

Row Labels	Count of Department
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	982
m6	3
m7	1
n10	1
n6	1
n9	1
Grand Total	7167

As you can see above, I also tried to use slicer in the pivot tables to give us a clear picture of the count of members on each position in each department. Chart 1 is another example, where we can see that the total count of members at c9 position counting for all departments in 232.

And if we only want to see the number of members in c10 positions in finance and general management departments, we can use the slicer. That accounted for a total of 14 members (Chart 2).

Chart 3 shows us the total count of members in different positions.

PROJECT 05

IMDB Movie Analysis

PROJECT 05

IMDB Movie Analysis

CONTENTS

PROJECT DESCRIPTION
A brief overview of the project

APPROACH
Tools and Techniques Used

TECH-STACK USED
Software and versions used and its purpose

INSIGHTS
Key Findings and meaningful trends

RESULTS
Understanding of IMBD Movie Anlytics

PROJECT 05

IMDB Movie Analysis

IMDB MOVIES ANALYSIS

- **Objective:**
Investigate the factors that influence the success of a movie on IMDB, defined by high IMDB ratings.

- **Impact:**
Provide insights to movie producers, directors, and investors for making informed decisions to create successful movies.

- **Data Cleaning:**
Preprocess the dataset by handling missing values, removing duplicates, converting data types, and performing feature engineering.

- **Data Analysis:**
Explore relationships between movie ratings and factors such as genre, director, budget, year of release, and actors.
Use statistical methods to understand the impact of these factors on IMDB ratings.

- **Five 'Whys' Approach:**
Use this technique to dig deeper into observed patterns and uncover root causes.
Example: Analyzing why higher budgets correlate with higher ratings by exploring aspects like production quality and viewer experience.

- **Report and Data Story:**
Create a comprehensive report that narrates the data analysis journey.
Use visualizations to present findings and make them easily understandable.

PROJECT 05

IMDB Movie Analysis

MY APPROACH

- Download the dataset
- Understand the data
- Data cleaning – removing null and duplicate columns , deleting all the unnecessary columns, and removing special characters.
- Using Excel and statistics formulas to solve the problems
- Create visuals to gain insights from it

TECH- STACK USED

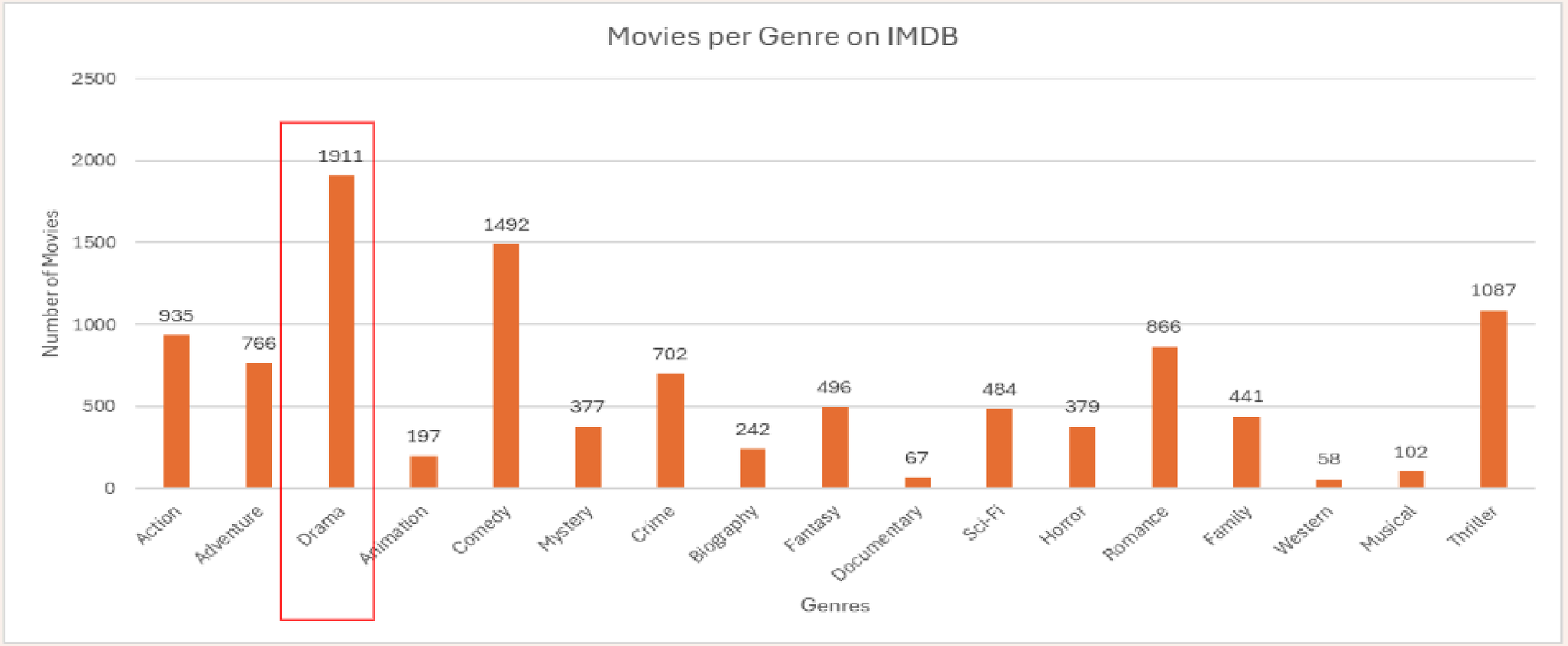
- Microsoft Excel (2019) : To analyse the data and create visuals.
- Microsoft PowerPoint : To create report

PROJECT 05

IMDB Movie Analysis

TASK A - MOVIE GENRE ANALYSIS:

ANALYSE THE DISTRIBUTION OF MOVIE GENRES AND THEIR IMPACT ON THE IMDB SCORES



Drama genre has the highest count of movies, with an average of 6.79 IMDB Score. The lowest is by Western genre, with only 58 movies and an average of 6.7 IMDB Score.

PROJECT 05

IMDB Movie Analysis

TASK A- MOVIE GENRE ANALYSIS:
ANALYSE THE DISTRIBUTION OF MOVIE GENRES AND THEIR IMPACT ON THE IMDB SCORES

GENRE	No_of_movies	Mean_imdb	Median_imdb	Mode_imdb	Max_imdb	Min_imdb	StdDev_imdb	Var_imdb
Action	935	6.285989305	6.3	6.6	9	2.1	1.038357736	1.078186788
Adventure	766	6.454960836	6.6	6.6	8.9	2.3	1.116926308	1.247524378
Drama	1911	6.789115646	6.9	6.7	9.3	2.1	0.891064898	0.793996652
Animation	197	6.700507614	6.8	7.3	8.6	2.8	0.993627525	0.987295659
Comedy	1492	6.183310992	6.3	6.3	8.8	1.9	1.039919012	1.081431552
Mystery	377	6.469496021	6.5	6.6	8.6	3.1	1.007391835	1.014838309
Crime	702	6.548148148	6.6	6.6	9.3	2.4	0.984105199	0.968463042
Biography	242	7.140082645	7.2	7	8.9	4.5	0.71009671	0.504237338
Fantasy	496	6.285080645	6.4	6.7	8.9	2.2	1.140414241	1.30054464
Documentary	67	7.011940299	7.2	6.6	8.5	1.6	1.199939694	1.439855269
Sci-Fi	484	6.327272727	6.4	7	8.8	1.9	1.16718415	1.362318841
Horror	379	5.903957784	5.9	6.2	8.6	2.3	0.991023285	0.982127152
Romance	866	6.426212471	6.5	6.5	8.5	2.1	0.968996249	0.938953731
Family	441	6.2	6.3	5.4	8.6	1.9	1.169576458	1.367909091
Western	58	6.765517241	6.8	6.8	8.9	4.1	0.998516746	0.997035693
Musical	102	6.550980392	6.7	7.1	8.5	2.1	1.143535	1.307672297
Thriller	1087	6.372309108	6.4	6.5	9	2.7	0.969078327	0.939112803

Mathematical and arithmetical functions used:

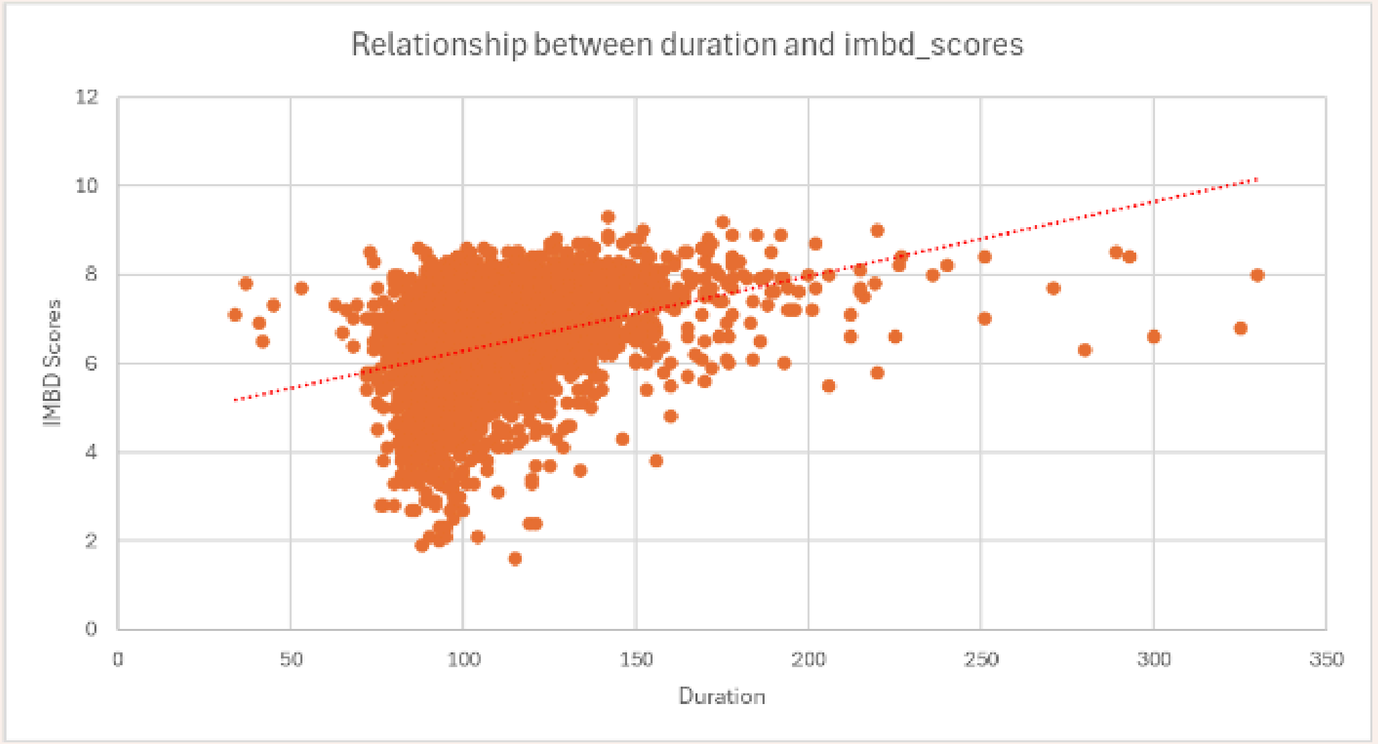
1. Countif
2. Avergeif
3. Median
4. Maxif
5. Minif
6. Std.dev
7. Search
8. Variance

PROJECT 05

IMDB Movie Analysis

TASK B - MOVIE DURATION ANALYSIS:

