Data Analytics

Project Portfolio

By Batul Hussain

LinkedIn

Professional Background

Hello, there!

My name is Batul Hussain and I have completed my BTech in Electronics & Telecommunication from Bhilai Institute of Technology, Durg, Chhatisgarh in the First (Honours) Division.

I am currently working as a Coding and Math instructor for the last 2 years and have completed the Data Analytics Virtual Internship project hosted on Trainity platform.

About Me:

Technical Skills:

- MySQL
- Microsoft Excel
- Data Visualization
- Python Programming Language
- Statistics
- Machine Learning (Basic)

Soft Skills:

- Storytelling
- Research
- Presentation & Report

Table of Content

Project 01:

Data Analytics Process-Application in Real-Life Scenario Case Study

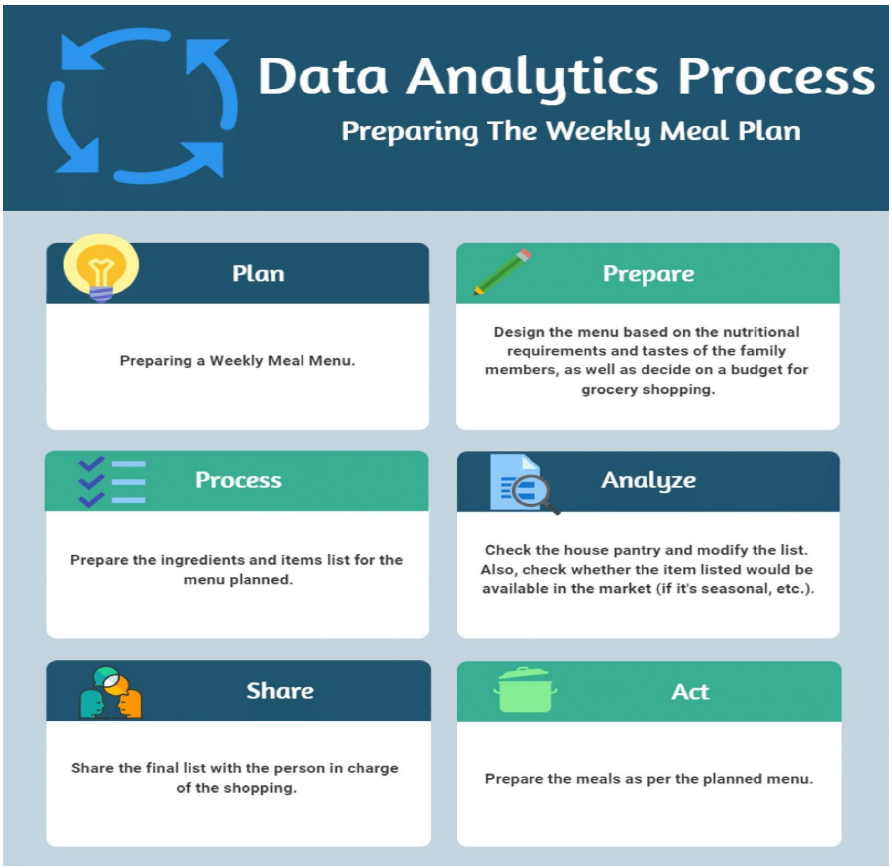
Description:

The task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.

What is Data Analytics?

The science of evaluating raw data such that inferences can be drawn from it is known as data analytics. Techniques for data analytics can highlight metrics and trends that might otherwise be buried in the sea of data. When procedures are optimised, a business or system's overall efficiency can be raised.

Steps of Data Analytics:



Project 02:

Instagram User Analytics

Description:

In order to get business insights for the marketing, product, and development teams, we track how consumers connect with and interact with our digital product (software or mobile application).

Teams from throughout the company utilise this information to develop new marketing campaigns, choose which features to include in apps, gauge the performance of the apps by looking at user interaction, and generally improve the user experience while assisting in business expansion.

As a member of Instagram's product team, the product manager has asked us to share our perspective on the queries posed by the management team.

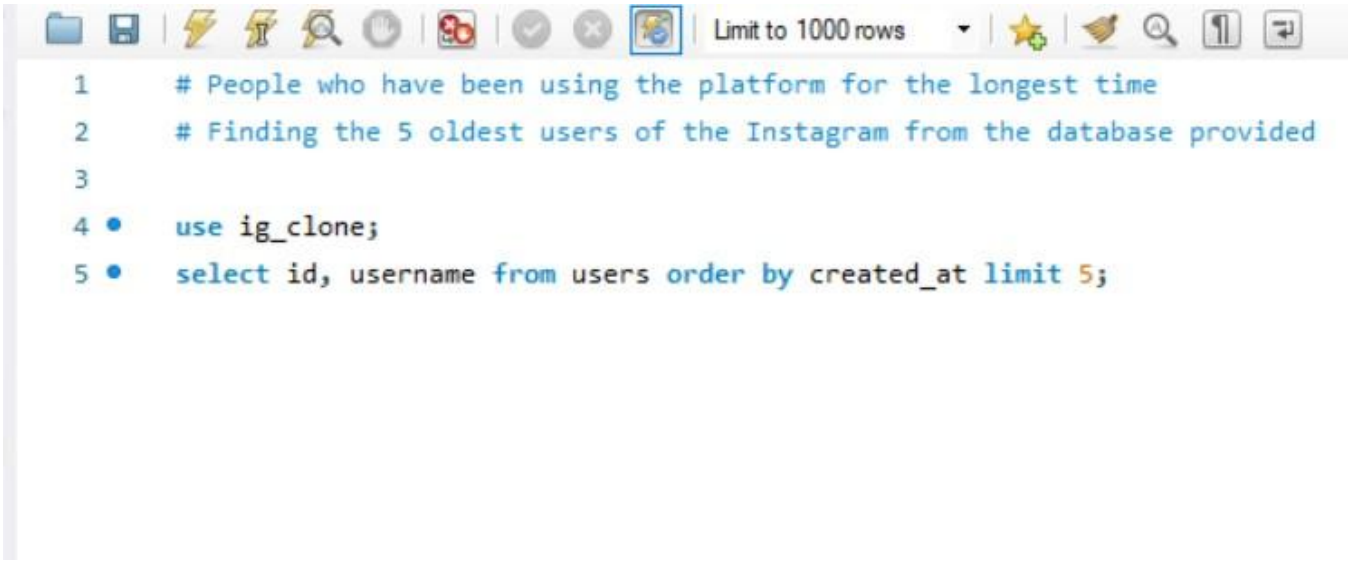
Approach:

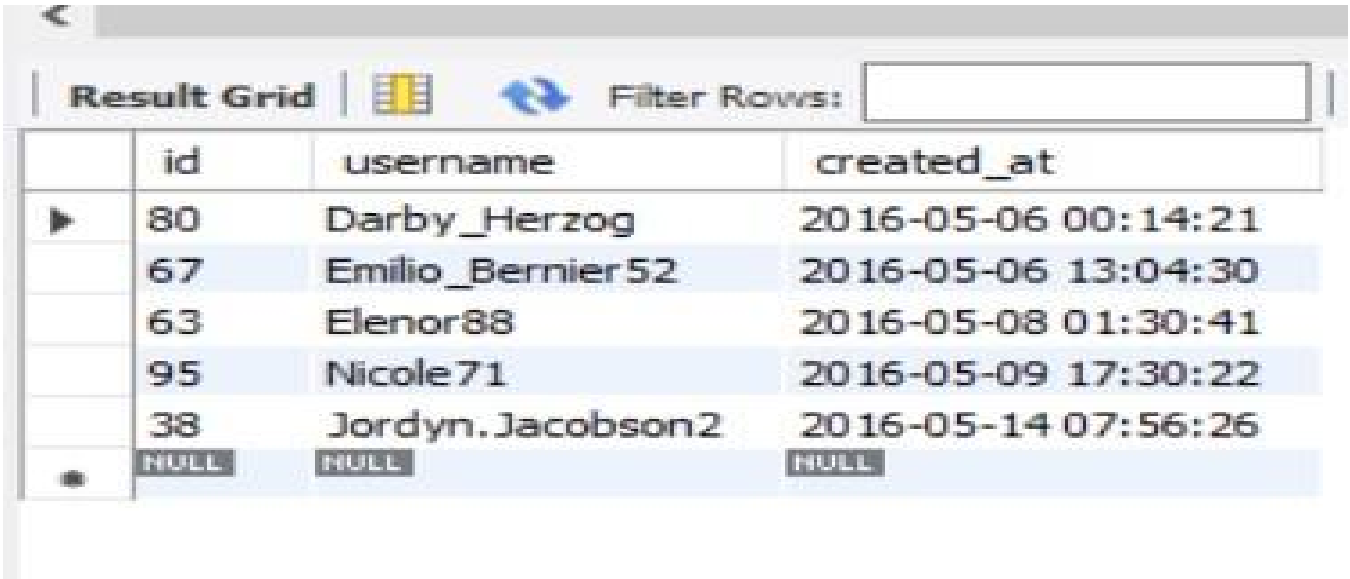
- Creating the relational database of the user's data.
- Going through all the tables and figuring out the relationship between them.
- Going through the values of each table and getting the gist of it.
- Running appropriate SQL queries based on the questions asked on the database
- and figuring out insights from it.

Insights:

Marketing Metrics:

1. **Rewarding Most Loyal Users:**





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

2. **Remind Inactive Users to Start Posting:**

```
# Remind Inactive Users to Start Posting
# Finding the users who have never posted a single photo on Instagram

select id, username from users where id not in (select user_id from photos);
```

id	username
5	Aniya_Hackett
7	Kasandra_Homenick
14	Jadyn81
21	Rocio33
24	Maxwell.Halvorson
25	Tierra.Trantow
34	Pearl7
36	Ollie_Ledner37
41	Mckenna17
45	David.Osinski47
49	Morgan.Kassulke
53	Linnea59
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
68	Franco_Keebler64
71	Nia_Haag
74	Hulda.Macejkovic
75	Leslie67
76	Janelle.Nikolaus81
80	Darby_Herzog
81	Esther.Zulauf61
83	Bartholome.Bernhard
89	Jessyca_West

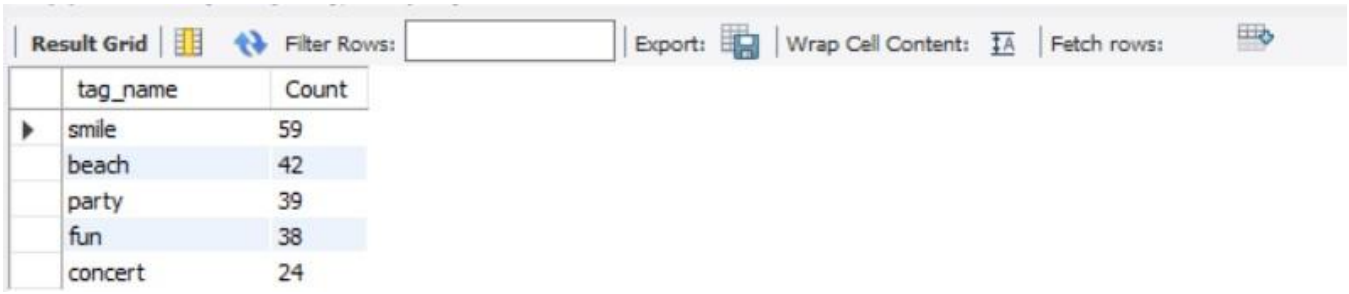
3. **Declaring Contest Winner: The team started a contest and the user who gets the most likes on a single photo will win the contest now they wish to declare the winner.**

```
#Declaring Contest Winner:
#Identifying the winner of the contest and provide their details to the team

#Step 1: Finding the maximum liked photo_id
select photo_id, count(user_id) as `No. of Likes` from likes group by photo_id order by `No. of Likes` desc limit 5;

#Step 2: Finding user_id corresponding to the photo_id
select user_id from photos where id='145';

#Step 3: Finding the username
select username from users where id=52;
```

Since Zack_Kemmer93 has the highest number of likes on his post he is the declared winner of the competition.

4. **Hashtag Researching: A partner brand wants to know, which hashtags to use in the post to reach the most people on the platform.**

```
#Hashtag Researching:
#Identifying and suggest the top 5 most commonly used hashtags on the platform

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



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

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

The aforementioned hashtags are the most used hashtags with the posts.

5. **Launch AD Campaign: The team wants to know, which day would be the best day to launch ADs.**

```
#Launch AD Campaign:
#Finding what day of the week do most users register on.

select dayname(created_at) as Day, count(*) as Total
from users
group by Day
order by Total desc;
```

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

Day	Total
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Since Thursday and Sunday are the days when most accounts are registered, it is suggested to launch the ad campaign on the aforementioned days.

Investor Metrics:

1. **User Engagement: Are users still as active and post on Instagram or they are making fewer posts**

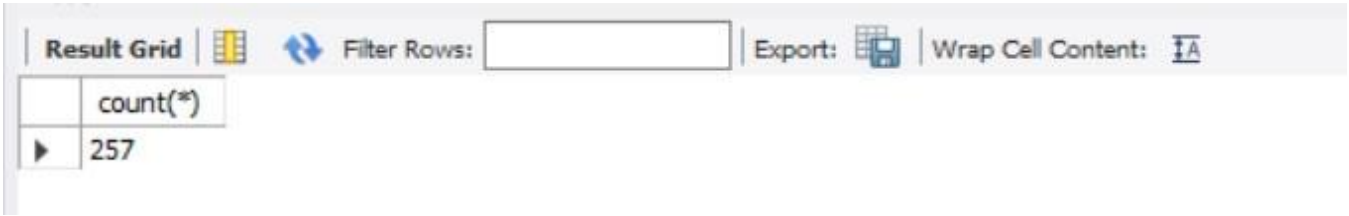
```
#User Engagement:
#Finding how many times does average user posts on Instagram.

select ((select count(*) from photos) / (select count(*) from users)) as `Average Photos Posted Per User`;
```

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

	Average Photos Posted Per User
▶	2.5700

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

```
#Bots & Fake Accounts:
# Finding no. of users who have liked all the photos
#Step 1: Finding all the photos posted
select count(*) from photos;

#Step 2: Finding users who liked the same number of photos as Step 1

select count(photo_id) as Total, user_id, users.username from likes
join
users on users.id=likes.user_id
group by user_id order by Total desc
limit 15
;
```



The total number of photos posted is 257. Now to figure out the bot/fake accounts we will find out the accounts which liked every posted photo.

Project 03:

Operation Analytics and Investigating Metric Spike

Description:

Operation analytics is the analysis performed for a company's whole end-to-end operations. This helps the business identify the areas where it needs to make improvements. You collaborate closely with the operations team, the support team, the marketing team, etc. and assist them in drawing conclusions from the data they gather.

Being one of the most crucial components of a business, this form of analysis is also utilised to forecast the general upward or downward trend in a company's fortune. Better automation, improved communication among cross-functional teams, and more efficient workflows are the results.

Investigating metric spikes is a crucial component of operational analytics since a data analyst needs to be able to answer queries like, "Why is there a decline in daily engagement?" or at least help other teams answer these questions. Why have sales decreased? Etc. Daily answers to questions like these are required, so it is crucial to look into metric increases.

We are required to perform operational analytics for Case Study II and investigate the Metric Spike for Case Study II.

Approach

- Creating the relational database and inserting values into it for Case Study-I.
- Uploading given CSV files for creating a relational database for Case Study-II
- Going through the values of each table and getting the gist of it.
- Running appropriate SQL queries based on the questions asked on the database and figuring out insights from it.

Insights:

CASE STUDY-I: Operational Analytics

The table schema is as follows:

- Table: job_data
- job_id: the unique identifier of jobs
- actor_id: the unique identifier of actors
- event: decision/skip/transfer
- language: the language of the content
- time_spent: time spent reviewing the job in seconds
- org: organization of the actor
- ds: date in the yyyy/mm/dd format.

The queries and observations for each question asked are given below.

1. **Creation of the table and insertion of values in it.**

```
use job_data;
create table job_data(
job_id int,
actor_id int,
event varchar(255),
language varchar(255),
time_spent int,
org varchar(255),
ds date);
```

