Data Analysis Portfolio

Professional Background

As a final year B.Tech student in Computer Science and Engineering, I have developed a strong foundation in the principles and practices of programming, data analysis, and machine learning. I have gained proficiency in several technical skills, including Python, machine learning, data science, Tableau, and data analysis.

I have completed numerous projects that have enhanced my skills in these areas, including developing a sentiment analysis model, building a recommendation system, and creating a predictive model for customer churn.

In addition, I have completed internships with Solar Secure and Suvidha Foundation, where I gained valuable experience in data science and machine learning. During my time at these organizations, I worked on several projects, including developing a fraud detection system and building a predictive model for employee attrition.

As a fresher, I am eager to learn and grow in my field. My ultimate goal is to become a data analyst, and I am committed to working hard and gaining the skills and experience necessary to achieve this goal.

Table of Contents

Table of Contents (Cont..)

Data Analytics Process

Description

We use Data Analytics in everyday life without even knowing it.
Your task is to give the example(s) of such a real-life situation where we use DataAnalytics and link it with the data analytics process.

Data Analytics Process

Design

1. Plan: The hospital sets a goal to improve patient outcomes and reduce costs. They identify key metrics to track, such as patient satisfaction, readmission rates, and length of stay. These metrics are important indicators of the quality of care and the efficiency of the hospital's operations. The hospital also identifies specific areas of focus, such as reducing readmission rates for certain conditions or improving patient satisfaction scores.

2. Prepare: The hospital collects and organizes data from various sources, such as electronic health records, claims data, and patient surveys. The data is collected from different departments and systems within the hospital, such as the emergency department, inpatient units, and billing systems. This data is then cleaned, transformed and integrated to make sure it is accurate, consistent and complete. The data is also de-identified to protect patient privacy and compliance with regulations.

3. Process: The hospital uses statistical methods to process the data and identify patterns and trends. They may use techniques such as machine learning and natural language processing to analyze patient data, identify risk factors and predict outcomes. For example, they may use machine learning algorithms to identify patients at high risk of readmission, or natural language processing to extract information from unstructured data such as physician notes.

4. Analyze: The hospital interprets the results of the data analysis and draws conclusions about patient outcomes and costs. They may use visualizations such as heat maps, scatter plots and dashboards to communicate the results to stakeholders. This can help identify areas where the hospital is performing well and areas where improvement is needed. They also conduct statistical analysis to identify the most significant factors impacting patient outcomes and costs.

5. Share: The hospital shares the results of the data analysis with relevant stakeholders, such as doctors, nurses, and administrators. They also present recommendations for how the hospital can improve patient outcomes and reduce costs. These recommendations may include changes to protocols, implementation of new technology, or improvements in communication.

6. Act: The hospital implements the recommendations from the data analysis, such as changing protocols, implementing new technology, and improving communication. They also continue to track the key metrics to measure the success of the actions taken. They also conduct ongoing monitoring of the data and make adjustments as needed to ensure that the desired outcomes are being achieve.

Data Analytics Process

Conclusions

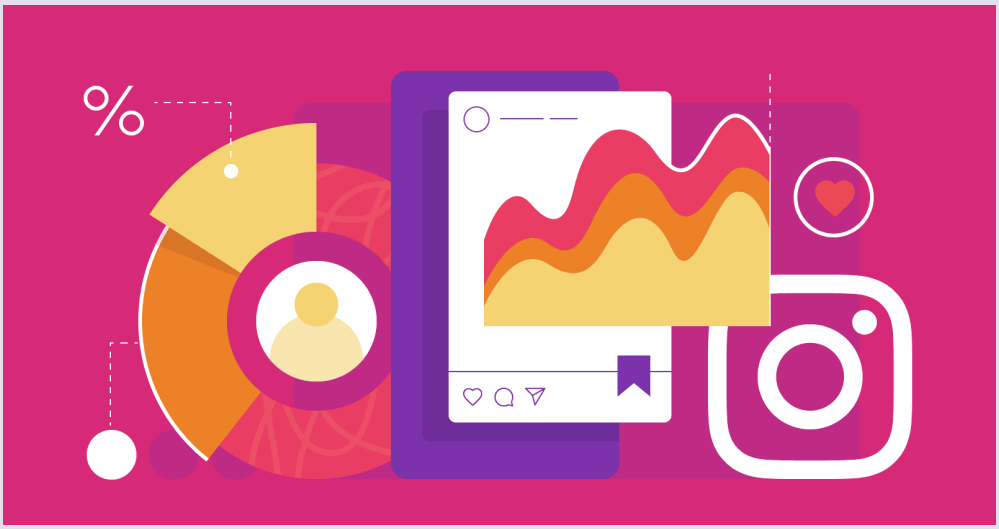
By following this process, the hospital can gain insight into patient outcomes and costs, and use this information to make informed decisions that can improve patient outcomes and reduce costs. This can lead to better care for patients, higher patient satisfaction, and cost savings for the hospital.

Hence, we have seen how we can use the 6 steps of Data Analytics while making any decision in real life scenarios.

The 6 steps used to take decisions in real life scenarios are:-

- **Plan**
- **Prepare**
- **Process**
- **Analyze**
- **Share**
- **Act**

Instagram User Analytics



Description

User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.

These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow.

You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

Instagram

Instagram User Analytics

The Problem

A) Marketing: The marketing team wants to launch some campaigns, and they need your help with the following

- **Rewarding Most Loyal Users: People who have been using the platform for the longest time. Your Task: Find the 5 oldest users of the Instagram from the database provided**

- **Remind Inactive Users to Start Posting: By sending them promotional emails to post their 1st photo. Your Task: Find the users who have never posted a single photo on Instagram**

- **Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner. Your Task: Identify the winner of the contest and provide their details to the team**

- **Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform. Your Task: Identify and suggest the top 5 most commonly used hashtags on the platform**

- **Launch AD Campaign: The team wants to know, which day would be the best day to launch ADs. Your Task: What day of the week do most users register on? Provide insights on when to schedule an ad campaign**

Instagram User Analytics

The Problem (Cont..)

B) Investor Metrics: Our investors want to know if Instagram is performing well and is not becoming redundant like Facebook, they want to assess the app on the following grounds

- **User Engagement: Are users still as active and post on Instagram or they are making fewer posts Your Task: Provide how many times does average user posts on Instagram. Also, provide the total number of photos on Instagram/total number of users**

- **Bots & Fake Accounts: The investors want to know if the platform is crowded with fake and dummy accounts Your Task: Provide data on users (bots) who have liked every single photo on the site (since any normal user would not be able to do this).**

Instagram User Analytics

Design

Steps taken to load the data into the data base

- **Using the 'create db' function of MySQL create a data base**
- **Then add tables and column names**
- **Then add the values into them using the 'insert into' function of MySQL**
- **By using the 'select' command we can query the desired output**

Software used for querying the results

--> MySQL Workbench 8.0 CE

Instagram

Instagram User Analytics

Findings - I

Query :-

```
-- 5 oldest users of the Instagram from the database provided
  SELECT * FROM ig_clone.users order by created_at asc limit 5;
```

Output/Result

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

Instagram

Instagram User Analytics

Findings - II

Query:-

```
-- the users who have never posted a single photo on Instagram

select username,id from ig_clone.users where not (users.id
=any(select user_id from ig_clone.photos)) order by id;
```

Output/Result

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

Instagram

Instagram User Analytics

Findings - III

Query:-

```
select photo_id,Count(user_id) as likes from likes group by photo_id order by
likes desc;


-- from this we get that most likes photo_id is 145
 select User_id from ig_clone.photos where id=145;

-- winner of the contest
select username from users where id=52;
```

Output/Resul

	username
▶	Zack_Kemmer93

Instagram

Instagram User Analytics

Findings - IV

Query:-

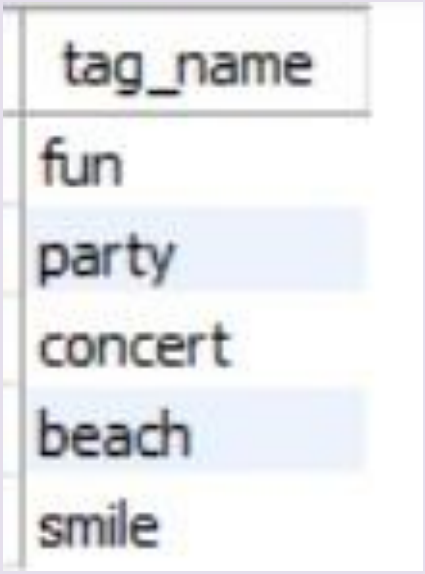
```
-- first identify top 5 tag_id

select tag_id ,count(tag_id)as TagTimes from ig_clone.photo_tags group
by tag_id order by TagTimes desc limit 5;


-- the top 5 most commonly used hashtags on the platform

Select tag_name from ig_clone.tags where id in (21,20,17,13,18);
```

Output/Result



tag_name
fun
party
concert
beach
smile

Instagram

Instagram User Analytics

Findings – V

Query:-

```
-- numbers of user_register with respect to weekday name

select dayname(Created_at)as dayname,count(username)as users_register from ig_clone.users group by dayname order by users_register desc;

-- Based on this we predict Thursday and sunday would be the best day to launch ADs.
```

Output/Result

dayname	users_register
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Instagram

Instagram User Analytics

Findings - VI

Query:-

-- user_id with number of photos uploaded

```
select user_id,count(id) as photos_uploaded from ig_clone.photos group by user_id ;

--These many times does average user posts on Instagram.

Select count(id)/count(distinct user_id) as avg_times from ig_clone.photos; -- the total number of photos on Instagram/total number of users select (select count(id) from ig_clone.photos)/(select count(id) from ig_clone.users) ;
```

Output/Result

total_photos_divide_total_photos
2.57

Instagram

Instagram User Analytics

Findings – VII

Query:-

```
-- users (bots) who have liked every single photo on the site

select user_id from ig_clone.likes where photo_id=all(select id from ig_clone.photos);
```

Output/Result

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

Instagram User Analytics

Analysis

From the Insights we can answer all the questions of market team
and investor team.

A. Market Team

1. Darby_Herzog,Emilio_Bernier52,Elenor88,Nicole71,Jordyn.Jacobson2 account with these username are the Most Loyal Users.

2. There are 26 users on the platform that are inactive and have not posted a single photo.

3. Zack_Kemmer93 is the username of the user who won the contest with 48 likes on his single photo.

4. Fun, Party, Concert, Beach, Smile hashtags to use in the post to reach the most people on the platform.

5. Thursday and Sunday would be the best day to launch ADs.

B. Investor Metrics

1. 3.4730 times does average user posts on Instagram. And the total number of photos on Instagram/total number of users is 2.5700. Based on this we can say Instagram is performing well.

2. There is many bot users.

Instagram User Analytics

Analysis (Cont...)

Using the 5 Whys approach I am finding the root cause of the following:-

- **Why did the Marketing team wanted to know the most inactive users?**
- **--> So, they can reach out to those users via mail and ask them What's keeping them away from using the Instagram.**

- **Why did the Marketing team wanted to know the top 5 #hashtags used?**
- **--> May be the tech team wanted to add some filter features for photos and videos posted using the top 5 mentioned #hashtags**

- **Why did the Marketing team wanted to know on which day of the week the platform had the most new users registered?**
- **--> So, that they can run more Ads of various brands during such days and also get profit from it**

- **Why did the Investors wanted to know about the average posts per user has on Instagram?**
- **--> It is a fact that every brand or social platform is determined by the user engagement on such platforms, also investors wanted to know whether the platform has the right and authenticated user base. It also helps the tech team determine how to handle such traffic on the platform with the latest tech without disrupting the smooth and efficient functioning of the platform**

- **Why did the Investors wanted to know the count of BOTS and Fake accounts if any?**
  --> So that the Investors are assured that they are investing into an Asset and not a Future Liability

Instagram

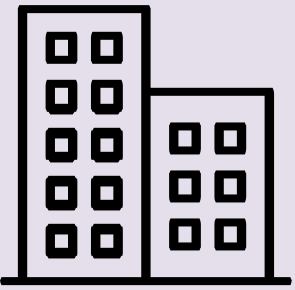
Instagram User Analytics

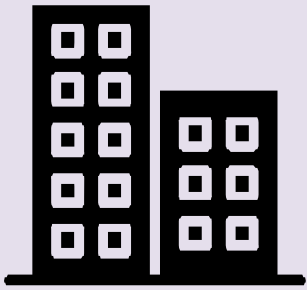
Conclusion

In conclusion, I would like to conclude that not only Instagram but many other social media and commercial firms use such Analysis to find the insights from their customer data which in turn help the firms to find the customers who will be an Asset to the firm in the future and not some Liability.

Such Analysis and sorting of the customer base is done at an weekly, monthly, quarterly or yearly basis as per the needs of the business firms so as to maximize their profits in future with minimal cost to the company

Operation Analytics and Investigating Metric Spike





Description

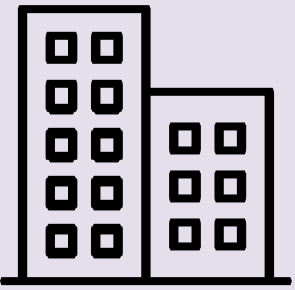
Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they collect.

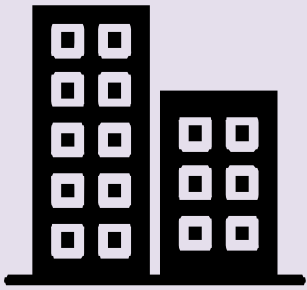
Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that its very important to investigate metric spike.

You are working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which you must derive certain insights out of it and answer the questions asked by different departments.

Operation Analytics and Investigating Metric Spike





The Problem

Case Study 1 (Job Data)

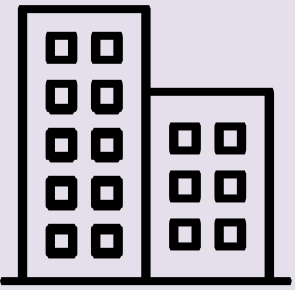
- **Number of jobs reviewed: Amount of jobs reviewed over time.**
- **Your task: Calculate the number of jobs reviewed per hour per day for November 2020?**

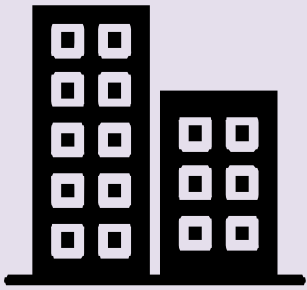
- **Throughput: It is the no. of events happening per second.**
- **Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?**

- **Percentage share of each language: Share of each language for different contents.**
- **Your task: Calculate the percentage share of each language in the last 30 days?**

- **Duplicate rows: Rows that have the same value present in them.**
- **Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?**

Operation Analytics and Investigating Metric Spike





The Problem(Cont...)

Case Study 2 (Investigating metric spike)

- **User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.**
- **Your task: Calculate the weekly user engagement?**

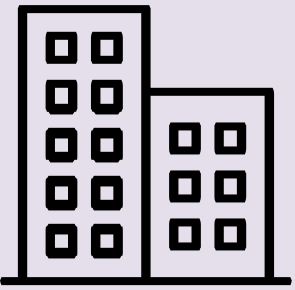
- **User Growth: Amount of users growing over time for a product.**
- **Your task: Calculate the user growth for product?**

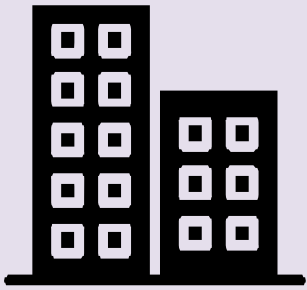
- **Weekly Retention: Users getting retained weekly after signing-up for a product.**
- **Your task: Calculate the weekly retention of users-sign up cohort?**

- **Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.**
- **Your task: Calculate the weekly engagement per device?**

- **Email Engagement: Users engaging with the email service.**
- **Your task: Calculate the email engagement metrics?**

Operation Analytics and Investigating Metric Spike





Design

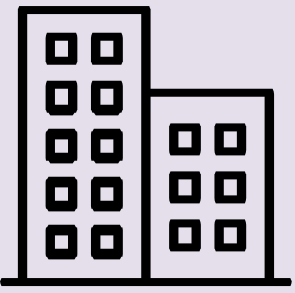
Steps taken to load the data into the data base

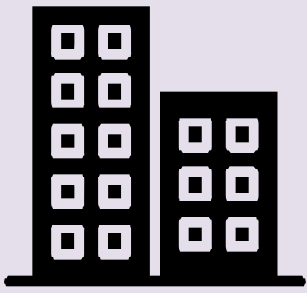
- **Using the 'create db' function of MySQL create a data base**
- **Then add tables and column names**
- **Then add the values into them using the 'insert into' function of MySQL**
- **By using the 'select' command we can query the desired output**

Software used for querying the results

--> MySQL Workbench 8.0 CE

Operation Analytics and Investigating Metric Spike





Job Data

Findings -1

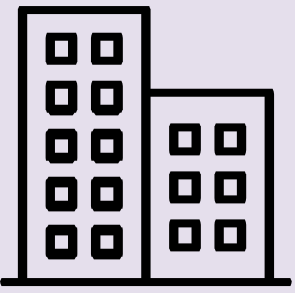
Query:-

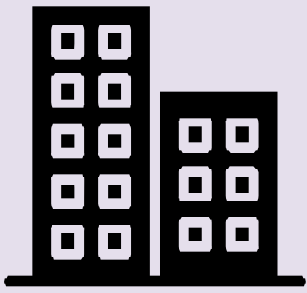
```
1. -- Number of jobs reviewed per hour per day
SELECT
DATE(ds) AS date,
COUNT(job_id) AS jobs_reviewed,
SUM(time_spent) / 3600 AS total_hours,
COUNT(job_id) / SUM(time_spent) OVER (PARTITION BY DATE(ds))
* 3600 AS
jobs_per_hour_per_day
FROM
job_data
WHERE
DATE(ds) BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY
DATE(ds)
```