ANALYSE THE DISTRIBUTION OF MOVIE DURATIONS AND IDENTIFY THE RELATIONSHIP BETWEEN MOVIE DURATION AND IMDB SCORE



Operations	Values
Mean	109.808505
Median	105
Mode	101
Standard De	22.763201
Variance	518.16332

Mathematical and arithmetical functions used:

1. Mean
2. Median
3. Mode
4. Std.dev
5. Variance

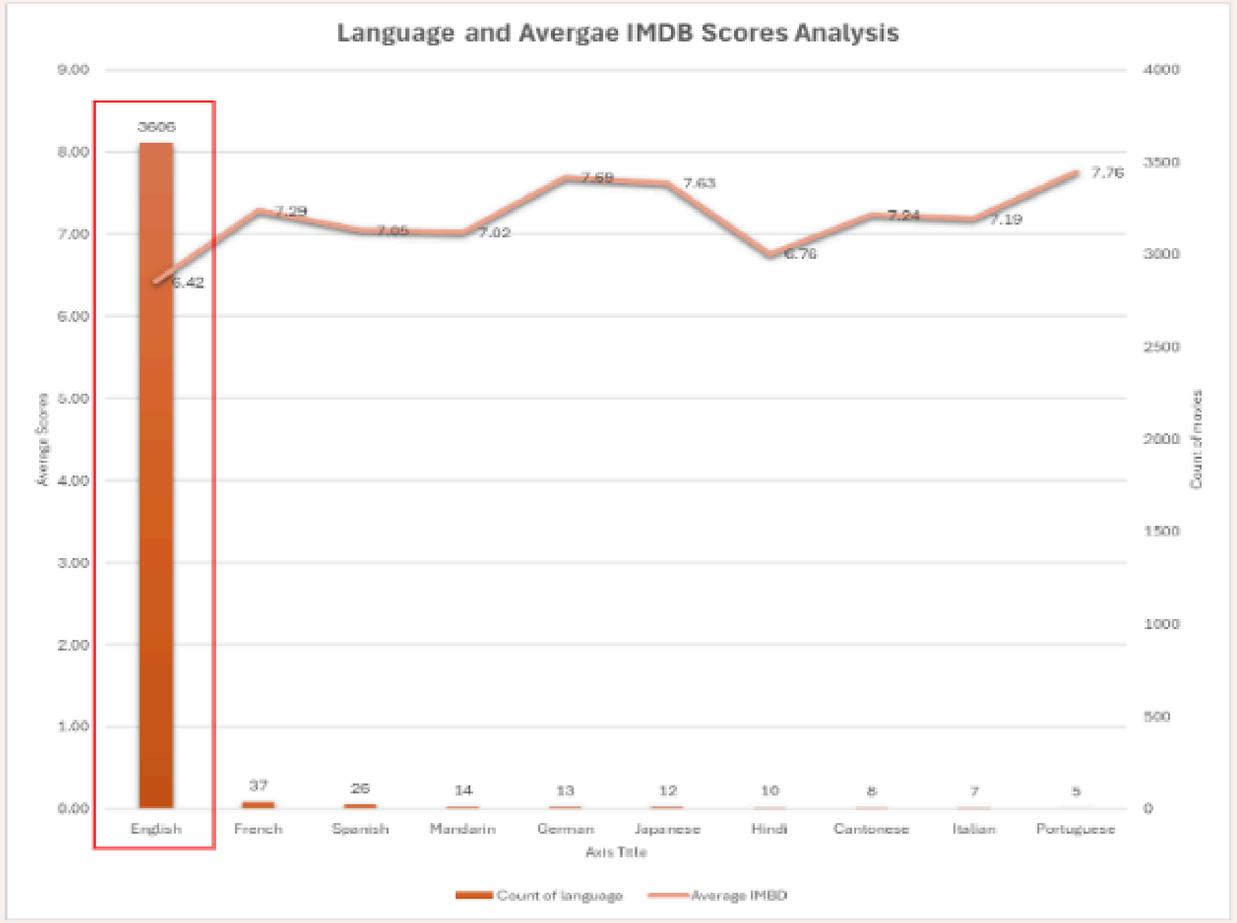
From the scatter plot we can see that the duration of 80 to 130 has got the maximum of films and the IMDB score lies between 4.5 –8.5.Shortest movie duration is of 34 min whereas longest movie duration is of 330 minutes.

PROJECT 05

IMDB Movie Analysis

TASK C LANGUAGE ANALYSIS:

DETERMINE THE MOST COMMON LANGUAGES USED IN MOVIES AND ANALYSE THEIR IMPACT ON THE IMDB SCORE USING DESCRIPTIVE STATISTICS



Here, we are looking at the top 10 languages with the highest count of movies. Majority of movies are in English with an average of 6.42 score , followed by French (7.29 score average).

Also, apart from these languages, Korean, German, Japanese, French, Cantonese, Spanish, Italian, and Mandarin have higher average IMDb ratings.

This is due to consistent audience and fewer movies in these language

PROJECT 05

IMDB Movie Analysis

TASK C - LANGUAGE ANALYSIS:

DETERMINE THE MOST COMMON LANGUAGES USED IN MOVIES AND ANALYSE THEIR IMPACT ON THE IMDB SCORE USING DESCRIPTIVE STATISTICS

Row Labels	Count of language	Average IMBD	Median IMBD	Std Dev.
English	3606	6.421436495	6.5	1.052498903
French	37	7.286486486	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.021428571	7.25	0.765786244
German	13	7.692307692	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.2375	7.3	0.440575922
Italian	7	7.185714286	7	1.155318962
Portuguese	5	7.76	8	1.05750842

Mathematical and arithmetical functions used:

1. Countif
2. Avergeif
3. Median
4. Std.dev

Also, we can see that average movie ratings are consistent across languages ranging from 6.4 –8.

Higher standard deviation leads to more variability. On the other hand, variance gives us an idea of how spread the data is across mean.

PROJECT 05

IMDB Movie Analysis

TASK D : DIRECTOR ANALYSIS:

IDENTIFY THE TOP DIRECTORS BASED ON THEIR AVERAGE IMDB SCORE AND ANALYSE THEIR CONTRIBUTION TO THE SUCCESS OF MOVIES USING PERCENTILE CALCULATIONS.

Top 5 directors with the highest average IMDB Scores and percentile

director_name	Average IMDB Score	Percentile
Charles Chaplin	8.6	0.999
Tony Kaye	8.6	0.999
Alfred Hitchcock	8.5	0.998
Damien Chazelle	8.5	0.998
Majid Majidi	8.5	0.998

Top 5 directors with the highest number of movies

Row Labels	Count of movie_title2
Steven Spielberg	25
Woody Allen	19
Clint Eastwood	19
Ridley Scott	16
Martin Scorsese	16

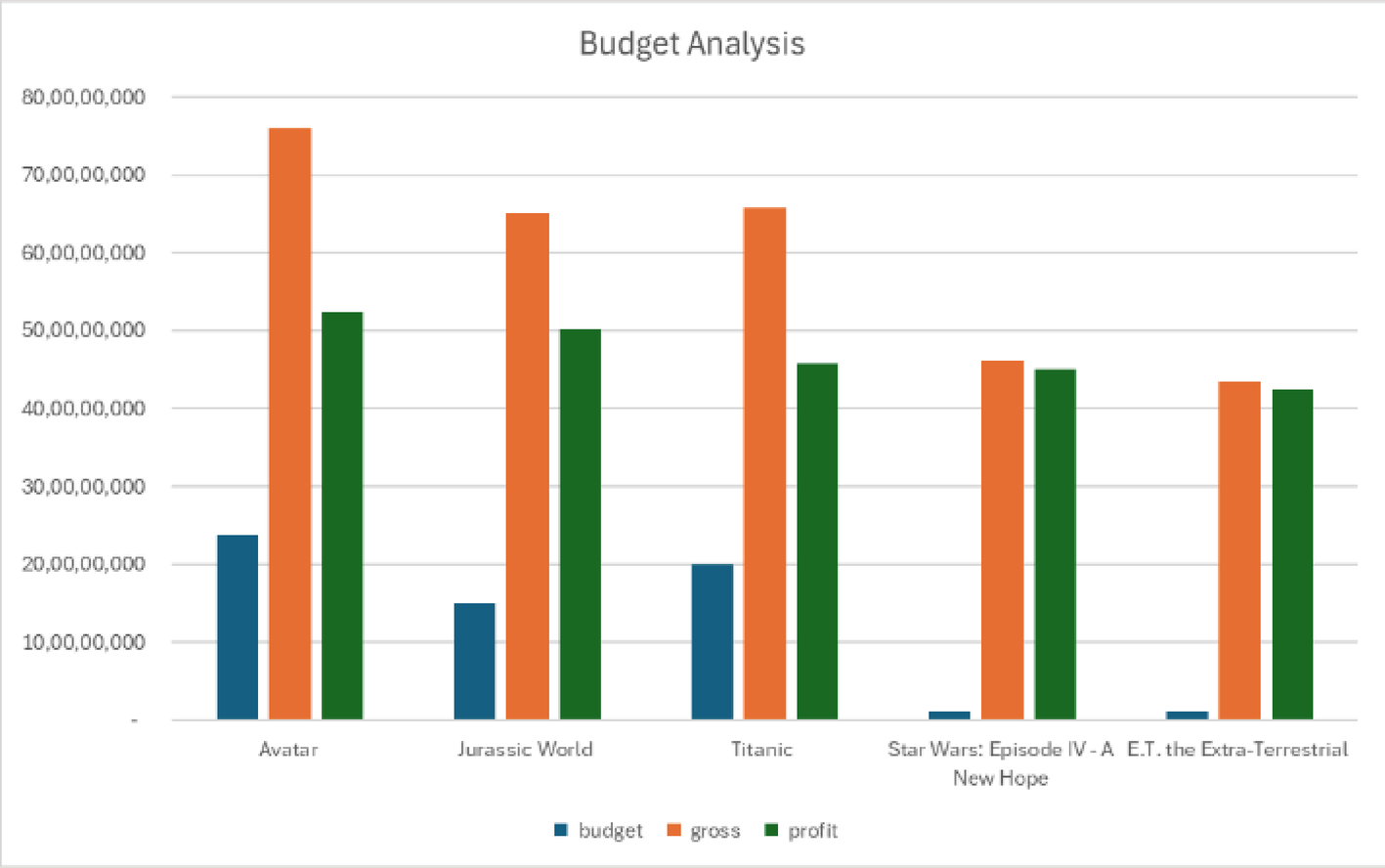
As we can see from the above, Charles Chaplin and Tony Kaye have the highest average IMDB scores. On the other hand, Steven Spielberg holds the record of most movies made with a total of 25 movies.
Now, here the percentile refers to each director's average IMDb score when it is compared against a common benchmark, to know their relative position in the dataset.

PROJECT 05

IMDB Movie Analysis

TASK E- BUDGET ANALYSIS:

ANALYSE THE CORRELATION BETWEEN MOVIE BUDGETS AND GROSS EARNINGS AND IDENTIFY THE MOVIES WITH THE HIGHEST PROFIT MARGIN.