```
INSERT INTO job_data (ds, job_id, actor_id, `event`, `language`, time_spent, org)
VALUES
('2020-11-30',21,1001,'skip','English',15,'A'),
('2020-11-30',22,1006,'transfer','Arabic',25,'B'),
('2020-11-29',23,1003,'decision','Persian',20,'C'),
('2020-11-28',23,1005,'transfer','Persian',22,'D'),
('2020-11-28',25,1002,'decision','Hindi',11,'B'),
('2020-11-27',11,1007,'decision','French',104,'D'),
('2020-11-26',23,1004,'skip','Persian',56,'A'),
('2020-11-25',20,1003,'transfer','Italian',45,'C'),
('2020-11-26',12,1006,'decision','Arabic',55,'B'),
('2020-11-27',11,1009,'skip','French',52,'F'),
('2020-11-28',30,1009,'decision','English',69,'F'),
('2020-11-29',29,1003,'skip','English',58,'C'),
('2020-11-30',13,1005,'skip','French',78,'D'),
('2020-12-01',14,1009,'decision','French',55,'F'),
('2020-12-01',15,1010,'skip','Spanish',65,'E'),
('2020-12-02',13,1010,'transfer','Spanish',25,'E'),
('2020-12-03',17,1010,'decision','English',69,'E'),
('2020-12-04',18,1365,'decision','Hindi',52,'F'),
('2020-12-05',19,1222,'skip','Hindi',83,'I'),
('2020-12-06',52,1222,'decision','Hindi',55,'I'),
('2020-12-07',89,1223,'transfer','Spanish',105,'J'),
('2020-12-08',89,1656,'decision','Spanish',102,'L'),
('2020-12-09',89,1525,'skip','Spanish',23,'L'),
('2020-12-04',18,1201,'skip','Hindi',88,'I'),
('2020-12-05',18,1254,'transfer','Hindi',66,'I'),
('2020-12-06',5,1111,'decision','Arabic',56,'K'),
('2020-12-07',5,1121,'skip','Arabic',58,'L'),
('2020-12-08',5,1110,'transfer','Arabic',25,'I'),
('2020-12-09',10,2566,'skip','Punjabi',125,'A'),
('2020-12-10',10,452,'decision','Punjabi',152,'B'),
('2020-12-10',10,5252,'transfer','Punjabi',36,'C');
```

2. **The number of jobs reviewed: Calculating the number of jobs reviewed per hour per day for November 2020.**

```
110    # No. of jobs reviewed per hour per day for November 2020
111
112 •  select (b.jobs / b.`time`)/3600 as job_reviewed
113    from
114    (select count(job_id) as jobs, sum(time_spent) as `time` from job_data
115    where ds>='2020-11-01' and ds<='2020-11-30') as b;
116
117
118
119
```

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

job_reviewed
0.00000592

3. **Throughput: It is the no. of events happening per second. Calculating the 7-day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?**

```
# 7 day rolling average of throughput
select
    ds,
    a.jobs,
    avg(a.jobs) over (order by ds rows between 6 preceding and current row) as `7DayRollingAvg`
from
    (select ds, count(job_id) as jobs from job_data group by ds) as a
group by ds;
```

Result Grid | Filter Rows:

ds	jobs	7DayRollingAvg
2020-11-25	1	1.0000
2020-11-26	2	1.5000
2020-11-27	2	1.6667
2020-11-28	3	2.0000
2020-11-29	2	2.0000
2020-11-30	3	2.1667
2020-12-01	2	2.1429
2020-12-02	1	2.1429
2020-12-03	1	2.0000
2020-12-04	2	2.0000
2020-12-05	2	1.8571
2020-12-06	2	1.8571
2020-12-07	2	1.7143
2020-12-08	2	1.7143
2020-12-09	2	1.8571
2020-12-10	2	2.0000

The 7-day rolling avg would be preferred otherwise it would be difficult to detect the trend over time due to the frequent fluctuation of values.

4. **Percentage share of each language: Share of each language for different contents in the last 30 days.**

```
93
94   # Percentage share of each language in the last 30 days
95 • select
96       `language`,
97       z.count,
98       (count / sum(z.count) over (order by `language` rows between unbounded preceding and unbounded following)) *100 as percentage_share
99   from
100      (select `language`, count(`language`) as count from job_data where ds>='2020-11-01' and ds<='2020-11-30' group by `language`) as z
101  group by
102      `language`
103      ;
```

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

language	count	percentage_share
Arabic	2	15.3846
English	3	23.0769
French	3	23.0769
Hindi	1	7.6923
Italian	1	7.6923
Persian	3	23.0769

5. **Duplicate rows: Displaying duplicates in the table.**

Limit to 100 rows

```
1    # Duplicate rows
2 •  select
3        ds, job_id, actor_id, `event`, `language`, time_spent, org
4    from
5        job_data
6    group by
7        ds, job_id, actor_id, `event`, `language`, time_spent, org
8    having
9        count(ds) >1
10       and count(job_id) >1
11       and count(actor_id) >1
12       and count(`event`) >1
13       and count(`language`) >1
14       and count(time_spent) >1
15       and count(org) >1;
```

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

ds	job_id	actor_id	event	language	time_spent	org

CASE STUDY-II: Investigating Metric Spike

The table schema is as follows:

- Table-1: users
    - This table includes one row per user, with descriptive information about that user's account.
- Table-2: events
    - This table includes one row per event, where an event is an action that a user has taken. These events include login events, messaging events, search events, events logged as users progress through a signup funnel, and events around received emails.
- Table-3: email_events
    - This table contains events specific to the sending of emails. It is similar in structure to the events table above.

1. **User Engagement: Calculating the weekly user engagement. (To measure the activeness of a user. Measuring if the user finds quality in a product/service.)**

```
1    #User Engagement: To measure the activeness of a user.
2    #Weekly User Engagement
3 •  select extract(week from occurred_at) as week_num, count(distinct user_id) as no_of_users
4    from events
5    group by week_num order by week_num;
```

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

week_num	no_of_users
17	356
18	669
19	705
20	736
21	705
22	746
23	803
24	843
25	825
26	860
27	912
28	890
29	900
30	993
31	846
32	832
33	862
34	842
35	73

2. **User Growth: Calculating the user growth for the product. (Amount of users growing over time for a product.)**

```
#User Growth: Amount of users growing over time for a product.
select
    year_no,
    week_no,
    active_users,
    sum(active_users) over (order by year_no, week_no rows between unbounded preceding and current row) as cum_active_users_per_week
from
    (select
        extract(year from activated_at) as year_no,
        extract(week from activated_at) as week_no,
        count(distinct user_id) as active_users
    from users
    where state='active'
    group by year_no, week_no
    order by year_no, week_no) as a;
```

year_no	week_no	active_users	cum_active_users_per_week
2013	27	52	1183
2013	28	72	1255
2013	29	67	1322
2013	30	67	1389
2013	31	67	1456
2013	32	71	1527
2013	33	73	1600
2013	34	78	1678
2013	35	63	1741
2013	36	72	1813
2013	37	85	1898
2013	38	90	1988
2013	39	84	2072
2013	40	87	2159
2013	41	73	2232
2013	42	99	2331
2013	43	89	2420
2013	44	96	2516
2013	45	91	2607
2013	46	88	2695
2013	47	102	2797
2013	48	97	2894
2013	49	116	3010
2013	50	124	3134
2013	51	102	3236
2013	52	47	3283
2014	0	83	3366

3. Weekly Retention: Calculate the weekly retention of users-sign up cohort. (Users getting retained weekly after signing up for a product.)

```
# Weekly retention of users-sign up cohort?
SELECT distinct user_id, COUNT(user_id), SUM(CASE WHEN retention_week = 1 Then 1 Else 0 END) as retention_per_week
FROM
    (
    SELECT a.user_id, a.signup_week, b.engagement_week, b.engagement_week - a.signup_week as retention_week
    FROM(
        (SELECT distinct user_id, extract(week from occurred_at) as signup_week FROM events WHERE event_type = 'signup_flow' AND event_name = 'complete_signup')a
    LEFT JOIN
        (SELECT distinct user_id, extract(week from occurred_at) as engagement_week FROM events WHERE event_type = 'engagement'
        )b
    ON a.user_id = b.user_id
        )
)d
GROUP BY user_id
ORDER BY user_id
LIMIT 25;
```

user_id	COUNT(user_id)	retention_per_week
11768	1	0
11770	1	0
11775	2	1
11778	3	0
11779	1	0
11780	2	1
11785	1	0
11787	3	1
11791	2	1
11793	1	0
11795	2	1
11798	6	1
11799	10	1
11801	1	0
11804	2	1
11806	1	0
11809	1	0
11811	2	1
11813	6	0
11816	3	0
11818	2	1
11820	4	1
11823	3	1
11824	1	0
11825	1	0

4. **Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.**

```
#Weekly Engagement Per Device

select
    extract(year from occurred_at) as `year`,
    extract(week from occurred_at) as `week`,
    device,
    count(distinct user_id) as no_of_users
from
    events
where event_type='engagement'
group by `year`, `week`, device
order by `year`, `week`, device
limit 25;
```

year	week	device	no_of_users
2014	17	acer aspire desktop	5
2014	17	acer aspire notebook	11
2014	17	amazon fire phone	1
2014	17	asus chromebook	9
2014	17	dell inspiron desktop	12
2014	17	dell inspiron notebook	19
2014	17	hp pavilion desktop	8
2014	17	htc one	7
2014	17	ipad air	9
2014	17	ipad mini	10
2014	17	iphone 4s	13
2014	17	iphone 5	33
2014	17	iphone 5s	18
2014	17	kindle fire	1
2014	17	lenovo thinkpad	45
2014	17	mac mini	3
2014	17	macbook air	25
2014	17	macbook pro	72
2014	17	nexus 10	4
2014	17	nexus 5	19
2014	17	nexus 7	9
2014	17	nokia lumia 635	10
2014	17	samsumg galaxy tablet	5
2014	17	samsung galaxy note	4
2014	17	samsung galaxy s4	28

5. **Email Engagement: Calculating the email engagement metrics. (Users engaging with the email service.)**