Output /Result

date	jobs_reviewed	total_hours	jobs_per_hour_per_day
2020-11-25	1	0.0125	80.0000
2020-11-26	1	0.0156	64.2857
2020-11-27	1	0.0289	34.6154
2020-11-28	2	0.0092	327.2727
2020-11-29	1	0.0056	180.0000
2020-11-30	2	0.0111	480.0000

Operation Analytics and Investigating Metric Spike





Job Data

Findings -2

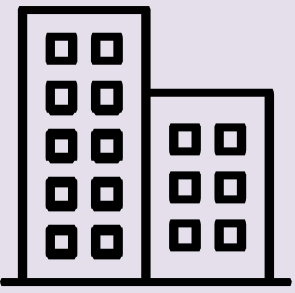
Query:-

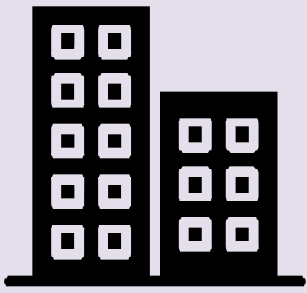
```
SELECT
 date,
 throughput,
 AVG(throughput) OVER (ORDER BY date ROWS BETWEEN 6
PRECEDING AND CURRENT ROW) AS
7_day_rolling_average
FROM (
 SELECT
 DATE(ds) AS date,
 COUNT(job_id) AS jobs_reviewed,
 SUM(time_spent) / 3600 AS total_hours,
 COUNT(job_id) / SUM(time_spent) * 3600 AS throughput
 FROM
 job_data
 WHERE
 DATE(ds) BETWEEN '2020-11-01' AND '2020-11-30'
 GROUP BY
 DATE(ds)
)sub
```

Output /Result

date	throughput	7_day_rolling_average
2020-11-25	80.0000	80.00000000
2020-11-26	64.2857	72.14285000
2020-11-27	34.6154	59.63370000
2020-11-28	218.1818	99.27072500
2020-11-29	180.0000	115.41658000
2020-11-30	180.0000	126.18048333

Operation Analytics and Investigating Metric Spike





Job Data

Findings -

Query:-

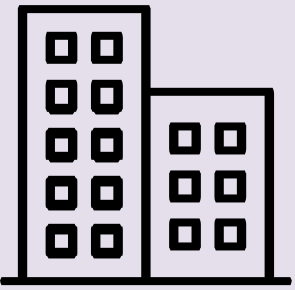
```
SELECT  language,
SUM(time_spent) AS total_time_spent,
(SUM(time_spent)  /  SUM(total_time_spent_30_days))  *  100  AS
percentage_share
FROM
( SELECT
ds,
language,
time_spent,
SUM(time_spent)  OVER  (ORDER  BY  ds  ROWS  BETWEEN  29
PRECEDING AND CURRENT ROW) AS
total_time_spent_30_days
FROM
job_data
) sub
GROUP BY
language;
```

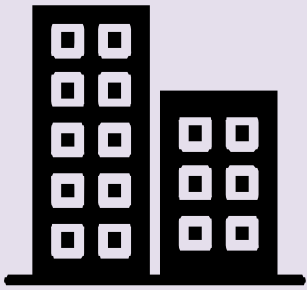
Output /Result

C. Percentage share of each language:

language	total_time_spent	percentage_share
Italian	45	100.0000
Persian	98	16.7235
French	104	50.7317
Hindi	11	4.6218
English	15	5.4945
Arabic	25	8.3893

Operation Analytics and Investigating Metric Spike





Job Data

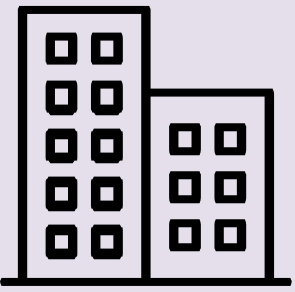
Findings - IV

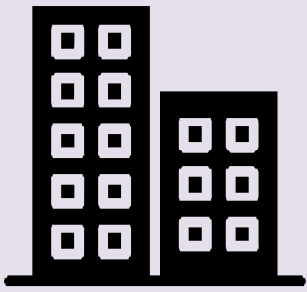
Qurey:-

```
SELECT column1, column2, ..., columnN, count(*)
FROM table_name
GROUP BY column1, column2, ..., columnN
HAVING count(*) > 1;
```

To display duplicates from a table, one approach would be to first identify the columns that contain the duplicate values, and then use SQL commands such as SELECT and GROUP BY to aggregate the data and find any duplicates.

Operation Analytics and Investigating Metric Spike





Investigating Metric Spike

Findings - 1

Query:-

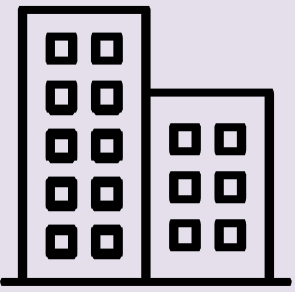
```
SELECT user_id, weekofyear(occurred_at) AS week_of_year, SUM(CASE
WHEN event_type = 'engagement' THEN 1 ELSE 0 END) OVER
(PARTITION BY user_id, weekofyear(occurred_at)) AS
engagement_count FROM events;
```

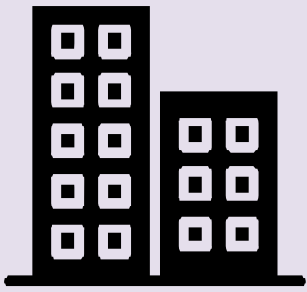
Output /Result

. User Engagement:

user_id	week_of_year	engagement_count
4	20	4
4	20	4
4	20	4
4	20	4
4	21	8
4	21	8
4	21	8
4	21	8
4	21	8

Operation Analytics and Investigating Metric Spike





Investigating Metric Spike

Findings - 2

Query:-

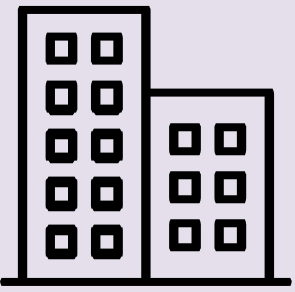
```
SELECT
 date,
 daily_user_count,
 SUM(daily_user_count) OVER (ORDER BY row_num) AS running_user_count
FROM( SELECT
 DATE(created_at) AS date,
 COUNT(DISTINCT user_id) AS daily_user_count,
 ROW_NUMBER() OVER (ORDER BY DATE(created_at)) AS row_num
 FROM
 users
 GROUP BY
 DATE(created_at)
)sub;
```

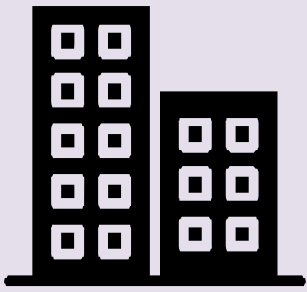
Output /Result

B. User Growth:

date	daily_user_count	running_user_count
2013-01-01	13	13
2013-01-02	11	24
2013-01-03	14	38
2013-01-04	11	49
2013-01-05	3	52
2013-01-06	4	56
2013-01-07	13	69
2013-01-08	13	82
2013-01-09	11	93

Operation Analytics and Investigating Metric Spike





Investigating Metric Spike

Findings - 3

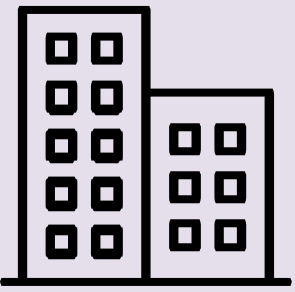
Query:-

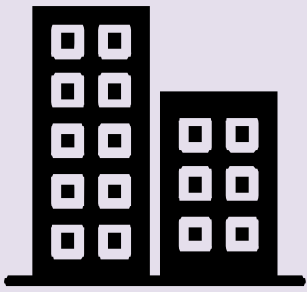
```
SELECT
 yearweek(users.created_at)AS date,
 COUNT(DISTINCT users.user_id) AS Weekly_signup,
 SUM( case WHEN events.event_type = 'engagement' THEN 1 ELSE 0 END)
OVER (PARTITION BY
yearweek(events.occurred_at)) AS Retained_user
 FROM
 users join events
 on users.user_id=events.user_id
 GROUP BY
 yearweek(users.created_at)
 order by yearweek(users.created_at);
```

Output /Result

date	Weekly_signup	Retained_user
201301	9	5
201302	13	2
201303	18	10
201304	14	10
201305	21	11

Operation Analytics and Investigating Metric Spike





Investigating Metric Spike

Findings - 4

Query:-

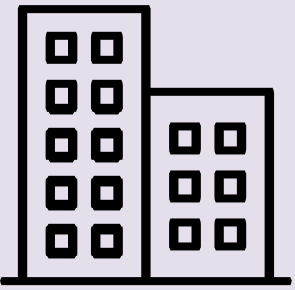
```
-- the weekly engagement per device
SELECT
 user_id,
 device,
 weekofyear(occurred_at) AS week_of_year,
 SUM(CASE WHEN event_type = 'engagement' THEN 1 ELSE 0 END) OVER
(PARTITION BY user_id,
weekofyear(occurred_at)) AS engagement_count
FROM
events;
```

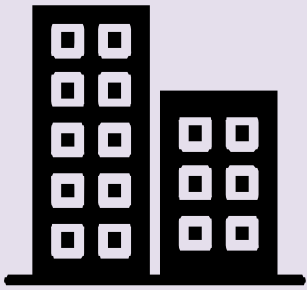
Output /Result:-

Weekly Engagement:

user_id	device	week_of_year	engagement_count
4	lenovo thinkpad	20	4
4	lenovo thinkpad	20	4
4	lenovo thinkpad	20	4
4	lenovo thinkpad	20	4
4	samsung galaxy s4	21	8
4	samsung galaxy s4	21	8
4	samsung galaxy s4	21	8
4	samsung galaxy s4	21	8

Operation Analytics and Investigating Metric Spike





Investigating Metric Spike

Findings - V

Query:-

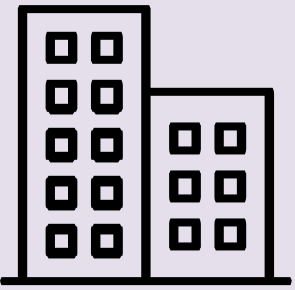
```
-- email engagement metrics
WITH base_query AS ( SELECT user_id, yearweek(occurred_at) AS week,
action FROM email_events WHERE action IN ('sent_weekly_digest',
'email_open') ), digests_sent AS ( SELECT user_id, week, COUNT(CASE
WHEN action = 'sent_weekly_digest' THEN 1 END) AS digests_sent FROM
base_query GROUP BY user_id, week ), emails_opened AS ( SELECT
user_id, week, COUNT(CASE WHEN action = 'email_open' THEN 1 END) AS
emails_opened FROM base_query GROUP BY user_id, week ) SELECT
digests_sent.week, SUM(digests_sent.digests_sent) AS total_digests_sent,
SUM(emails_opened.emails_opened) AS total_emails_opened,
SUM(emails_opened.emails_opened) / SUM(digests_sent.digests_sent) AS
email_engagement_rate FROM digests_sent JOIN emails_opened ON
digests_sent.user_id = emails_opened.user_id AND digests_sent.week =
emails_opened.week GROUP BY digests_sent.week ORDER BY
digests_sent.week;
```

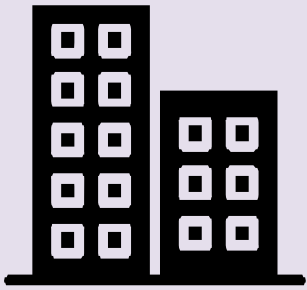
Output /Result:-

Email Engagement:

week	total_digests_sent	total_emails_opened	email_engagement_rate
201417	908	310	0.3414
201418	2602	912	0.3505
201419	2665	972	0.3647
201420	2733	1004	0.3674
201421	2822	1014	0.3593
201422	2911	987	0.3391
201423	3003	1075	0.3580
201424	3105	1155	0.3720
201425	3207	1096	0.3418

Operation Analytics and Investigating Metric Spike





Analysis

Using the Why's approach I am trying to find more insights

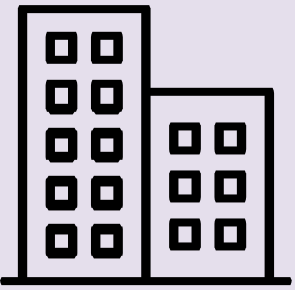
- **Why there is a difference of values between the number of distinct jobs reviewed per day and number of non-distinct jobs reviewed per day?**
- **----> May be due to repeated values in two or more rows or the dataset consisted of duplicate rows**

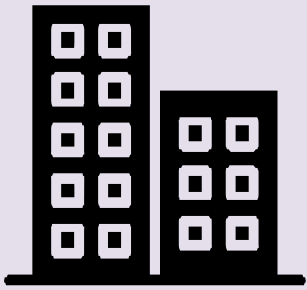
- **Why one shall use 7 day rolling average for calculating throughput and not daily metric average?**
- **----> For calculating the throughput we will be using the 7-day rolling because 7-day rolling gives us the average for all the days right from day 1 to day 7 Whereas daily metric gives us average for only that particular day itself.**

- **Why is it that percentage share of all other languages is 12.5% but that of language = 'Persian' is 37.5?**
- **-----> In such cases there are two chances i.e. either there were duplicate rows having language as 'Persian' or there were really two or more unique people who were speaking in Persian language**

- **Why do we need to look for duplicate rows in an dataset?**
- **----> Duplicates have a direct influence of the Analysis going wrong and may led to wrong Business Decision leading to loss to the company or any entity; so to avoid these one must look for duplicates and remove them where necessary**

Operation Analytics and Investigating Metric Spike





Analysis (Cont...)

I have used the Why's approach to gain few more insights:-

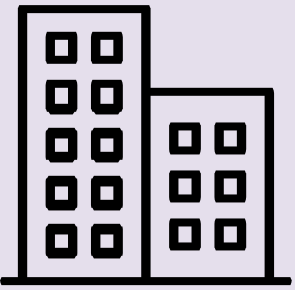
- **Why is the weekly user engagement so less in the beginning and then got increased?**
- **-----> It is a fact that for any new product or service launched, during it's initial period in the market it is less known to all people only some people use the product and based on their experience the product/service engagement increases or decreases depending on whether the consumer experience was good or bad. In this case since the user engagement increased after 2-3 weeks of the launch means that the consumer had a good experience with the product/service**

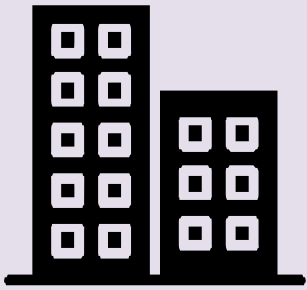
- **Why is weekly retention so important?**
- **---> Weekly retention helps the firms to convince and help those visitors who just complete the sign-up or leave the sign-up process in between, such visitors may become customers in future if they are guided and convinced properly**

- **Why is weekly engagement per device plays an important role?**
- **----> Based on the reviews from users weekly engagement per device helps the firms on which devices they must focus more and which devices need more improvements so they also get a good review in users weekly engagement per device**

- **Why is Email Engagement plays an important role?**
- **----> Email Engagement helps the firms to decide the discounts and offers on specific products. In this case the email_opening_rate is 33.58 i.e. out of the 100 mails send only 34 mails were opened and the email_clicking_rate is 14.789 i.e. out of 100 mails opened only 15 mails were clicked for more details regarding the discount/product details. This means that the current firm needs to have some more catchy line for mails also the firm needs to do rigorous planning and deciding content before sending the mails**

Operation Analytics and Investigating Metric Spike





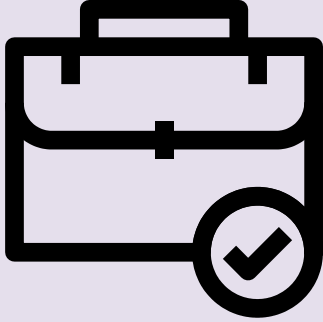
Conclusion

In Conclusion , I would like to conclude that Operation Analytics and Investigating Metric Spike are very necessary and they must be done on daily, weekly, Monthly, Quarterly or Yearly basis based on the Business needs of the firm.

Also, any firm/entity must focus on the Email Engagement with the customers; the firm must use catchy headings along with reasonable discounts and coupons so as to increase their existing customer base

Also any firm must have a separate department(if possible) so as to hear out to the problems of those Visitors who had left the Sign-up Process in between, the firm must guide them so as to convert them from Visitors to Customers

Hiring Process Analytics



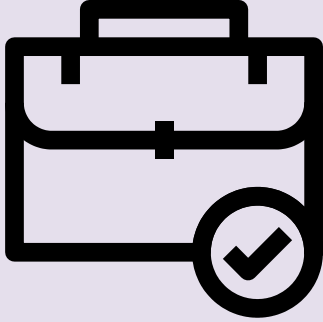
Description

Hiring process is the fundamental and the most important function of a company. Here, the MNCs get to know about the major underlying trends about the hiring process. Trends such as- number of rejections, number of interviews, types of jobs, vacancies etc. are important for a company to analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!

Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.

You are working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked you to answer certain questions making sense out of that data.

