

The background of the slide is a dark blue, abstract digital pattern. It consists of numerous thin, glowing blue lines and small squares that create a sense of depth and movement, resembling a data stream or a complex network. The lines and squares are arranged in a way that suggests a three-dimensional space, with some elements appearing closer and others further away.

Data Analysis Portfolio

Prepared by:
Mohammed sabir

Professional BackGround:

- Currently in final year pursuing B.Tech-C.S.E with a specialization in Artificial Intelligence and Machine Learning. I have a strong academic background in various computer languages like Java, C, and Python(numpy, pandas) and SQL . Proficient in Microsoft Office including Advanced Excel and Microsoft Power BI.
- I can effectively create Dashboards to facilitate end-to-end business decision-making.
- Adept at communication, and stakeholder management.
- Actively looking for opportunities in the field of Data Analytics, Artificial Intelligence and Machine Learning where I can leverage my current skillset.

Table Of Contents:

1.	Professional BackGround -----	2
2.	Project 1 -----	4-5
3.	Project 2 -----	6-14
4.	Project 3 -----	15-24
5.	Project 4 -----	25-31
6.	Project 5 -----	32-37
7.	Project 6 -----	38-43
8.	Project 7 -----	44-49
9.	Project 8 -----	50-55
10.	Learnings -----	56
11.	Project Links -----	57-58

Project 1: Data Analytics Process

- Description: We use Data Analytics in every aspect of life without even knowing it.
- E.g. Going to a war
- Plan: Before engaging in war, the first step we take is to strategically assess the strengths & weaknesses of our own military forces, as well as those of our adversaries.
- Prepare: After the plan has been prepared, we need to find a way to overcome our weaknesses and also outlining the specific measures we will implement to enhance our capabilities.
- Process: Then after preparation we focus on providing advanced weapons and training to inexperienced soldiers in order to enhance their skills and capabilities.
- Analyze: After all the preceding steps, we analyze our readiness for war, assessing our position, potential and also forecasting the likelihood of victory or defeat and we refrain from fighting if victory is uncertain.
- Share: Here we share our analysis with the commander to determine the most optimal solution.
- Act: Then you finally go for a war.

Conclusion Of Project 1:

- In Project 1 we got the basic understanding about what is actual data analytics with real life example it was designed in such a way that layman can also understand what data analytics is.

Project 2: Instagram User Analytics

- **Project Description:** Analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow, these insights will help the team with the future direction of the Instagram App.
- **Approach:** First we create a database `ig_clone` in MySQL workbench then after that we create particular tables according to the requirement in the database `ig_clone` then we insert the values in each of the table, after all this is done we use MySQL commands to extract valuable insights from database.
- **Tech-Stack Used:** We use MySQL and MySQL Workbench to create and extract insights from Instagram database.

Loyal User Reward

Code:

```
select*from users  
order by created_at  
limit 5;
```

5 Loyal user reward
Output/Result:

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

Inactive User Engagement

Code:

```
select username  
from users  
left join photos on users.id=photos.user_id  
where photos.id is null;
```

Inactive User Engagement,
Output/Result:

Aniya_Hackett
Bartholome.Bernhard
Bethany20
Darby_Herzog
David.Osinski47
Duane60
Esmeralda.Mraz57
Esther.Zulauf61
Franco_Keebler64
Hulda.Macejkovic
Jaclyn81
Janelle.Nikolaus81
Jessyca_West
Julien_Schmidt

Kassandra_Homenick
Leslie67
Linnea59
Maxwell.Halvorson
Mckenna17
Mike.Auer39
Morgan.Kassulke
Nia_Haag
Ollie_Ledner37
Pearl7
Rocio33
Tierra.Trantow

Contest Winner Declaration

Code:

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

users_id	username	photo_id	image_url	total_likes
52	Zack_Kemmer93	145	https://jarret.name	48

Hashtag Research

Code:

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

Five most commonly used hashtags on the platform
Output/result:

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

Ad Campaign Launch

Code:

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

day_of_week	user_registered
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

User Engagement

- Code: Calculate the average number of posts per user on Instagram

```
select user_id,count(*) as post_count  
from photos  
group by user_id  
order by user_id;
```

user_id	post_count
86	9
87	4
88	11
92	3
93	2
94	1
95	2
96	3
97	2
98	1
99	3
100	2

user_id	post_count
42	3
43	5
44	4
46	4
47	5
48	1
50	3
51	5
52	5
55	1
56	1
58	8
59	10
60	2
61	1
62	2
63	4
64	5
65	5
67	3
69	1
70	1
72	5
73	1
77	6
78	5
79	1
82	2
84	2
85	2

user_id	post_count
1	5
2	4
3	4
4	3
6	5
8	4
9	4
10	3
11	5
12	4
13	5
15	4
16	4
17	3
18	1
19	2
20	1
22	1
23	12
26	5
27	1
28	4
29	8
30	2
31	1
32	4
33	5
35	2
37	1
38	2
39	1
40	1

- Code: Total number of photos on Instagram divided by the total number of users.

```
select(select count(*)  
from photos)/(select count(*) from users) as division_result;
```

Output/Result:

division_result
2.5700

Bots & Fake Accounts

Code:

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

user_id	username	total_likes
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

PROJECT 3: OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

Project Description:

- In this project our role is of lead data analyst at a company like Microsoft we have to perform operation analytics on the given data , our goal is to derive valuable insights from the given data ,these analysis helps identify areas for improvement within the company.
- One of the key aspects of Operational Analytics is investigating metric spikes. This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales

CASE STUDY 1: JOB DATA ANALYSIS

First of all we create a database named job_data then we create a table named jobs after the table is created we export the data from csv file into MYSQL.