```
57  #Email Engagement Metrics
58  select
59      (sum(case when e='opened_email' then 1 else 0 end) / sum(case when e='sent_email' then 1 else 0 end))*100 as opening_email_rate,
60      (sum(case when e='clicked_email' then 1 else 0 end) / sum(case when e='sent_email' then 1 else 0 end))*100 as clicking_email_rate
61  from
62      (select *,
63      case
64      when action in ('sent_weekly_digest', 'sent_reengagement_email') then 'sent_email'
65      when action in('email_open') then 'opened_email'
66      when action in ('email_clickthrough') then'clicked_email'
67      end as e
68      from email_events) a;
```

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

opening_email_rate	clicking_email_rate
33.5834	14.7899

Opening EMail Rate = No. of emails opened/ No.of emails sent

Clicking Email Rate= No. of emails clicked / No. of emails sent.

Project 04:

Hiring Process Analytics

Description:

The hiring process is the foundational and crucial part of a business. The MNCs learn about the key underlying trends relating to the hiring process here. Before hiring freshmen or anybody else, a corporation should consider trends such as the number of rejections, interviews, sorts of jobs, openings, etc. Hence, there is a chance for Data Analyst employment here as well!

We have been given a dataset of a company where the details about people who registered for a particular post in a department of that company are provided. We are required to use our knowledge of statistics and use different formulas in Excel and draw necessary conclusions about the company.

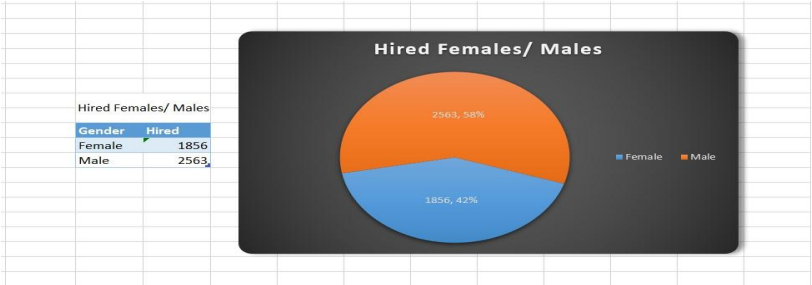
Approach:

- Creating the given data into a table.
- Looking for duplicates and removing them.
- Finding the outliers and deciding what to do with them.
- Running appropriate Excel commands based on the questions asked and figuring out insights from them.

Insights:

1. **Hiring: Process of intaking people into an organization for different kinds of positions.**

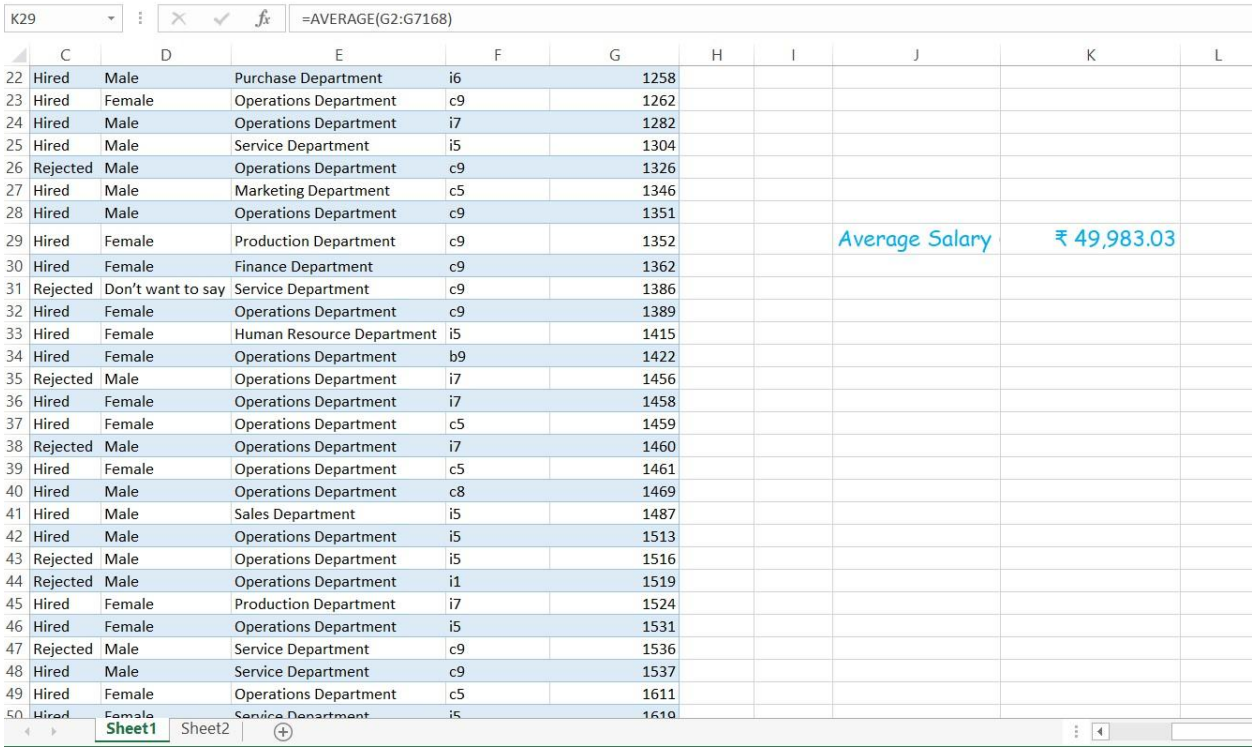
   Finding how many males and females have been hired.



- Formula used for counting Females hired: COUNTIFS(D:D, "Female", C:C, "Hired")
- Formula used for counting Males hired: =COUNTIFS(D:D, "Male", C:C, "Hired")
- The percentage of female hiring is 42% whereas for males it is 58%.

2. **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.**

   Finding the average salary offered in this company.



	C	D	E	F	G
22	Hired	Male	Purchase Department	i6	1258
23	Hired	Female	Operations Department	c9	1262
24	Hired	Male	Operations Department	i7	1282
25	Hired	Male	Service Department	i5	1304
26	Rejected	Male	Operations Department	c9	1326
27	Hired	Male	Marketing Department	c5	1346
28	Hired	Male	Operations Department	c9	1351
29	Hired	Female	Production Department	c9	1352
30	Hired	Female	Finance Department	c9	1362
31	Rejected	Don't want to say	Service Department	c9	1386
32	Hired	Female	Operations Department	c9	1389
33	Hired	Female	Human Resource Department	i5	1415
34	Hired	Female	Operations Department	b9	1422
35	Rejected	Male	Operations Department	i7	1456
36	Hired	Female	Operations Department	i7	1458
37	Hired	Female	Operations Department	c5	1459
38	Rejected	Male	Operations Department	i7	1460
39	Hired	Female	Operations Department	c5	1461
40	Hired	Male	Operations Department	c8	1469
41	Hired	Male	Sales Department	i5	1487
42	Hired	Male	Operations Department	i5	1513
43	Rejected	Male	Operations Department	i5	1516
44	Rejected	Male	Operations Department	i1	1519
45	Hired	Female	Production Department	i7	1524
46	Hired	Female	Operations Department	i5	1531
47	Rejected	Male	Service Department	c9	1536
48	Hired	Male	Service Department	c9	1537
49	Hired	Female	Operations Department	c5	1611
50	Hired	Female	Service Department	i5	1619

- Formula Used: =AVERAGE(G2:G7168)
- This is an average salary that includes both Hired and Rejected Employees.

3. **Class Intervals: The class interval is the difference between the upper-class limit and the lower-class limit.**

   Finding the class intervals for the salary offered in the company.

Department	Post Name	Offered Salary	Offered Salary
Service Department	i5	100	400000
Service Department	m6	800	300000
Marketing Department	c9	1007	200000
General Management	i7	1022	99967
Operations Department	c9	1027	99953

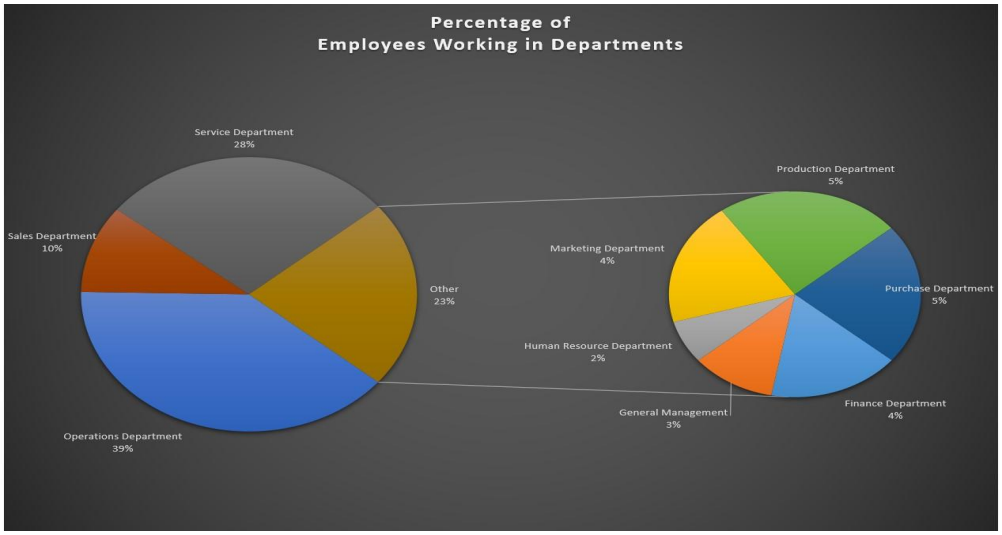
Minimum Salary	1007
Maximum Salary	99967
Class Interval	98960

Minimum Salary	100
Maximum Salary	400000
Class Interval	399900

- The class interval is calculated by finding out the minimum and maximum salary offered by using min() and max() and finding the difference between them.
- The class interval with outliers comes out to be 399900 units.
- Then, we found the outliers by sorting the data. 100, 800,200000,300000, and 400000 are taken as outliers and are removed.
- The class interval after removing the outliers comes to 98960 units.

4. **Drawing a Pie Chart / Bar Graph (or any other graph) to show the proportion of people working in different departments.**

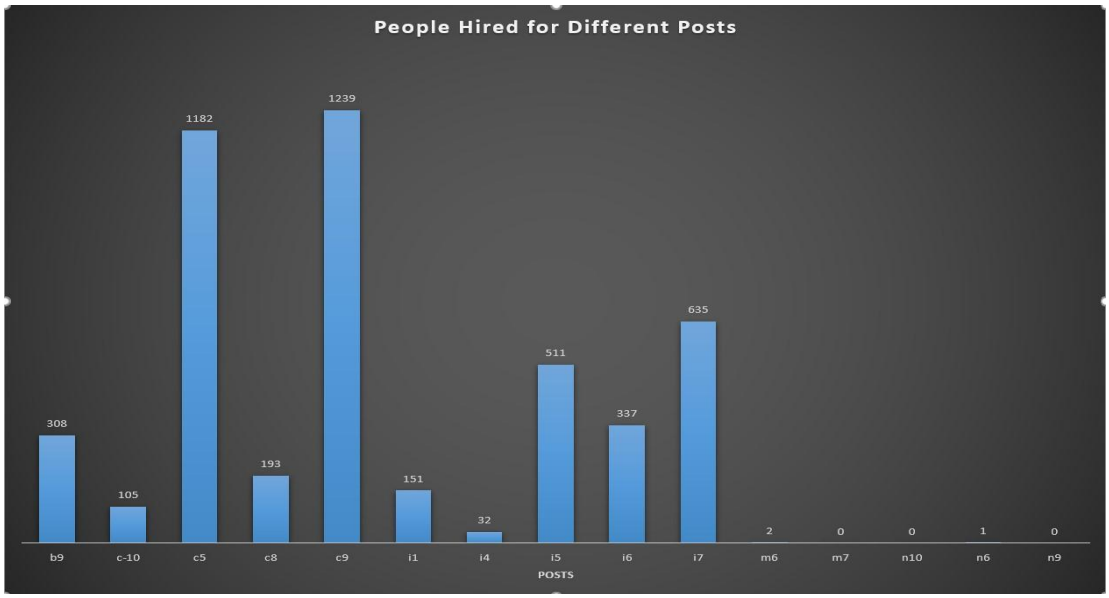
Department	No.of Employees
Finance Department	176
General Management	113
Human Resource Department	70
Marketing Department	202
Operations Department	1843
Production Department	246
Purchase Department	230
Sales Department	485
Service Department	1332



- The formula used for finding and sorting unique departments for the given data: =SORT((UNIQUE(FILTER(B:B,B:B<>""))))
- The formula used for finding the number of hired employees working in respective departments: =COUNTIFS(A:A,"Hired", B:B," Financial Department")

5. **Representing different post tiers using chart/graph.**

Post Name	People Hired
b9	308
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
m7	0
n10	0
n6	1
n9	0



- The formula used to find unique posts and filter the missing data: =SORT(UNIQUE(FILTER(F2:F7164,F2:F7164<>"-")))
- The formula used to find the number of employees hired for respective posts: =COUNTIFS(F:F,I28,C:C, "Hired")

Project 05:

IMDB Movie Analysis

Description:

We have a dataset with various columns of different IMDB Movies. We are required to frame the problem we want to shed some light on. We are required to provide a detailed report for the below data record mentioning the answers to the questions that follow.

We can do this by asking the following 'Whats?':