Correlation Coefficient Formulae Used:
"= Correl()"

Result: 0.096 > Weak relation

5 movies with the highest profits are:

1. Avatar
2. Jurassic World
3. Titanic
4. Star wars
5. Extra Terrestrial

PROJECT 05

IMDB Movie Analysis

INTERPRETATIONS AND RECOMMENDATIONS

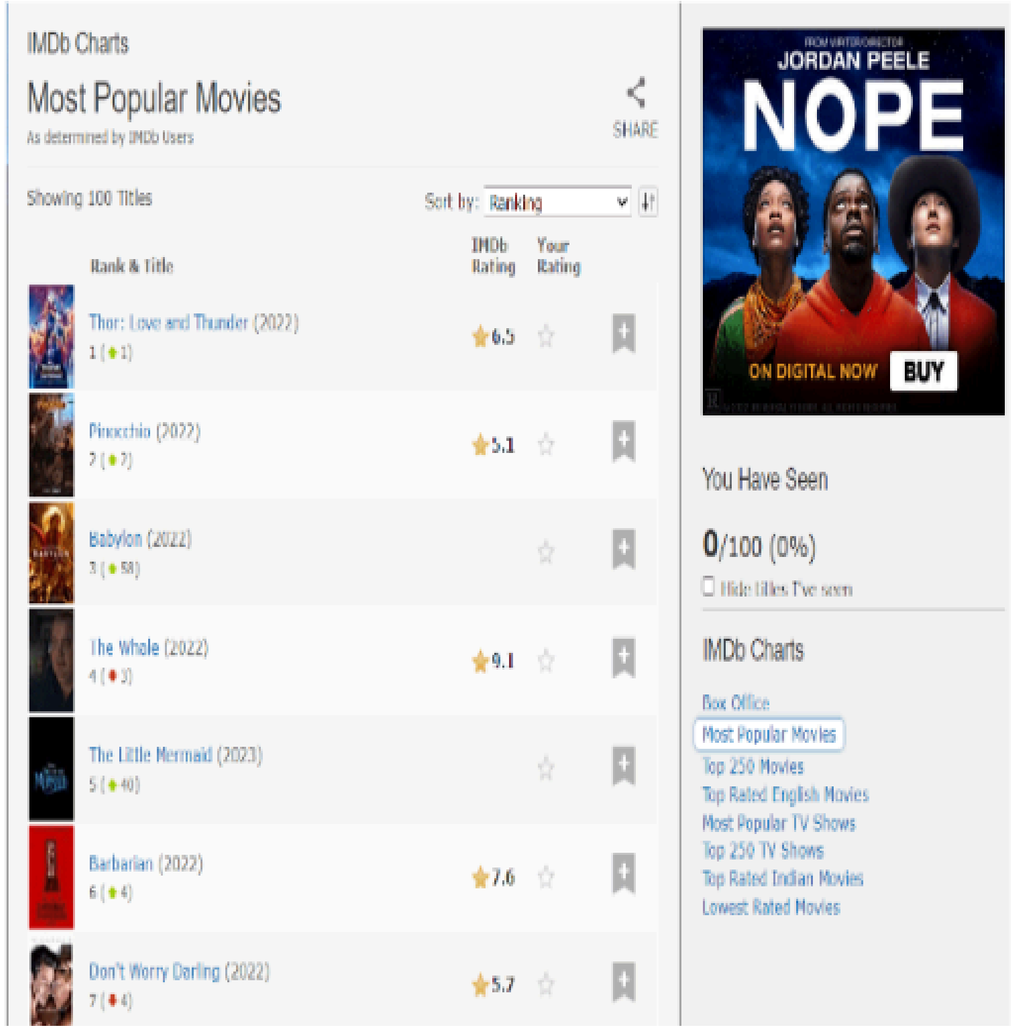
1. Genre Focus
   - Prioritize Drama and Biography genres.
   - Increase production in high-scoring, underrepresented genres like Western.

2. Optimize Duration
   - Aim for movie durations between 80 to 130 minutes.
   - Avoid extreme durations unless justified by content.

3. Language Diversification
   - Produce more movies in high-rated languages (e.g., German, Japanese, French).
   - Enhance non-English movies with subtitles and dubbing.

4. Director Collaboration
   - Partner with directors known for high IMDB scores.
   - Implement successful directors' styles and methods.

5. Budget Management
   - Balance movie budgets efficiently.
   - Study high-profit movies for factors contributing to success beyond budget.



PROJECT 06

Bank Loan Case Study

PROJECT 06

Bank Loan Case Study

This project aims to analyse loan application data using Exploratory Data Analysis (EDA) to identify factors influencing loan defaults among urban customers. The objectives are to predict potential defaults to mitigate financial risks and ensure capable applicants are not wrongly rejected. By identifying patterns and key factors, the company can make more informed loan approval decisions, balancing business growth with risk management.

- **Objectives**

1. Predict potential loan defaults to mitigate financial risks.
2. Ensure capable applicants are not wrongly rejected.
3. Identify key factors and patterns influencing loan defaults.

The provided dataset contains information about loan applications at the time of application and includes two types of scenarios:

1. Clients with payment difficulties: those who had late payments exceeding X days on at least one of the first Y installments of the loan in our sample.
2. All other cases: those where the payments were made on time.

- When a client applies for a loan, there are four possible outcomes:
- Approved
- Cancelled
- Refused
- Unused Offer

PROJECT 06

Bank Loan Case Study

MY APPROACH

1. Imported the loan application dataset into Microsoft Excel 2022 and conducted initial data cleaning.

2. Identified and imputed missing data using `COUNT`, `ISBLANK`, `IF`, `AVERAGE`, and `MEDIAN`, visualizing with bar charts.

3. Detected outliers using `QUARTILE`, `IQR`, and conditional formatting, visualizing with box plots.

4. Assessed data imbalance with `COUNTIF` and `SUM`, using pie and bar charts for visualization.

5. Performed univariate, segmented univariate, and bivariate analyses with pivot tables and filters, visualizing results with histograms, bar charts, box plots, scatter plots, and heatmaps; calculated and ranked correlations using `CORREL`.

TECH- STACK USED

Microsoft Excel 2022:

- Utilized for data cleaning, analysis, and visualization.
- Functions and features used: COUNT, ISBLANK, IF, AVERAGE, MEDIAN, QUARTILE, IQR, COUNTIF, SUM, CORREL, pivot tables, filters, sorting, conditional formatting, bar charts, pie charts, histograms, box plots, scatter plots, heatmaps.

PROJECT 06

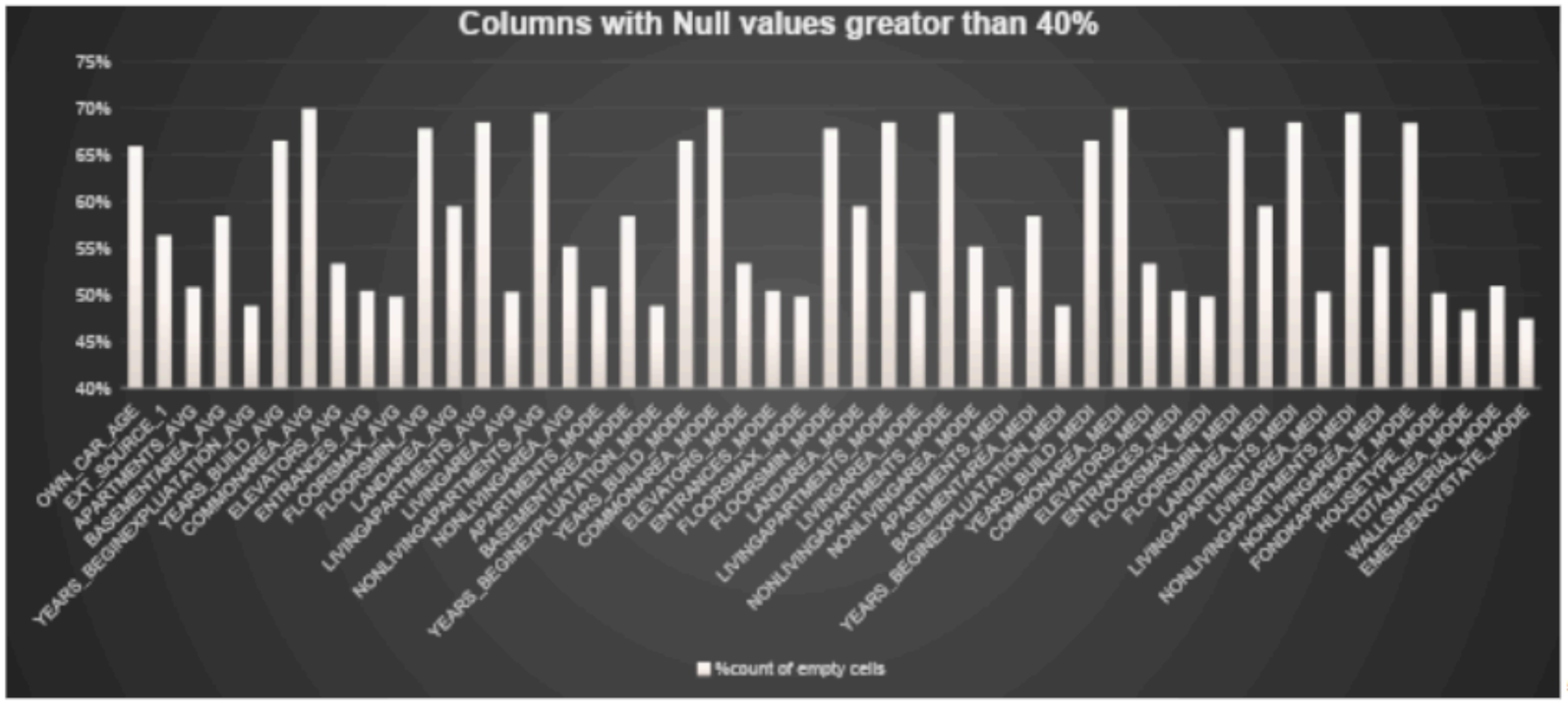
Bank Loan Case Study

Task A – Missing and Null values
File 1 – Application Data

Columns : 123

Rows: 5000

Columns with null value > 40% = 49



PROJECT 06

Bank Loan Case Study

Columns to Drop

- OWN_CAR_AGE
- EXT_SOURCE_1
- APARTMENTS_AVG
- BASEMENTAREA_AVG
- YEARS_BEGINEXPLUATATION_AVG
- YEARS_BUILD_AVG
- COMMONAREA_AVG
- ELEVATORS_AVG
- ENTRANCES_AVG
- FLOORSMAX_AVG
- FLOORSMIN_AVG
- LANDAREA_AVG
- LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG
- NONLIVINGAPARTMENTS_AVG
- NONLIVINGAREA_AVG
- APARTMENTS_MODE
- BASEMENTAREA_MODE
- YEARS_BEGINEXPLUATATION_MODE
- YEARS_BUILD_MODE
- COMMONAREA_MODE
- ELEVATORS_MODE
- ENTRANCES_MODE
- FLOORSMAX_MODE
- FLOORSMIN_MODE
- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MODE
- NONLIVINGAPARTMENTS_MODE
- NONLIVINGAREA_MODE
- APARTMENTS_MEDI
- BASEMENTAREA_MEDI
- YEARS_BEGINEXPLUATATION_MEDI
- YEARS_BUILD_MEDI
- COMMONAREA_MEDI
- ELEVATORS_MEDI
- ENTRANCES_MEDI
- FLOORSMAX_MEDI
- FLOORSMIN_MEDI
- LANDAREA_MEDI
- LIVINGAPARTMENTS_MEDI
- LIVINGAREA_MEDI
- NONLIVINGAPARTMENTS_MEDI
- NONLIVINGAREA_MEDI
- FONDKAPREMONT_MODE
- HOUSETYPE_MODE
- TOTALAREA_MODE
- WALLSMATERIAL_MODE
- EMERGENCYSTATE_MODE