In case study 1 we will be working with a table named jobs

TASK (A): JOBS REVIEWED OVER TIME

- Your Task: Write an SQL query to calculate the number of jobs reviewed per hour for each day in November 2020.

Code:

```
SELECT job_id,ds,
COUNT(*) AS jobs_reviewed,
SUM(time_spent) AS total_time_spent,
COUNT(*) / (SUM(time_spent) / 3600) AS jobs_review_per_hour
FROM jobs
GROUP BY job_id , ds
order by ds;
```

OUTPUT:

job_id	ds	jobs_reviewed	total_time_spent	jobs_review_per_hour
20	11/25/2020	1	45	80.0000
23	11/26/2020	1	56	64.2857
11	11/27/2020	1	104	34.6154
23	11/28/2020	1	22	163.6364
25	11/28/2020	1	11	327.2728
23	11/29/2020	1	20	180.0000
21	11/30/2020	1	15	240.0000
22	11/30/2020	1	25	144.0000

- INSIGHTS:
- Job_id 11 has highest total time spent and jobs reviewed per hour is least when compared to all others.
- Job_id 25 has least total time spent and jobs reviewed per hour is highest when compared to all others.
- Total time spent and jobs review per hour are inversely proportional to each other.
- On 28th November no.of jobs review per hour are highest and on 27th November lowest no.of jobs review per hour

TASK B:THROUGHPUT ANALYSIS

Write an SQL query to calculate the 7-day rolling average of throughput. Additionally, explain whether you prefer using the daily metric or the 7-day rolling average for throughput, and why.

CODE:

```
SELECT job_id,  
AVG(SUM(time_spent))  
OVER (ROWS BETWEEN 6 PRECEDING AND CURRENT ROW)  
AS rolling_avg_throughput  
FROM jobs  
GROUP BY job_id  
ORDER BY job_id;
```

OUTPUT:

job_id	rolling_avg_throughput
11	50.6000
20	49.6667
21	15.0000
22	20.0000
23	46.0000
25	37.2500

- INSIGHTS:
job_id 11 has the highest rolling avg throughput
- Job_is 21 has the lowest rolling avg throughput

❖ TASK 3: LANGUAGE SHARE ANALYSIS

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

❖ CODE:

```
select language,  
count(*) as lang_count,  
total_count_per_lang,  
((lang_count/total_count_per_lang)*100)*30 as percent_lang_share  
from(  
select language,  
count(*) as lang_count,  
sum(count(*)) over() as total_count_per_lang  
from jobs  
group by language ) as total_lang_count  
group by language,total_count_per_lang  
order by language;
```

language	lang_count	total_count_per_lang	percent_lang_share
Arabic	1	8	375.0000
English	1	8	375.0000
French	1	8	375.0000
Hindi	1	8	375.0000
Italian	1	8	375.0000
Persian	1	8	1125.0000

INSIGHTS:

Persian language has the highest percentage share.

◆ TASK 4: Duplicate Rows Detection

◆ Your Task: Write an SQL query to display duplicate rows from the jobs table.

CODE:

```
select job_id,language,  
count(*) as duplicate_count  
from jobs  
group by job_id,language  
having count(*)>1;
```

◆ Having is used to apply conditions to the result set after grouping has been performed

INSIGHTS:

Persian language and job_id 23 are the duplicates which are repeated 3 times

OUTPUT:

job_id	language	duplicate_count
23	Persian	3

CASE STUDY 2: INVESTIGATING METRIC SPIKE

In this case we will be dealing with three tables, we will create a database named metric, then we will export data from csv files to the tables in mysql.

Output:

week_num	weekly_user_engagement
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302

Task A: Weekly User Engagement

Your Task: Write an SQL query to calculate the weekly user engagement.

Code:

```
select extract(week from occurred_at) as week_num,  
count(DISTINCT user_id) as weekly_user_engagement  
from events  
where event_type='engagement'  
group by week_num  
order by week_num;
```

week_num	weekly_user_engagement
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104

Insights:

Week number 30 has the highest weekly user engagement

Week number 35 has the lowest weekly user engagement

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

```
select year,week_num,user_num,sum(user_num)
over (order by year,week_num) as user_data
from(
select extract(year from created_at) as year,
extract(week from created_at) as week_num,
count(distinct user_id) as user_num
from users
where not state='active'
group by year,week_num
order by year,week_num) as table_data;
```

year	week_num	user_num	user_data
2013	0	23	23
2013	1	30	53
2013	2	48	101
2013	3	36	137
2013	4	30	167
2013	5	48	215
2013	6	38	253
2013	7	42	295
2013	8	34	329
2013	9	43	372
2013	10	32	404
2013	11	31	435
2013	12	33	468
2013	13	39	507
2013	14	35	542
2013	15	43	585

year	week_num	user_num	user_data
2013	16	46	631
2013	17	49	680
2013	18	44	724
2013	19	57	781
2013	20	39	820
2013	21	49	869
2013	22	54	923
2013	23	50	973
2013	24	45	1018
2013	25	57	1075
2013	26	56	1131
2013	27	52	1183
2013	28	72	1255
2013	29	67	1322
2013	30	67	1389

- Insights:
- From 2013 to 2014 there is continuous growth in user number and user data
- In the last week of 2014 the user number is very low but the user data is high
- On week 33 of 2014 it has highest number of users
- On week 35 of 2014 it has lowest number of users