Hiring Process Analytics



The Problem

- **Hiring: Process of intaking of people into an organization for different kinds of positions.**
- **Your task: How many males and females are Hired ?**

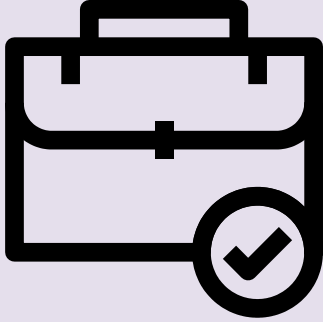
- **Average Salary: Adding all the salaries for a select group of employees and then dividing the sum by the number of employees in the group.**
- **Your task: What is the average salary offered in this company ?**

- **Class Intervals: The class interval is the difference between the upper class limit and the lower class limit.**
- **Your task: Draw the class intervals for salary in the company ?**

- **Charts and Plots: This is one of the most important part of analysis to visualize the data.**
- **Your task: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?**

- **Charts: Use different charts and graphs to perform the task representing the data.**
- **Your task: Represent different post tiers using chart/graph?**

Hiring Process Analytics



Design

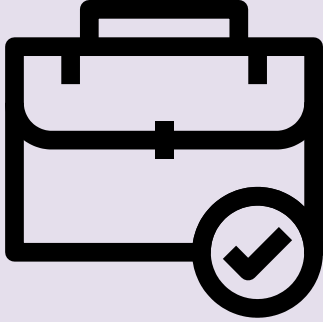
Before starting the actual analysis I have:-

- **Firstly I made a copy of the raw data where I can perform the Analysis so that what ever changes I made it will not affect the original data**
- **Secondly I looked for blank spaces and NULL values if any.**
- **Then I had imputed the numerical blank and NULL cells with mean of the column(if no outliers existed for that particular column) or with median (if outliers existed for that column)**
- **Then I looked for if any outliers exists and replaced them with the median of the particular column where the outlier existed**
- **Then for blank cells of categorical variables I had replaced with the variable with the highest count**
- **Then I looked for duplicate rows and removed them if any**
- **Then I removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis**

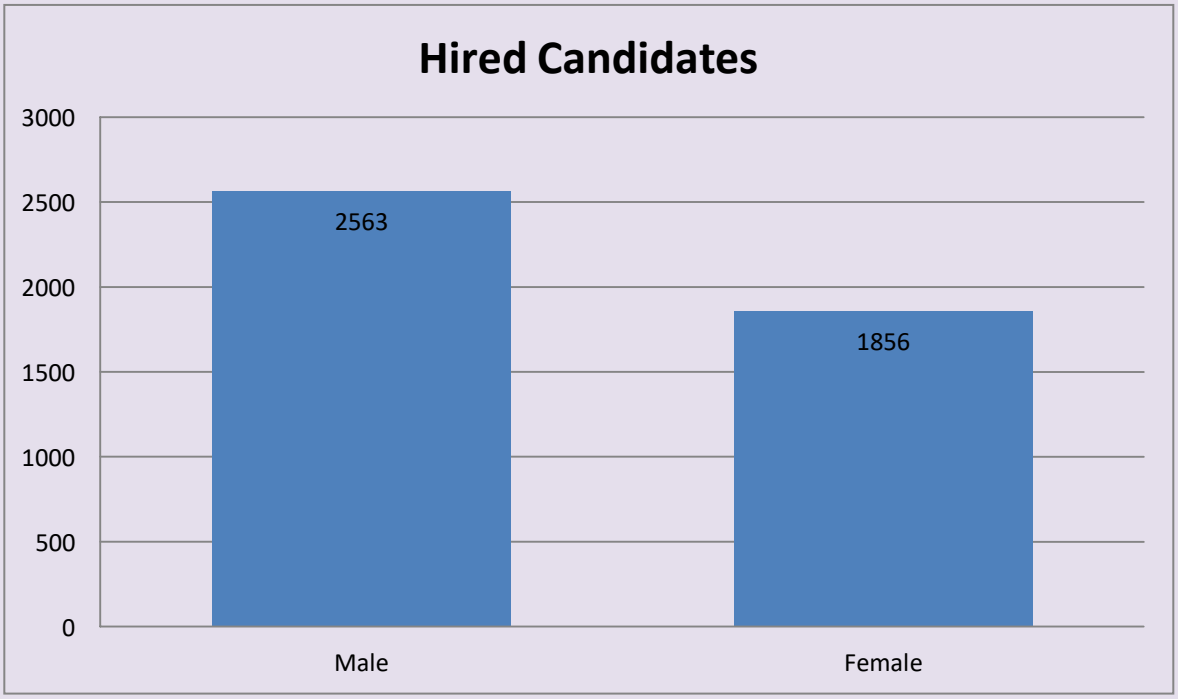
Software used for doing the overall Analysis:-

----> Microsoft Excel

Hiring Process Analytics



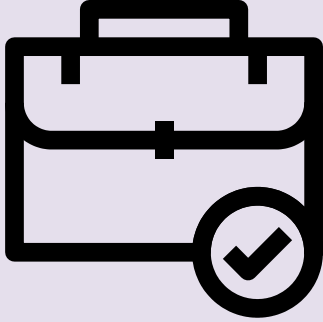
Findings - I



From the above table and bar plot I have inferred that:-

Out of 4697 employees hired, 2563 are males and 1856 are females.

Hiring Process Analytics



Findings - II

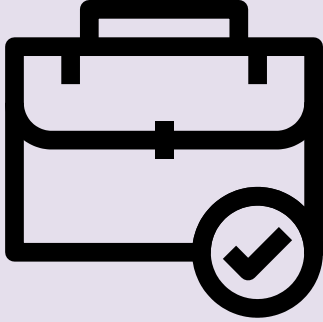
To find the average salary offered in this company:-
1. First, we need to remove the outliers i.e. to remove the salaries below 1000 and above 100000

2. Then using the formula
=AVERAGE(entire_column_of_salary_after_removing_outliers)

Output/Result

49983.03223

Hiring Process Analytics



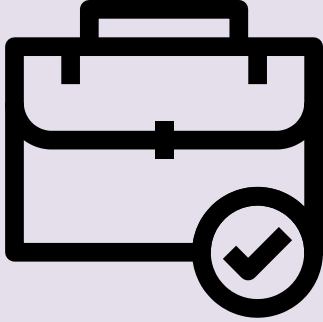
Findings - 3

Class Intervals:

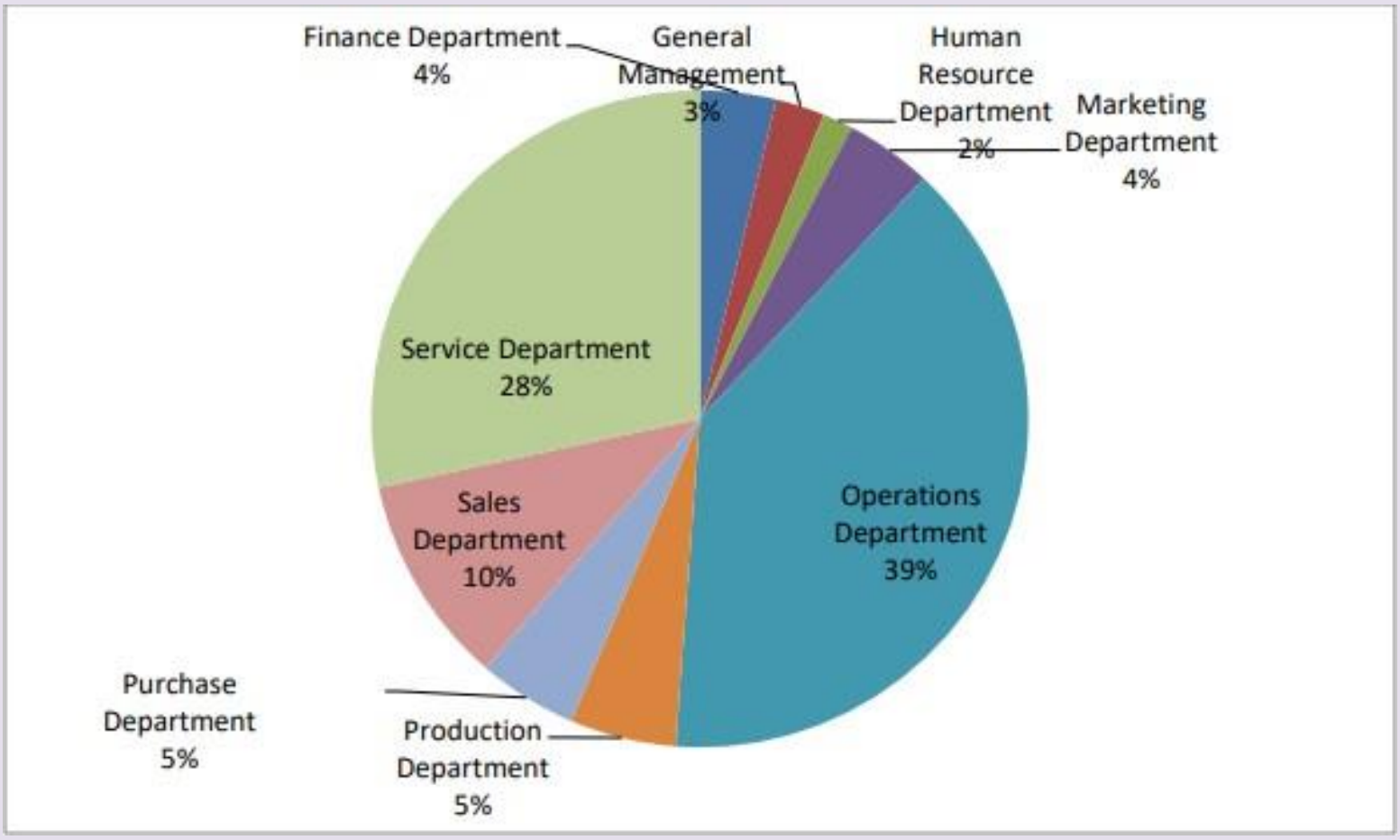
The class intervals for salary in the company are as follows:

0-20000
20001-40000
40001-60000
60001-80000
80001-100000
100001-120000
120001-140000
140001-160000
160001-180000
180001-200000

Hiring Process Analytics



Findings - 4

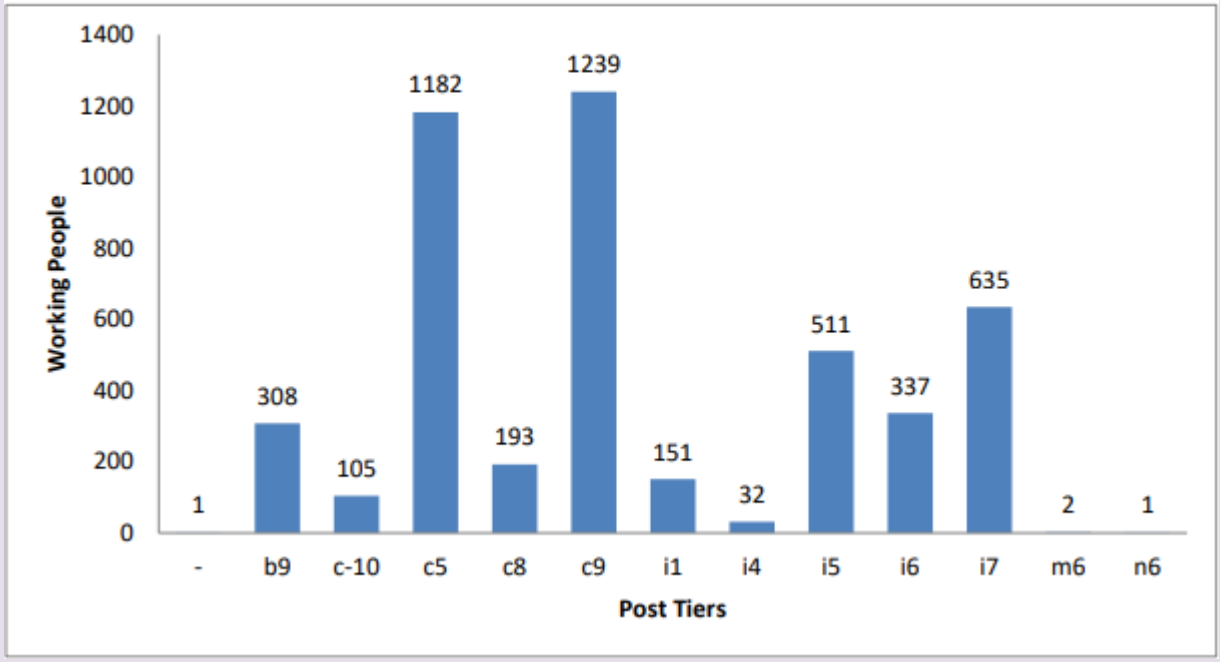


From the above table, pie chart I have inferred that the Highest number of people were working in the Operations Department i.e. 1843 which accounts for almost 39% of the total workforce of the company

Hiring Process Analytics

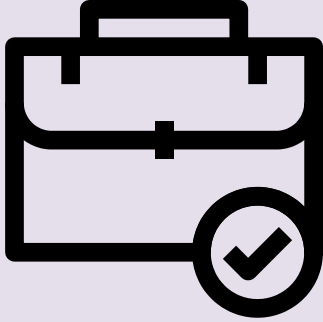
Findings - 5

Post Tiers: The Bar Graph shows the distribution of employees according to their post tiers.



By above bar chart I can say Most employees are in c9, followed by c5.

Hiring Process Analytics

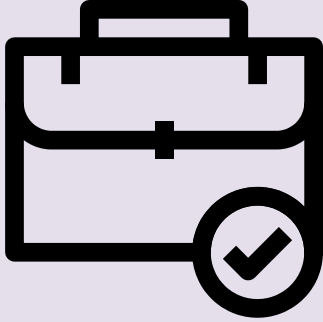


Analysis

The analysis provided valuable insights into the hiring process of the company. It was found that the gender distribution was not balanced, there are more males hired than females. The Operations Department had the highest number of working employees, indicating that the company may need to focus on hiring more employees in other departments. The average salary offered was 49983.03, which is a good benchmark for the company to maintain. The class intervals for salary were also identified, which can be useful in future analysis. Finally, the distribution of employees according to their post tiers was represented using a Bar Graph, which provided a clear picture of the number of employees in each tier. Most employees are in c9, followed by c5.

Overall, this analysis can be useful for the hiring department to make informed decisions about their recruitment process and improve the hiring process in the future.

Hiring Process Analytics



Conclusion

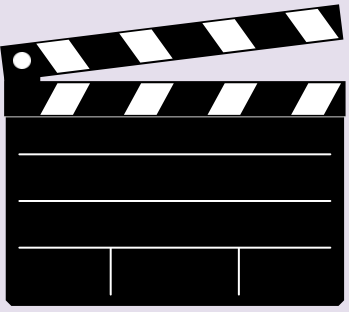
In the conclusion part, I would like to conclude that Hiring Process Analytics plays an important part for all the companies and firms to decide the job openings for the near future.

Hiring Process Analytics is done on monthly, quarterly or yearly basis as per the needs and policies of the companies

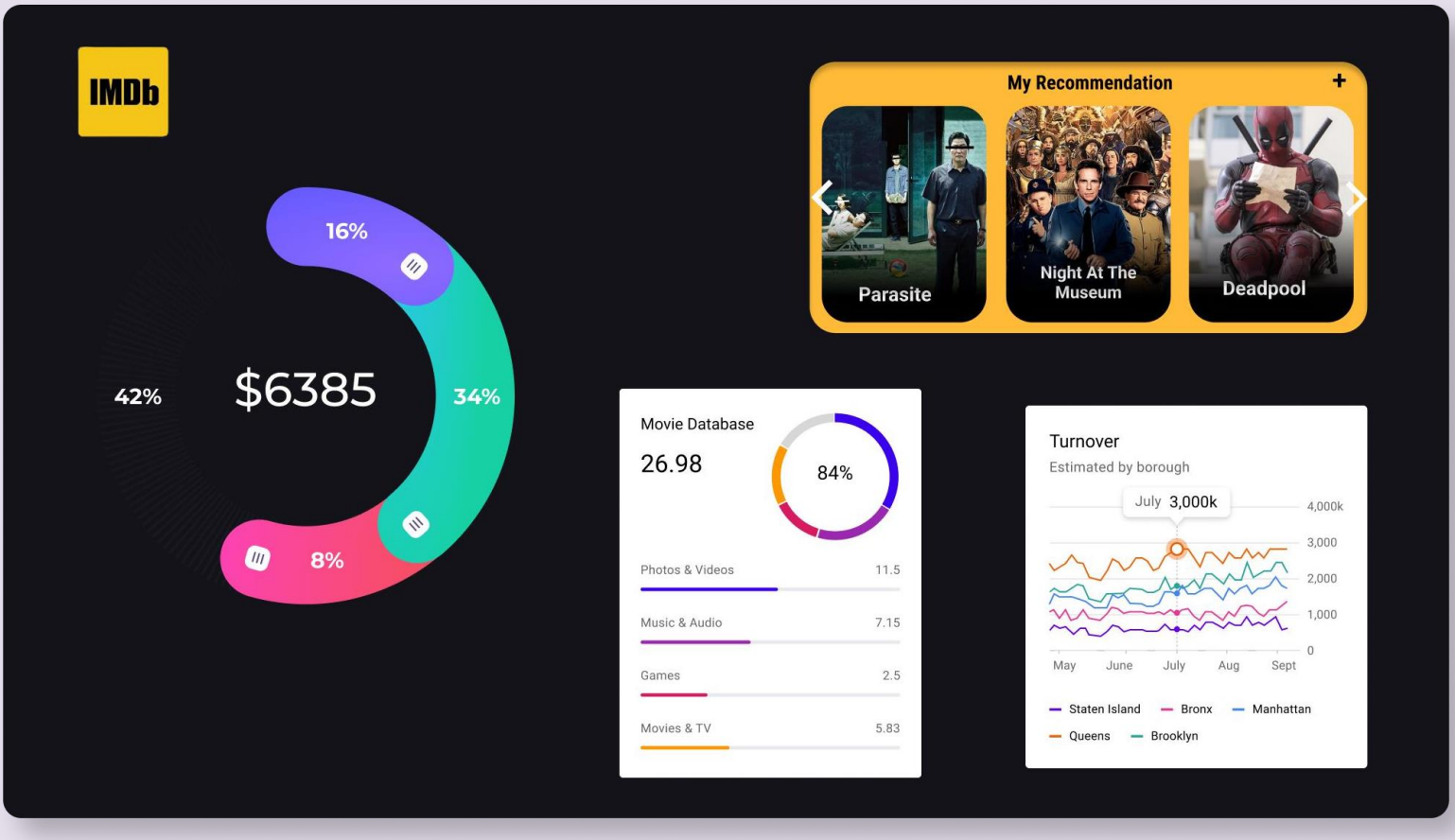
For any company the Operations Department has the highest number of workforce due to the workload on this department as this department acts as a central hub for all the executive tasks carried out

For any company there will some employees who have high salary packages compared to other employees, and this is due to the fact that they have some special skills and years of experience in their particular field of work

Hiring Process Analytics helps the company to decide the salaries for new freshers joining the company; also it tells requirement of workforce by each department; it also helps the company decide the appraisals and increment for it's current employess



IMDB Movie Analysis

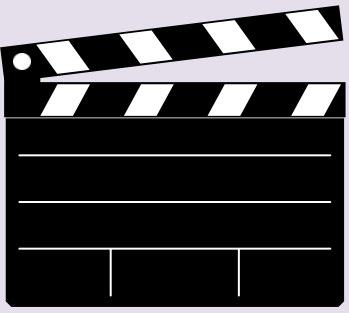


Description

You are provided with dataset having various columns of different IMDB Movies. You are required to Frame the problem.