- What do we see happening?
- What is our hypothesis for the cause of the problem? (this will be
- broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

Approach:

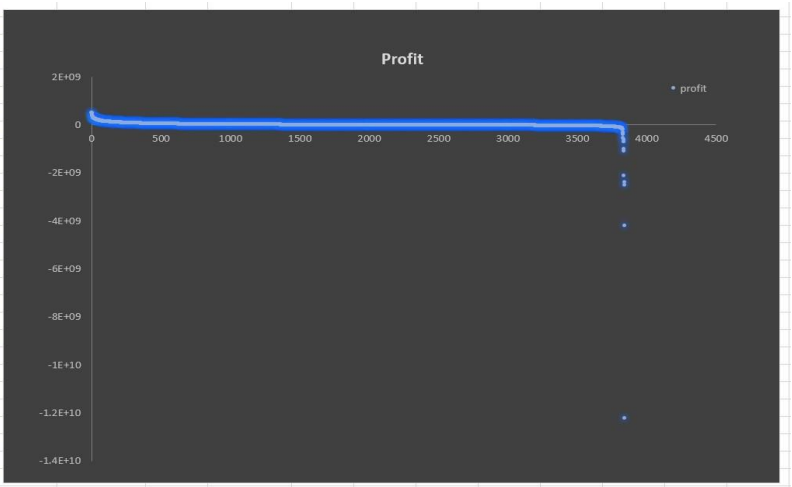
- Downloading and uploading the given CSV file in the MS Excel workbook using the Data tab.
- Clean the data:
    - Remove the irrelevant columns entirely.
    - Remove the rows with any blank or null value.
    - Remove the duplicate rows from the data.
- Running appropriate Excel commands based on the questions asked and figuring out insights from them.

Insights:

1. **Movies with the highest profit:**

- Create a new column called profit which contains the difference between the two columns: gross and budget.
- Sort the column using the profit column as a reference.
- Plot profit (y-axis) vs budget (x-axis).
- Observe the outliers using the appropriate chart type.

movie_title	profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255
DeadpoolÂ	305024263
The Hunger Games: Catching FireÂ	294645577
Jurassic ParkÂ	293784000
Despicable Me 2Â	292049635
American SniperÂ	291323553
Finding NemoÂ	286838870
Shrek 2Â	286471036
The Lord of the Rings: The Return of the KingÂ	283019252
Star Wars: Episode VI - Return of the JediÂ	276625409
Forrest GumpÂ	274691196
Star Wars: Episode V - The Empire Strikes BackÂ	272158751
Home AloneÂ	267761243
Star Wars: Episode III - Revenge of the SithÂ	267262555
Spider-ManÂ	264706375
MinionsÂ	262029560
The Sixth SenseÂ	253501675
JawsÂ	252000000
FrozenÂ	250736600



profit
-12213298588
-4199788333
-2499804112
-2397701809
-2127109510

Outliers

- ☐ The profit column is created by finding the difference between the gross and budget column and sort in descending order.
- ☐ The top 5 movies with the highest profits are Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope, and E.T. the Extra-Terrestrial.
- ☐ By plotting an x-y scatter plot, the 5 outliers are observed as shown above (highlighted in yellow).

2. **Top 250 Movies:**

- Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score).
- For all of these movies, make sure that the num_voted_users is greater than 25,000.
- 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.

IMDB_Top_250(All Languages):

- ☐ Firstly using a filter(), movie_titles are selected where num_voted_users is greater than 25000. The formula used is FILTER(A:A, B:B>25000).
- ☐ Then using index() and match(), the imdb_score and language of corresponding movies are extracted. The formula used is: =INDEX(D:D,MATCH(G3,A:A,0)), INDEX(C:C,MATCH(G3,A:A,0)) respectively.
- ☐ After that, the movies are ranked using the rank.eq() in descending order of the imdb_score. Along with rank.eq(), countif() is also used to get unique values of ranks. The formula used is: =RANK.EQ(H3,H:H)+COUNTIF(H3:H3,H3)-1
- ☐ Finally using filter() and sortby(), IMDB_Top_250 movies are selected. The formula used is: =FILTER(SORTBY(G3:I2609,I3:I2609,1),SORTBY(I3:I2609,I3:I2609,1)<251)

IMDB_Top_250	IMDB_Score	Language	Rank
The Shawshank RedemptionÂ	9.3	English	1
The GodfatherÂ	9.2	English	2
The Dark KnightÂ	9	English	3
The Godfather: Part IIÂ	9	English	4
The Lord of the Rings: The Return of the KingÂ	8.9	English	5
Schindler's ListÂ	8.9	English	6
Pulp FictionÂ	8.9	English	7
The Good, the Bad and the UglyÂ	8.9	Italian	8
InceptionÂ	8.8	English	9
The Lord of the Rings: The Fellowship of the RingÂ	8.8	English	10
Fight ClubÂ	8.8	English	11
Forrest GumpÂ	8.8	English	12
Star Wars: Episode V - The Empire Strikes BackÂ	8.8	English	13
The Lord of the Rings: The Two TowersÂ	8.7	English	14
The MatrixÂ	8.7	English	15
GoodfellasÂ	8.7	English	16
Star Wars: Episode IV - A New HopeÂ	8.7	English	17
One Flew Over the Cuckoo's NestÂ	8.7	English	18
City of GodÂ	8.7	Portuguese	19
Seven SamuraiÂ	8.7	Japanese	20
InterstellarÂ	8.6	English	21
Saving Private RyanÂ	8.6	English	22
Se7enÂ	8.6	English	23
The Silence of the LambsÂ	8.6	English	24
Spirited AwayÂ	8.6	Japanese	25
American History XÂ	8.6	English	26
The Usual SuspectsÂ	8.6	English	27
Modern TimesÂ	8.6	English	28
The Dark Knight RisesÂ	8.5	English	29
GladiatorÂ	8.5	English	30
Terminator 2: Judgment DayÂ	8.5	English	31
Django UnchainedÂ	8.5	English	32
The DepartedÂ	8.5	English	33
The Lion KingÂ	8.5	English	34
The Green MileÂ	8.5	English	35
The PrestigeÂ	8.5	English	36
The PianistÂ	8.5	English	37
Apocalypse NowÂ	8.5	English	38
Raiders of the Lost ArkÂ	8.5	English	39
PsychoÂ	8.5	English	40
Back to the FutureÂ	8.5	English	41
AlienÂ	8.5	English	42

IMDB_Top_250(Top_Foreign_Lang_Film)

- ☐ In the above-generated array, a filter is applied to the language column from which “English” is deselected to give the Top_Foreign_Lang_Film in the IMDB_Top_250 movies.

  Top 250 Movies:

IMDB_Top_250	IMDB_Score	Language	Rank
The Good, the Bad and the UglyÂ	8.9	Italian	8
City of GodÂ	8.7	Portuguese	19
Seven SamuraiÂ	8.7	Japanese	20
Spirited AwayÂ	8.6	Japanese	25
The Lives of OthersÂ	8.5	German	45
Children of HeavenÂ	8.5	Persian	46
AmÃ©lieÂ	8.4	French	49
Baahubali: The BeginningÂ	8.4	Telugu	50
Princess MononokeÂ	8.4	Japanese	53
Das BootÂ	8.4	German	57
OldboyÂ	8.4	Korean	59
A SeparationÂ	8.4	Persian	61
MetropolisÂ	8.3	German	74
DownfallÂ	8.3	German	76
The HuntÂ	8.3	Danish	82
Howl's Moving CastleÂ	8.2	Japanese	97
Pan's LabyrinthÂ	8.2	Spanish	99
IncendiesÂ	8.2	French	101
The Secret in Their EyesÂ	8.2	Spanish	103
The Sea InsideÂ	8.1	Spanish	137
Tae Guk Gi: The Brotherhood of WarÂ	8.1	Korean	138
AkiraÂ	8.1	Japanese	140
Elite SquadÂ	8.1	Portuguese	141
Amores PerrosÂ	8.1	Spanish	150
The CelebrationÂ	8.1	Danish	151
My Name Is KhanÂ	8	Hindi	193
PersepolisÂ	8	French	195
Central StationÂ	8	Portuguese	201
Waltz with BashirÂ	8	Hebrew	204
A Fistful of DollarsÂ	8	Italian	205
HeroÂ	7.9	Mandarin	221
Crouching Tiger, Hidden DragonÂ	7.9	Mandarin	237
Letters from Iwo JimaÂ	7.9	Japanese	240
AmourÂ	7.9	French	244
Veer-ZaaraÂ	7.9	Hindi	247
The ChorusÂ	7.9	French	249

3. **Best Directors:**

   Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director.

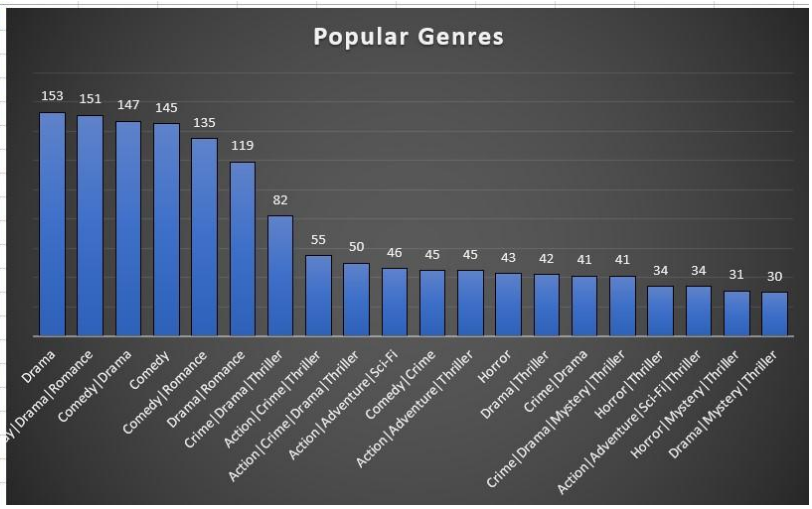
   In case of a tie in IMDb score between two directors, sort them alphabetically.

Top 10 Directors	Mean 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

- ☐ Firstly, a pivot table is created and director_name is added in the rows labels along with the mean(average) of imdb_score in the values section.