◆ Task C: Weekly Retention Analysis

Code:

```
WITH cte1 AS ( SELECT DISTINCT      user_id,
EXTRACT(WEEK FROM occurred_at) AS sign_up_week
FROM events
WHERE event_type = 'signup_flow'
AND event_name = 'complete_signup'
AND EXTRACT(WEEK FROM occurred_at) = 18),
cte2 AS ( SELECT DISTINCT      user_id,
EXTRACT(WEEK FROM occurred_at) AS engagement_week
FROM events
WHERE event_type = 'engagement')
```

```
SELECT  COUNT(DISTINCT user_id) AS total_engaged_users,
SUM(CASE WHEN retention_week = 0 THEN 1 ELSE 0 END) AS retained_users
FROM (
SELECT
a.user_id,
a.sign_up_week,
b.engagement_week,
COALESCE(b.engagement_week - a.sign_up_week, 0) AS retention_week
FROM  cte1 a
LEFT JOIN  cte2 b ON a.user_id = b.user_id
ORDER BY  a.user_id) sub_quer;
```

total_engaged_users	retained_users
163	163

◆ Task D: Weekly Engagement Per Device

Your Task: Write an SQL query to calculate the weekly engagement per device.

Code:

```
with cte as (select extract(year from occurred_at)||'-'||  
extract(week from occurred_at) as weeknum,  
device, count(distinct user_id) as user_count  
from events  
where event_type = 'engagement'  
group by weeknum, device  
order by weeknum)  
select weeknum,device,user_count  
from cte;
```

weeknum	device	user_count
1	acer aspire desktop	198
1	acer aspire notebook	338
1	amazon fire phone	89
1	asus chromebook	355
1	dell inspiron desktop	360
1	dell inspiron notebook	677
1	hp pavilion desktop	339
1	htc one	196
1	ipad air	478
1	ipad mini	292
1	iphone 4s	409
1	iphone 5	1025
1	iphone 5s	626
1	kindle fire	205
1	lenovo thinkpad	1309
1	mac mini	150

◆ Task E: Email Engagement Analysis


Your Task: Write an SQL query to calculate the email engagement metrics.

Code:

```
select
100*sum(case when email_cat = 'email_open' then 1 else 0 end)/
sum(case when email_cat = 'email_sent' then 1 else 0 end) as email_open_rate,
100*sum(case when email_cat = 'email_clicked' then 1 else 0 end)/
sum(case when email_cat = 'email_sent' then 1 else 0 end) as email_click_rate
from(select*,
Case when action in ('sent_weekly_digest','sent_reengagement_email')
then 'email_sent'
when action in ('email_open') then 'email_open'
when action in ('email_clickthrough') then 'email_clicked'
end as email_cat
from `email events`)sub;
```

Output:

email_open_rate	email_click_rate
33.5834	14.7899



PROJECT 4: HIRING PROCESS ANALYTICS

DESCRIPTION: AS A DATA ANALYST IN MULTINATIONAL COMPANY LIKE GOOGLE
OUR TASK IS TO ANALYZE COMPANY'S HIRING PROCESS DATA AND DRAW
MEANINGFUL INSIGHTS FROM IT.

WE HAVE USED EXCEL TO SOLVE THE GIVEN TASKS AND EXTRACT THE INSIGHT

Handling Missing Data

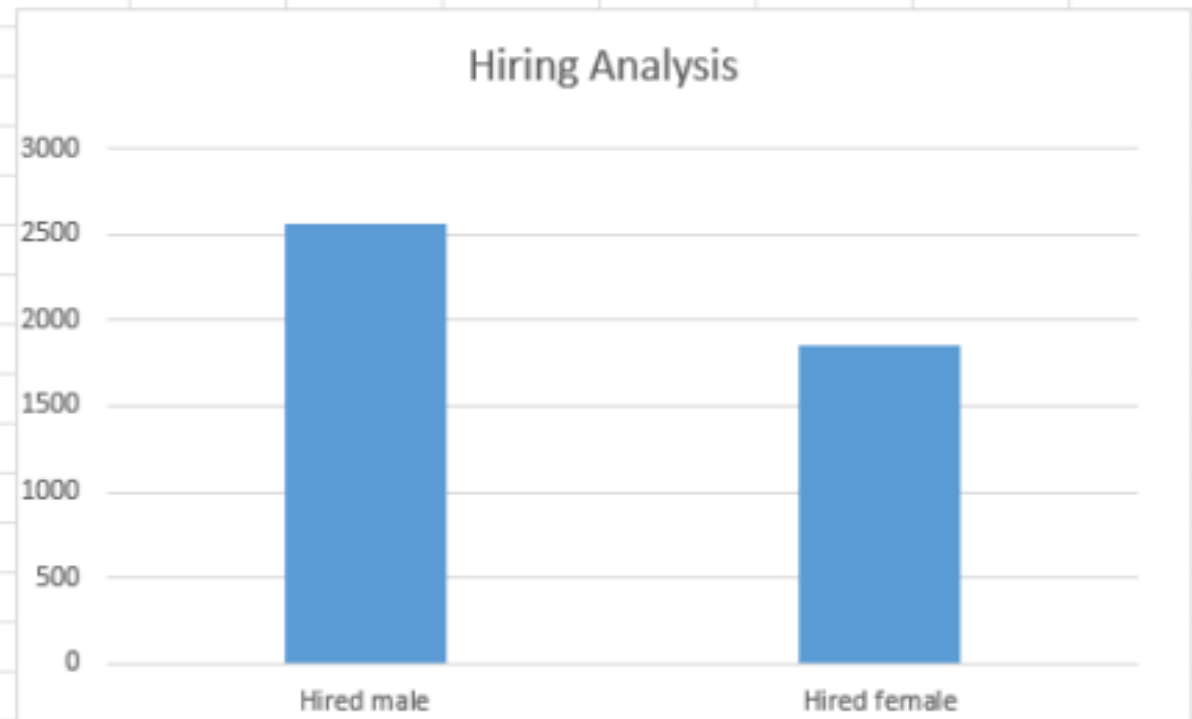
- ▶ Handling missing data involves managing and addressing the absence of values in a dataset. It is a crucial step in data analysis and machine learning as missing data can adversely affect the validity and reliability of results. The goal is to choose appropriate strategies to either remove or fill in missing values, ensuring that the analysis or model is based on as complete and accurate data as possible.