For this task, you will need to define a problem you want to shed some light on.

Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.



IMDB Movie Analysis

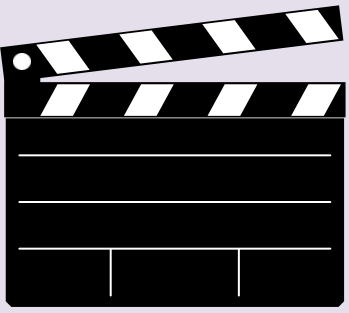
The Problem

- **Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.**
- **Your task: Find the movies with the highest profit?**

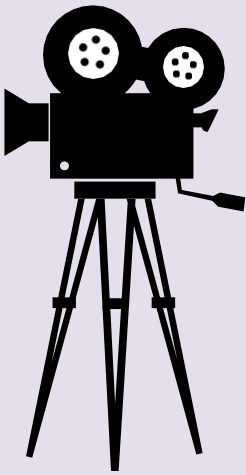
- **Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.**
- **Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!**
- **Your task: Find IMDB Top 250**

- **Best Directors: Group the column using the director_name column. Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.**
- **Your task: Find the best directors**

- **Popular Genres: Perform this step using the knowledge gained while performing previous steps.**
- **Your task: Find popular genres**



IMDB Movie Analysis



The Problem (Cont...)

- **Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.**

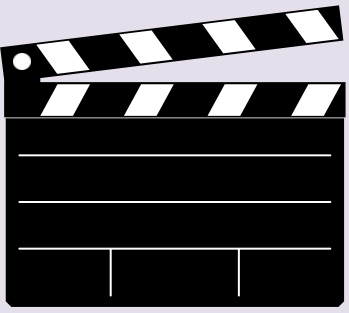
- **Append the rows of all these columns and store them in a new column named Combined.**

- **Group the combined column using the actor_1_name column.**

- **Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.**

- **Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.**

- **Your task: Find the critic-favorite and audience-favorite actors**



IMDB Movie Analysis

Design

The project will begin with cleaning the data, dropping unnecessary columns, and removing null values. Next, the analysis will focus on finding movies with the highest profit by creating a new column called "profit" and sorting the dataset by this column. Outliers will be identified using appropriate chart types.

Cleaning the data is one of the most important steps before moving forward with analysis. We dropped some columns that were not useful for our analysis and removed null values by droping, imputation with median and replacement .We also created a new column called profit which contained the difference between gross and budget.
Remaining columns are:
'director_name', 'num_critic_for_reviews', 'gross', 'genres', 'actor_1_name', 'movie_title', 'num_voted_users', 'num_user_for_reviews', 'language', 'budget', 'title_year', 'imdb_score'

The analysis was performed using Microsoft Excel 2021. Pivot Table are used to filter and sort the data. Excel charts and graphs were also used to represent the data visually. For Data Cleaning ,Jupyter notebook is used with python along with pandas ,NumPy, Seaborn and matplotlib pakages.

IMDB Movie Analysis

Findings - I

To find the movies with the highest profit: -

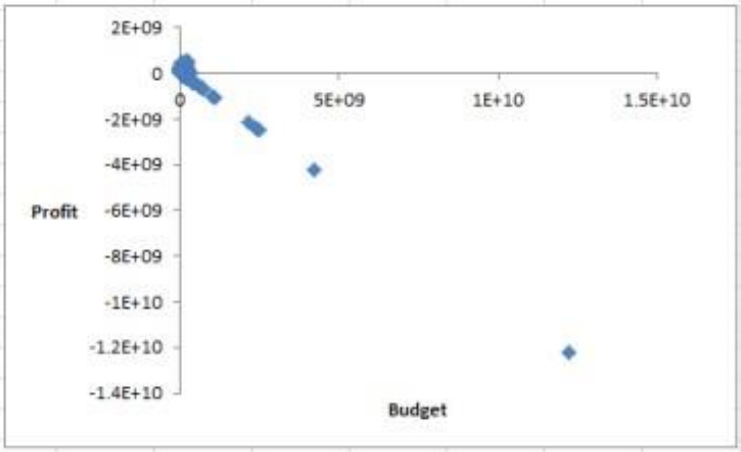
1. First we need to subtract the budget value from the gross value to getthe profit.

2. Then, by using the scatter plot option we will plot values of profit(y_axis)and budget(x_axis)

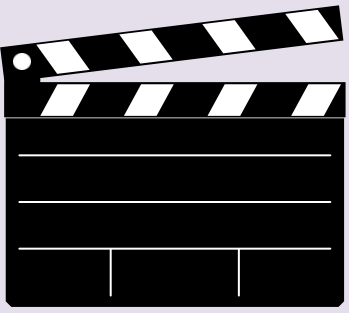
3. Then with the help of graph we will be finding the outliers

We found the movies with the highest profit by sorting the dataset based on the profit column.

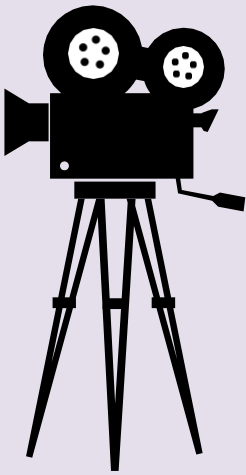
budget	title_year	imdb_score	Profit
237000000	2009	7.9	523505847
300000000	2007	7.1	9404152
245000000	2015	6.8	-44925825
250000000	2012	8.5	198130642
263700000	2012	6.6	-1.91E+08
258000000	2007	6.2	78530303
260000000	2010	7.8	-59192738
250000000	2015	7.5	208991599
250000000	2009	7.5	51956980
250000000	2016	6.9	80249062
209000000	2006	6.1	-8930592
200000000	2008	6.7	-31631573
225000000	2006	7.3	198032628
215000000	2013	6.5	-1.26E+08
225000000	2013	7.2	66021565
225000000	2008	6.6	-83385977
220000000	2012	8.1	403279547



Avatar is the movie with Highest Profit .



IMDB Movie Analysis



Findings - II

To find the IMDB Top 250 we will:-

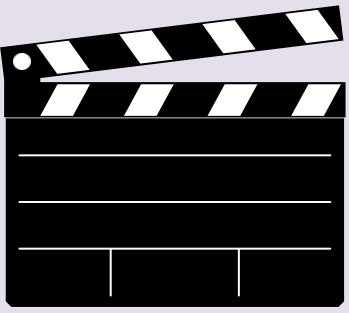
1. **First we will filter out those rows whose num_voted_users > 25000 using thesort and filter option**

2. **Then we will arrange the dataset on the basis of imdb_score in descendingorder**

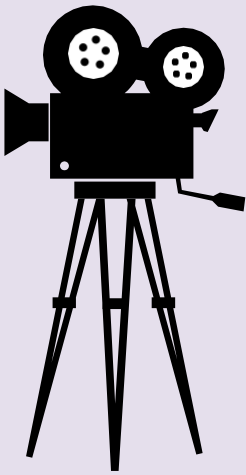
3. **Then we will select only the top 250 rows for the further analysis**

4. **Then we will create a new column rank using the RANK() function**
5. **Then we will filter out (unselect 'English') from the language column and wewill get the desired output**

Top - 5 IMDB Movies all languages



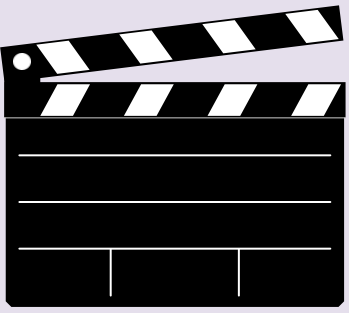
IMDB Movie Analysis



Output sample:-

Rank	IMDB_Top_250
1	The Shawshank Redemption
2	The Godfather
3	The Dark Knight
4	The Godfather: Part II
5	The Lord of the Rings: The Return of the King
6	Schindler's List
7	Pulp Fiction
8	The Good, the Bad and the Ugly
9	Inception
10	The Lord of the Rings: The Fellowship of the Ring
11	Fight Club
12	Forrest Gump
13	Star Wars: Episode V - The Empire Strikes Back
14	The Lord of the Rings: The Two Towers
15	The Matrix
16	Goodfellas
17	Star Wars: Episode IV - A New Hope
18	One Flew Over the Cuckoo's Nest
19	City of God
20	Seven Samurai
21	Interstellar
22	Saving Private Ryan
23	Se7en
24	The Silence of the Lambs
25	Spirited Away

From the above table I have inferred that 'The Shawshank Redemption' had the highest IMDB ratings



IMDB Movie Analysis

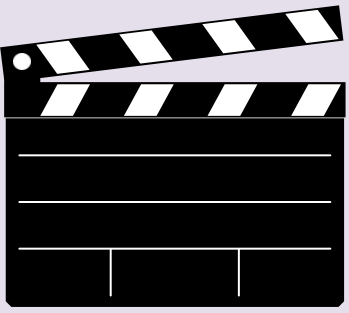
Findings - III

To find the best top 10 directors on the basis of mean of imdb_scorewe will:-

- **First select the imdb_score column of the cleaned dataset**
- **Then we will click on pivot table**
- **We will add director_name into the series section of the pivot table**
- **Then we will add average imdb_score into the values section of the pivot table**
- **Then we will first sort the data on the basis of average of imdb_score in descending order and then on the basis of directorname alphabetically**

Director_name	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
S.S. Rajamouli	8.4

From the above table I have inferred that Charles Chaplin and Tony Kaye had the highest mean of IMDB Score i.e. 8.6



IMDB Movie Analysis

Findings - IV

To find the Popular Genres we will:-

1. **First select the genres column of the cleaned dataset**

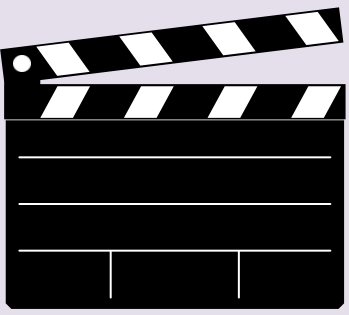
2. **Then we will go for the pivot table option**

3. **Then we will Select the genres name as row labels**

4. **Then we will the values as the count of the avg of imdb_score and then sort it in descending order .**

Most Popular Genres	Average of imdb_score
Crime|Drama|Fantasy|Mystery	8.5
Adventure|Animation|Drama|Family|Musical	8.5
Adventure|Animation|Fantasy	8.4
Action|Adventure|Drama|Fantasy|War	8.4
Adventure|Drama|Thriller|War	8.4
Documentary|Drama|Sport	8.3
Documentary|War	8.3
Biography|Drama|History|Music	8.3
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
Adventure|Drama|War	8.25

In the top 10 according to mean of imdb_score Drama is the most common ,Hence Drama is #1 in term of popular Genres.



IMDB Movie Analysis

Findings - V

Movies Extraction of all 3 actors

Meryl Streep	Leonardo DiCaprio	Brad Pitt
It's Complicated	Titanic	The Curious Case of Benjamin Button
The River Wild	The Great Gatsby	Troy
Julie & Julia	Inception	Ocean's Twelve
The Devil Wears Prada	The Revenant	Mr. & Mrs. Smith
Lions for Lambs	The Aviator	Spy Game
Out of Africa	Django Unchained	Ocean's Eleven
Hope Springs	Blood Diamond	Fury
One True Thing	The Wolf of Wall Street	Seven Years in Tibet
The Hours	Gangs of New York	Fight Club
The Iron Lady	The Departed	Sinbad: Legend of the Seven Seas
A Prairie Home Companion	Shutter Island	Interview with the Vampire: The Vampire Chronicles
	Body of Lies	The Tree of Life
	Catch Me If You Can	The Assassination of Jesse James by the Coward Robert Ford
	The Beach	Babel
	Revolutionary Road	By the Sea
	The Man in the Iron Mask	Killing Them Softly
	J. Edgar	True Romance
	The Quick and the Dead	
	Marvin's Room	
	Romeo + Juliet	
	The Great Gatsby	

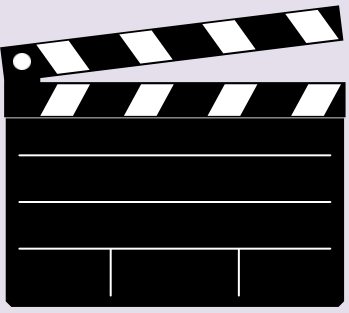
From this we can say Leonardo Dicaprio play maximum lead role in movies out of all these actors.

To find the critic-favorite and audience-favorite actors we will:-

1. **First three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors from the actor_1_name column**

2. **Then we will append the above 3 created columns into 1 column combine**

3. **Then we will group the 3 columns of critic-favorite and audience-**



IMDB Movie Analysis

favoriteactors

4. Then using the pivot table we will find the average, sum and count of critic favorite and audience-favorite actors

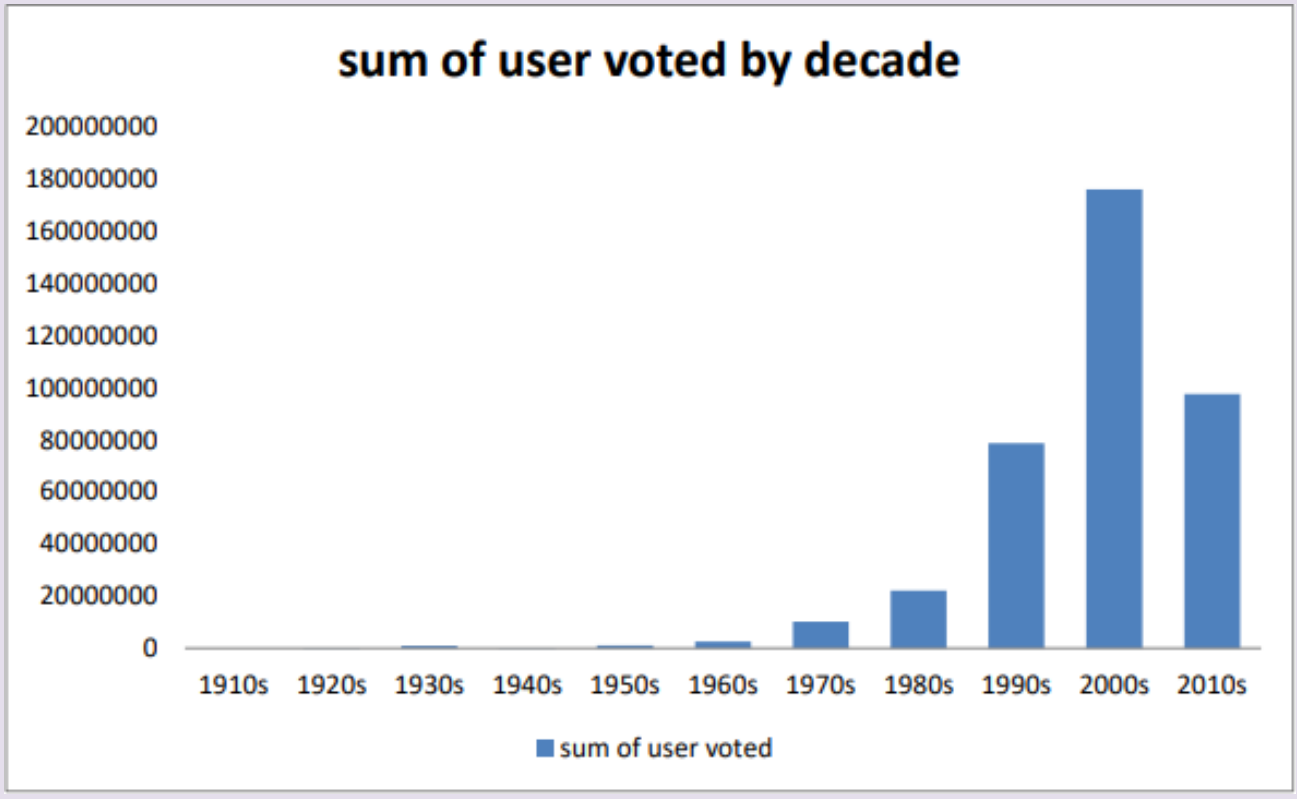
Mean	Brad Pitt	Leonardo DiCaprio	Meryl Streep
Mean of num_user_for_reviews	742.3529412	914.4761905	297.1818182
Mean of num_critic_for_reviews	245	330.1904762	181.4545455
		Maximum Mean	

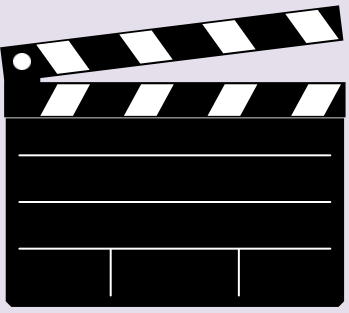
Based on mean of num_critic_for_reviews and mean of num_user_for_reviews we find that Leonardo DiCaprio is the criticfavorite and audience-favorite actor.

IMDB Movie Analysis

Findings - VI

Below is the bar chart showing the number of users voted by decade .We can clearly see that maximum numbers of user voted in the decade 2000s.