PROJECT 06

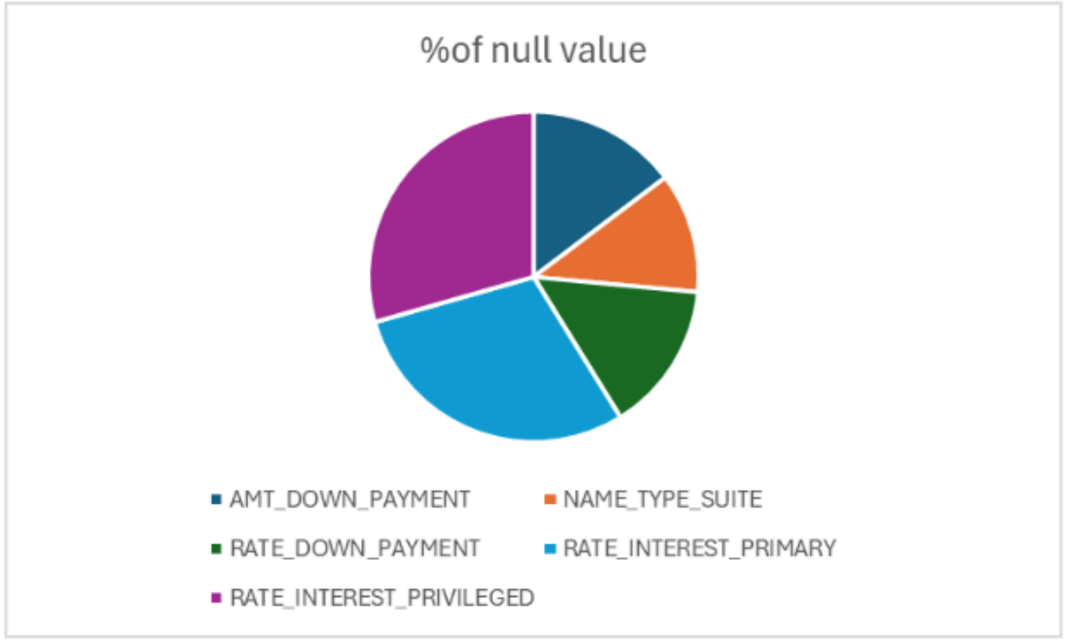
Bank Loan Case Study

Task A – Missing and Null values
File 2 – Previous Application Data

Columns : 37

Rows: 5000

Columns with null value > 40% = 5



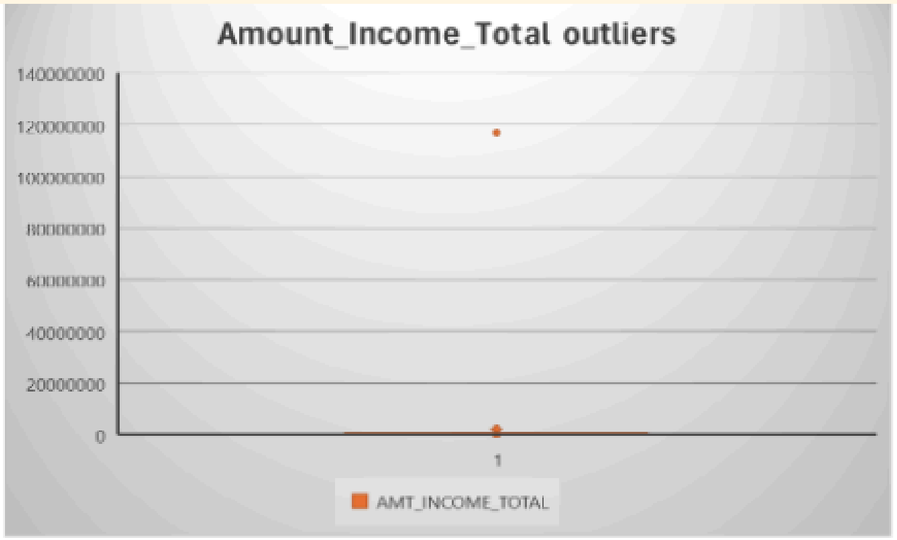
Columns to Drop

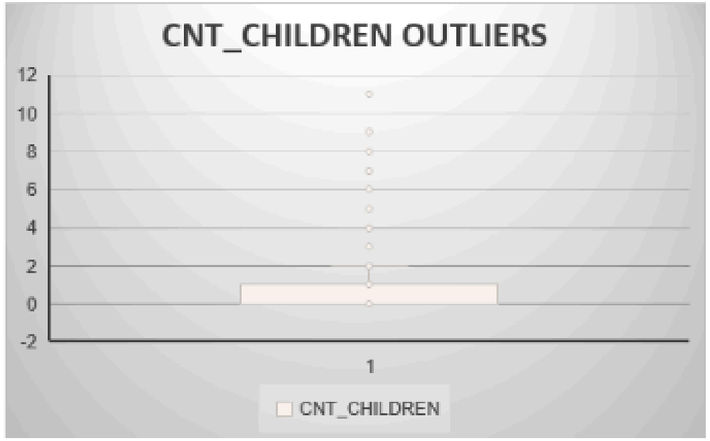
- AMT_DOWN_PAYMENT
- NAME_TYPE_SUITE
- RATE_DOWN_PAYMENT
- RATE_INTEREST_PRIMARY
- RATE_INTEREST_PRIVILEGED

PROJECT 06

Bank Loan Case Study

Task B – Outliers

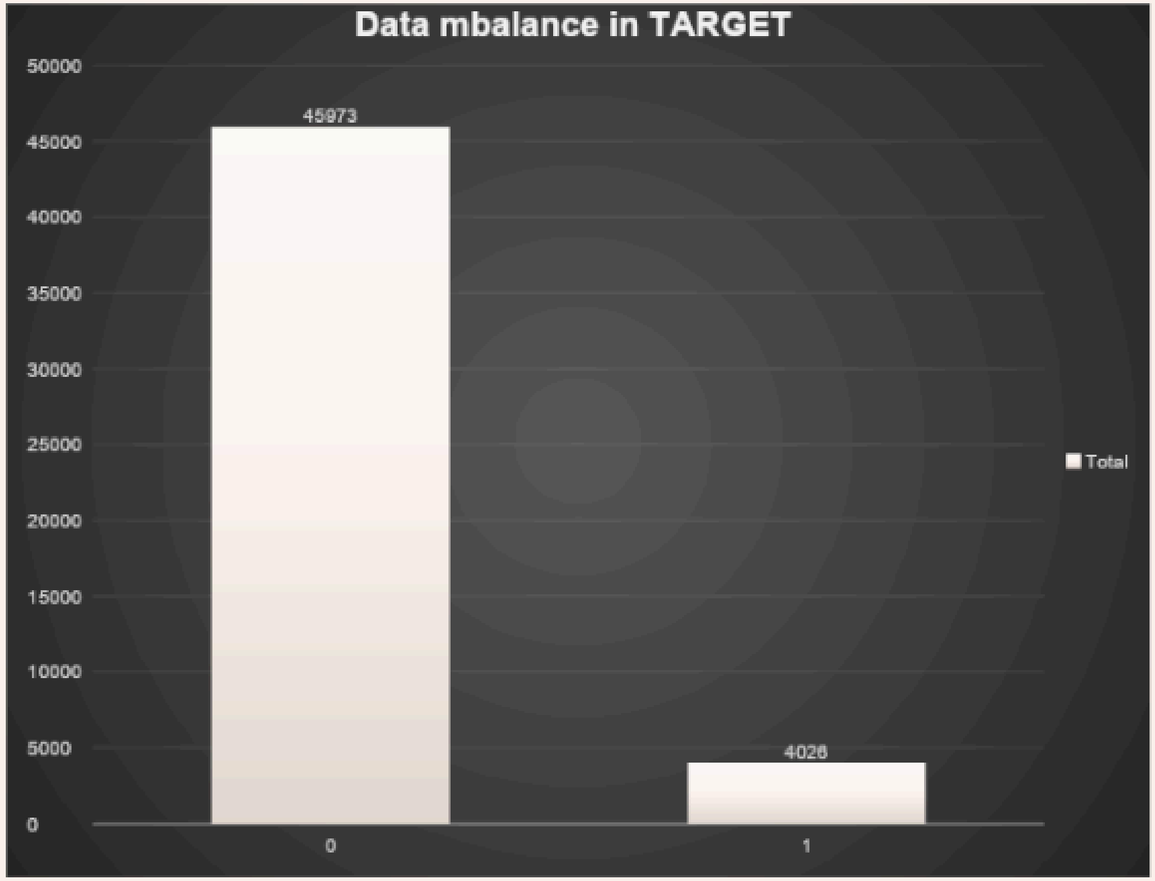




PROJECT 06

Bank Loan Case Study

Task C – Data Imbalance



We see that we have :

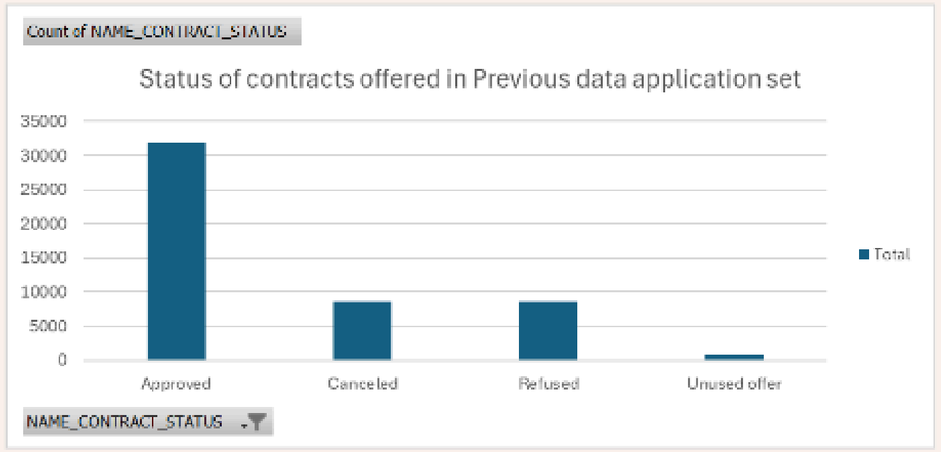
- 91% as loan re-payers
- 9% as Defaulters

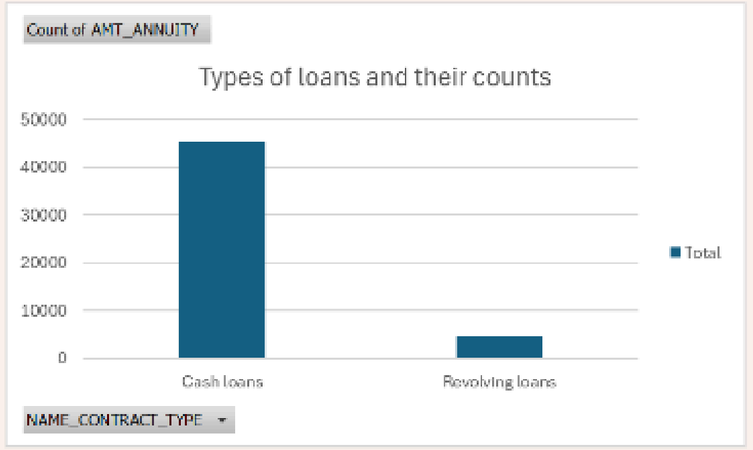
Which gives us a clear indication that the
da ta is highly imbalanced