Data Analytics Tasks:

- ▶ Task A: Hiring Analysis
- ▶ The hiring process involves bringing new individuals into the organization for various roles.
- ▶ Your Task: Determine the gender distribution of hires. How many males and females have been hired by the company?

Count of event_name	Column Labels		
Row Labels	Female	Male	Grand Total
Hired	1856	2563	4419
Grand Total	1856	2563	4419

Hired male	2563
Hired female	1856



- **INSIGHTS:**
- With the help of pivot table we got to know the number of males and females hired
- More number of males are hired when compared to female

Task B: Salary Analysis

- ▶ The average salary is calculated by adding up the salaries of a group of employees and then dividing the total by the number of employees.
- ▶ Your Task: What is the average salary offered by this company? Use Excel functions to calculate this.

Average salary			49983.02902	

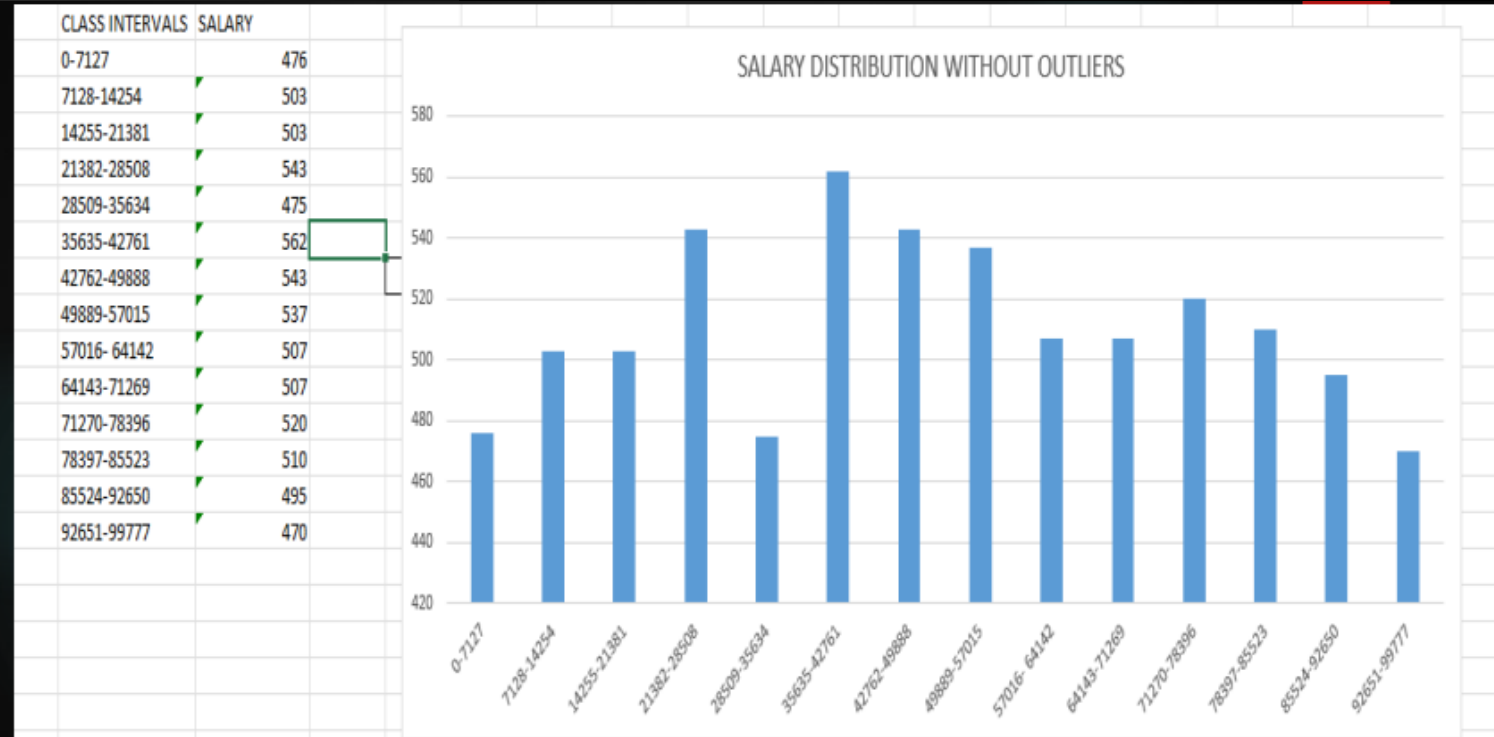
- Insights:
- We have calculated average salary using average function
- Average is 49983 after removing outliers

Task C: Salary Distribution

- ▶ Class intervals represent ranges of values, in this case, salary ranges. The class interval is the difference between the upper and lower limits of a class.
- ▶ Your Task: Create class intervals for the salaries in the company. This will help you understand the salary distribution.