IMDB Movie Analysis

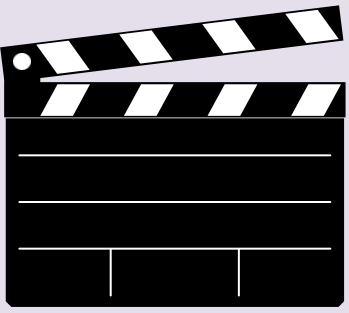
Analysis

Using the Why's approach I am trying to find some useful insights

- **Why is it that the Most rated IMDB movie and the highest profit movie not the same?**
- **-----> Maybe, due to fact that during the IMDB rating only recognized and people who know how to vote on IMDB have the access to the IMDB portal. On the other hand the profit is calculated on the basis of the tickets sold in theatres worldwide.**

- **Why there are more number of votes during the decade 2001-2010?**
- **-----> The period 2001-2010 saw many scientific advancements and computer graphics advancement, also during this interval there was a splendid increase in the production of movies all over the world, so huge number of movies were produced and released during this decade. Also before 2000 there were no laws around the world that had a separate ministry/board/committee from the Government side that looked into the matters of film production and release**

- **Why is it that only movies having language as 'English' are the top 5 ranked movies on the basis of IMDB?**
- **----->** **Movies having language as English were having country of origin as USA; Also it is a well known fact that USA economy was robust during those days. So the social media investors looked for directors made movies so as to gain some financial gains**

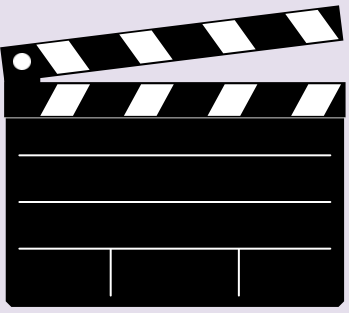


IMDB Movie Analysis

Analysis (Cont..)

- **Why is it that only Drama and Comedy had the highest popularity?**
- **----> Most of people all over the world are stressed with their work life so they need a relaxing refreshment and not some action or horror type thing. So people prefer watching movies that were of Comedy or Drama genre or both. But, most of them preferred Comedy genre films**

- **Why is it that there were more number of votes for the decade 2001-2010 than compared to 2011-2020, though there was advancement in graphics and animation during 2011-2020?**
- **----> It is a fact that there was a great and immense growth of technology not only in the graphics and animation sector but in all aspects of life; Also it was during this interval VPN was introduced; VPN led to piracy (illegal distribution of film) due to which most of people avoided going to theatres.**



IMDB Movie Analysis

Conclusion

In Conclusion, I would like to conclude that IMDB Movie Analysis or any such analysis is done not only by Movie makers before movie production, but it is also done by various investors, stake-holders, theatre outlet owners.

Normal people would not mind to do such analysis but such analysis plays an crucial part during the pre-production phase of the movies and also during the post-production phase

Also, it is not necessary that the movie with the highest IMDB rating will have the highest profit.
Profit is calculated truly on the basis on the number of tickets sold by theatres all over the world

Most of the people are tired with their daily lives and they prefer movies with Comedy/ Drama genre or both, and they would not go for movies with Action/Horror genre

So, directors and production team must keep in mind the above points and shall do the pre-production analysis before the commencement of filming

Bank Loan Case Study

Description

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- **If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.**
- **If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.**

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1. **Approved: The company has approved loan application**
2. **Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.**
3. **Refused: The company had rejected the loan (because the client does not meet their requirements etc.).**
4. **Unused Offer: Loan has been cancelled by the client but on different stages of the process.**

Bank Loan Case Study

The Problem

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

It aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics – understanding the types of variables and their significance should be enough).

Bank Loan Case Study

Design

Data understanding: Understand the dataset and the variables involved.

Data cleaning: Identify missing data and outliers and use appropriate methods to deal with them. Data exploration: Perform univariate, segmented univariate, and bivariate analyses to identify patterns and correlations.

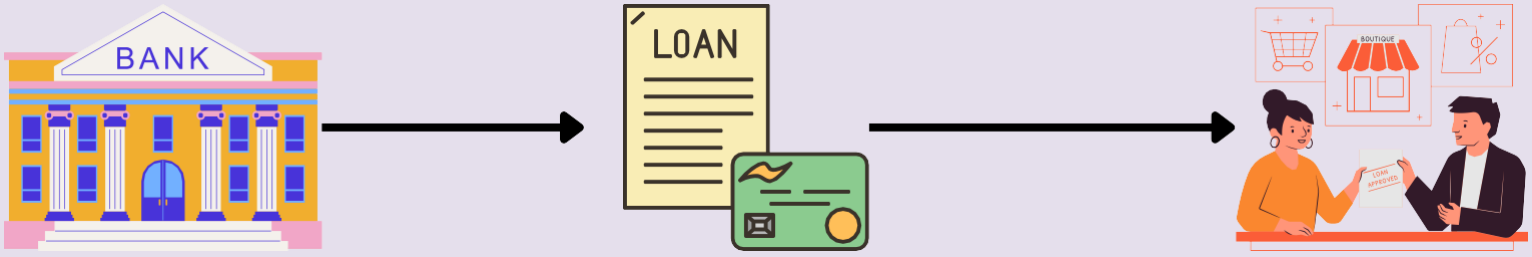
Visualization: Use appropriate graphs and charts to summarize and visualize the insights. Report: Prepare a report summarizing the key insights and recommendations for the company.

Data Cleaning: The dataset provided contains missing data. To deal with the missing data, we have used the following approach: Identify the columns with missing values. Analyze the nature of the missing data, i.e. whether it is missing at random or not. For columns with missing values, replace them with an appropriate value. For example, for numerical variables, we have used the mean or median, while for categorical variables, we have used the mode.

Outlier Detection: Outliers can have a significant impact on the analysis, and therefore it is essential to identify them. We have identified outliers using the following approach: Analyze the distribution of each variable and identify the values that lie outside the expected range. For numerical variables, we have used box plots to identify outliers. For categorical variables, we have used bar charts to identify values that have an unusually high frequency.

Data Imbalance: Data imbalance is a common issue in classification problems, where one class has significantly fewer data points than the other. We have identified data imbalance using the following approach: Analyze the distribution of the target variable and identify the ratio of data imbalance.

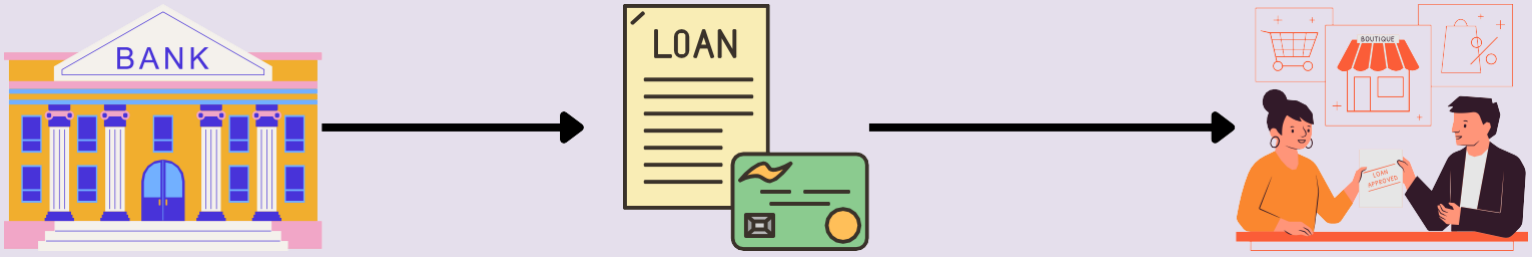
Data Exploration: We have performed univariate, segmented univariate, and bivariate analyses to explore the data and identify patterns and correlations. The following are the key insights from our analysis: Univariate analysis: We have analyzed the distribution of



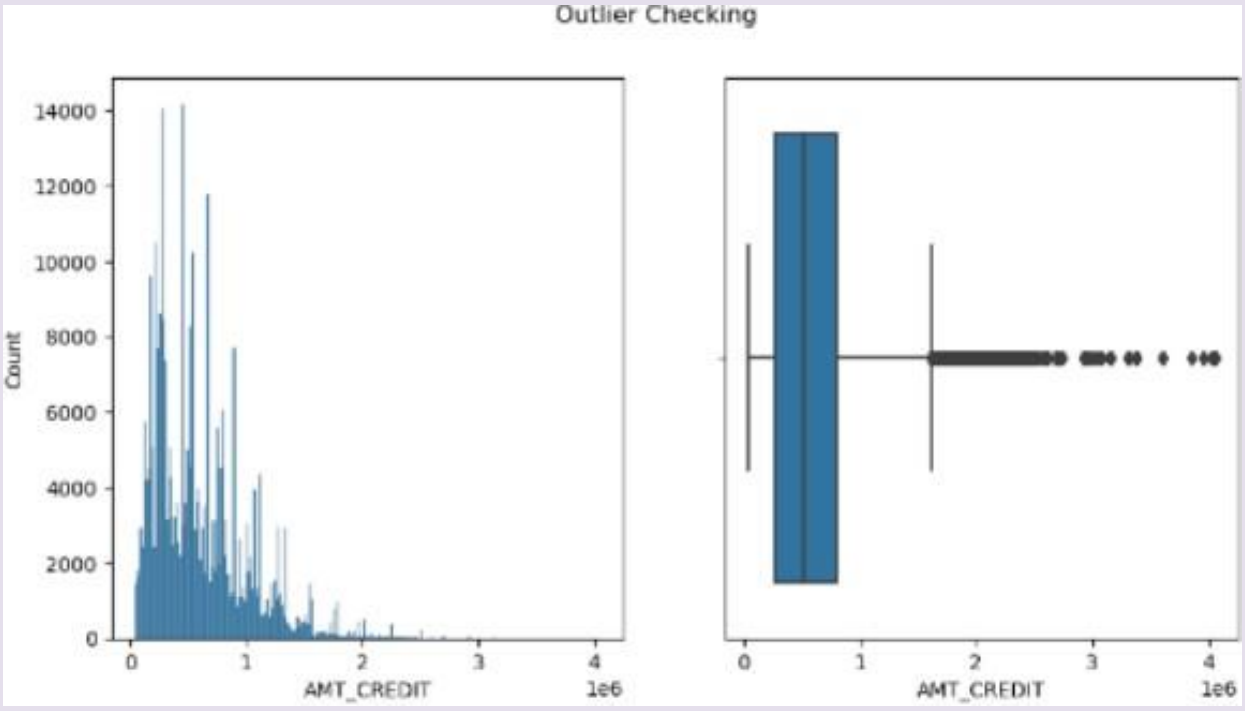
Bank Loan Case Study

each variable and identified their mean, median, and mode values. We have also identified the range, quartiles, and standard deviation for numerical variables. We have used appropriate graphs such as histograms, box plots, and bar charts to visualize the distribution of each variable. Segmented univariate analysis: We have segmented the data based on the target variable and performed univariate analysis for each segment. We have identified the key differences between the segments and the variables that have a significant impact on the target variable. Bivariate analysis: We have analyzed the correlation between pairs of variables using correlation matrices and scatter plots. We have identified the top 10 correlations for the Client with payment difficulties and all other cases (target variable). Visualization: We have used appropriate graphs and charts to summarize and visualize the insights. The following are some of the key graphs and charts we have used: Histograms: To visualize the distribution of numerical variables. Box plots: To identify outliers in numerical variables. Bar charts: To visualize the frequency of categorical variables. Scatter plots: To visualize the correlation between pairs of variables

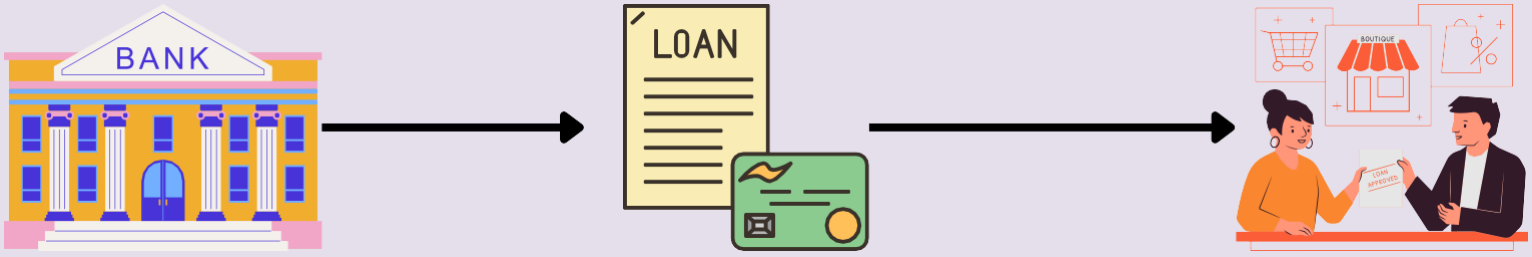
The analysis was performed using Python Jupyter Notebook.Various libraries such as Pandas , NumPy , Seaborn and Matlplotlib are used for analysis and visualization ..



Bank Loan Case Study



This shows outliers are present in data.



Bank Loan Case Study

The ratio of data imbalance.

```
round(df1.CODE_GENDER.value_counts()/len(df1)*100,2)
```

```
F      65.83
M      34.16
XNA     0.00
Name: CODE_GENDER, dtype: float64
```

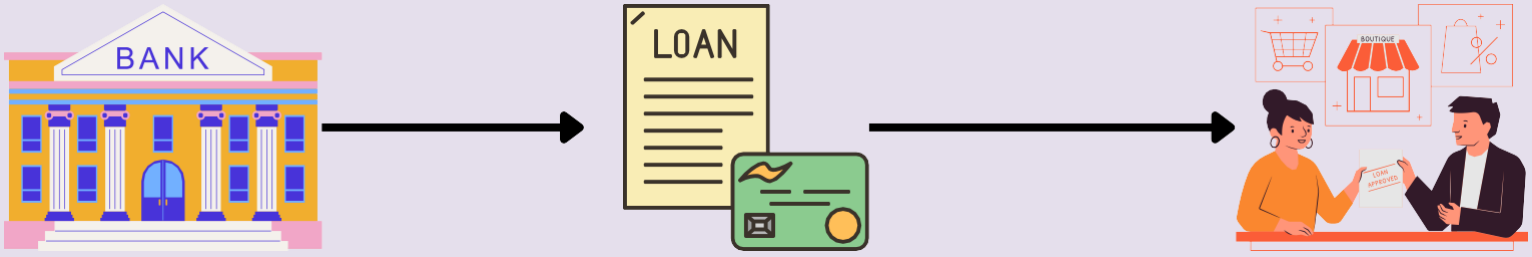
```
round(df1.TARGET.value_counts()/len(df1)*100,2)
```

```
0    91.93
1     8.07
Name: TARGET, dtype: float64
```

```
round(df1.NAME_INCOME_TYPE.value_counts()/len(df1)*100,2)
```

```
Working                 51.63
Commercial associate    23.29
Pensioner               18.00
State servant            7.06
Unemployed               0.01
Student                  0.01
Businessman              0.00
Maternity leave          0.00
Name: NAME_INCOME_TYPE, dtype: float64
```

This shows the percentage of data imbalance in various columns.

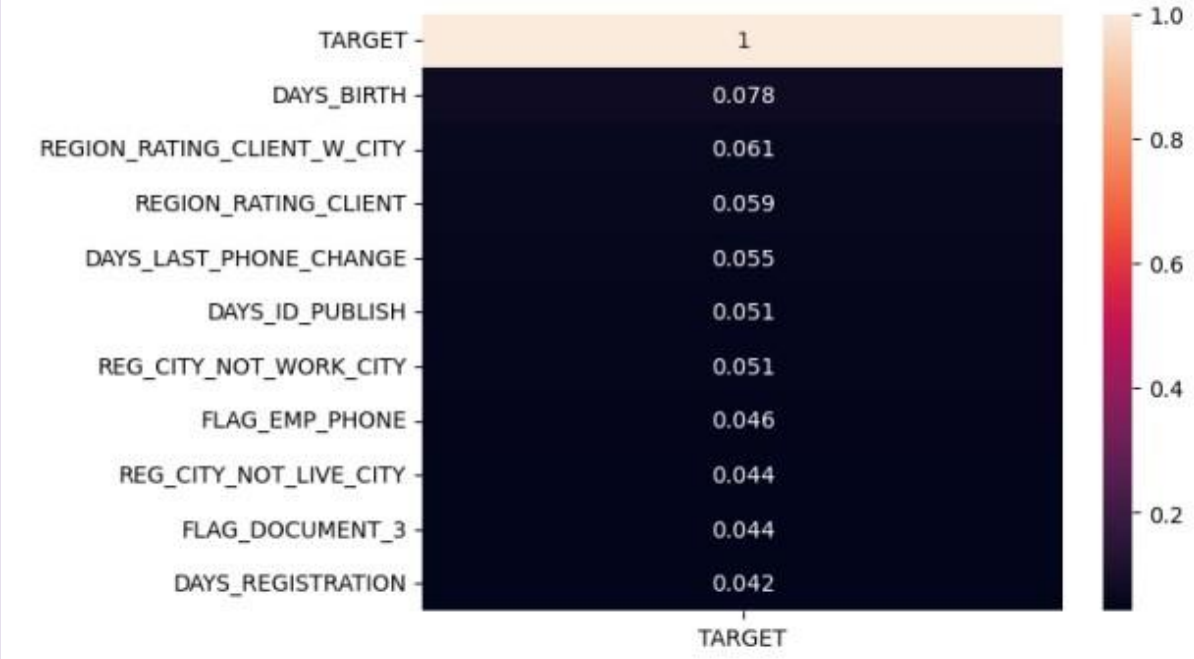


Bank Loan Case Study

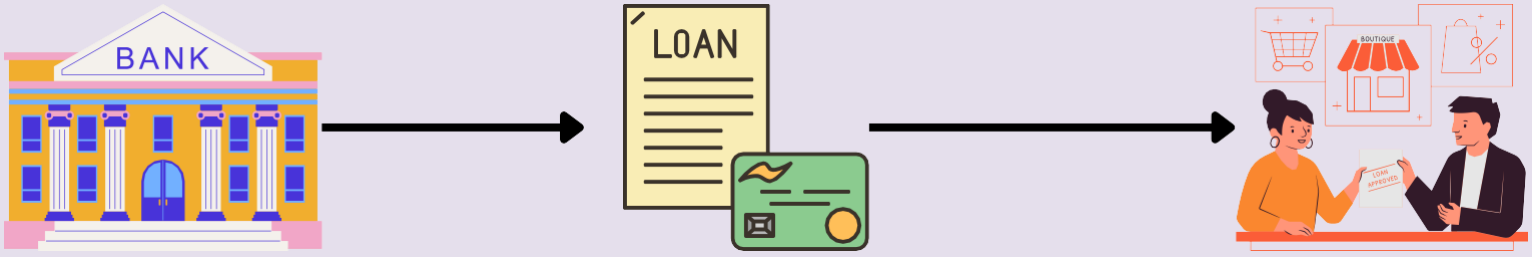
Top 10 Correlations with TARGET variable

```
corr = df1.corr()[['TARGET']].sort_values(by='TARGET', ascending=False).head(11)
sns.heatmap(corr, annot=True)
```