- ☐ Then, the director_name is sorted in ascending order followed by sorting of the mean values in descending order.
- ☐ The top 10 values were selected and presented as a bar chart for better understanding.

4. Popular Genres:

Genre	Count
Drama	153
Comedy|Drama|Romance	151
Comedy|Drama	147
Comedy	145
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46
Comedy|Crime	45
Action|Adventure|Thriller	45
Horror	43
Drama|Thriller	42
Crime|Drama	41
Crime|Drama|Mystery|Thriller	41
Horror|Thriller	34
Action|Adventure|Sci-Fi|Thriller	34
Horror|Mystery|Thriller	31
Drama|Mystery|Thriller	30



- ☐ A pivot table is created and genre is added to the row labels and count of the genre as the values.
- ☐ Then it is sorted in descending order, top 20 genres were selected and plotted for better understanding.
- ☐ Drama, Comedy, Romance, Crime and Thriller are the top 5 popular genres.

5. Find the critic-favourite and audience-favourite actors:

- 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 the number of voted users over decades using a bar chart.

- Create a column called decade which represents the decade to which every movie belongs.
- Sort the column based on the column decade, group it by decade and find the sum of users who voted in each decade. Store this in a new data frame called df_by_decade.

User & Critic Favourite Actor:

☐ First using sort() and filter(), movie titles of Brad Pitt, Leonardo DiCaprio and Meryl Streep were selected from the clean data and combined into a single column called combined.

The formulas used were:

SORT(FILTER($C:$C,$B:$B="Brad Pitt"))

SORT(FILTER($C:$C,$B:$B="Leonardo DiCaprio"));

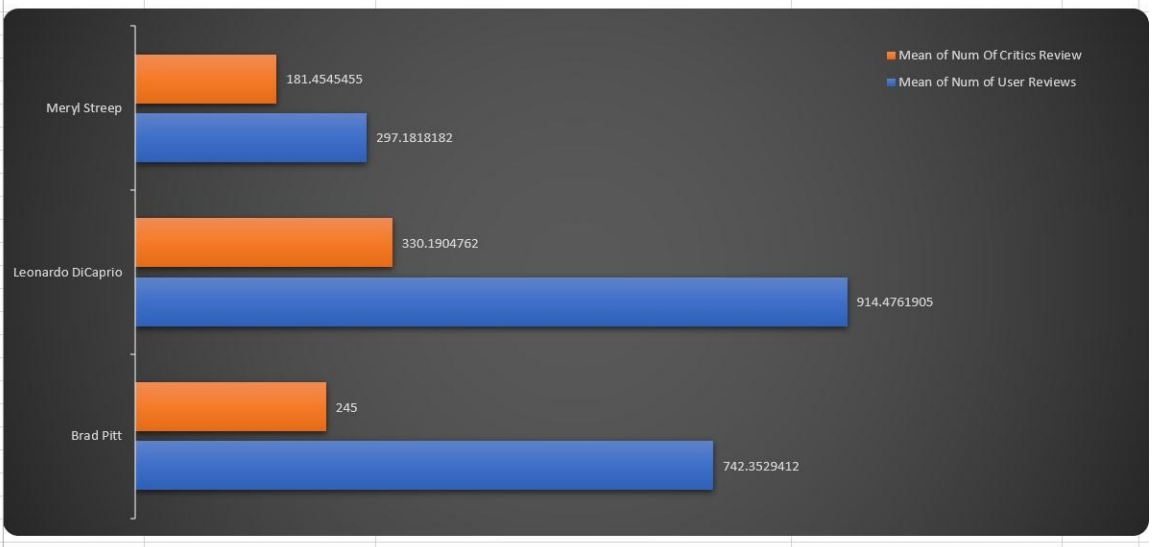
SORT(FILTER($C:$C, $B:$B="Meryl Streep"));

Meryl_Streep	Leo_Caprio	Brad_Pitt
A Prairie Home CompanionÂ	Blood DiamondÂ	BabelÂ
Hope SpringsÂ	Body of LiesÂ	By the SeaÂ
It's ComplicatedÂ	Catch Me If You CanÂ	Fight ClubÂ
Julie & JuliaÂ	Django UnchainedÂ	FuryÂ
Lions for LambsÂ	Gangs of New YorkÂ	Interview with the Vampire: The Vampire ChroniclesÂ
One True ThingÂ	InceptionÂ	Killing Them SoftlyÂ
Out of AfricaÂ	J. EdgarÂ	Mr. & Mrs. SmithÂ
The Devil Wears PradaÂ	Marvin's RoomÂ	Ocean's ElevenÂ
The HoursÂ	Revolutionary RoadÂ	Ocean's TwelveÂ
The Iron LadyÂ	Romeo + JulietÂ	Seven Years in TibetÂ
The River WildÂ	Shutter IslandÂ	Sinbad: Legend of the Seven SeasÂ
	The AviatorÂ	Spy GameÂ
	The BeachÂ	The Assassination of Jesse James by the Coward Robert FordÂ
	The DepartedÂ	The Curious Case of Benjamin ButtonÂ
	The Great GatsbyÂ	The Tree of LifeÂ
	The Great GatsbyÂ	TroyÂ
	The Man in the Iron MaskÂ	True RomanceÂ
	The Quick and the DeadÂ	
	The RevenantÂ	
	The Wolf of Wall StreetÂ	
	TitanicÂ	

Ocean's TwelveÂ		
Seven Years in TibetÂ		
Sinbad: Legend of the Seven SeasÂ		
Spy GameÂ		
The Assassination of Jesse James by the Coward Robert FordÂ		
The Curious Case of Benjamin ButtonÂ		
The Tree of LifeÂ		
TroyÂ		
True RomanceÂ		
Blood DiamondÂ		
Body of LiesÂ		
Catch Me If You CanÂ		
Django UnchainedÂ		Brad Pitt
Gangs of New YorkÂ		Leonardo DiCaprio
InceptionÂ		Meryl Streep
J. EdgarÂ		
Marvin's RoomÂ		
Revolutionary RoadÂ		
Romeo + JulietÂ		
Shutter IslandÂ		
The AviatorÂ		
The BeachÂ		
The DepartedÂ		
The Great GatsbyÂ		
The Man in the Iron MaskÂ		
The Quick and the DeadÂ		
The RevenantÂ		
The Wolf of Wall StreetÂ		
TitanicÂ		
Blood DiamondÂ		
Body of LiesÂ		
Catch Me If You CanÂ		
Django UnchainedÂ		
Gangs of New YorkÂ		

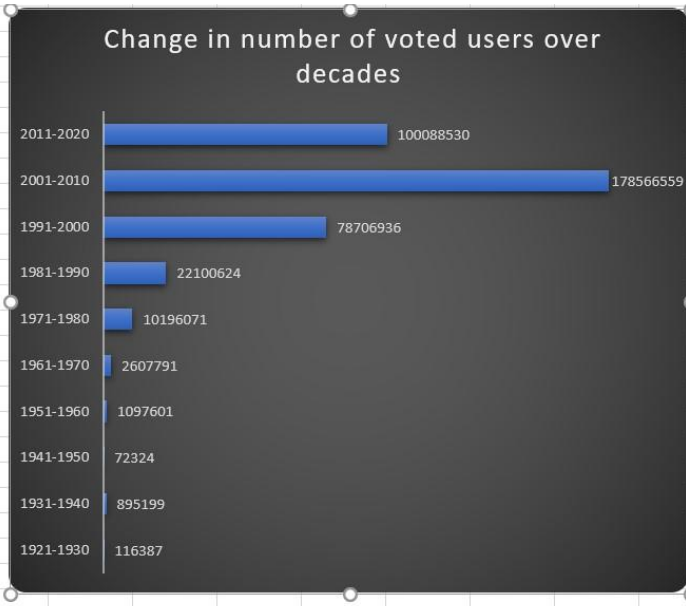
- Then a pivot table was created to find out the mean of num_critic_for_reviews and num_user_for_reviews and it was observed that Leonardo DiCaprio was both a critic and an audience favourite.

Lead Actor	Mean of Num of User Reviews	Mean of Num Of Critics Review
Brad Pitt	742.3529412	245
Leonardo DiCaprio	914.4761905	330.1904762
Meryl Streep	297.1818182	181.4545455
Grand Total	**716.1836735**	**267.244898**



- Another pivot table was created from the clean data to figure out the change in the number of voted users over decades using the group properties in the title_year using the starting value as 1921 and using 10 as the by-value.

Title_Year	df_by_decade	Sum of num_voted_users
1920s	1921-1930	116387
1930s	1931-1940	895199
1940s	1941-1950	72324
1950s	1951-1960	1097601
1960s	1961-1970	2607791
1970s	1971-1980	10196071
1980s	1981-1990	22100624
1990s	1991-2000	78706936
2000s	2001-2010	178566559
2010s	2011-2020	100088530



Project 06:

Bank Loan Case Study

Description:

This case study intends to provide an example of how EDA might be used in a real-world corporate setting. In this case study, we will learn the fundamentals of risk analytics in banking and financial services and how data is applied to reduce the risk of financial loss while lending money to clients. In this case study, we will use EDA to understand how consumer and loan attributes influence the default tendency.

Our goal is to detect the patterns and trends that show whether a client has trouble paying their instalments. We will also try to find out the variables that are strong predictors of loan default.

Business Understanding:

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

Approach:

- We have been given 2 datasets.
    - ‘application_data.csv’ contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
    - ‘previous_application.csv’ contains information about the client’s previous loan data. It contains the data on whether the previous application had been approved, cancelled, refused or unused offer.
- We imported these datasets using Python libraries such as numpy, pandas, matplotlib and seaborn.

- Then we used various functions to understand the data, worked with the null values and decided how to deal with them, explored outliers and used charts and graphs for visualizations.

Dealing with null values and missing data:

Identify the missing data and use an appropriate method to deal with it. (Remove columns/or replace them with an appropriate value).

1. In the application dataset, there are 307511 rows and 122 features while in the previous application dataset, we have 1670214 rows and 37 feature columns.
2. On analysis, we found that there are 41 features with null values of more than 50%. Most of the features are related to the housing information of the loan applicant. Hence we can drop these.