INSIGHTS: Salary Distribution

- We are having 14 class intervals
- Most of the employees have salary between 35635-42761
- And least number of employees have salaries between 92651-99777
- most of them have salaries near to average or less

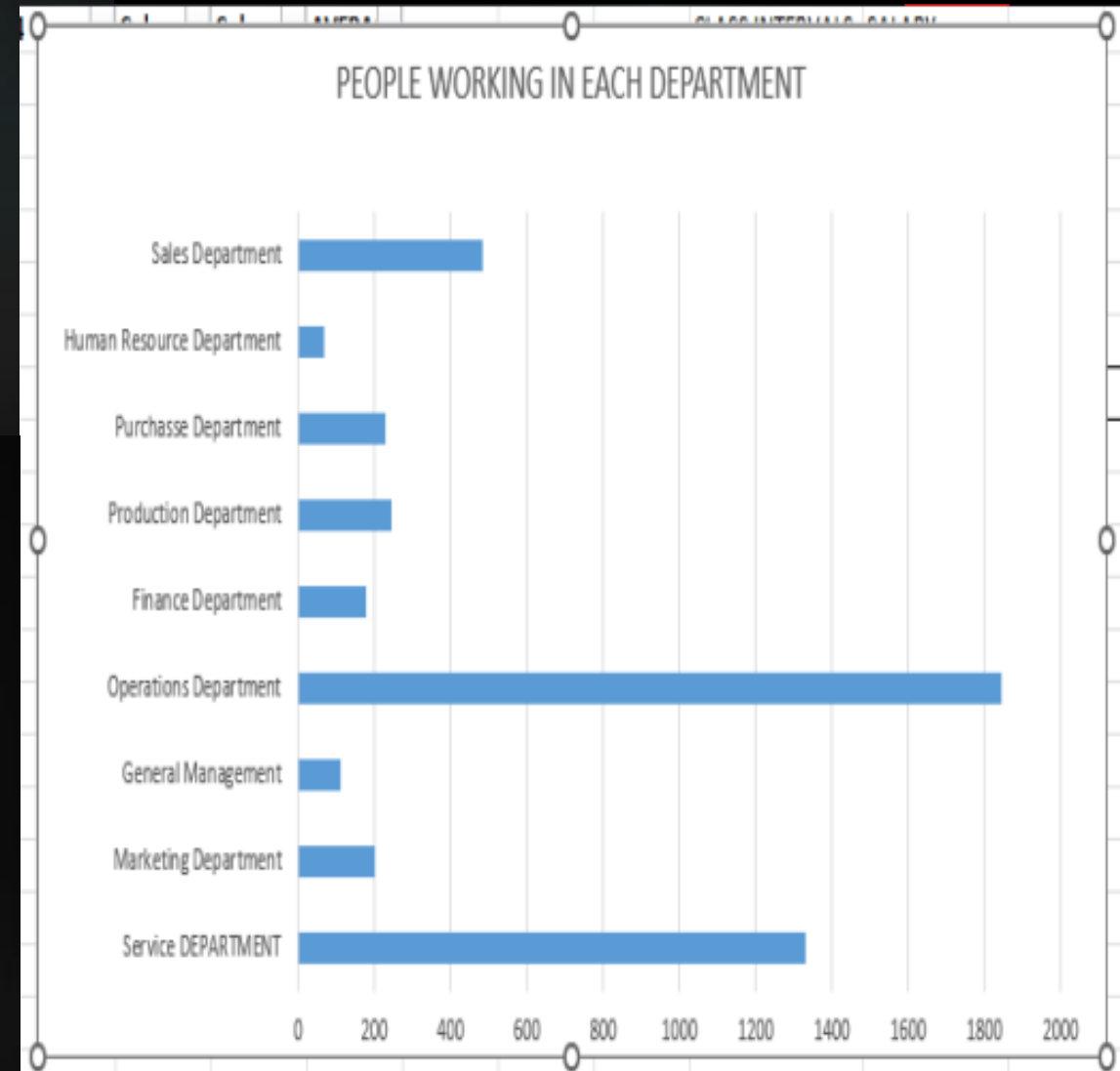


Task D: Department Analysis

- ▶ Visualizing data through charts and plots is a crucial part of data analysis.
- ▶ Your Task: Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

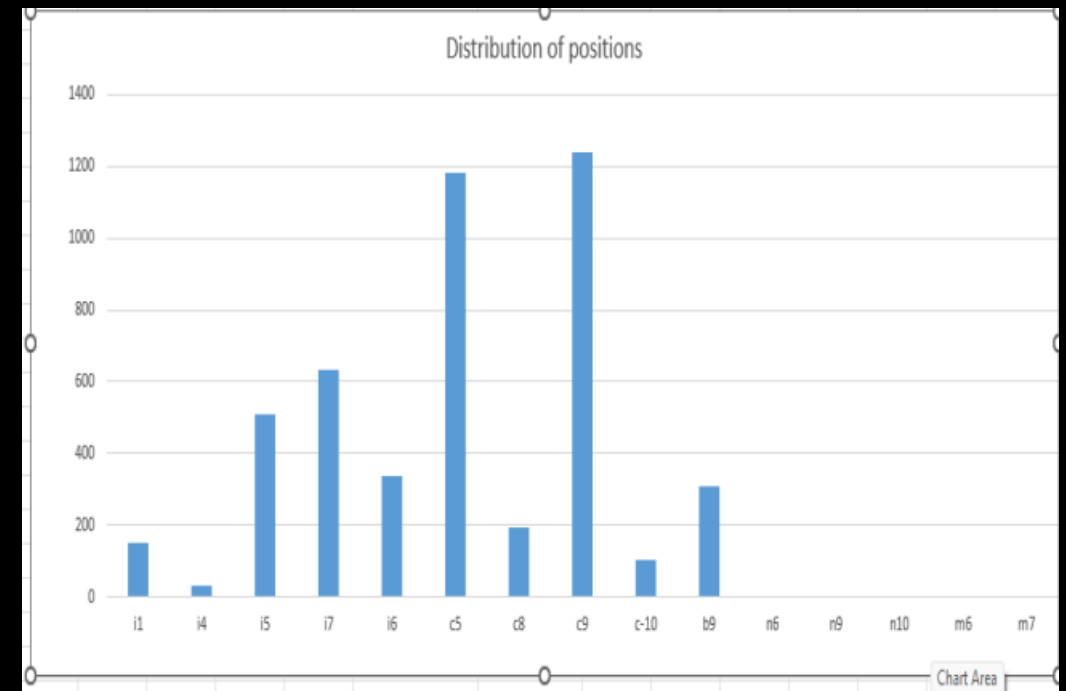
Insights:

- ▶ Operation department has the highest number of employees working
- ▶ Human Resource Department has the lowest number of employees working
- ▶ There is huge difference between the number of employees in Operation Department when compared to other departments.



Task E: position tier analysis

- ▶ Position Tier Analysis: Different positions within a company often have different tiers or levels.
- ▶ Your Task: Use a chart or graph to represent the different position tiers within the company. This will help you understand the distribution of positions across different tiers.



Insights:

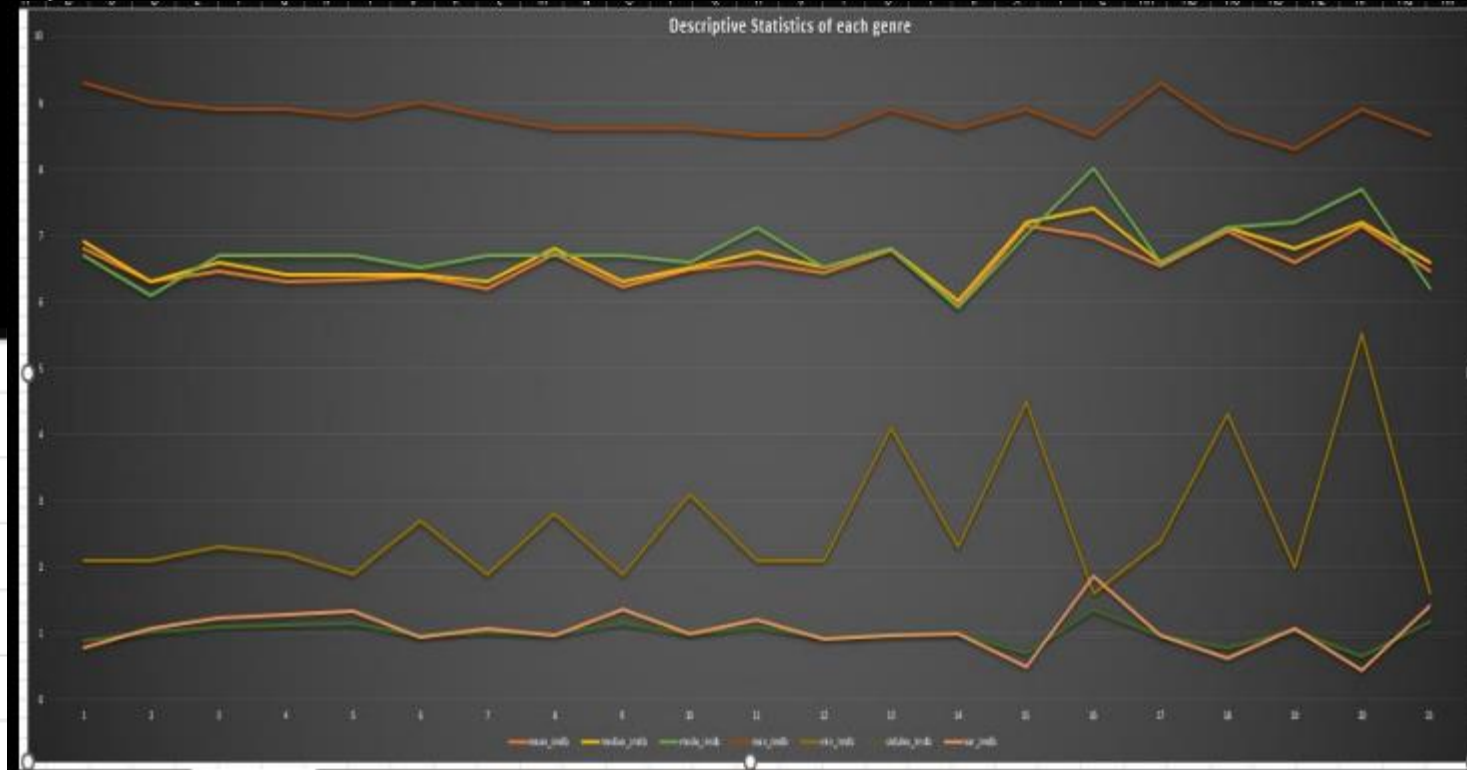
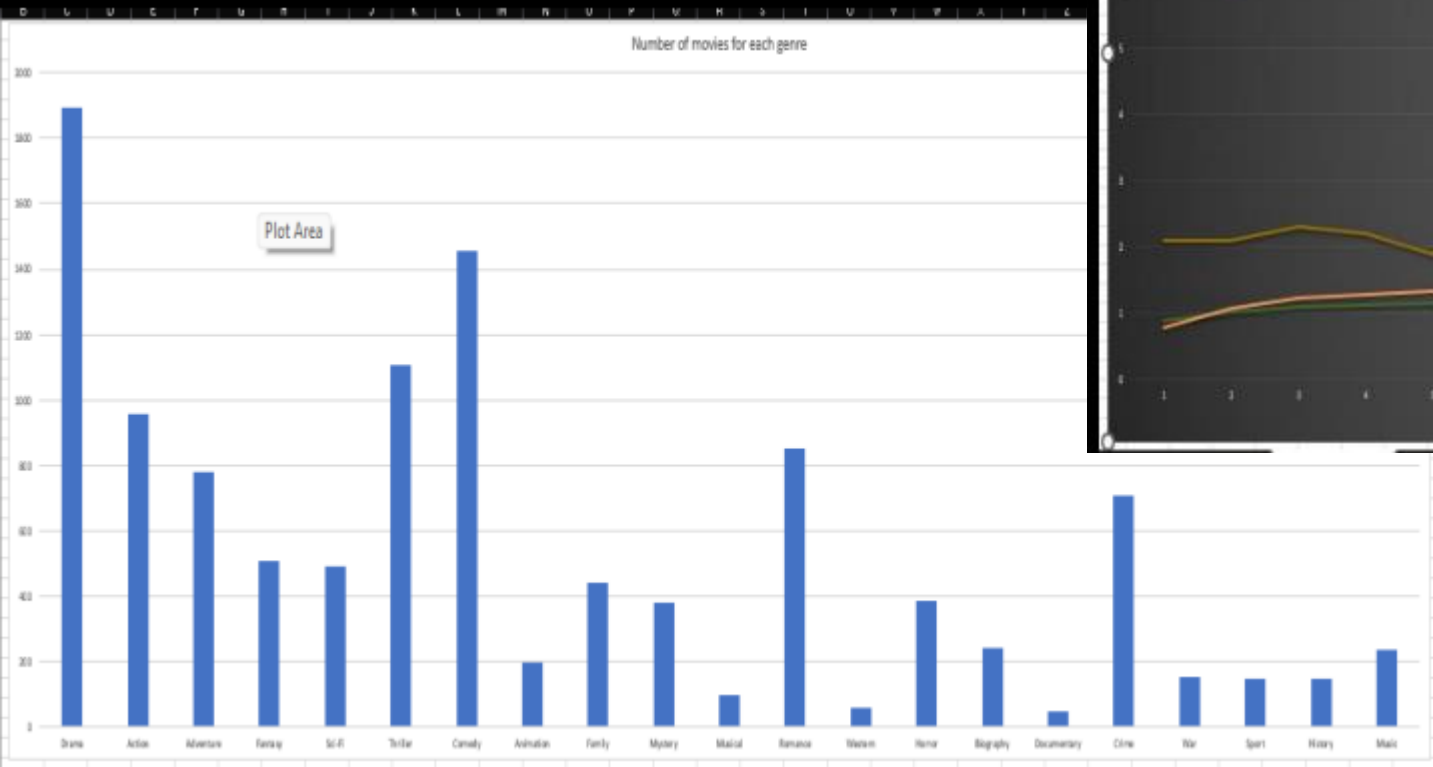
- ▶ C9 has the highest number of position distribution
- ▶ N9,n10,m7 doesn't have any position distribution