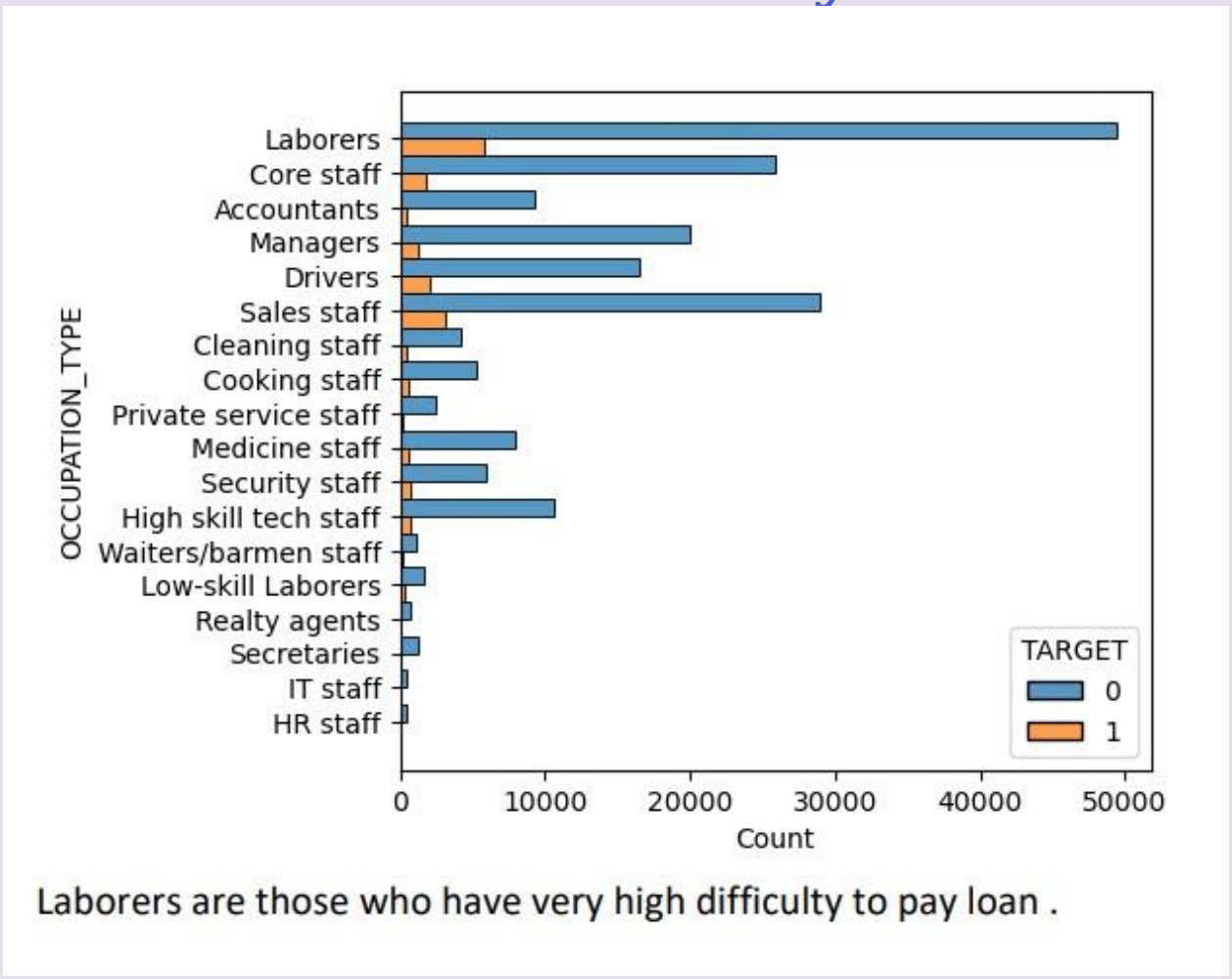
```
<AxesSubplot:>
```



This shows that columns Days_Birth and REGION_RATING_CLIENT_W_CITY are most related with Target.



Bank Loan Case Study



After Laborers Core staff are the ones who have very high difficulty To pay loan.

Bank Loan Case Study

Analysis

Using the Why's approach I am trying to find some more useful insights

- **Why is it that the target_variable is of so much importance?**
- **---> In this dataset target_variable represents whether the client had some payment issues(1) or the client didn't had some payment issues(0); It is important because the target_variable decides whether the bank should increase/decrease it's interest rates on various loans given by the bank; Also in this case almost 92% of the clients didn't had any payment issues and only 8% of them had payment issues, this tells that bank's credit score is good and it has very less or no Non-preforming Accounts.**

- **Why is it that proportion of Female clients more than that of the Male clients?**
- **---> In countries like India especially there have been laws made by the Government for Women who want to establish their own Start-up, Business or their own classes, catering services, etc.; These laws offer loans to women clients at a relatively low interest rates; Also in some cases people purposely use their retired/household mother or household wife so that they can get some sort of concession i.e. low interest rates while applying for Home loans**

- **Why should bank prefer other Housing type clients though House/Apartments Housing type clients have the highest proportion of non-defaulters?**
- **----> Cause people in other groups like Municipal Apartment, Co-op Apartment, Rented Apartment, with Parents are in the search of their own house of their own name plate; Also now a days in India the joint family system is declining and the future generations opt to live in their own 1/2 BHK's rather than living together will all family members in big Family Apartments**

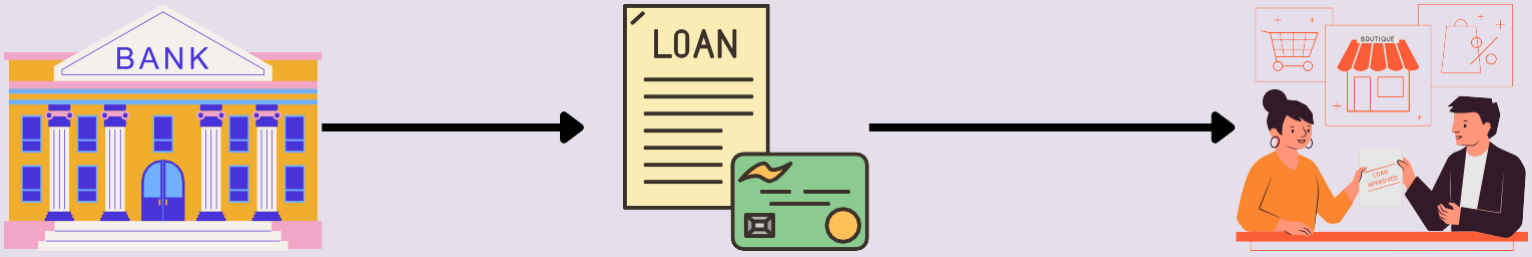
Bank Loan Case Study

Analysis (Cont..)

- **Why should bank opt for working class clients more than the state-government class clients though state-government employees enjoy a lot of benefits and regular salary?**
- **----> It is true that state government employee enjoy a lot of benefits but they also get housing allowances greater than that of working class and in some cases they even get an apartment to live with their families as long as they work for the state government; On the other hand the working class don't enjoy such housing allowances or get very less of it, also the working class don't get an apartment to live in for their entire professional life(i.e. until retirement) and so working class opt for purchasing their own house by taking house loan**

- **Why should Bank not go for approving loans to 'Laborers' occupation_type clients though they have the highest non-defaulters count?**
- **-----> Laborers take only personal loans for marriage or house repair purpose and their loan amount is also less and the interest on such loans is also less as compared to home loan, car loan, etc. which in turn will cause less profits to the bank.**

- **Why is it that females with low income group have the lowest count of defaulters?**
- **----> Females belonging to such groups take loan of small amounts just for starting their own start-ups, business or catering/ parlor services and they usually enjoy benefit from government schemes for such purpose**



Bank Loan Case Study

Conclusion

In conclusion, I would like to conclude the following:-

- **The proportion/percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92%**
- **The Bank generally lends more loan to Female clients as compared to Males clients as the count of Female clients in the defaulter's list is less than that of Males. Still Bank can look for more Male clients if their credit amount is satisfied**
- **Also the clients who belong to Working class tend to pay their loans on time followed by the clients who fall under Commercial Associate**
- **Clients having Education status like Secondary/ Higher Secondary or more tend to pay loan on time so bank can prefer lending loans to clients having such Education Status**
- **Clients taking loan for purchasing New Home i.e. clients taking Home Loans or purchasing New Car i.e. Car Loans and clients who have a income type as State Servant tend to pay their loans on time and hence Bank should prefer clients having such background**
- **The Bank should be more cautious when lending money to clients with Repairs purpose because they have high count of Defaulters along with High count of Defaulters**

Analyzing the Impact of Car Features on Price and Profitability

Description

The automotive industry has been rapidly evolving with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars. This project aimed to analyze the relationship between a car's features, market category, and pricing, and identify which features and categories are most popular among consumers and most profitable for the manufacturer. The goal was to develop a pricing strategy that balances consumer demand with profitability and identifies which product features to focus on in future product development efforts.

Analyzing the Impact of Car Features on Price and Profitability

The Problem

You need to build an interactive dashboard in Excel from the tasks given below:

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

- Task 1.A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.
- Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

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

- Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

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

- Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

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

Analyzing the Impact of Car Features on Price and Profitability

- Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.
- Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.

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

- Task 5.A: Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.
- Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

Analyzing the Impact of Car Features on Price and Profitability

Design

Data:
The dataset used in this project contained information on over 11,000 car models and their specifications, including details on the car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP). The dataset was cleaned and prepared before analysis to ensure accurate and reliable results.
Data Cleaning: Python and its libraries, such as Pandas, NumPy, Matplotlib, Seaborn, and YData Profiling, were used for the data cleaning process inside Jupyter Notebook.
Steps Performed to clean the data are:
Steps performed:
step1: importing necessary libraries
step2: loading csv file
step3: analyzing columns
step4: cleaning the DataFrame
1)droping duplicate rows
2)check dtypes
3)check null values
4)fill null values
step5: Save file after data cleaning

Analyzing the Impact of Car Features on Price and Profitability

Design(cont...)

Building Dashboard:

First we load the Cleaned_Car_data.csv file in Tableau
Then we create following Worksheets:
Task1: For Model Building I created a parameter to filter the model for each feature and then create a calculated field Car_Features.

```
CASE [Features/Pricing]
WHEN 'City Mpg' THEN [City Mpg]
WHEN 'Engine Cylinders' THEN [Engine Cylinders]
WHEN 'Engine HP' THEN [Engine HP]
WHEN 'highway MPG' THEN [highway MPG]
WHEN 'Msrp' THEN [Msrp]
WHEN 'Number of Doors' THEN[Number of Doors]
WHEN 'Popularity' THEN [Popularity]
END
```

Task2 Pivot Table: For this another Calculated field is created named Market Category Segment to Segment Each category .
Task 2 Bar Chart: Same Market Category Segment is used to create Bar Chart.
Task 3 :We use same Parameter Features/Pricing to show how different features affect the price.
Task 4:Here we Created a calculated field named Profitabilty
As Profitabilty= [Msrp]*[Popularity] and use this to build the Regression Model and filter the Model using feature(model).
Task 5:Year is changed to date type measure. Then we use Parameter Feature/Pricing to show the trend of car_features and price with line chart.
After that all the worksheets are combined to Build Dashboard.

Analyzing the Impact of Car Features on Price and Profitability

<u>Findings</u> - 1

Trend Lines Model

A polynomial trend model of degree 3 is computed for Msrp given Car_Features. The model may be significant at p <= 0.05.

Model formula: Features/Pricing*(Car_Features^3 + Car_Features^2 + Car_Features + intercept)

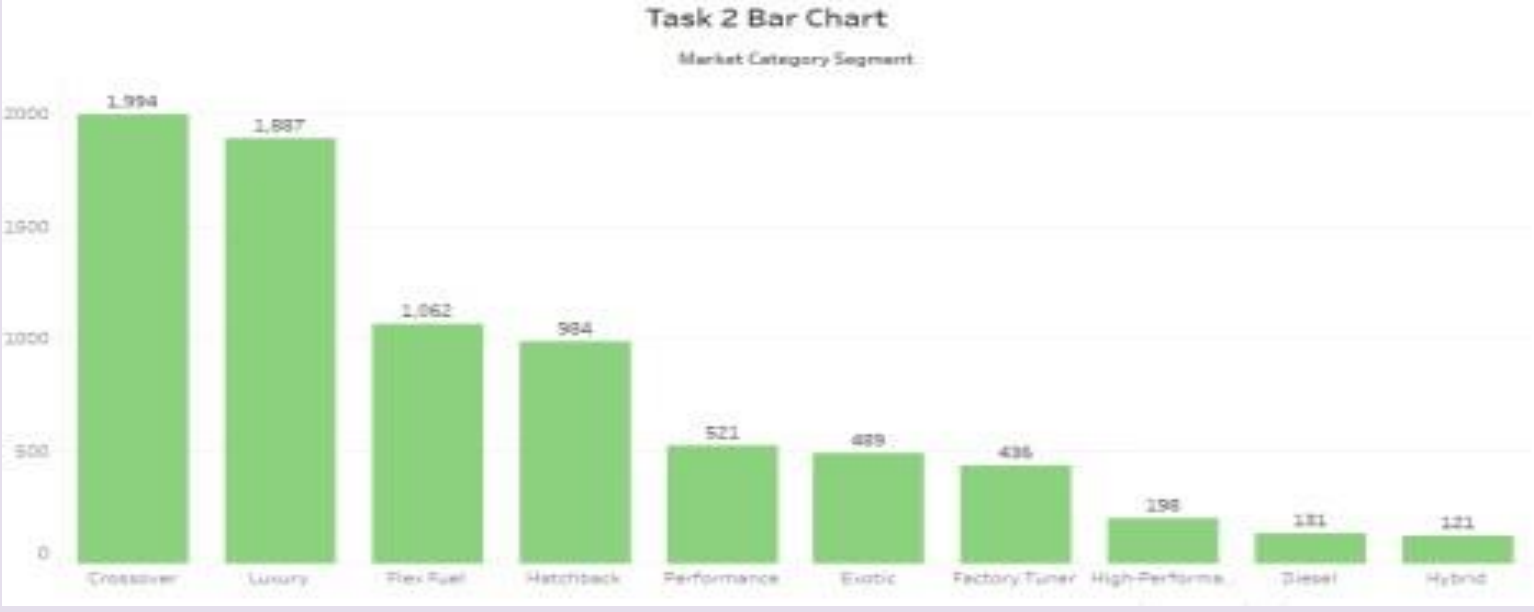
For Engine Cylinders: Msrp = 583.318*Car_Features^3 + -6953.02*Car_Features^2 + 26788*Car_Features + -5440.43

Analyzing the Impact of Car Features on Price and Profitability

Finding -2

Pivot Table Market Segment

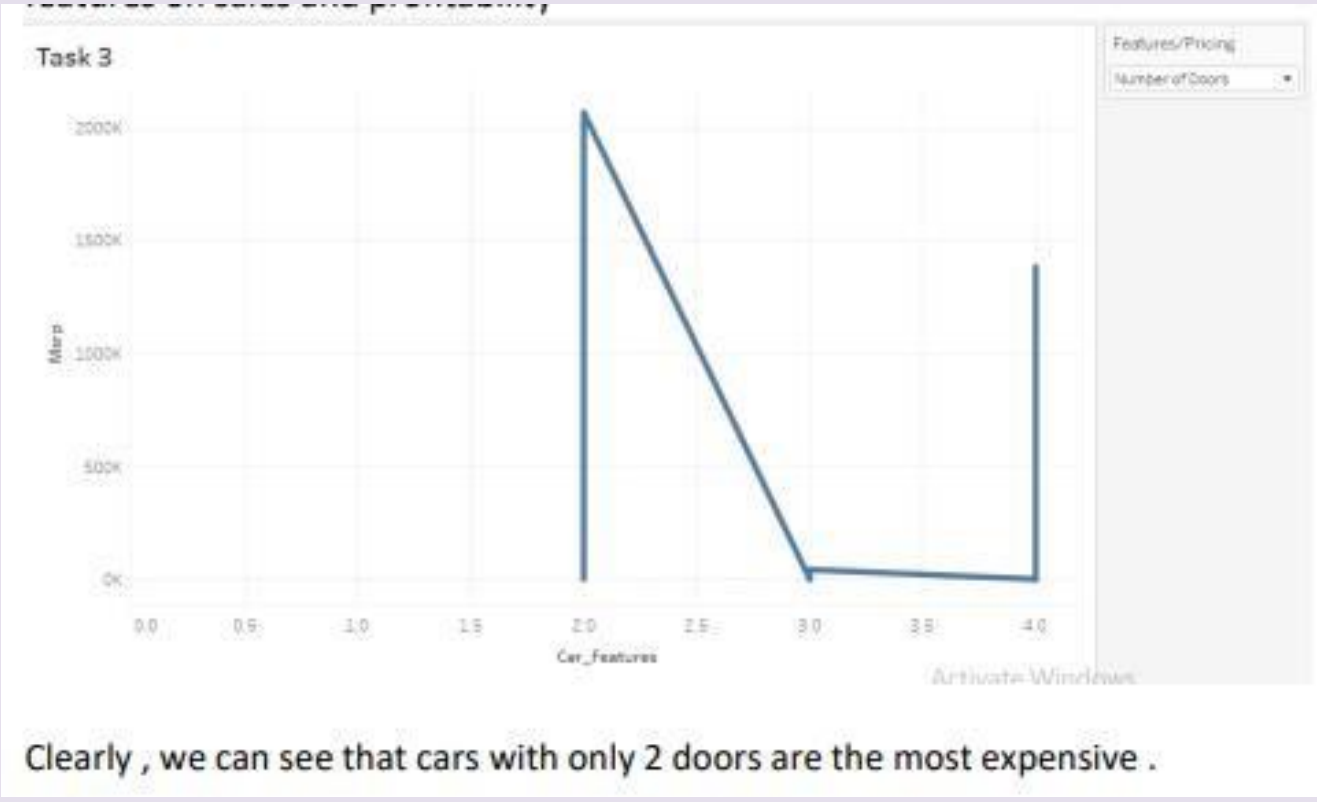
Market Category S..	
Crossover	1,994
Diesel	131
Exotic	489
Factory Tuner	436
Flex Fuel	1,062
Hatchback	984
High-Performance	198
Hybrid	121
Luxury	1,887
Other	3,376
Performance	521