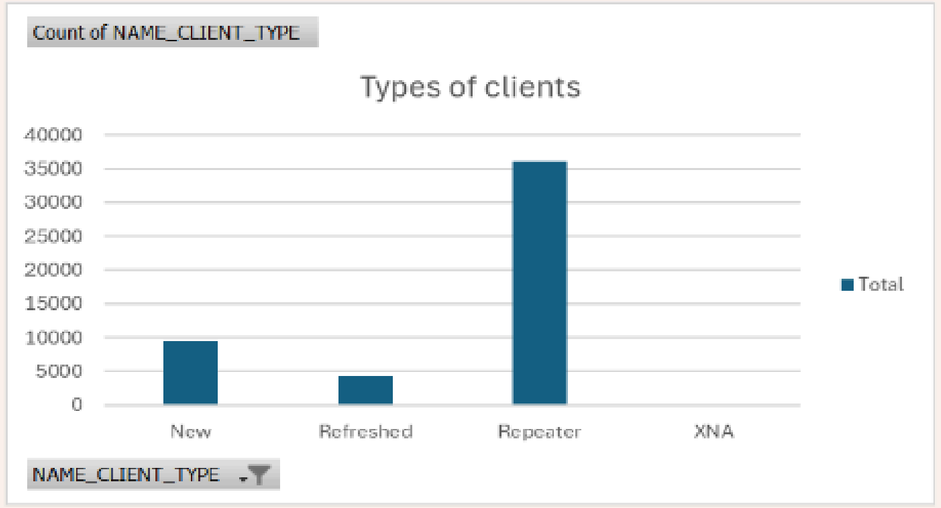
PROJECT 06

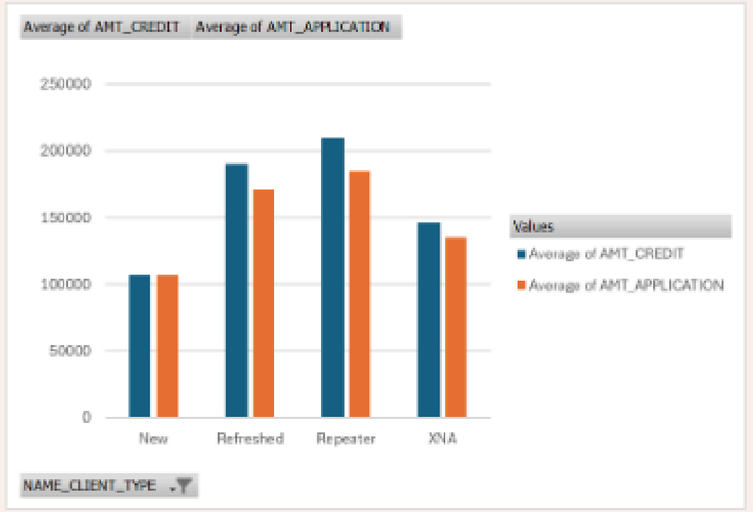
Bank Loan Case Study

Task D – Univariate, Segmented Univariate, and Bivariate Analysis:









PROJECT 06

Bank Loan Case Study

Task E – Correlations

Target 1 Correlations

	Correlations					
CNT_CHILDREN	1	0.01	0.00	-0.33	-0.24	0.03
AMT_INCOME_TOTAL	0.01	1	0.07	-0.02	-0.03	-0.04
AMT_CREDIT	0.00	0.07	1	0.06	-0.07	-0.10
YEARS_BIRTH	-0.33	-0.02	0.06	1	0.62	-0.02
YEARS_EMPLOYED	-0.24	-0.03	-0.07	0.62	1	0.03
REGION_RATING_CLIENT	0.03	-0.04	-0.10	-0.02	0.03	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	YEARS_BIRTH	YEARS_EMPLOYED	REGION_RATING_CLIENT

Target 0 Correlations

CNT_CHILDREN	1	0.010	0.005	-0.329	-0.242	0.026
AMT_INCOME_TOTAL	0.010	1	0.069	-0.016	-0.032	-0.038
AMT_CREDIT	0.005	0.069	1	0.059	-0.068	-0.101
YEARS_BIRTH	-0.329	-0.016	0.059	1	0.622	-0.017
YEARS_EMPLOYED	-0.242	-0.032	-0.068	0.622	1	0.035
REGION_RATING_CLIENT	0.026	-0.038	-0.101	-0.017	0.035	1
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	YEARS_BIRTH	YEARS_EMPLOYED	REGION_RATING_CLIENT

PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Overview

This project investigates how various car features influence pricing and profitability in the automotive industry. With shifts in consumer preferences towards sustainability and technology, understanding these dynamics is crucial for manufacturers.

Business Problem

The primary question is: **How can car manufacturers optimize pricing and product development to maximize profitability while addressing consumer demand?** This involves identifying key features that drive consumer interest and understanding their impact on pricing strategies.

PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Data Sources

The analysis utilizes the "Car Features and MSRP" dataset, featuring:

- **Observations**: 11,159 entries
- **Variables**: 16 attributes capturing car features such as engine power, fuel efficiency, and manufacturer information.

This dataset was sourced from Kaggle.

Data Cleaning and Preprocessing

Key steps included:

- **Removal of Duplicates**
- **Handling Missing Values**
- **Standardization** of categorical variables

These measures ensured a clean dataset suitable for analysis.

Assumptions

The dataset is assumed to be representative of market trends, and consumer preferences have not dramatically shifted since its last update.

PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Analytical Methods

The project employs:

- **Descriptive Statistics**
- **Regression Analysis**
- **Data Visualization**

Reasoning Behind Methods

These methods help understand relationships within the dataset and predict pricing based on key features, facilitating informed decision-making.

Challenges Encountered

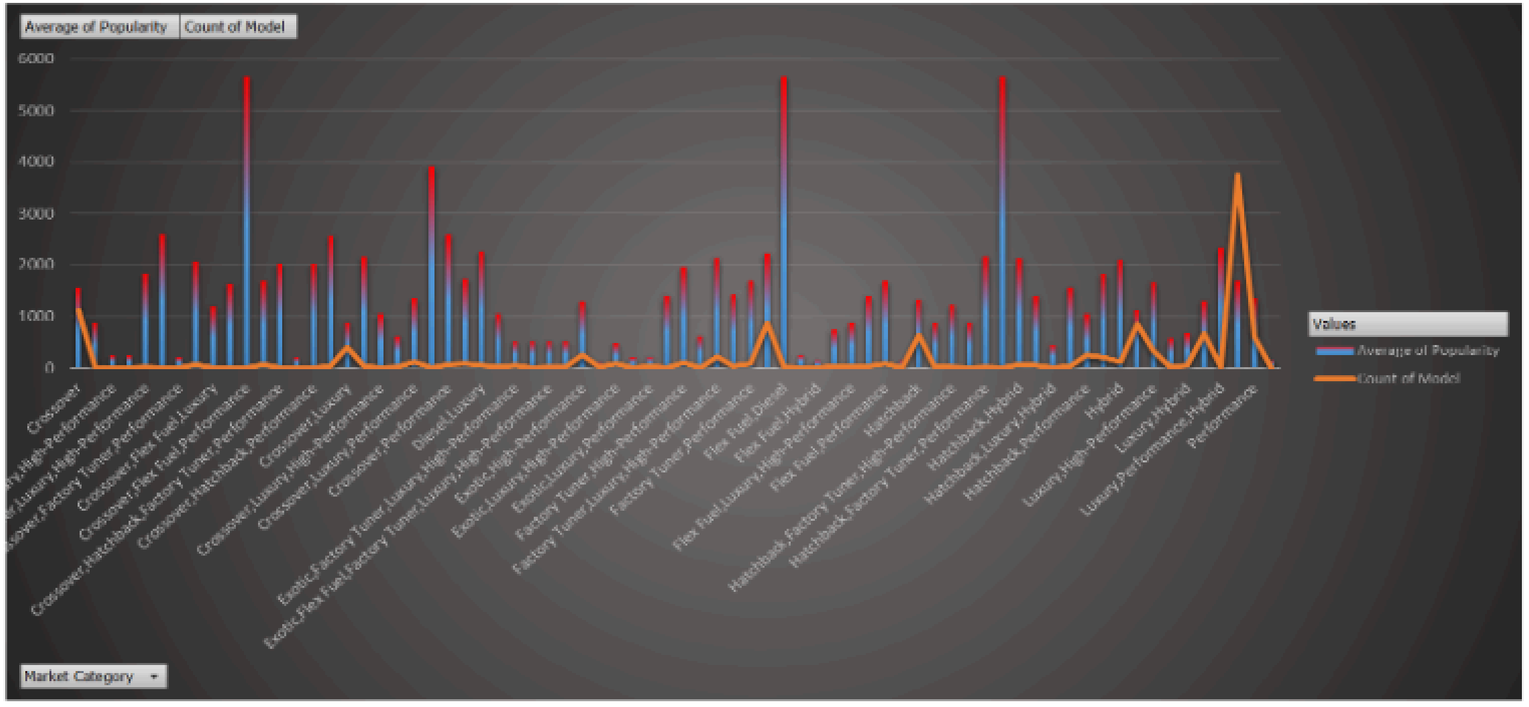
- **Data Age**: The dataset may not fully reflect current market conditions.
- **Biases in Popularity Rankings**: Potential biases could affect analysis outcomes.

PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Task 1: Market Category Analysis

1. **High Popularity**: Categories like "Luxury" and "High-Performance" have higher average popularity, indicating strong consumer preference.
2. **Model Count Variation**: Categories with many models (e.g., "Crossover") may attract more consumer interest, suggesting diversity boosts popularity.
3. **Strategic Focus**: Manufacturers should enhance features and marketing in high-popularity categories to maximize profitability.



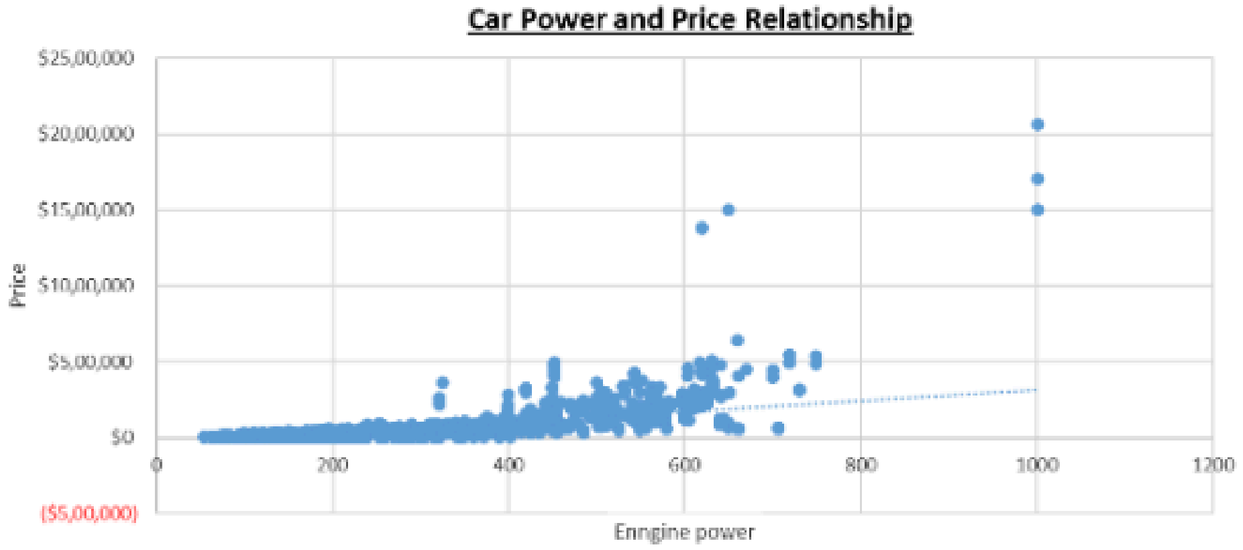
PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Task 2: Engine Power and Price Relationship

The scatter plot illustrates the relationship between engine power and car price:

1. **Trend Observation**: As engine power increases, car prices generally show a slight upward trend, indicating that more powerful cars tend to be priced higher.
2. **Data Distribution**: Most data points cluster at lower engine power levels (below 400 HP), with very few high-powered cars (above 600 HP) commanding significantly higher prices.
3. **Outliers**: There are a few outlier points with exceptionally high prices, suggesting that certain high-performance models can dramatically increase average price.