3. EXT_SOURCE_1 was dropped earlier because of the high null values percentage. EXT_SOURCE_2 and EXT_SOURCE_3 are not highly correlated with our TARGET. Hence, dropped these features as well.
4. We observed that only FLAG_DOCUMENT_3 has been submitted during the application, the rest of the other flags (FLAGS_X) were dropped.
5. There is almost no correlation between 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL' with the "TARGET" column and hence were dropped.
6. Now, we were left with 73 features with null values less than 50% for the application dataset.
7. For the previous application dataset:
    a. RATE_INTEREST_PRIVILEGED and RATE_INTEREST_PRIMARY can be dropped as the null value percentage is more than 99%.
    b. AMT_DOWN_PAYMENT and RATE_DOWN_PAYMENT are also dropped because the percentage of null value is more than 50%.
    c. The rows of PRODUCT_COMBINATION and AMT_CREDIT with null values can be removed as the percentage of null values is less than 2%.
    d. SK_ID_PREV, WEEKDAY_APPR_PROCESS_START,HOUR_APPR_PROCESS_START, FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY, NFLAG_INSURED_ON_APPROVAL can be dropped as well (not needed for analysis).

   e. DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION, SELLERPLACE_AREA can also be dropped.
   f. We are now left with 30 features for further analysis.
8. We imputed the null values of all the numerical features with the median (as the outliers exist) and with the mode for the categorical features.
9. For ORGANIZATION_TYPE, XNA is quite high in number and we cannot be exactly sure of the value so we let it be the same.
10. For OCCUPATION_TYPE, 0.0 can be replaced with "Unknown" as the number is high and can make the analysis biased if substituted with any other value.

Identify the Outliers in the dataset.

Application Dataset:

1. Values after 2000000 are existing for AMT_GOODS_PRICE but they can be valid prices so cannot be considered an outlier.
2. AMT_INCOME_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high income compared to the others.
3. For CNT_FAM_MEMBERS, more than 5 members in a family is possible so cannot be considered outliers.

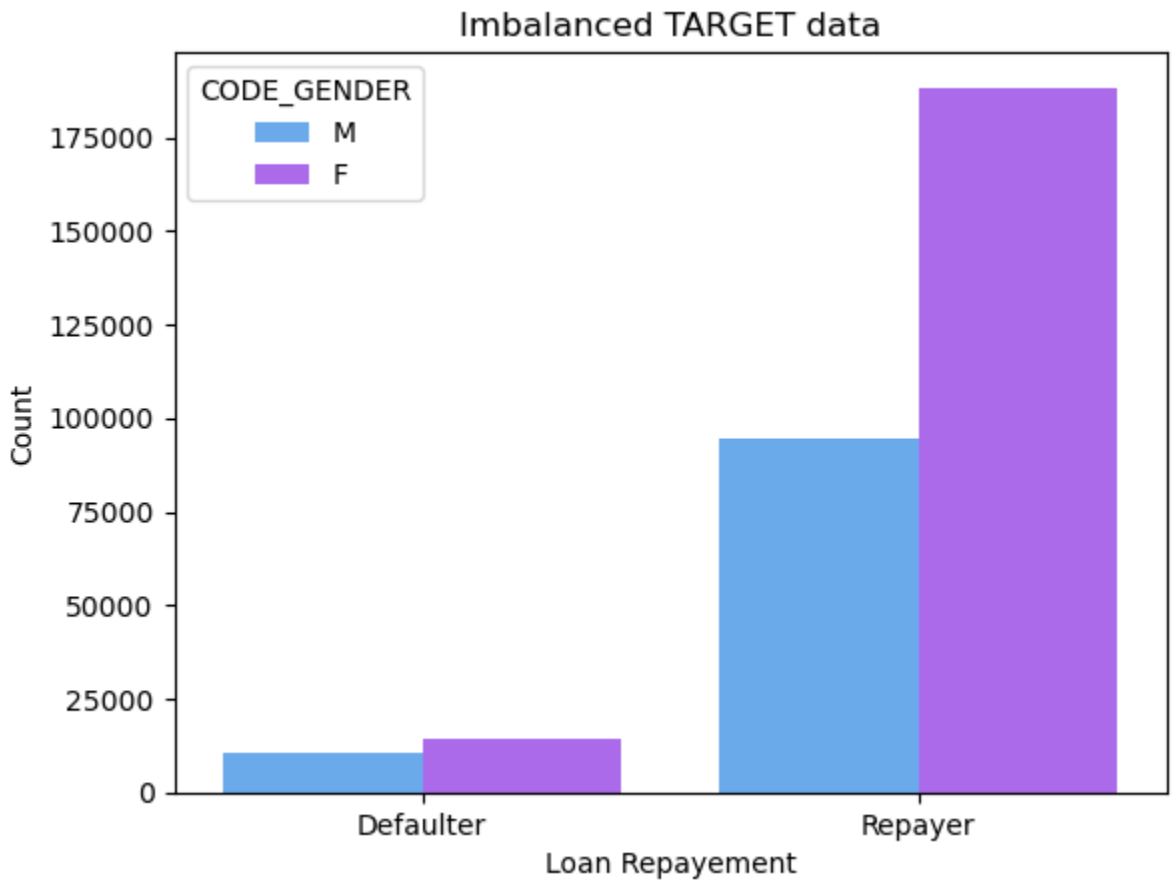
Previous Application Dataset:

1. For AMT_GOODS_PRICE, values above 3000000 can be dropped as outliers.
2. For CNT_PAYMENT, there is not much difference between the maximum value and 99 percentile values so we can leave the data as it is.

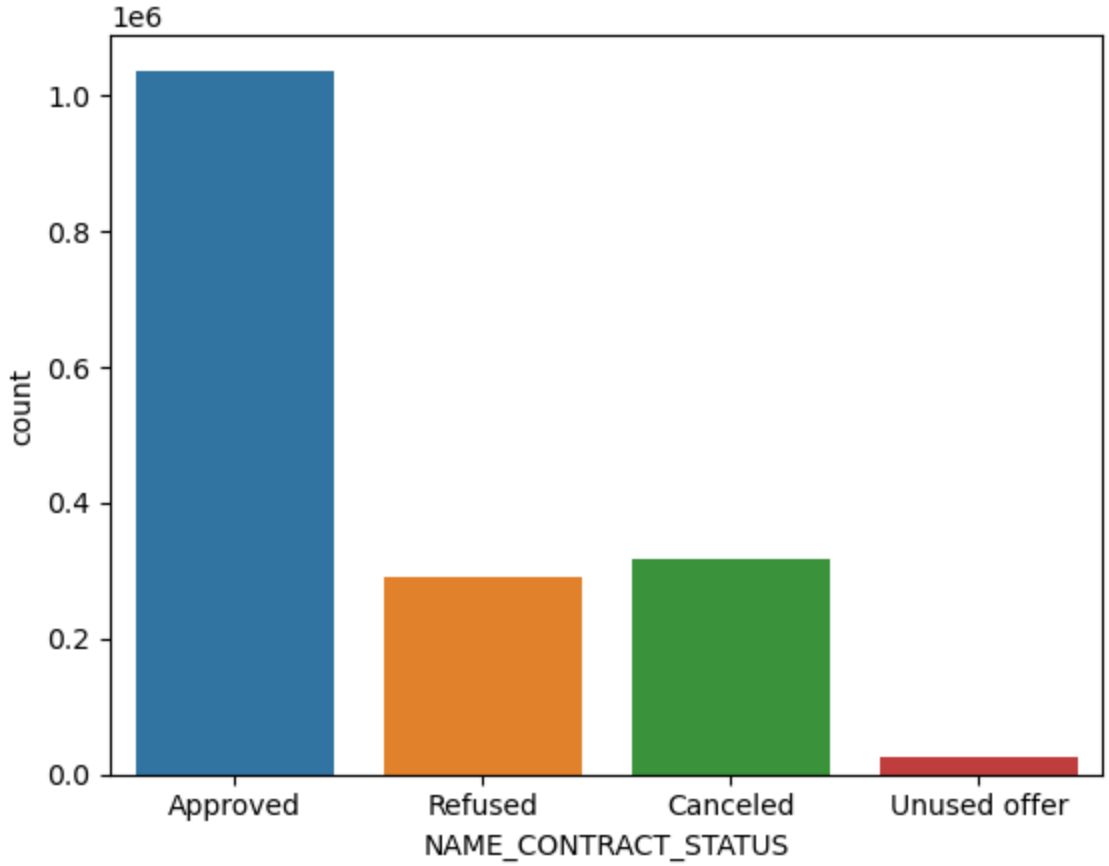
Data Imbalance:

Application Dataset:

- The data set is highly imbalanced.
- Data for loan repayer is far more than defaulters.
- Data on females is more than the males.
- There is 1 defaulter for every 11 loan repayers.



Previous Application Dataset:



- Approved loan data is higher.
- Refused and Canceled loan data is almost the same.
- Unused offer loan data os the lowest

Insights:

Univariate, segmented univariate, bivariate analysis:

Application Dataset:

Characteristics of Majority of Applicants:

1. Applied for Cash Loan
2. Female Applicants
3. Does not own realty or cars.
4. Were Unaccompanied during the application.
5. Income category is Working followed by Commercial Associates.
6. Secondary/ Secondary Special level educated.
7. Married, with 2 family members and no children.
8. Living in apartments/houses.
9. Occupation is mostly Unknown followed by Laborers.
10. On weekdays most applications process was started.
11. Organisation type is Business Entity Type 3 and Unknown.
12. More than 50% of applicants have an income range of 100K-200K.
13. AMT_CREDIT_RANGE is majorly less than 300K.
14. Majority of applicants are of age 40 and above.
15. AMT_ANNUITY offered is mostly less than 100000.
16. Highest AMT_CREDIT value is offered to people with Academic Degrees.
17. Highest AMT_CREDIT value is offered to MANAGERS, IT STAFF and HR STAFF

Characteristics of Majority of Defaulters:

1. Applied for Cash Loan
2. Male Applicants
3. Does not have cars or realty.
4. Were accompanied by Other_B.
5. Were unemployed and on maternity leave.
6. Education Level: Lower Secondary

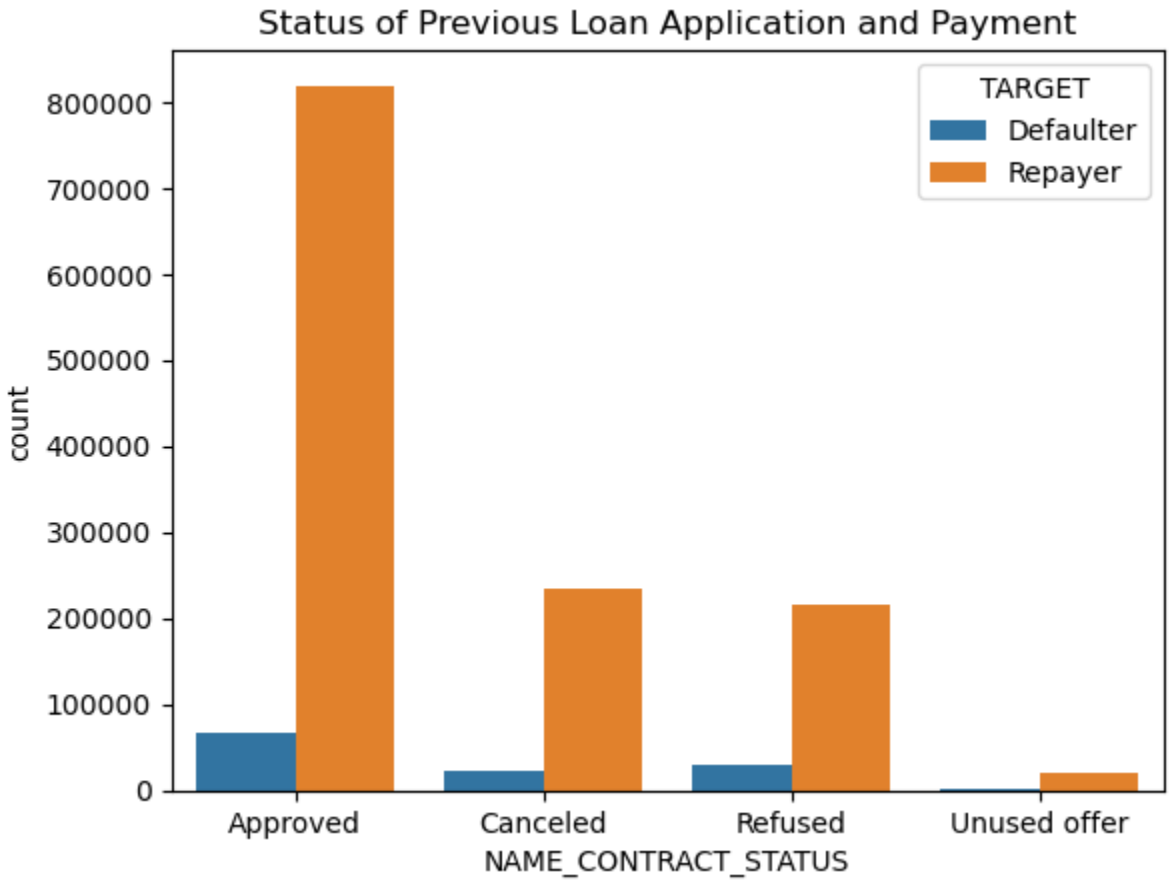
7. People who had civil marriages followed by singles.
8. Living in a rented apartment followed by parents.
9. Occupation Type: Low-skill Laborers followed by Drivers and Waiters/Barmen staff.
10. Application process started on Tuesdays.
11. Organisation type Transport: Type 3 followed by Industry: type 13
12. 6 or more family members.
13. 6 or more children.
14. The default rate is high among people with an income range of less than 300K. People with an income range of more than 500K have fewer chances of default.
15. The default rate is higher for AMT_CREDIT_RANGE 200K to 600K. The lowest default range can be seen for AMT_CREDIT_RANGE less than 100K and 900K-1M.
16. The default rate is higher in people of the age group 20-40.
17. Default rate is higher in the range of 200000-400000 of AMT_ANNUITY.
18. Businessmen have higher incomes but the data for them is very less. Also, they have the least probability of default.
19. Defaulter rate is higher were people with a high-income range(900K to 1M).
20. People between the age of 30-50 tends to default more.

Previous Application Dataset:

1. Consumer and cash loan data are almost similar.
2. We have very fewer data for revolving loans.
3. The loan purpose for the majority of applications was XAP, followed by XNA.
4. Largest number of consumer loans have been approved.
5. There seem to be no cancelled loans in the consumer loan category.
6. More cash loans have been refused than consumer loans.

Merged Dataset:

1. Status of previous loan applications and payments



2. Applicants whose prior loan applications were approved are more likely to make their current loan payment on time than applicants whose prior loan applications were refused.
3. 7.5% of loan applicants who had previously been approved defaulted on their current loans.
4. 88%% of those who had been previously rejected for loans were able to make their current loan payments.
5. HC, Limit and SCO are the common reasons for loan rejection.

Defaulters Characteristics:

a. Majority of defaulters previously refused to let the loan purpose be known. After that hobby and car, repairing is the purpose of the loans of the defaulter.