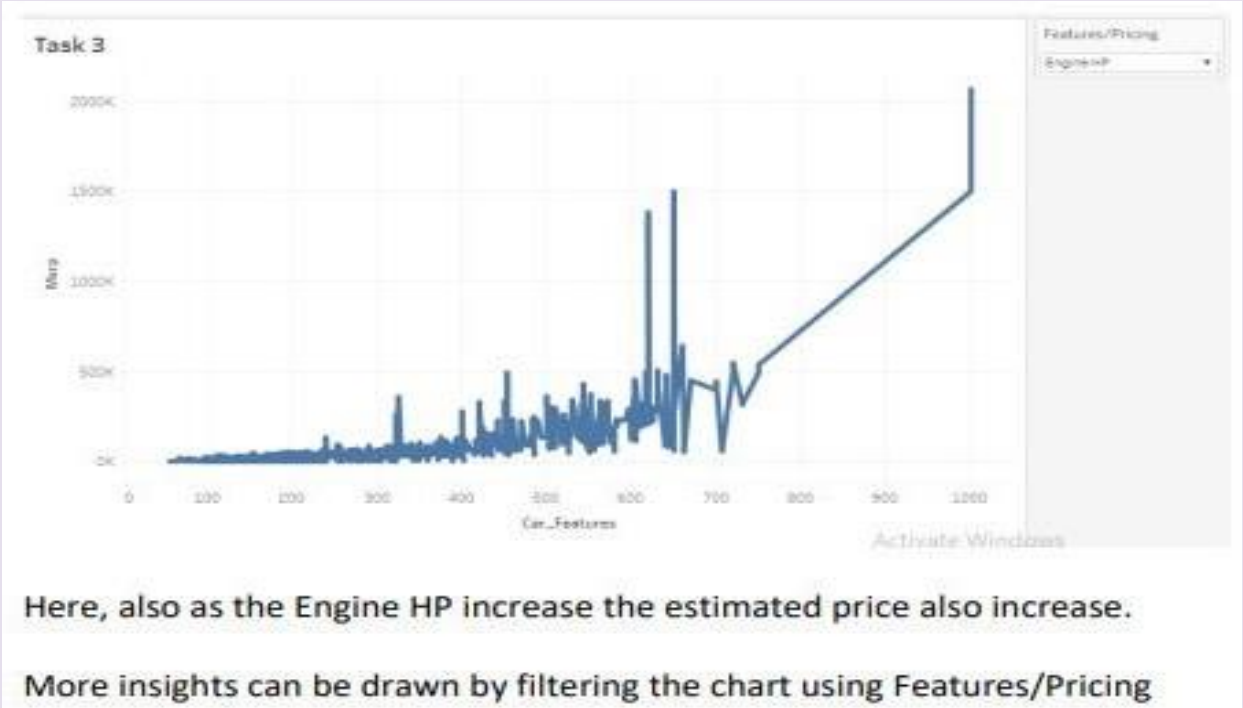
With this Bar chart we conclude that Crossover and Luxury are covering
most of the market segment.

Analyzing the Impact of Car Features on Price and Profitability

Findings -3



Clearly , we can see that cars with only 2 doors are the most expensive .

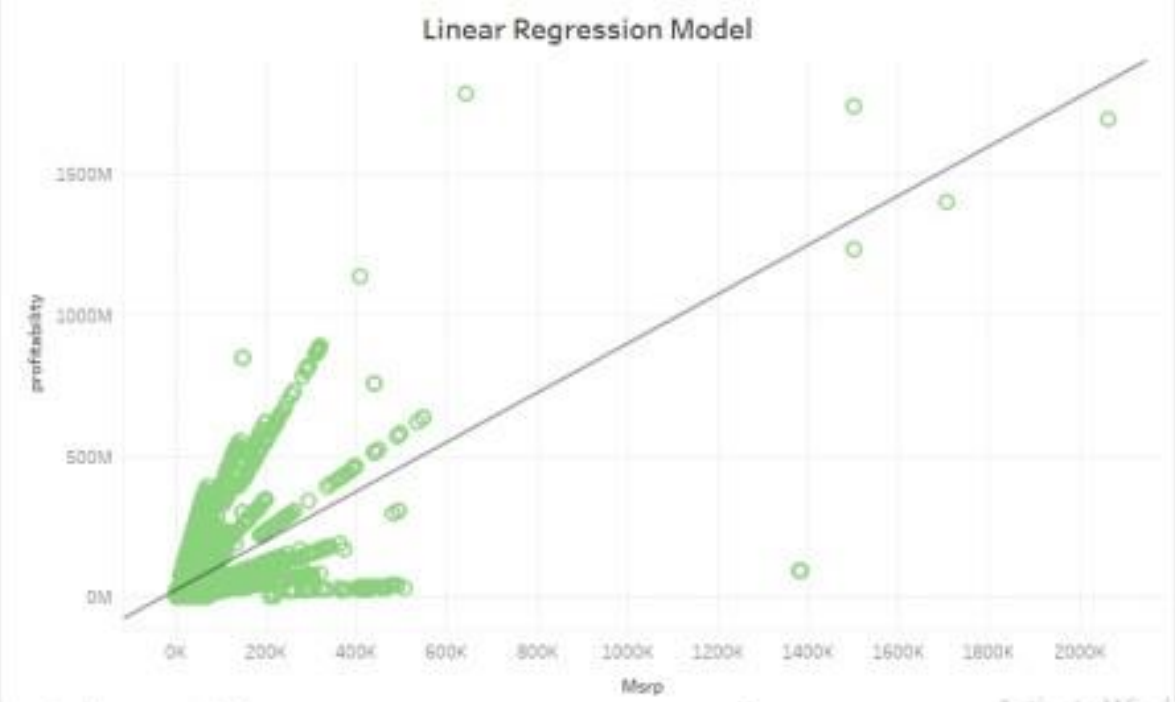


Here, also as the Engine HP increase the estimated price also increase.

More insights can be drawn by filtering the chart using Features/Pricing

Analyzing the Impact of Car Features on Price and Profitability

Findings -4



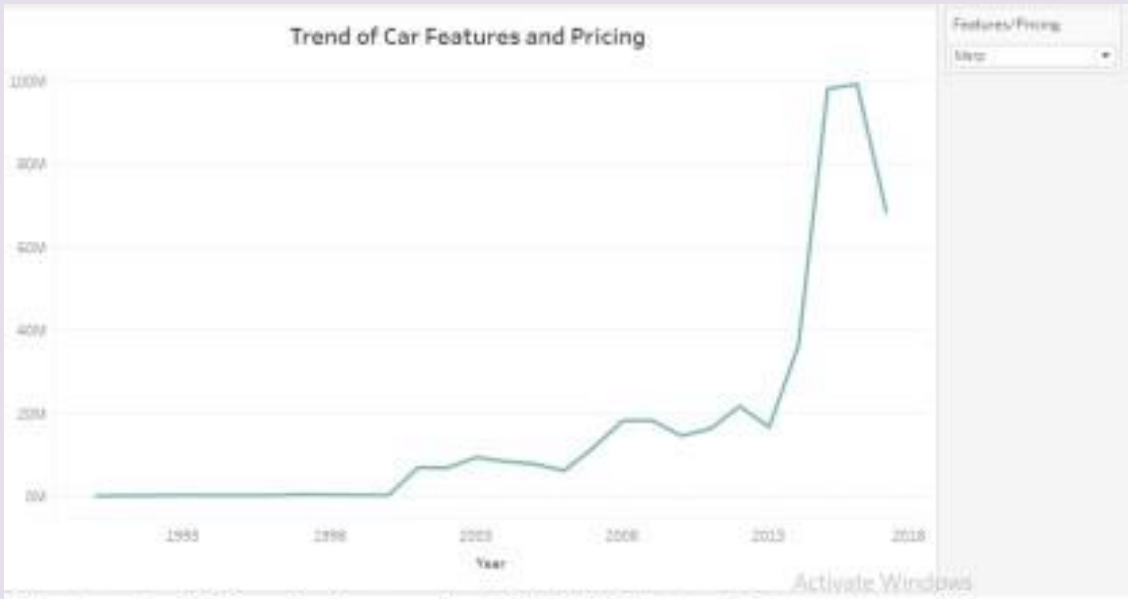
With this model we can say to get maximum profit. We can increase price along with demand or Popularity. With High Demand High Price Profit will be maximum.

Model Building:- Develop a model that relates price to demand and profitability: To do this, we can use a simple linear model that relates the price of a car to its demand and profitability:

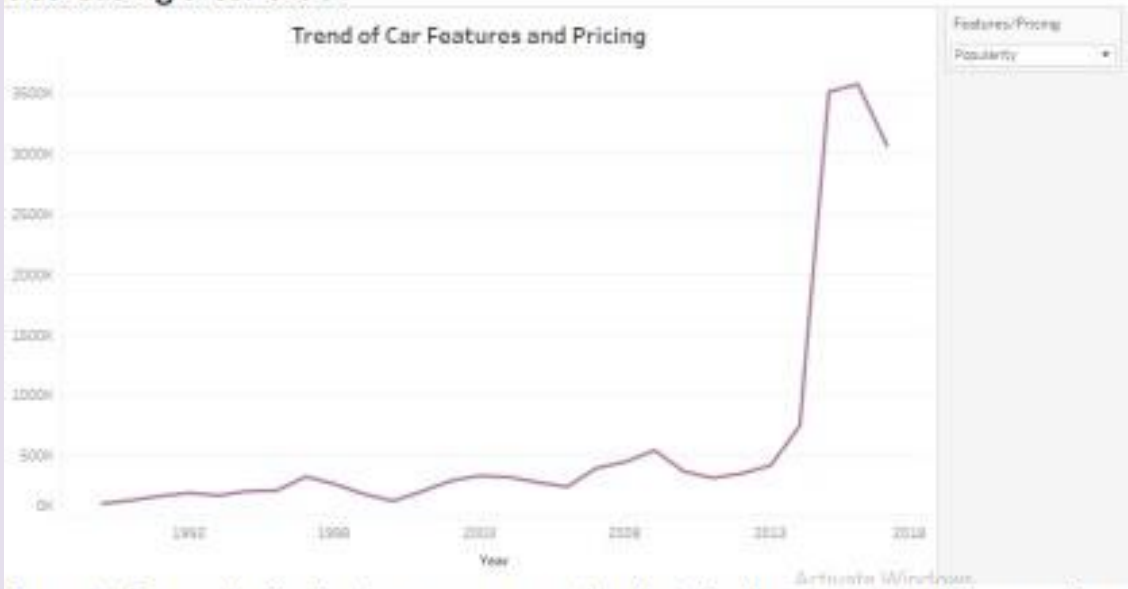
Profitability = Price * Popularity(Demand)

Analyzing the Impact of Car Features on Price and Profitability

Findings -5



The trend of Msrp is increasing till 2016 but unfortunately , it is decreasing after that .



Trend of popularity is also same may be that is the reason of decreasing trend of price after 2016.

Analyzing the Impact of Car Features on Price and Profitability

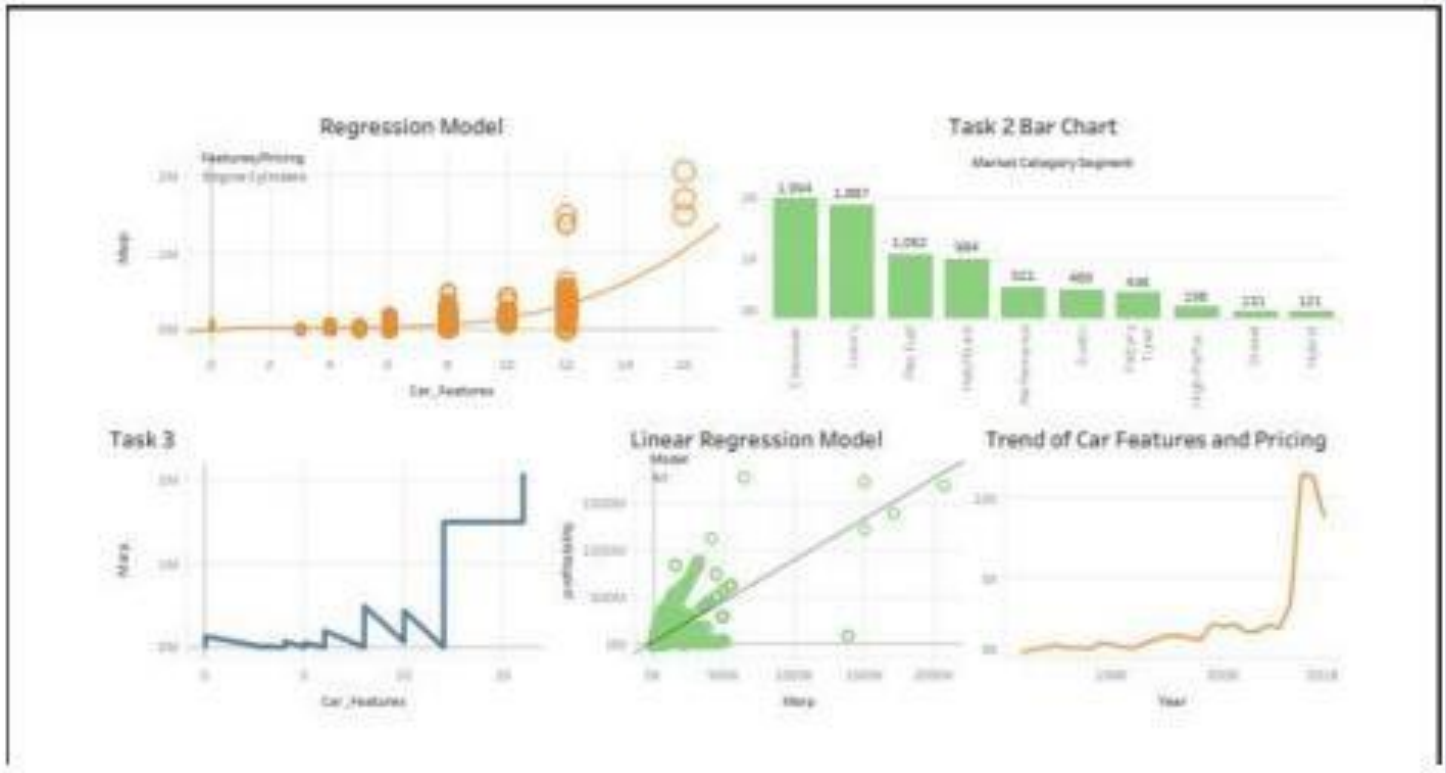
Analysis

This project aimed to analyze the impact of car features on price and profitability and develop a pricing strategy that balances consumer demand with profitability. The results showed that car features, market category, and manufacturing costs all play important roles in determining the optimal pricing strategy. The project provided valuable insights to a car manufacturer and helped them optimize their pricing and product development decisions to maximize profitability while meeting consumer demand.

Analyzing the Impact of Car Features on Price and Profitability

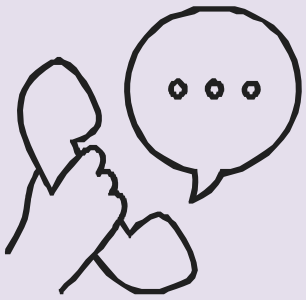
Conclusion:-

Final dashboard look like this:-



The results showed that car features, market category, and manufacturing costs all play important roles in determining the optimal pricing strategy. The project provided valuable insights to a car manufacturer and helped them optimize their pricing and product development decisions to maximize profitability while meeting consumer demand.





ABC Call Volume Trend Analysis

Description

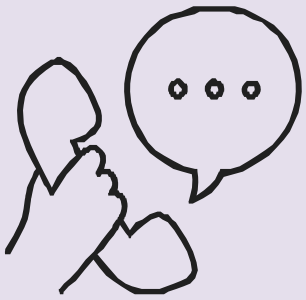
A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. Typically, these teams fulfil various roles and responsibilities such as: Customer experience programs (CX programs), Digital customer experience, Design and processes, Internal communications, Voice of the customer (VoC), User experiences, Customer experience management, Journey mapping, Nurturing customer interactions, Customer success, Customer support, Handling customer data, Learning about the customer journey.

Let's look at some of the most impactful AI-empowered customer experience tools you can use today:
Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, Intelligent Routing

In a Customer Experience team there is a huge employment opportunities for Customer service representatives A.k.a. call centre agents, customer service agents. Some of the roles for them include: Email support, Inbound support, Outbound support, social media support.

Inbound customer support is defined as the call centre which is responsible for handling inbound calls of customers. Inbound calls are the incoming voice calls of the existing customers or prospective customers for your business which are attended by customer care representatives. Inbound customer service is the methodology of attracting, engaging, and delighting your customers to turn them into your business' loyal advocates. By solving your customers' problems and helping them achieve success using your product or service, you can delight your customers and turn them into a growth engine for your business.