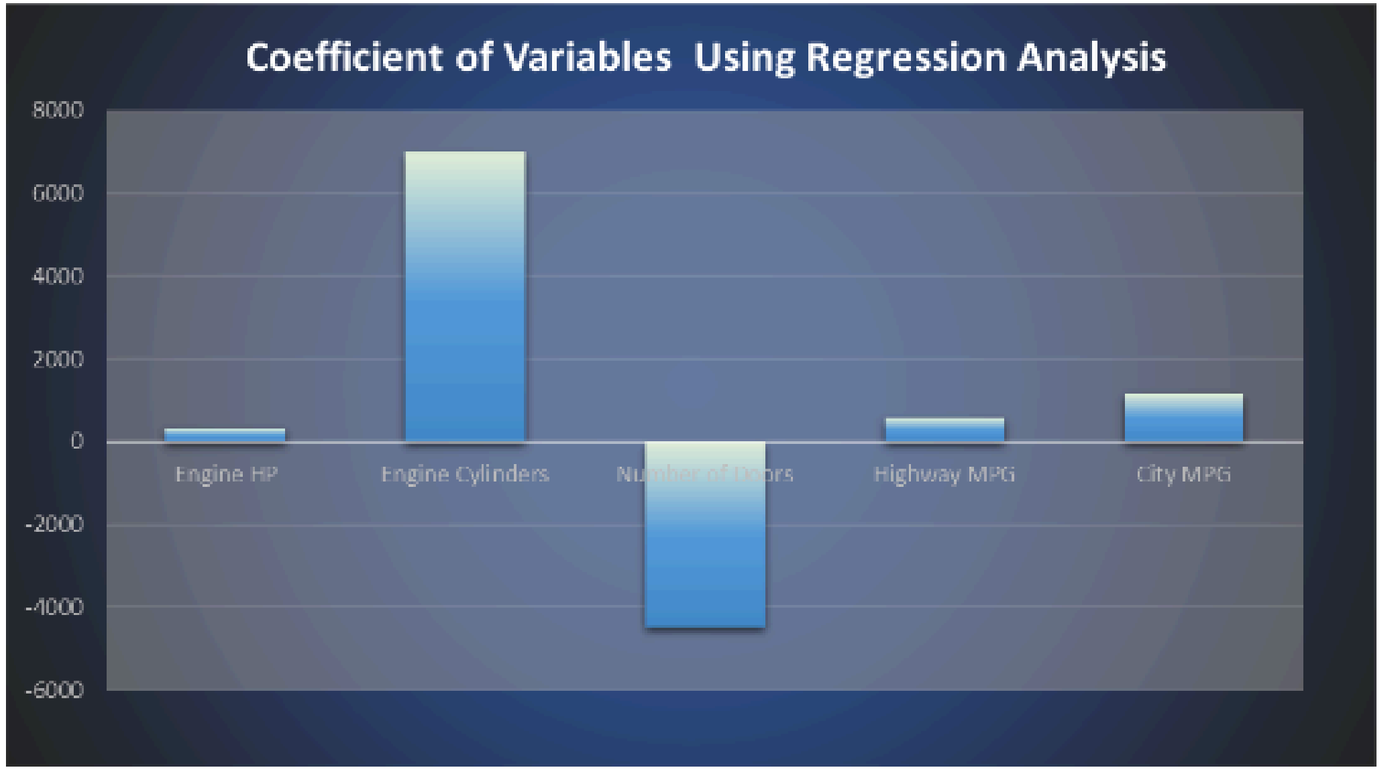
PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Task 3: Regression Analysis of Car Features

The bar chart displays the coefficients of various variables identified through regression analysis, indicating their impact on car price:

1. **Most Influential Variables**:
   - **Engine Cylinders**: Shows the highest positive coefficient, suggesting that an increase in the number of cylinders significantly raises the car price.
   - **Engine HP**: While influential, its impact is less pronounced compared to engine cylinders.
2. **Moderate Impact**:
   - **Number of Doors**: Has a positive coefficient, indicating that more doors may contribute to a higher price, but the effect is smaller than that of engine cylinders and horsepower.
3. **Negative Impact**:
   - **Highway MPG and City MPG**: Both have negative coefficients, suggesting that higher fuel efficiency may correlate with lower prices, possibly indicating a market preference for more powerful vehicles over fuel-efficient ones.



PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Task 4: Manufacturer Pricing Comparison

The horizontal bar chart illustrates the average MSRP (Manufacturer's Suggested Retail Price) for different car manufacturers:

1. **High-End Manufacturers**:
   - Brands like **Rolls-Royce**, **McLaren**, and **Maybach** show significantly higher average prices, indicating their positioning in the luxury market.
2. **Mid to Low-End Brands**:
   - Manufacturers such as **Volkswagen**, **Subaru**, and **Tesla** have lower average MSRPs, reflecting a broader market appeal and potentially higher sales volume.
3. **Extreme Values**:
   - **Bugatti** and **Aston Martin** represent the upper echelon of pricing, catering to ultra-luxury consumers, while brands like **Acura** and **GMC** fall into more accessible pricing categories.
4. **Market Insights**:
   - The data suggests that manufacturers targeting luxury segments tend to command significantly higher prices, while those in the mainstream market offer more affordability.

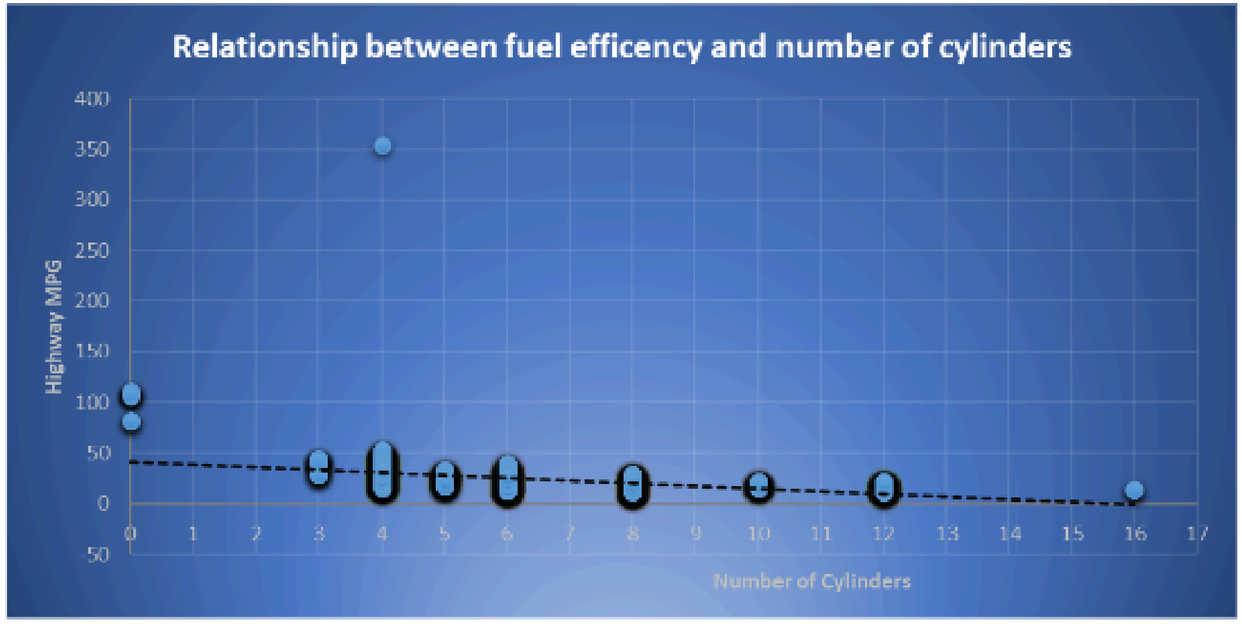
PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Task 5: Fuel Efficiency vs. Number of Cylinders

The scatter plot depicts the relationship between fuel efficiency (measured in Highway MPG) and the number of cylinders in cars:

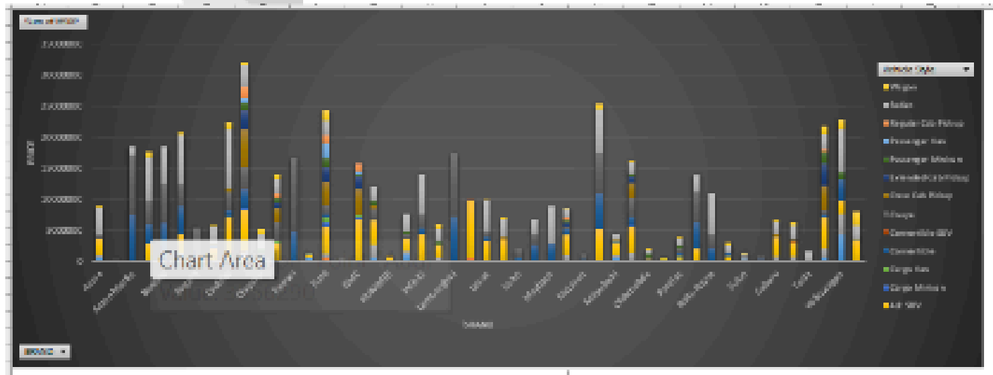
1. **Negative Correlation**:
   - There is a clear downward trend, indicating that as the number of cylinders increases, highway fuel efficiency tends to decrease. This suggests that more cylinders generally lead to higher fuel consumption.
2. **Low Fuel Efficiency for High Cylinder Counts**:
   - Vehicles with higher cylinder counts (8 or more) show significantly lower MPG, reinforcing the idea that larger engines are less efficient in terms of fuel economy.
3. **Outliers**:
   - A few points with unusually high MPG are present, likely indicating specialized vehicles or technologies that defy the general trend.

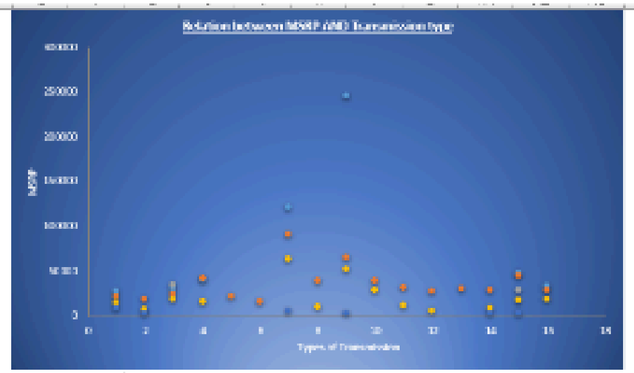


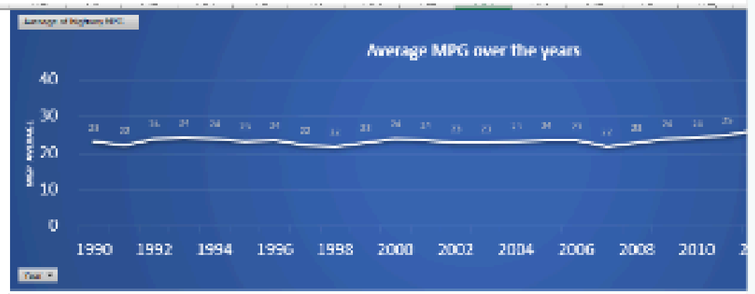
PROJECT 07

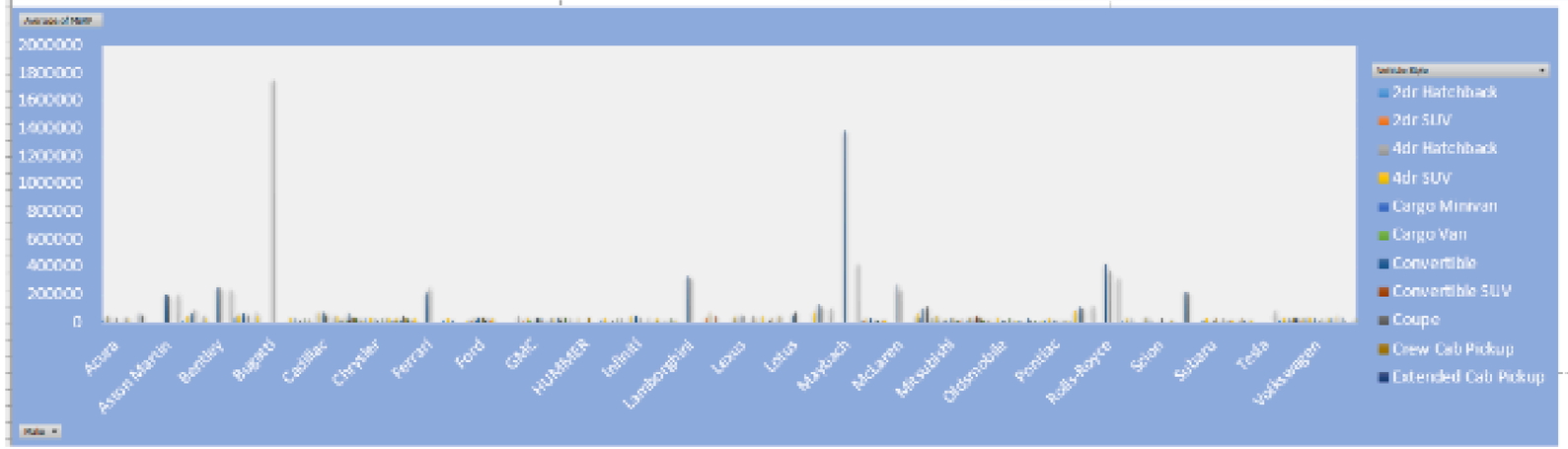
Analyzing the Impact of Car Features on Price and Profitability