b. The highest numbers of defaulters are for the insurance and vehicle previous loans.
c. For Cards, the default rate is the highest.
d. 12% of Walk-in applicants are defaulters.
e. 12.82% of defaulters are for AP+ (Cash loan)
f. In the seller industry, auto technology has the highest default rate. Tourism has the lowest.
g. NAME_YIELD_GROUP, for the majority of defaulters, is unknown.
h. The highest default percentage is for Cash Street.

TOP 10 CORRELATION for DEFAULTERS

Variable 1	Variable 2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.982783
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
REG_CITY_NOT_WORK_CITY	REG_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
TOTALAREA_MODE	FLOORSMAX_AVG	0.766951
TOTALAREA_MODE	FLOORSMAX_MEDI	0.765448
FLOORSMAX_MODE	TOTALAREA_MODE	0.763361
AMT_ANNUITY	AMT_GOODS_PRICE	0.752295

TOP 10 CORRELATION for Repayers:

Variable 1	Variable 2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.987022
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.776421
AMT_CREDIT	AMT_ANNUITY	0.771297
TOTALAREA_MODE	FLOORSMAX_AVG	0.754906
TOTALAREA_MODE	FLOORSMAX_MEDI	0.753490

Note: For a detailed analysis and all the visualisation please refer to the following link:

Bank Loan Case Study

Project 07:

Analyzing the Impact of Car Features on Price and Profitability

Description:

Over the last few decades, the automotive industry has experienced significant change, with an increasing emphasis on technological innovation, environmental sustainability, and fuel efficiency. This case study intends to understand the elements that drive customer demand for cars considering manufacturer competition and a changing consumer landscape.

Our goal is to provide valuable insights to car manufacturers and help them optimize their pricing and product development decisions to maximise profitability while meeting consumer demand.

Approach:

- We have a dataset containing information on various car models and their specifications, and is titled "Car Features and MSRP". It was collected and made available on Kaggle by Cooper Union, a private college in New York City.
- We imported this dataset using Python libraries such as numpy, pandas, matplotlib and seaborn and cleaned the data by removing the rows with missing values.
- The missing values in “Market Category” was replaced with “Unknown” as the missing data was huge.
- Then using pivot tables and pivot charts in Excel, we found the insights for the questions posed and created an interactive dashboard.

Insights:

Q1. How does the popularity of a car model vary across different market categories?

Most Popular Market Category:

- Hatchback, Flex Fuel
- Flex Fuel, Diesel
- Crossover, Flex Fuel, Performance

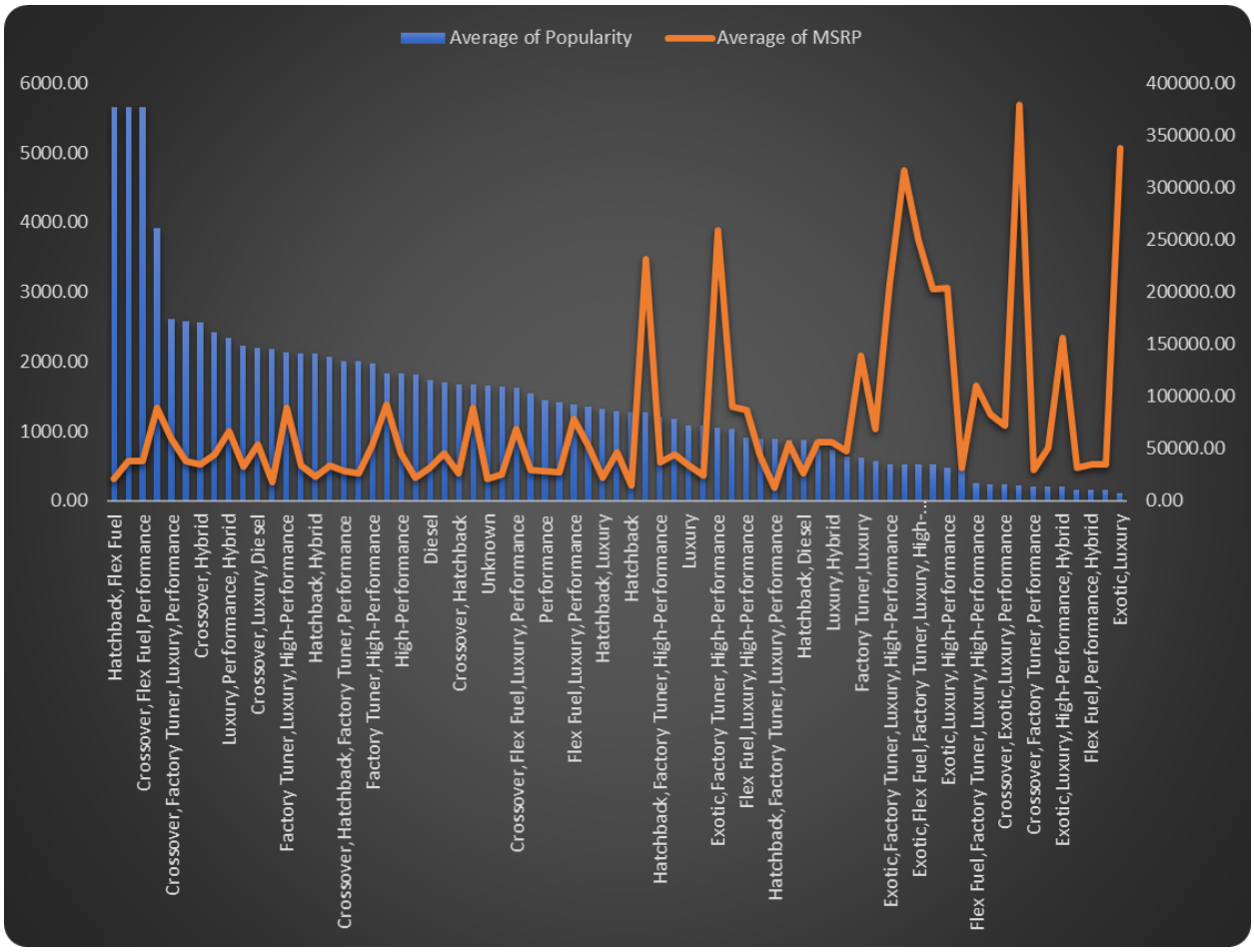
Least Popular Market Category:

- Performance, Hybrid
- Flex Fuel, Performance, Hybrid
- Flex Fuel, Hybrid
- Exotic, Luxury

Task 1. A: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

Market Category	Count of Model	Average of Popularity	Sum of Popularity
Unknown	3362	1658.682629	5576491
Crossover	1068	1539.475655	1644160
Flex Fuel	855	2225.71345	1902985
Luxury	815	1084.21227	883633
Luxury,Performance	659	1293.062215	852128
Hatchback	547	1279.113346	699675
Performance	503	1443.234592	725947
Crossover,Luxury	406	889.2142857	361021
Luxury,High-Performance	334	1668.017964	557118
Exotic,High-Performance	245	1270.326531	311230
Factory Tuner,Luxury,High-Performance	215	2133.367442	458674
Hatchback,Performance	198	1073.661616	212585
High-Performance	198	1823.378788	361029
Hybrid	121	2116.586777	256107
Crossover,Luxury,Performance	112	1349.089286	151098
Factory Tuner,High-Performance	104	1966.442308	204510
Diesel	84	1730.904762	145396
Flex Fuel,Performance	81	1702.358025	137891
Factory Tuner,Performance	81	1818.049383	147262
Exotic,Luxury,High-Performance	77	473.025974	36423
Crossover,Hatchback	72	1675.694444	120650
Crossover,Performance	69	2585.956522	178431
Crossover,Flex Fuel	64	2073.75	132720
Hatchback,Hybrid	64	2111.15625	135114
Exotic,Factory Tuner,Luxury,High-Performance	51	523.0196078	26674
Luxury,Hybrid	48	724.6875	34785

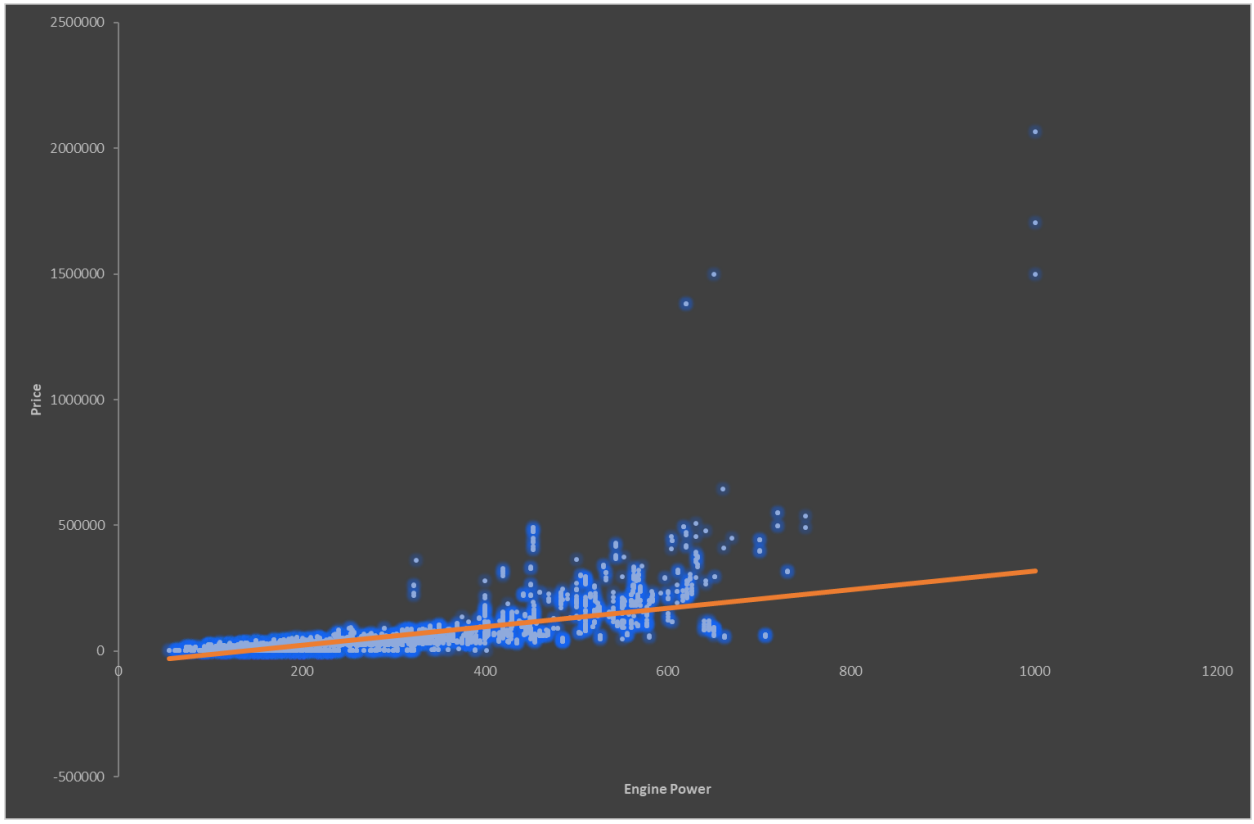
Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.



Q2. What is the relationship between a car's engine power and its price?

It is a positive linear relationship.

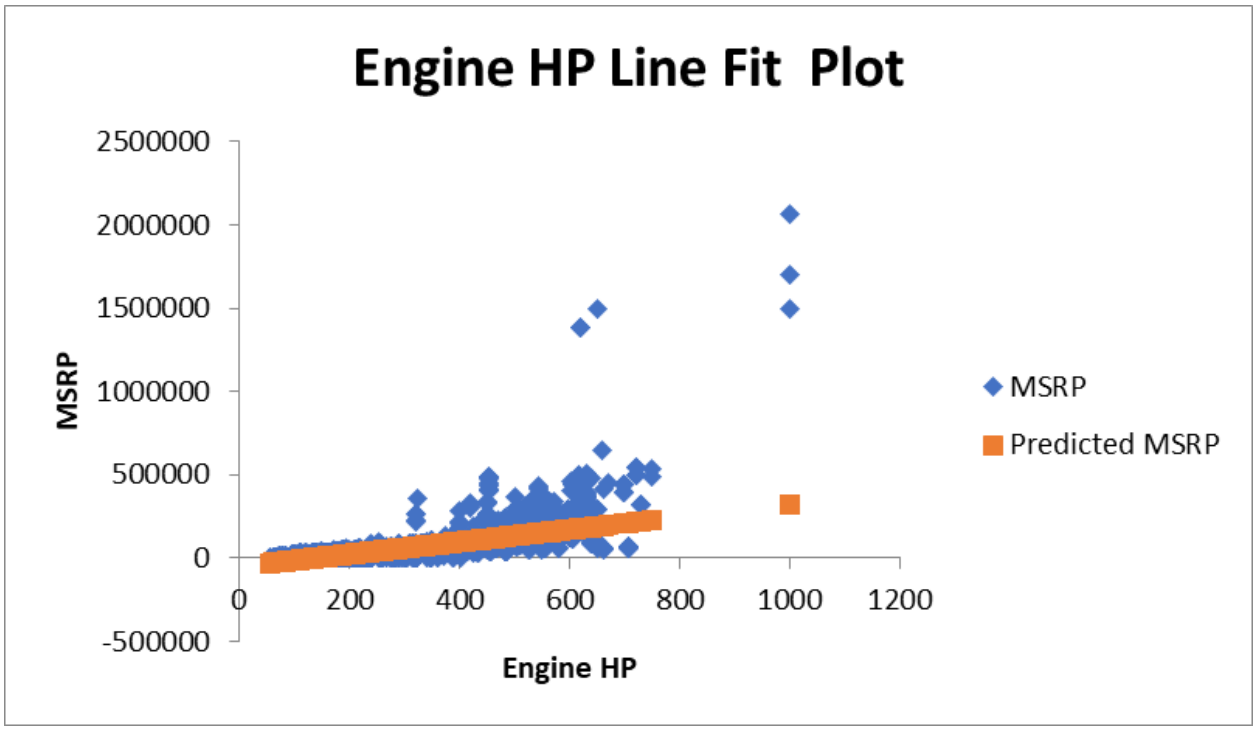
Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

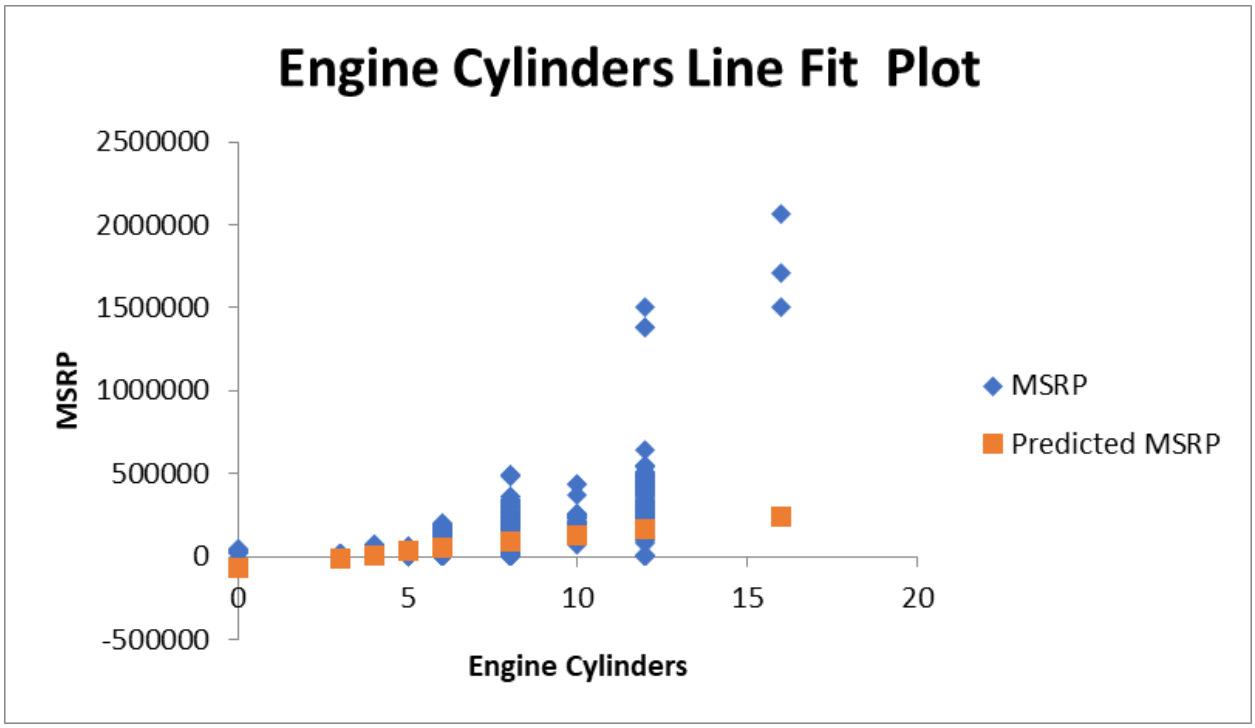


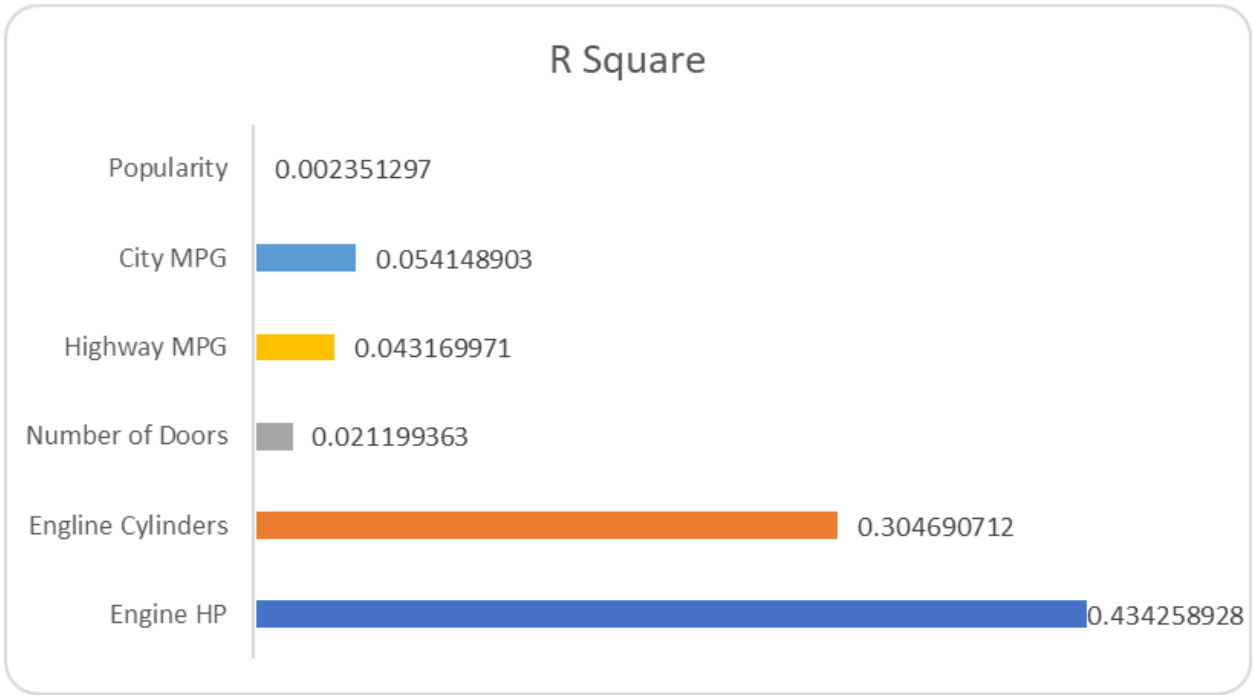
Q3. : Which car features are most important in determining a car's price?

Engine Horsepower and Engine Cylinders are most important.

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.




Engine Cylinders Line Fit Plot
2500000
2000000
1500000
1000000
500000
0
-500000
MSRP
0
5
10
15
20
Engine Cylinders
MSRP
Predicted MSRP

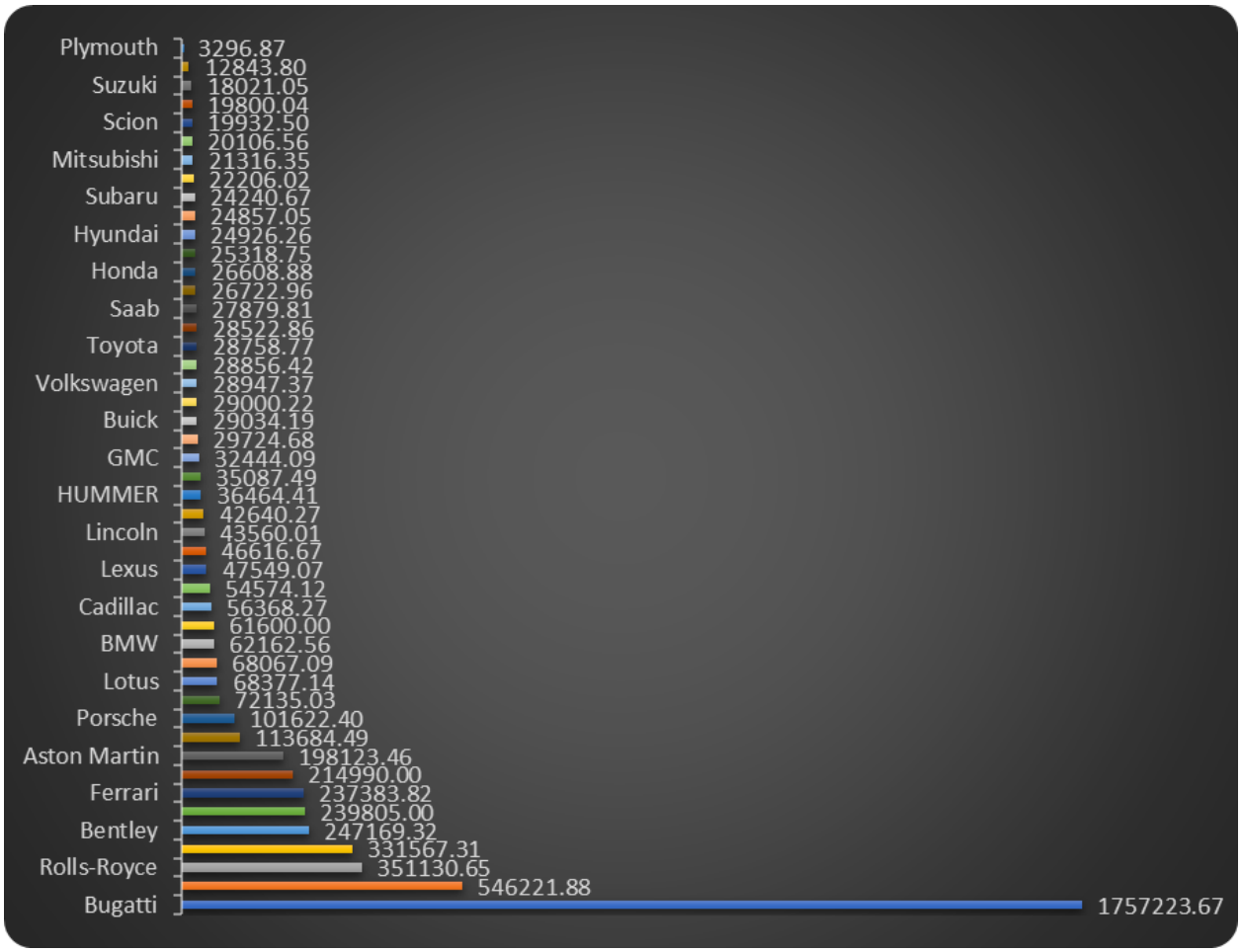

R Square
Popularity
0.002351297
City MPG
0.054148903
Highway MPG
0.043169971
Number of Doors
0.021199363
Engline Cylinders
0.304690712
Engine HP
0.434258928

Q4. How does the average price of a car vary across different manufacturers?

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

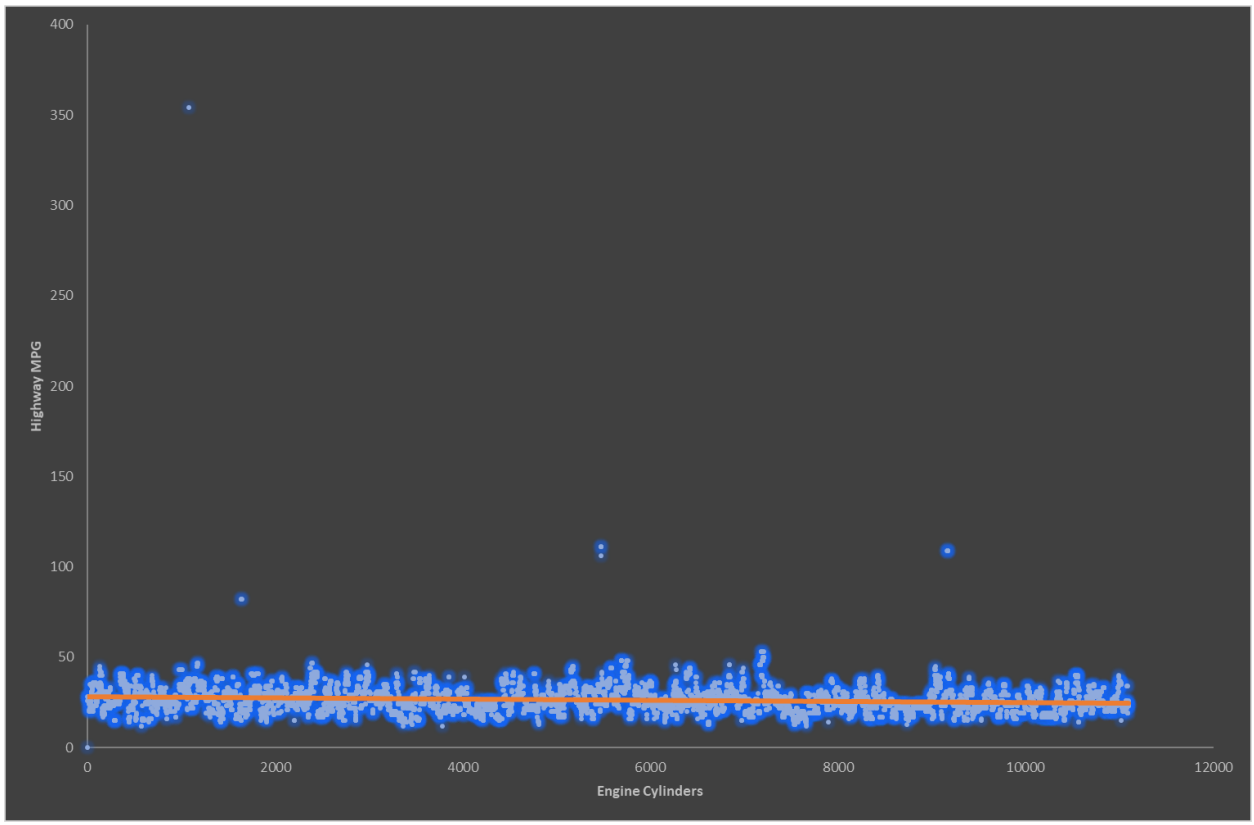
Manufacturer	Average of MSRP
Bugatti	1757223.67
Maybach	546221.88
Rolls-Royce	351130.65
Lamborghini	331567.31
Bentley	247169.32
McLaren	239805.00
Ferrari	237383.82
Spyker	214990.00
Aston Martin	198123.46
Maserati	113684.49
Porsche	101622.40
Mercedes-Benz	72135.03
Lotus	68377.14
Land Rover	68067.09
BMW	62162.56
Alfa Romeo	61600.00
Cadillac	56368.27
Audi	54574.12
Lexus	47549.07
Genesis	46616.67
Lincoln	43560.01
Infiniti	42640.27
HUMMER	36464.41
Acura	35087.49
GMC	32444.09
Volvo	29724.68
Buick	29034.19
Chevrolet	29000.22
Volkswagen	28947.37
Nissan	28856.42
Toyota	28758.77

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



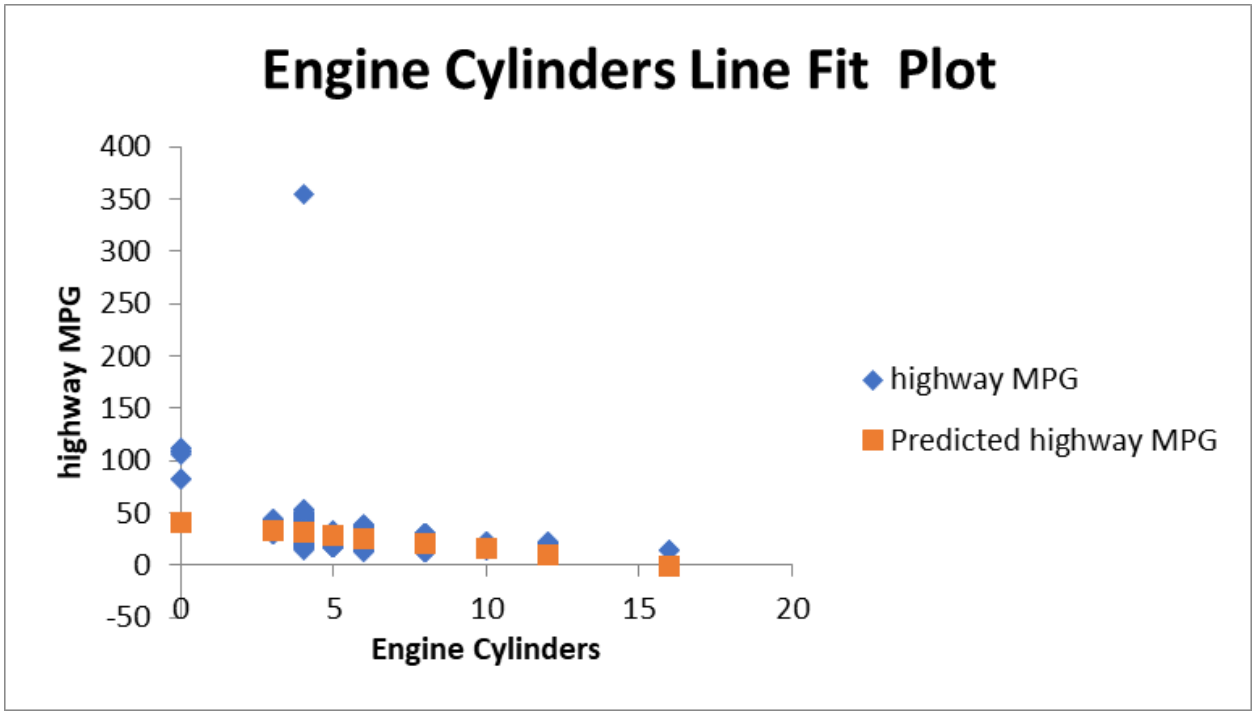
Q5. What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

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.



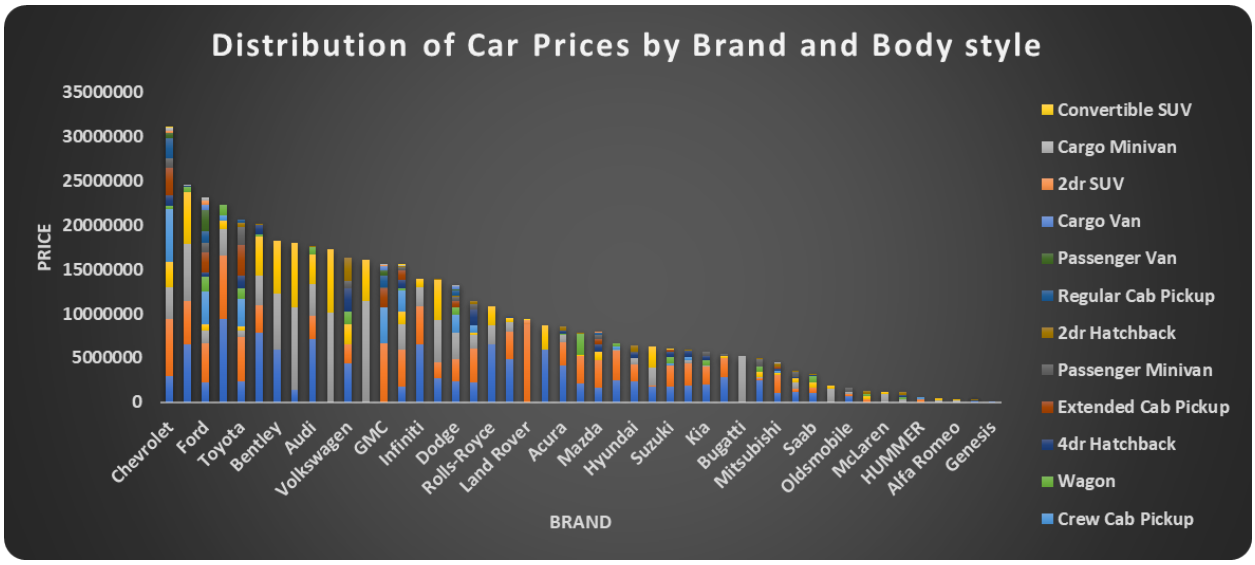
Task 5.B: Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

The correlation coefficient between the number of cylinders and highway MPG is **-0.614703148 (negative and strong)**