ABC Call Volume Trend Analysis

The Problem

Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

- **Show the total volume/ number of calls coming in via charts/ graphs [Number of calls v/s Time]. You can select time in a bucket form (i.e. 1-2, 2-3, ...)**

- **As you can see current abandon rate is approximately 30%. Propose a manpower plan required during each time bucket [between 9am to 9pm] to reduce the abandon rate to 10%. (i.e. You have to calculate minimum number of agents required in each time bucket so that at least 90 calls should be answered out of 100.)**

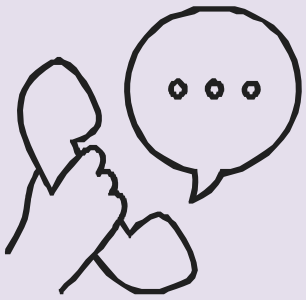
- **Let's say customers also call this ABC insurance company in night but didn't get answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose every 100 calls that customer made during 9 Am to 9 Pm, customer also made 30 calls in night between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

- **Now propose a manpower plan required during each time bucket in a day. Maximum Abandon rate assumption would be same 10%.**

Assumption: An agent work for 6 days a week; On an average total unplanned leaves per agent is 4 days a month; An agent total working hrs is 9 Hrs out of which 1.5 Hrs goes into lunch and snacks in the office. On average an agent occupied for 60% of his total actual working Hrs (i.e 60% of 7.5 Hrs) on call with customers/ users. Total days in a month is 30 days.





ABC Call Volume Trend Analysis

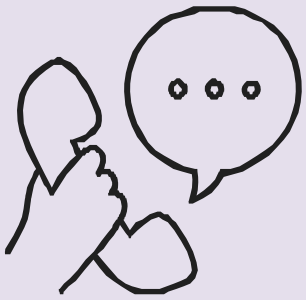
design

Steps Performed:
step1: Importing Necessary Libaries
step2: Loading csv file to DataFrame
step3:Data Cleaning
1)check dtypes
2)drop Duplicates
3)check null %
step4) EDA and Question's Answer

Tech-Stack Used:

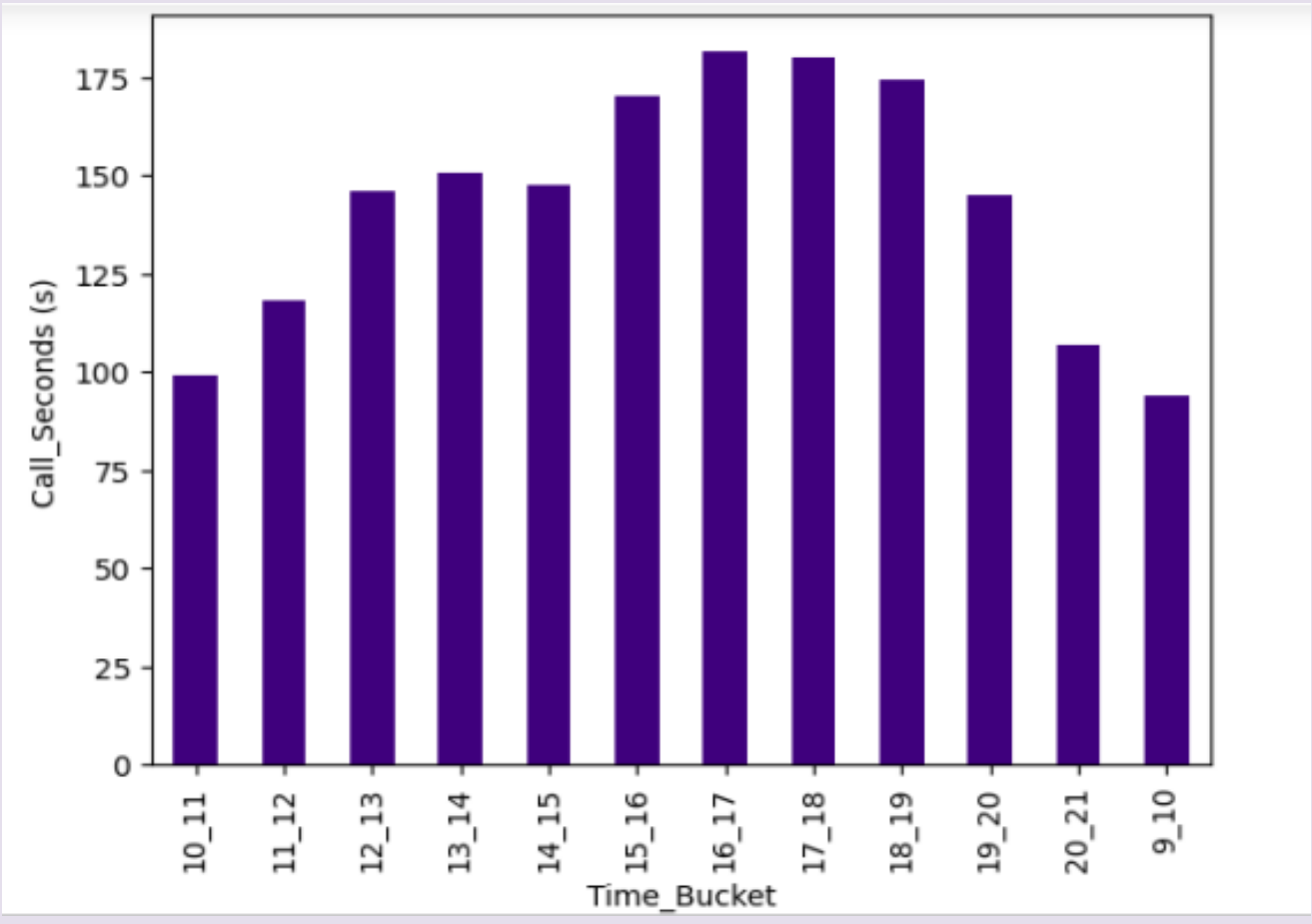
Jupyter Notebook:
The analysis was performed using Jupyter Notebook, with Python and its libraries, such as Pandas, NumPy, Matplotlib, and Seaborn, utilized to perform the analysis.





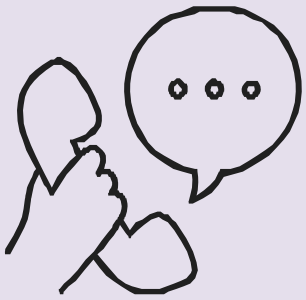
ABC Call Volume Trend Analysis

Findings-1



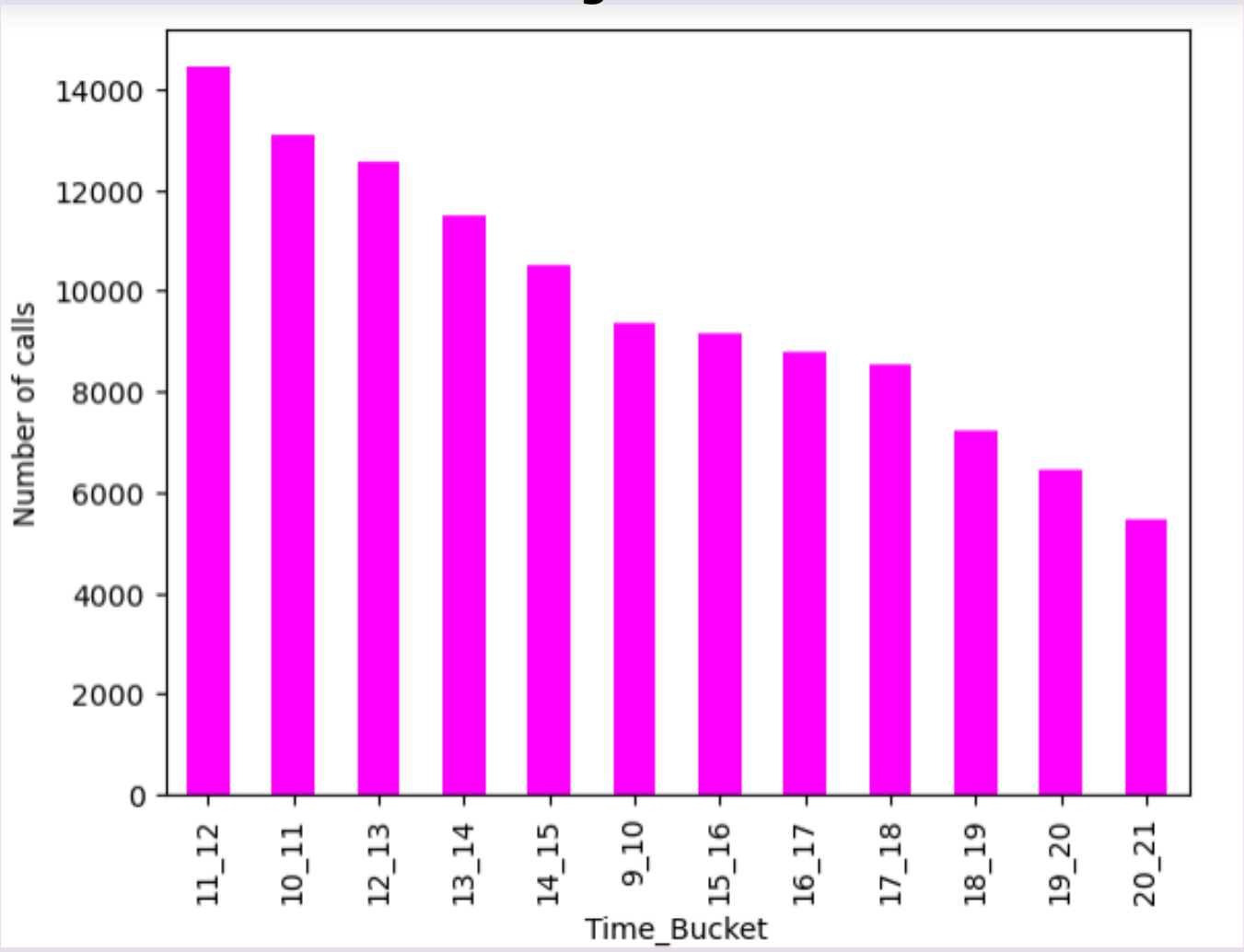
The average call time duration for all incoming calls received by agents is the highest in between 4 pm to 5 pm and from 5 pm to 6 pm
The average call time duration for all incoming calls received by agents is the least in between 9 am to 10 pm.





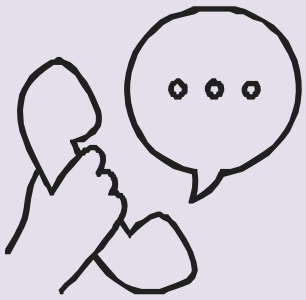
ABC Call Volume Trend Analysis

Findings:-2



The customers call the most in between 11 am to 12 noon.
The customers call the least in between 8 pm to 9 pm.





ABC Call Volume Trend Analysis

Findings:-3

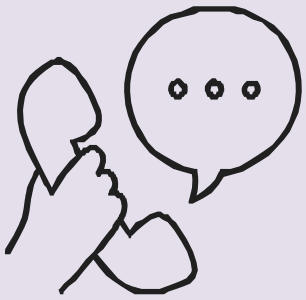
No of agents required in each time_bucket:-

Time_Bucket	
10_11	6
11_12	7
12_13	6
13_14	6
14_15	5
15_16	4
16_17	4
17_18	4
18_19	3
19_20	3
20_21	3
9_10	4

Name: Call_Seconds (s), dtype: int32

Approx. 29% of the calls are abandoned, 1% is transferred, while 70% of
the calls are answered from all incoming calls.
Total agents required to answer the 90% of the calls per day is 56.





ABC Call Volume Trend Analysis

Findings:-4

No of agents needed at night:

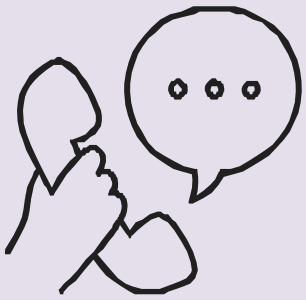
	Time_Bucket	num_of_agents_req
0	9pm-10pm	2
1	10pm-11pm	2
2	11pm-12am	1
3	12am-1am	1
4	1am-2am	1
5	2am-3am	1
6	3am-4am	1
7	4am-5am	1
8	5am-6am	2
9	6am-7am	2
10	7am-8am	2
11	8am-9am	3

There are least number of calls in the night so , the number of agents required to answer calls are also less.

The number of agents needed for night shift work are 17.

Number of calls in between 12am to 5am are least.





ABC Call Volume Trend Analysis

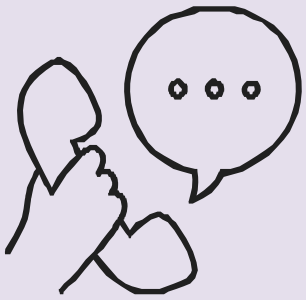
Analysis

Using the Why's approach I am trying to find some more insights:-

- **Why is that the average call answered were more in count in the time bucket of 10_11, 18_19, 19_20 and 20_21 as compared to other time buckets?**
- **---> Most of the customers are office people and they need to reach office by 10 AM or 11 AM, so these customers call during 10_11 time bucket i.e. while they in transit to office or have reached office and have some free time before they start their work; During the time bucket 18_19, 19_20 and 20_21 the customers have either left their office and reached home or they are in the transit to reach home and during these time period i.e. 6 Pm to 9 Pm people have free time where they can share their concern to the customer service. During these time buckets most of the calls are from individual people with small problems which can be resolved quickly**

- **Why is it that the time bucket 11_12 has the highest number of incoming calls but it does not have the highest number of average answered calls?**
- **---> Maybe there were more number of incoming calls in the time bucket 11_12 and there were not enough personnel to handle most of the queries of the customers during the 11_12 time bucket**





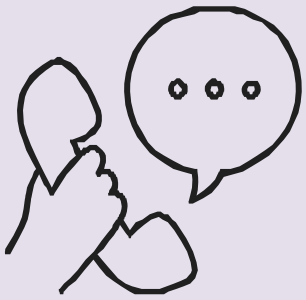
ABC Call Volume Trend Analysis

Analysis (Cont...)

- **Why is it that the total number of incoming calls reached it's peak value during the time bucket 11_12 and got decreased from time bucket 12_13 onwards?**
- **---> It is a general tendency of the customers(people) that they want their query/complaint get resolved on that particular day itself when they called the customer center; so most of the customers try to place their complaint/query before 12 Pm so that by the end of the day their complaint gets resolved depending upon the complexity of the problem faced by the customer**

- **Why is proportion if the monthly transfer rate is less than compared to monthly answered and abandon rate?**
- **---> In most of the customer service centers they have the dedicated toll free number of the particular problem faced by the customer, also there are skilled people at the call center who are well versed with the problems they come across while handling and guiding thousands of customers on daily basis; And so most of the calls gets answered by providing an solution to the query, some of the calls get abandon due to unavailability or shortage of the skilled person, and very few calls gets transferred from the junior level to senior level if the problem is too complex for the junor level expertise**



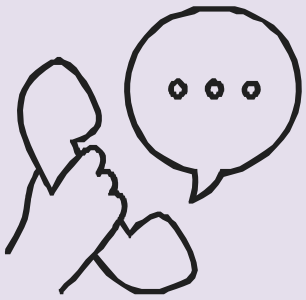


ABC Call Volume Trend Analysis

Analysis (Cont...)

- **Why is that one cannot provide the exact distribution of agents during the night time i.e. from 9 PM to 9 AM if the number of agents available during the night shift are already defined, so as to keep the abandon rate 10%?**
- **---> For this particular case, Since we have only 17 agents during night we need to distribute in an non analytical way i.e. the agents who work in 19_20, 20_21 time bucket to wait and work in 21_22 and 22_23 time buckets as well. Also agents who work during 9_10, 10_11 time bucket can be asked to work for 7_8 and 8_9 time bucket as well. he agents who work in the time bucket 1_2, 2_3, 3_4 and 4_5 can be asked to work in time buckets 6_7, 7_8 and 8_9 so as to keep the abandon rate at 10%. Also, the company needs to consider various factors like how far is the home of the agent if he/she is made to do night shift, Is the transport facility available during the night hours from the agent's home to company and many other factors and hence the exact distribution cannot be given using an analytical approach**





ABC Call Volume Trend Analysis

Conclusion

In conclusion, the proposed manpower plan can help ABC Insurance Company to reduce the abandon rate and improve customer experience. However, the actual implementation of the plan would require further analysis and consideration of additional factors such as agent availability, training, and performance metrics.

Overall, the results of an ABC call volume trend analysis can provide valuable insights for call center managers and help them to optimize their operations for improved customer satisfaction and efficiency.
In this project, I learned how analysts can make an impact in the customer service department and how companies deal with customers to provide them with the utmost satisfaction. I gained knowledge about IVR Duration, an AI tool that answers calls and identifies the customer's exact question and then transfers it to the right agent to answer their queries. Additionally, I learned about behavioral analytics

Appendix

---> Google Drive link for projects:

- **Data Analytics Process:-**

  https://drive.google.com/file/d/11WeMdQTD6AY6jQ7Q6Oy0RL1BixM94pU8/view?usp=sharing

- **Instagram User Analytics:-**

  https://drive.google.com/file/d/1TFZ4zaf0wdmv7-RN11JxcjMSFhpEPJRN/view?usp=sharing

- **Operation Analytics and Investigating Metric Spike Analysis:-**

  https://drive.google.com/file/d/1UinBZ1HrVe89EAc8nVryrd9ejh2QyHI6/view?usp=sharing

- **Hiring Process Analytics:-**

  https://drive.google.com/file/d/1oFlGc_htLyPSAQUfxpleZebRrVDF-hF1/view?usp=sharing

- **IMDB Movie Analysis-**
  https://drive.google.com/drive/folders/1FVB1yxKFk8PvHnRkjCByLLGoIbmandlF?usp=sharing

- **Bank Loan Case Study:-**

  https://drive.google.com/drive/folders/1nK0DYAWcLGNGo20v5BNCvIgZMvoH7pXo?usp=sharing

- **Analyzing the Impact of Car Features on Price and Profitability**

  https://drive.google.com/drive/folders/1A1AcBXPJJaqhBrE2eixvQemAOqBehb15?usp=sharing

- **ABC Call Volume Trend Analysis:-**

  https://drive.google.com/drive/folders/1aT6B0cTQYwvJNwRWerlQLxj1_SBkDqoJ?usp=sharing