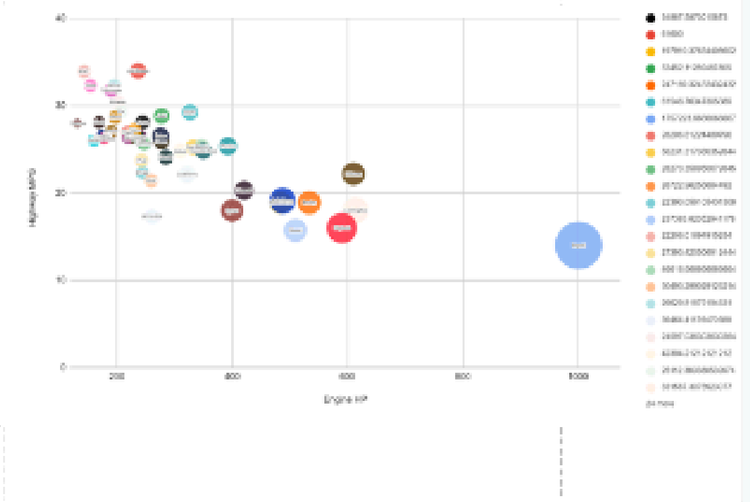
Dashboard











PROJECT 07

Analyzing the Impact of Car Features on Price and Profitability

Overall Recommendations

1. **Expand Luxury and SUV Offerings**:
   - Focus on developing high-end models and SUVs to align with growing consumer demand for luxury and diverse vehicle styles.
2. **Enhance Fuel Efficiency**:
   - Invest in research and development for fuel-efficient technologies to meet regulatory standards and consumer preferences for sustainability.
3. **Leverage Data-Driven Marketing**:
   - Utilize insights from the analysis to tailor marketing strategies that emphasize vehicle performance, fuel efficiency, and luxury features.
4. **Monitor Competitive Landscape**:
   - Regularly analyze competitor offerings and market trends to identify gaps and opportunities, ensuring a proactive approach to product development.
5. **Educate Consumers**:
   - Provide information on the benefits of advanced features (e.g., transmission options, fuel efficiency) to help consumers make informed purchasing decisions.
6. **Focus on Technological Innovation**:
   - Prioritize advancements in technology, particularly in transmission systems and fuel management, to differentiate products in a competitive market.
7. **Adapt to Consumer Preferences**:
   - Stay agile in responding to shifts in consumer preferences, particularly towards electric and hybrid vehicles, by expanding offerings in these categories.

PROJECT 08

ABC Call Volume Trend

PROJECT 08

ABC Call Volume Trend

Project Description

Overview

The project aimed to improve customer service at ABC Insurance Company by analyzing call center operations. Key focus areas included:

1. **Call Volume Analysis**: Identifying patterns and peak times for incoming calls.
2. **Call Duration Assessment**: Evaluating average call lengths to find efficiency gaps.
3. **Staffing Requirements**: Calculating the minimum number of agents needed to ensure calls are answered promptly.
4. **Night Shift Planning**: Addressing the lack of agent availability during nighttime hours.

Business Problem

1. **High Abandonment Rates**: Approximately **30%** of calls were abandoned, particularly during peak hours, leading to customer dissatisfaction.
2. **Suboptimal Staffing**: Inadequate staffing during high-demand periods resulted in longer wait times and unanswered calls.
3. **Variable Call Handling Times**: Inconsistent durations affected efficiency, causing delays for subsequent callers.
4. **Lack of Night Coverage**: Insufficient agent availability from **9 PM to 9 AM** led to missed customer interactions and poor service experiences.

PROJECT 08

ABC Call Volume Trend

Data Sources

- **Call Data**: Information on call volumes, durations, and statuses (answered, abandoned, transferred).
- **Operational Metrics**: Agent availability, handling times, and occupancy rates.

Data Cleaning and Preprocessing

- **Normalization**: Standardized call durations and categorized call statuses.
- **Missing Values**: Handled missing or inconsistent data entries.
- **Aggregation**: Summarized data into time buckets for analysis.

PROJECT 08

Approach

ABC Call Volume Trend

Analytical Methods

- **Descriptive Analytics**: Analyzed call volume trends and abandonment rates.
- **Predictive Modeling**: Estimated required staffing levels based on historical call data.
- **Visualizations**: Created charts and graphs to illustrate call patterns and agent requirements.

Challenges Encountered

- **Data Quality**: Inconsistent data entries and missing values required extensive cleaning.
- **Dynamic Call Patterns**: Variability in call volumes made it difficult to predict staffing needs accurately.
- **Resource Allocation**: Balancing agent workload while ensuring customer satisfaction posed logistical challenges.
- **Resistance to Change**: Implementing new staffing strategies and training programs met with some organizational pushback.

PROJECT 08

ABC Call Volume Trend

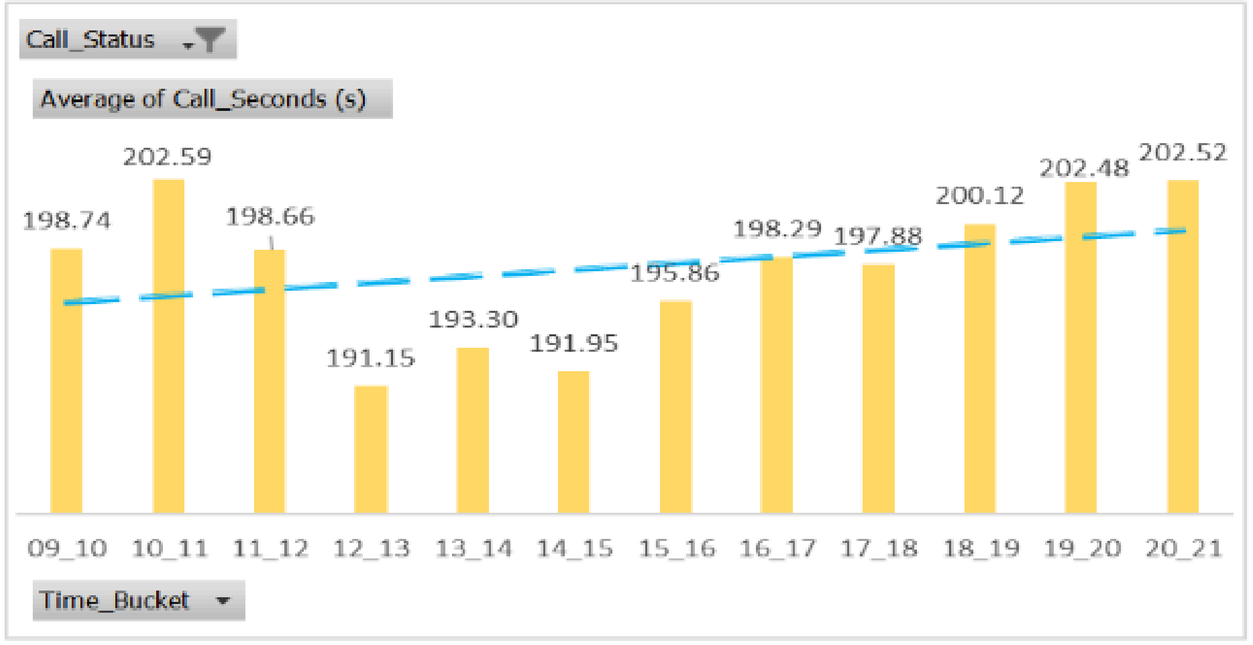
Task 1: **Average Call Duration:** Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket

Your Task: What is the average duration of calls for each time bucket?

1. Excel Functions Used: Pivot Table, Filter, Bar chart, Trendline
2. Insights and Interpretation : **Duration Range**: Average call durations range from **198.74 seconds** (09-10 bucket) to **202.52 seconds** (20-21 bucket), indicating longer calls in the evening.
3. **Trend**: The bar chart shows a general increase in call duration throughout the day, suggesting more complex inquiries as the day progresses.
4. **Actionable Insights**:
   a. **Staffing**: Allocate more agents during peak evening hours to manage longer calls effectively.
   b. **Training**: Provide targeted training to handle complex inquiries that contribute to longer call times.

Call_Status	(Multiple Items)

Time Bucket	Average of Call_Seconds (s)
09_10	198.74
10_11	202.59
11_12	198.66
12_13	191.15
13_14	193.30
14_15	191.95
15_16	195.86
16_17	198.29
17_18	197.88
18_19	200.12
19_20	202.48
20_21	202.52
Grand Total	**196.96**



PROJECT 08

ABC Call Volume Trend

Task 2: Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).