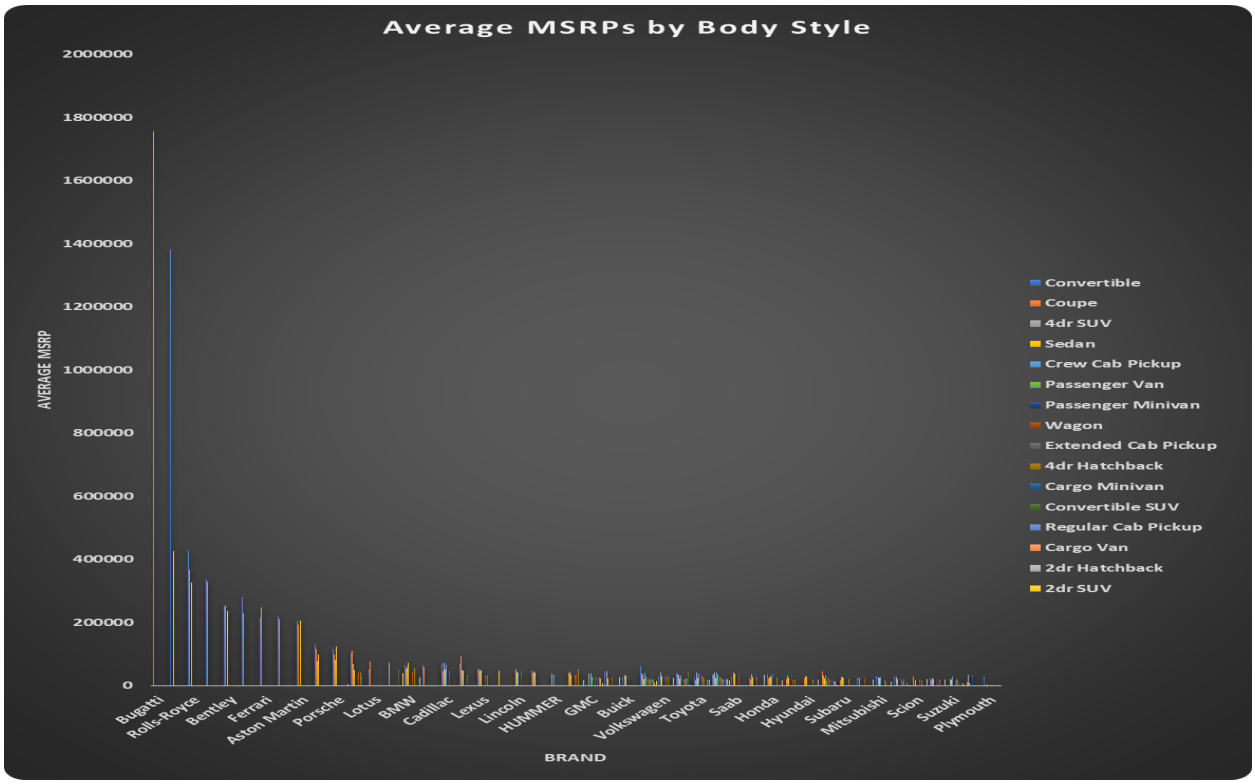


Building the Dashboard:

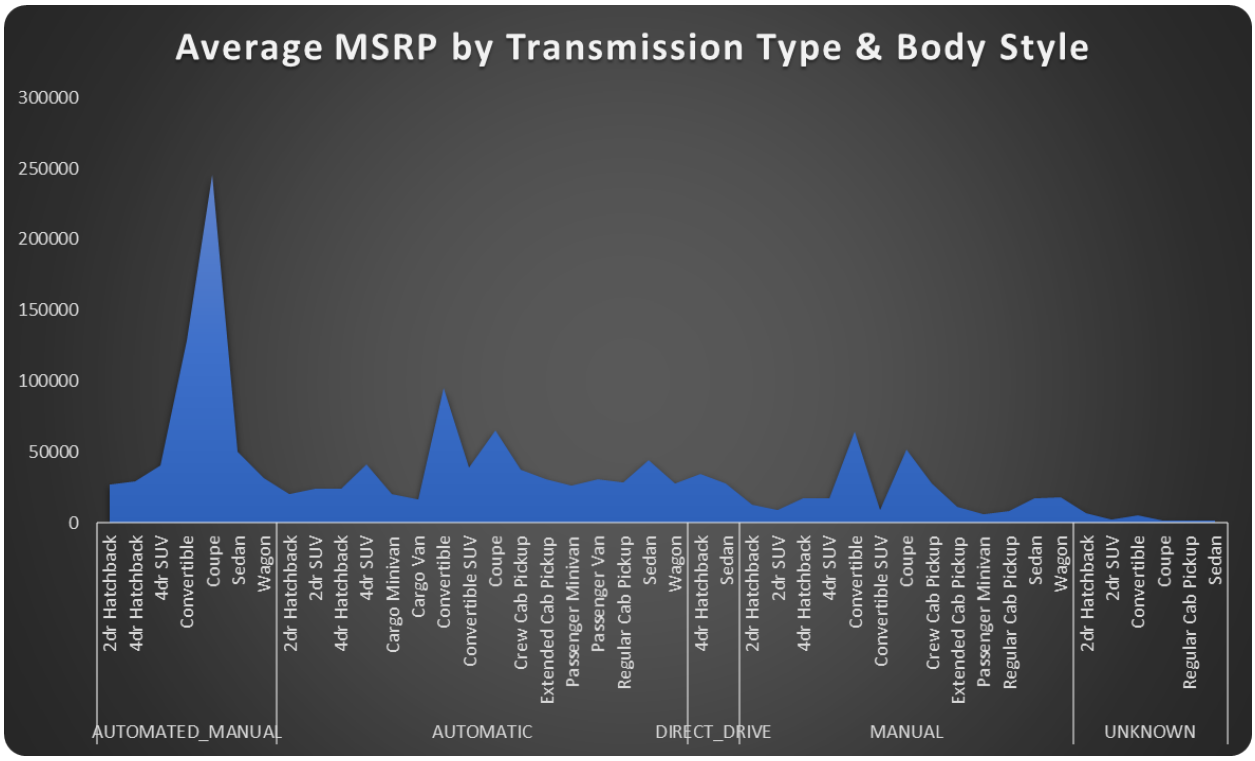
Task 1: How does the distribution of car prices vary by brand and body style?



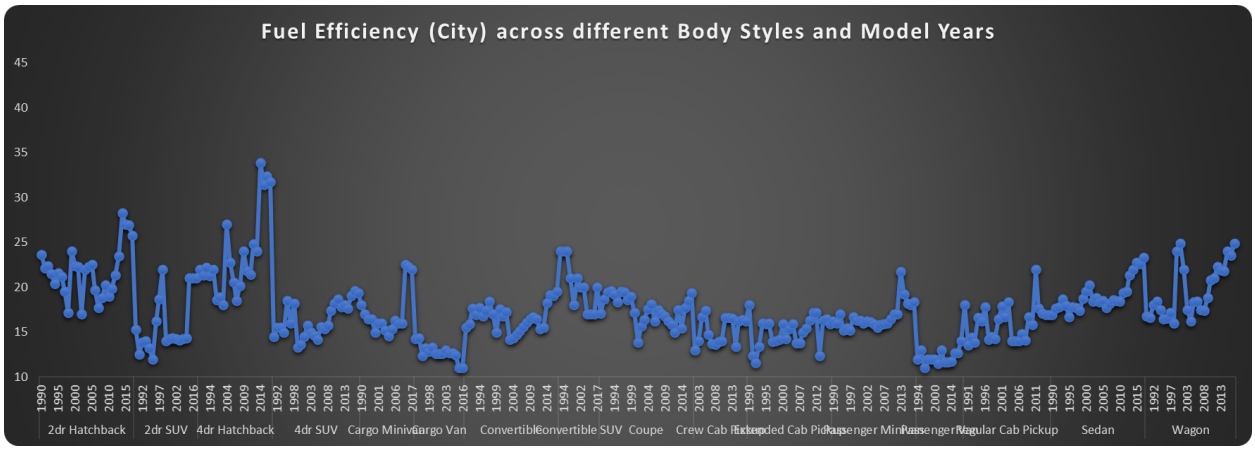
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

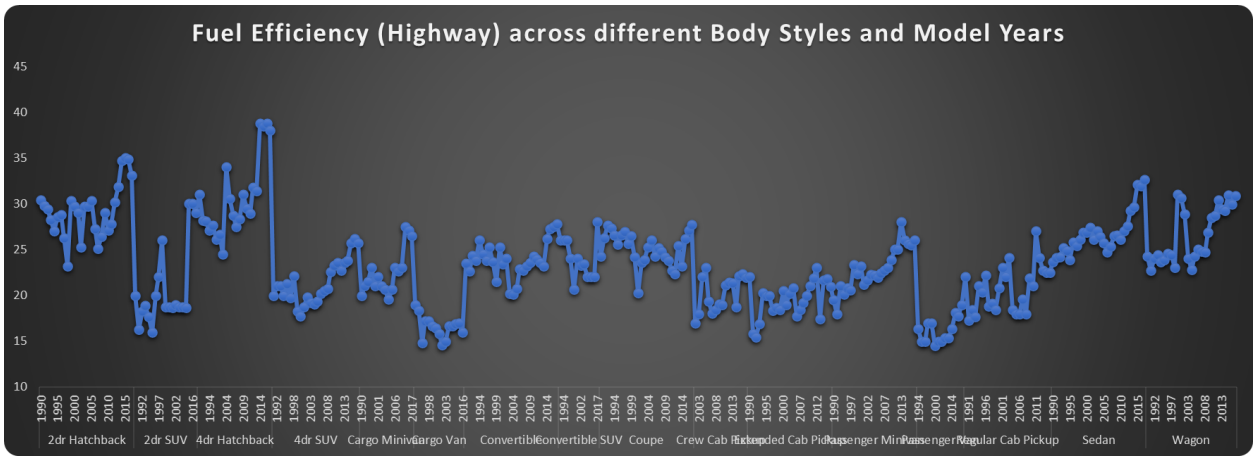


Task 3: How do the different features such as transmission type affect the MSRP, and how does this vary by body style?



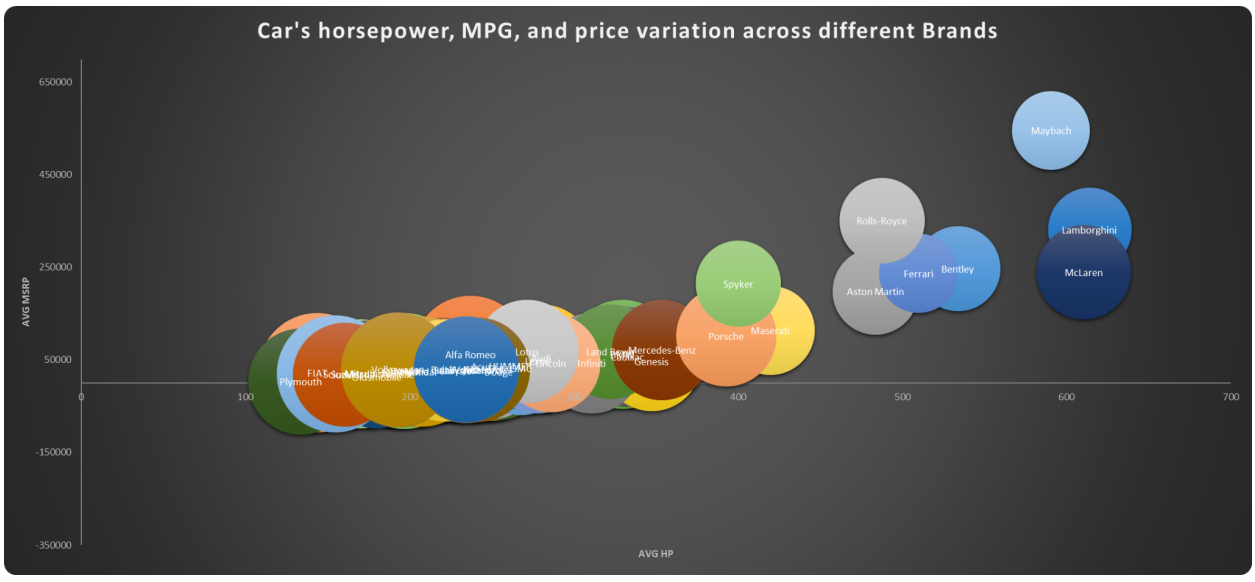
Task 4: How does the fuel efficiency of cars vary across different body styles and model years?





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

Brand Name	Average of Engine HP	Average of Avg MPG	Average of MSRP
Acura	245	24	35087
Alfa Romeo	237	29	61600
Aston Martin	484	16	198123
Audi	280	24	54574
Bentley	534	15	247169
BMW	330	25	62163
Bugatti	1001	11	1757224
Buick	220	23	29034
Cadillac	333	21	56368
Chevrolet	250	22	29000
Chrysler	229	22	26723
Dodge	254	20	24857
Ferrari	510	13	237384
FIAT	144	30	22206
Ford	250	21	28523
Genesis	347	21	46617
GMC	268	19	32444
Honda	197	28	26609
HUMMER	261	15	36464
Hyundai	205	26	24926
Infiniti	311	21	42640
Kia	208	26	25319
Lamborghini	614	15	331567
Land Rover	323	19	68067
Lexus	277	23	47549
Lincoln	286	21	43560
Lotus	272	22	68377
Maserati	420	17	113684
Maybach	591	13	546222
Mazda	170	25	20107
McLaren	610	19	239805



Note: For a detailed analysis and all the visualisation please refer to the following link:

Insights | Dashboard

Project 08:

ABC Call Volume Trend Analysis

Description:

We are provided with a dataset of a Customer Experience (CX) Inbound calling team for 23 days. Data includes Agent_Name, Agent_ID, Queue_Time [duration for which customers have to wait before they get connected to an agent], Time [time at which a call was made by a customer in a day], Time_Bucket [for easiness we have also provided you with the time bucket], Duration [duration for which a customer and executives are on call, Call_Seconds [for simplicity we have also converted those time into seconds], call status (Abandon, answered, transferred).

We are required to analyze the data for the questions asked and provide a detailed report.

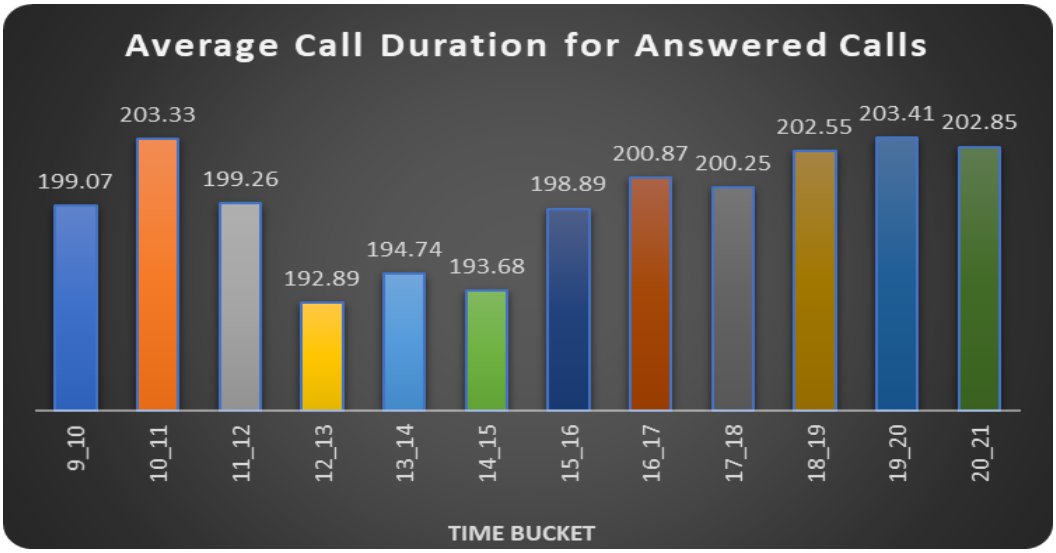
Approach:

- Download and upload the given CSV file in the MS Excel workbook using the Data tab.
- Clean the data:
- Remove the irrelevant columns entirely.
- Remove the rows with any blank or null value.
- Remove the duplicate rows from the data.
- Using pivot tables and pivot charts to find the insights.

Insights:

Q1. Calculate the average call time duration for all incoming calls received by agents (in each Time_Bucket).

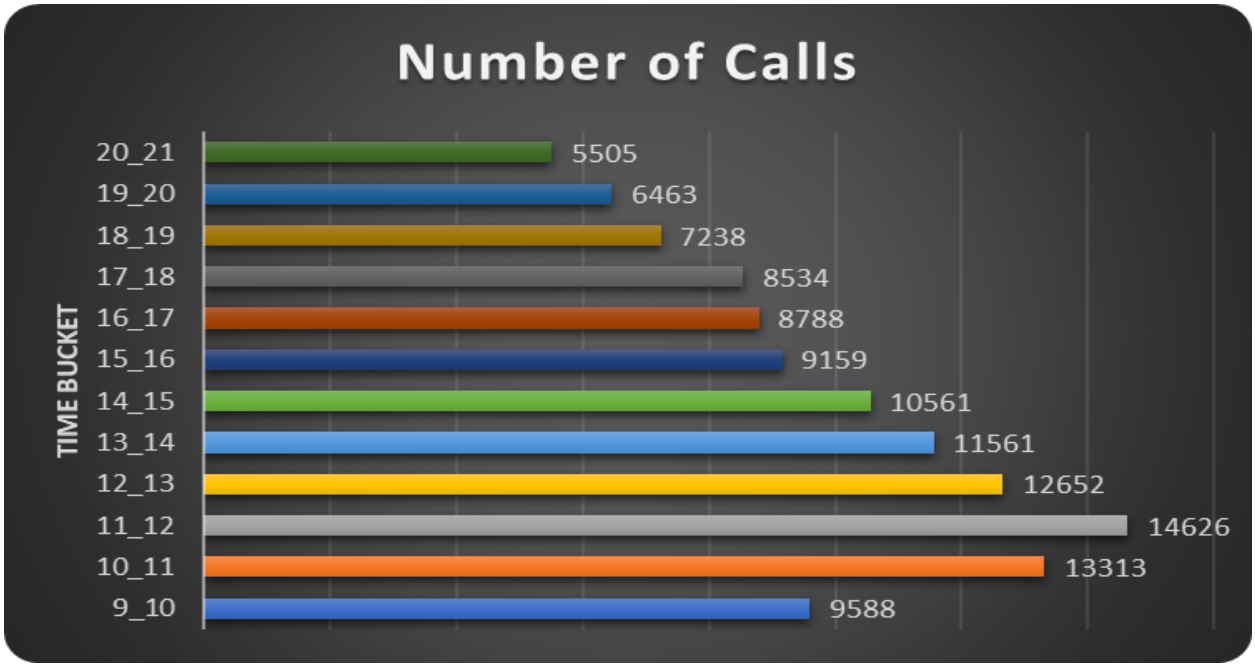
Call_Status	answered
Time Bucket	**Average of Call_Seconds (s)**
9_10	199.07
10_11	203.33
11_12	199.26
12_13	192.89
13_14	194.74
14_15	193.68
15_16	198.89
16_17	200.87
17_18	200.25
18_19	202.55
19_20	203.41
20_21	202.85
Grand Total	**198.62**



- The total average call time duration for all incoming calls received by agents is 198.62 seconds.
- The busiest time buckets if from 10 AM to 11 AM and 07 PM to 08 PM.
- The lowest average call time duration for the answered call is between 12 PM to 4 PM.

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

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



- Most calls are received between 11 AM to 12 PM.
- The least calls are received between 8 PM to 9 PM.

Q3. Let's say customers also call this ABC insurance company at night but didn't get an answer as there are no agents to answer, this creates a bad customer experience for this Insurance company. Suppose for every 100 calls that the customer made from 9 AM to 9 PM, the customer also made 30 calls in the night between interval [9 Pm to 9 Am] and the 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. The maximum abandon rate assumption would be the same 10%.

Average Call per Day Shift:

= Total No. of Calls (from Q2) / 23

= 117988 / 23

= ~5130 calls

Average Call per Night Shift:

= 30% of Average Call per Day Shift

= 30% * 5130

= ~1539 calls

Average Calls answered per second= 198.62

Call Status	No.of Calls	% No. of Calls	Average of Call_Seconds (s)
abandon	34403	29.16%	0.00
answered	82452	69.88%	198.62
transfer	1133	0.96%	76.15
Grand Total	**117988**	**100.00%**	**139.53**

Additional Hours(Night)

=Average Call per Night Shift * Average Calls answered per second* 0.9 (% answered calls) / 3600

= ~76 Hours

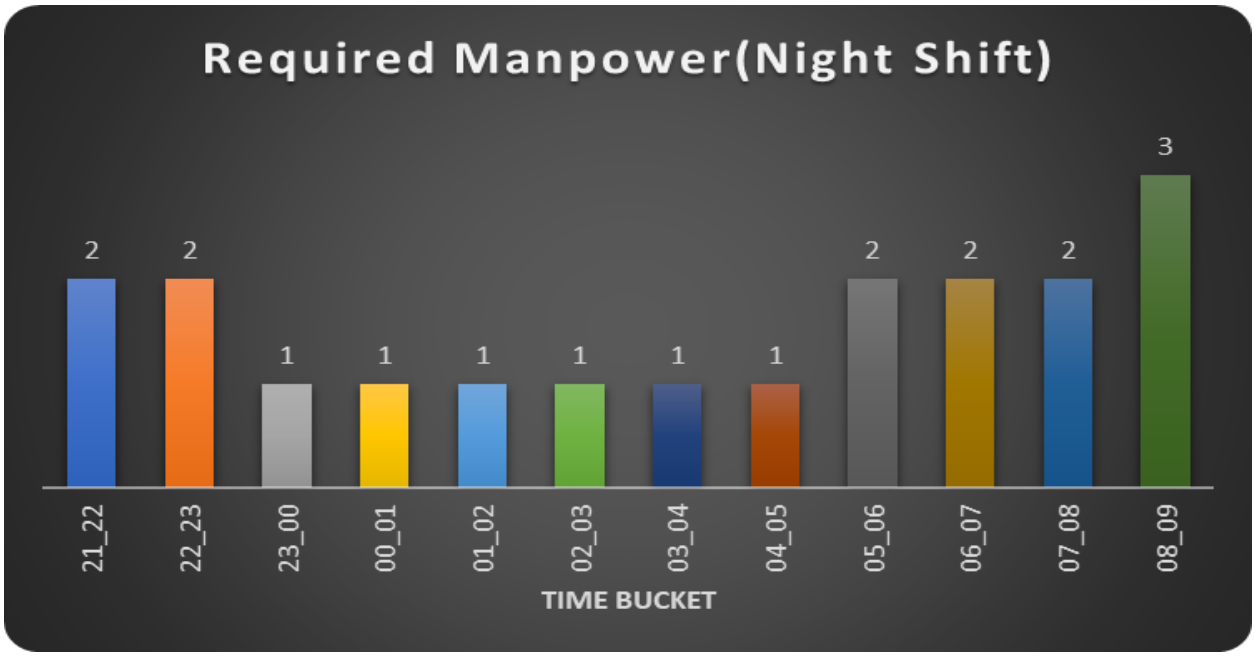
Additional Manpower:

= Additional Hours / Working Hours

= 76 / 4.5

= ~17

Night Timing	Call Distribution	% Call Distribution	Required Manpower (Actual)	Required Manpower
21_22	3	0.10	1.7	2
22_23	3	0.10	1.7	2
23_00	2	0.07	1.133333333	1
00_01	2	0.07	1.133333333	1
01_02	1	0.03	0.566666667	1
02_03	1	0.03	0.566666667	1
03_04	1	0.03	0.566666667	1
04_05	1	0.03	0.566666667	1
05_06	3	0.10	1.7	2
06_07	4	0.13	2.266666667	2
07_08	4	0.13	2.266666667	2
08_09	5	0.17	2.833333333	3
Total	30	1.00	17	19



- Total Manpower Required (Day & Night) is 83 [64+19]
- We have considered a whole number for the manpower because manpower cannot be in fractions.
- The work shifts can be manipulated to get maximum results with the minimum number of manpower (5 AM to 2 PM shift & 2 PM to 11 PM shift).

Note: For a detailed analysis and all the visualisation please refer to the following link:

Call Volume Trend Analysis

My Learnings:

I have learned a lot during the course of these projects. Starting from the basics to the advanced level concepts were integrated well with these and helped me with hands-on practice on real-life projects.

Some skills I have developed and polished during these projects are:

- Understanding the 6 steps of the Data Analytics Process.
- Basic and Advance MySQL concepts with use cases.
- Understanding the data given and framing the questions around the data.
- Business Case Studies and Report Making
- Analysing the data using Python programming languages and its libraries.
- Hand-On data visualization and deriving valuable insights from them.
- Creating interactive dashboards.
- Statistical concepts with their use cases.
- Concepts of Operation Analysis & Investigating Metric Spikes
- HR, Risk & Behavioral Analytics
- Researching, asking for help, and practising to understand the concepts.