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

Insights and Interpretation

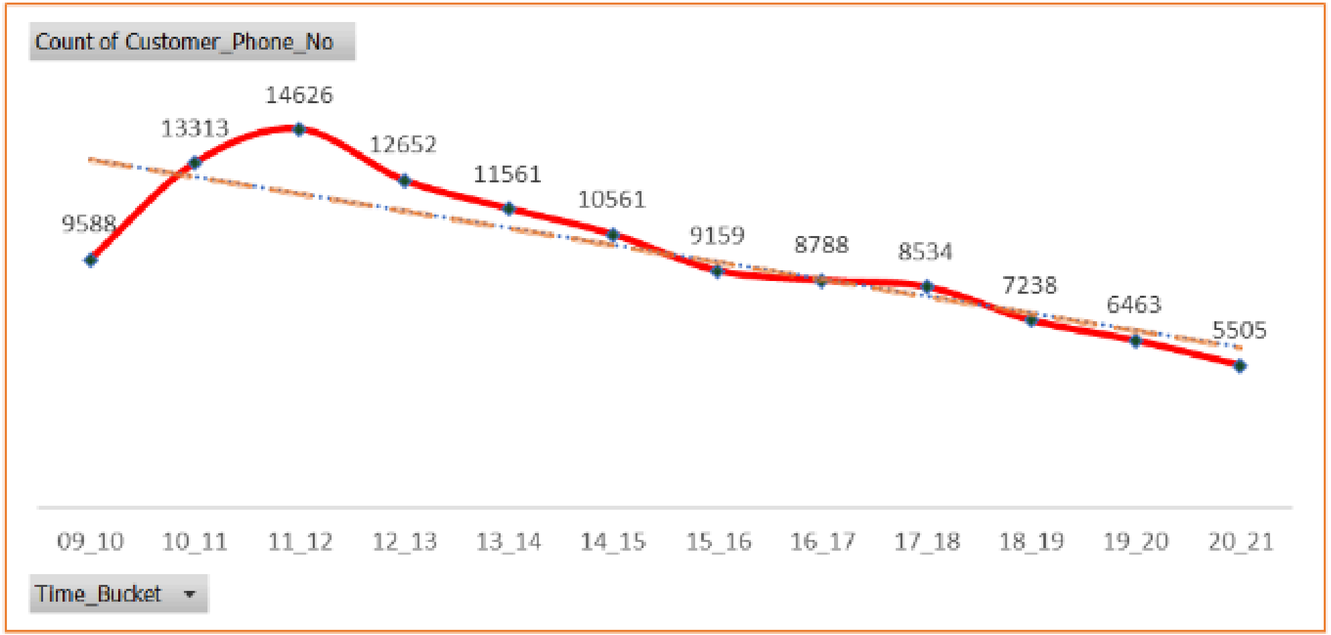
1. **Total Calls**: The total number of calls received is **117,988**, with varying volumes across different time buckets.
2. **Peak Call Times**: The highest call volume occurs in the **12-1 PM** bucket (**14,626 calls**), indicating this as a peak period for customer inquiries.
3. **Trend Observation**: The line chart shows a general decline in call volume after the peak, suggesting fewer customer interactions later in the day.

Actionable Insights:

- **Staffing Strategy**: Increase agent availability during peak hours (12-1 PM) to manage high call volumes effectively.
- **Monitoring Off-Peak**: Assess staff allocation during off-peak hours to optimize resources and reduce costs.

Excel Functions Used: Pivot Table, Filter, Line Chart , Trendline

Time Bucket	Count of Customer_Phone_No
09_10	9588
10_11	13313
11_12	14626
12_13	12652
13_14	11561
14_15	10561
15_16	9159
16_17	8788
17_18	8534
18_19	7238
19_20	6463
20_21	5505
Grand Total	**117988**



PROJECT 08

ABC Call Volume Trend

Task 2: **Call Volume Analysis:** Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).

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

1. **Total Calls**: **117,988**
   a. **Answered**: **82,452**
   b. **Abandoned**: 34,403
   c. **Transferred**: 1,133

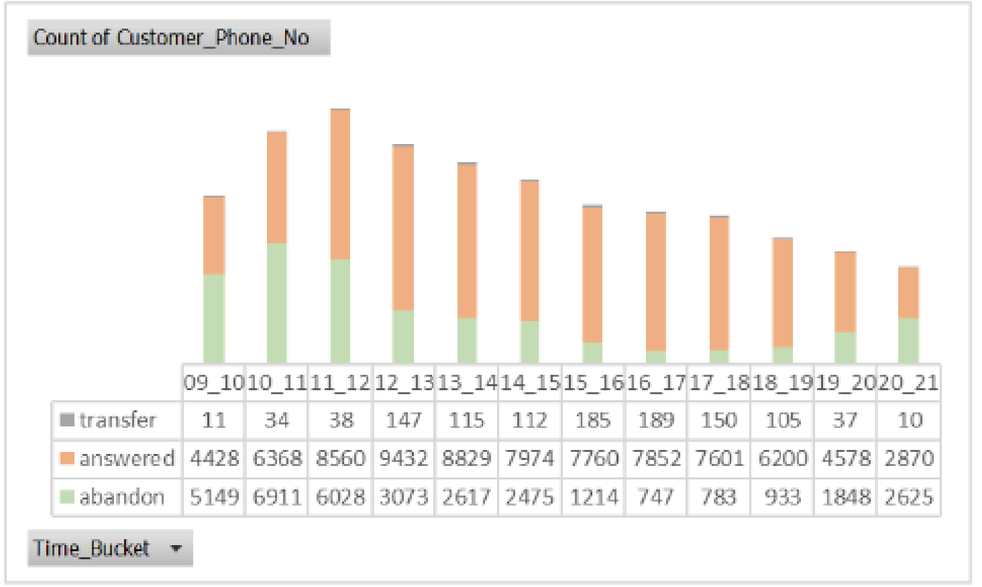
Key Insights

1. **High Abandonment Rates**: Significant abandonment in early morning buckets (e.g., **5,149** calls from 09-10).
2. **Peak Times**: Highest call volume at **12-1 PM** (14,626 total calls), indicating challenges in managing high demand.

Actionable Insights

- **Optimize Staffing**: Improve response times by increasing staffing during peak hours.
- **Investigate Abandonment**: Analyze causes of high abandonment rates for operational improvements.
- **Enhance Training**: Provide targeted training for agents during high-demand periods to improve efficiency.

Count of Custo	Call Status			
Time Bucket	abandon	answered	transfer	Grand Total
09_10	5149	4428	11	9588
10_11	6911	6368	34	13313
11_12	6028	8560	38	14626
12_13	3073	9432	147	12652
13_14	2617	8829	115	11561
14_15	2475	7974	112	10561
15_16	1214	7760	185	9159
16_17	747	7852	189	8788
17_18	783	7601	150	8534
18_19	933	6200	105	7238
19_20	1848	4578	37	6463
20_21	2625	2870	10	5505
Grand Total	**34403**	**82452**	**1133**	**117988**



PROJECT 08

ABC Call Volume Trend

Task 3: Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.

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

Row Labels	Count of Call_Status	Count of Call_Status
abandon	34403	29.16%
answered	82452	69.88%
transfer	1133	0.96%
Grand Total	117988	100.00%

Agent's activties	Duration
Total Call Incoming (9am-9pm)	117988
Number of Calls Handled	82452
Gap	35536
Working Hour of Each Agent	9
Average Call Handling Time(s)	196
Occupancy on Average	60%

- **Current Abandon Rate**: **30%**, goal to reduce to **10%** (90 calls answered per 100).
- **Agent Capacity**: Each agent can handle about **99 calls**.
- **Minimum Agents Required**: **1,590 agents** needed to meet demand and reduce abandonment.

Actionable Insights

- **Increase Staffing**: Allocate enough agents in each time bucket to reduce abandonment.
- **Shift Planning**: Implement shifts to cover peak times effectively.
- **Monitor Trends**: Continuously adjust staffing based on call volume data.

Lets Assume that an agent at the call centre works for 6 days a week.

So, on an average the total unplanned leaves per agent would be 4 days a month. Aditionally, an agent works for a total of 9 hours. During which they give about 1.5/2 hours for lunch and evening breaks at th ecall centre.

So we could say that they must be occupies by 60% of thei entire time at the centre, that is 7.5/8 hours is the time spent by them to talk to the customers and the clients.

Agent	Formulaes	Values
Call handling Capacity	$\frac{(\text{working time of agent in seconds})(\text{occupancy})}{(\text{Average Call Handling Time})}$	99.18367347
minimum agents required	$\frac{\text{Total Incoming Calls}}{\text{Call Handling Capacity}}$	1189.590947
Head Count Required	$\frac{\text{Minimum Agents Required}}{1 - \text{Shrinkage Percentage}}$	1586.121262
Man power in each time bucket	Head count/ 12	132.1767718

PROJECT 08

ABC Call Volume Trend

Task 4: Night Shift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am. The distribution of these 30 calls is as follows:

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

Total Incoming Calls in 9am to 9pm	117988
Given that calls betweem 9pm to 9 am is 30% of calls between 9am to 9pm	
Total Incoming calls in 9pm to 9am	35396
Call Handling Capacity	99.18367347
minimum agents required	356.877284
head count required	475.8363786
Man power in each time bucket	39.65303155

Agent	Formulaes	Values
Call handling Capacity	$\frac{(\text{working time/agent/month})(\text{occupancy})}{(\text{Average Call Handling Time})}$	99.18367347
minimum agents required	$\frac{\text{Total Incoming Calls}}{\text{Call Handling Capacity}}$	356.877284
Head Count Required	$\frac{\text{Minimum Agents Required}}{1 - \text{Shrinkage Percentage}}$	475.8363786
Man power in each time bucket	Head count/ 12	39.65303155

Time Bucket	
9_10	3540
10_11	3540
11_12	2360
12_1	2360
1_2	1180
2_3	1180
3_4	1180
4_5	1180
5_6	3540
6_7	4720
7_8	4720
8_9	5899
Grand Total	35396

1. Excel Functions Used: Pivot Table, Filter, Bar chart, Trendline
2. Insights and Interpretation :
3. **Night Call Volume**: Total of **35,396 calls** from **9 PM to 9 AM**.
4. **Agent Requirements**: Need **356 agents** to maintain a **10% abandonment rate**.
5. **Call Distribution**: Highest volume in early hours and peak at **5,899 calls** between **8-9 AM**.

Actionable Insights

- **Night Staffing**: Ensure adequate agents are available during peak night hours.
- **Monitor Patterns**: Analyze call trends to dynamically adjust staffing.
- **Training Focus**: Train agents for late-night calls to enhance service quality.

PROJECT 08

ABC Call Volume Trend

Overall Recommendations

1. **Optimize Staffing Levels**:
   - Increase agent availability during peak hours (especially 12-1 PM and 9-10 PM) to manage high call volumes effectively and reduce abandonment rates.
2. **Enhance Training Programs**:
   - Provide targeted training for agents to improve efficiency in handling calls, particularly during high-demand periods. Focus on common issues leading to longer call durations and high abandonment rates.
3. **Implement Dynamic Shift Scheduling**:
   - Use historical call volume data to create flexible shift schedules that align with expected demand, ensuring sufficient coverage during both daytime and nighttime hours.
4. **Monitor Call Trends**:
   - Regularly analyze call data to identify patterns and adjust staffing and training needs accordingly. This proactive approach will help in maintaining service quality and customer satisfaction.
5. **Improve Response Times**:
   - Develop strategies to streamline call handling processes, enabling agents to respond more quickly to customer inquiries and reducing overall call handling times.
6. **Conduct Post-Call Analyses**:
   - Investigate reasons for abandoned calls and customer complaints to make informed adjustments to processes and staffing, ensuring a better customer experience.
7. **Leverage Technology**:
   - Consider implementing call management software or AI tools to assist in routing calls and managing workloads, thereby improving efficiency and customer service

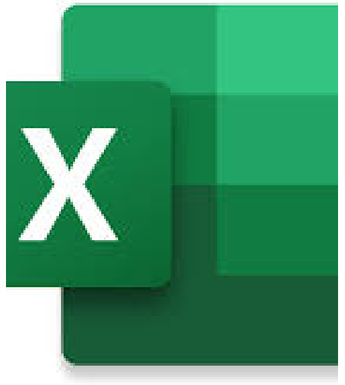
learning and REFLECTIONS

This course has made me confident in technical as well as soft skills. From improving speaking skills to making excel spreadsheets, I have come a long way only because of Trainity.

I am now confident in using Excel, Python, SQL, and data visualization techniques such as PowerBI.

I also want to thank the staff and teaching faculty for being the best guides all along the way. I gained a better understanding of descriptive and inferential statistics for data-driven decisions. Managing multiple projects taught me how to stay organized and meet deadlines effectively. I also deepened my knowledge in various domains, including healthcare, social media, hiring, and customer service analytics. My communication skills improved as I learned to present findings clearly to different audiences. Additionally, tackling real-world challenges enhanced my critical thinking and problem-solving abilities.

I recognized the value of teamwork in achieving project goals and embraced a mindset of continuous learning to stay updated with industry trends.



THANKS FOR WATCHING

npoorva2002@gmail.com

www.linkedin.com/in/poorvanimishnahar