IMDB Movie Analysis

- Project Description: We have been given data about movies with imdb ratings and all other information about movies , our role as a data analyst is to figure out “what factors influence the success of a movie on IMDB”
- Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.
- IMDB:
IMDb stands for the Internet Movie Database. It is an online database that provides information about films, television programs, video games, and streaming content

TASK A:Movie Genre Analysis

- ◆ A. Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.
- ◆ Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

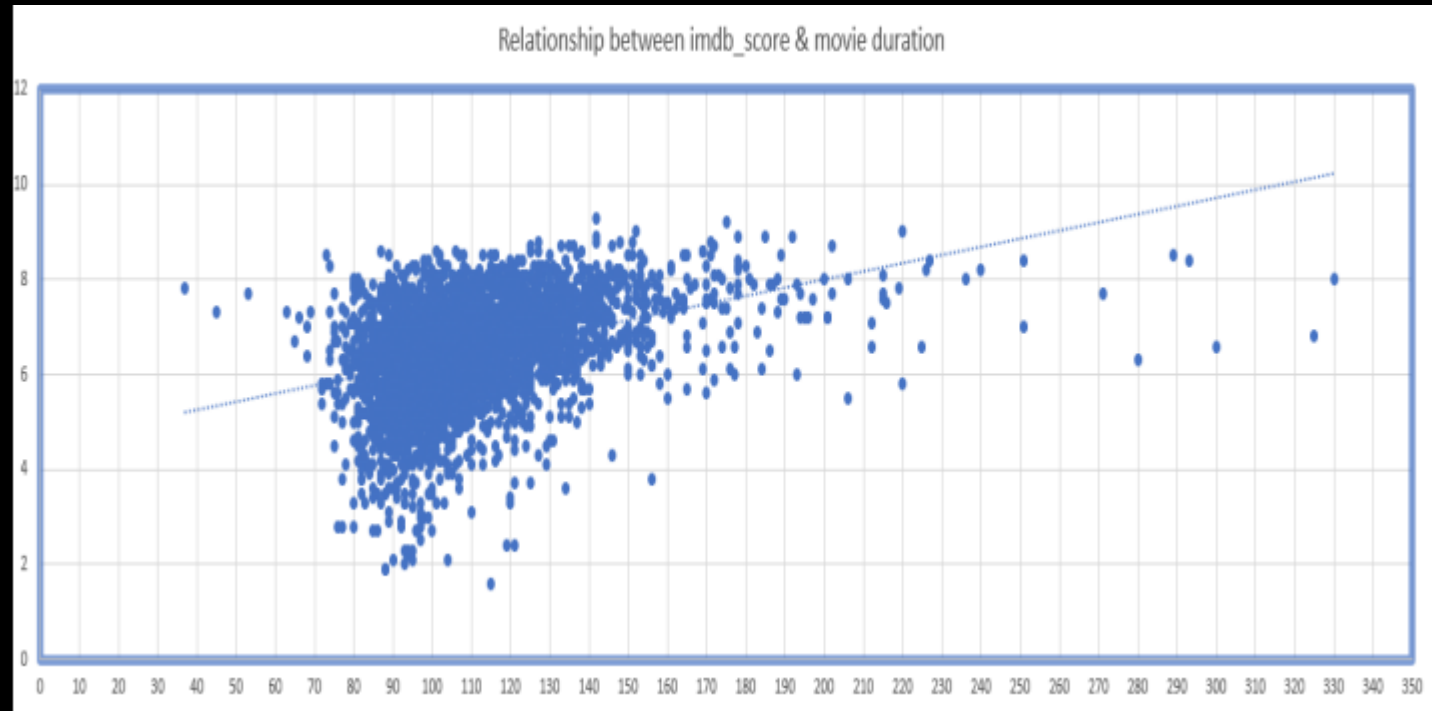


Task B: Movie Duration Analysis

- ◆ Analyze the distribution of movie durations and its impact on the IMDB score.
- ◆ Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Movie duration based on Imdb score

Movie Duration	
Mean	110.2131
Median	106
Mode	101
Standard Deviation	22.67561
Variance	514.1832



Language Analysis

- ◊ Situation: Examine the distribution of movies based on their language.
- ◊ Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Languages	count of movies	mean_imdb	median_imdb	Std.dev_imdb	var_imdb
English	3591	6.4265107	6.5	1.048170912	1.09866
Aboriginal	2	6.95	6.95	0.55	0.3025
Arabic	1	7.2	7.2	0	0
Aramaic	1	7.1	7.1	0	0
Bosnian	1	4.3	4.3	0	0
Cantonese	7	7.3428571	7.3	0.324509048	0.10531
Czech	1	7.4	7.4	0	0
Danish	3	7.9	8.1	0.43204938	0.18667
Dari	2	7.08125	7.4	0.709065186	0.50277
Dutch	3	7.5666667	7.8	0.329983165	0.10889
Filipino	1	6.7	6.7	0	0
French	34	7.3558824	7.3	0.51173935	0.26188
German	11	7.7636364	7.8	0.644237008	0.41504
Hebrew	1	8	8	0	0
Hindi	5	7.22	7.4	0.716658915	0.5136
Hungarian	1	7.1	7.1	0	0

Indonesian	2	7.9	7.9	0.3	0.09
Italian	7	7.1857143	7	1.069617517	1.14408
Japanese	10	7.66	8	0.939361485	0.8824
Kazakh	1	6	6	0	0
Korean	5	7.7	7.7	0.509901951	0.26
Mandarin	14	7.0214286	7.25	0.737930089	0.54454
Maya	1	7.8	7.8	0	0
Mongolian	1	7.3	7.3	0	0
None	1	8.5	8.5	0	0
Norwegian	4	7.15	7.3	0.497493719	0.2475
Persian	3	8.1333333	8.4	0.449691252	0.20222
Portuguese	5	7.76	8	0.875442745	0.7664
Romanian	1	7.9	7.9	0	0
Russian	1	6.5	6.5	0	0
Spanish	23	7.0826087	7.2	0.841660974	0.70839
Thai	3	6.6333333	6.6	0.368178701	0.13556
Vietnamese	1	7.4	7.4	0	0
Zulu	1	7.3	7.3	0	0

Director Analysis

- ◆ Influence of directors on movie ratings.
- ◆ Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

director_name	avg_imdb	percentile	count_movi
John Landis	8.7	0.999	3
Sam Mendes	8.6	0.999	8
Wes Anderson	8.6	0.999	7
Spike Lee	8.5	0.998	15
Peter Ho-Sun Chan	8.5	0.998	1
Paul Weitz	8.5	0.998	6
Tate Taylor	8.5	0.998	2
Bruce Dellis	8.433333333	0.997	1
Martin Scorsese	8.433333333	0.997	16
Trent Cooper	8.433333333	0.997	1
Dominic Sena	8.425	0.995	3
Nick Cassavetes	8.425	0.995	4
Howard Zieff	8.425	0.995	2
Damian Nieman	8.425	0.995	1
William Cottrell	8.425	0.995	1
Alan Parker	8.425	0.995	3
Paul Greengrass	8.425	0.995	7
Morten Tyldum	8.425	0.995	2
Len Wiseman	8.4	0.994	3
Terry George	8.4	0.994	1

Budget Analysis

- ◆ Explore the relationship between movie budgets and their financial success.
- ◆ Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

movie_title	top-10 movies
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Kinsey	449935665
Insidious:chapter 3	424449459
The Avengers	403279547
The Lion King	377783777
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061
The Hunger Games	329999255

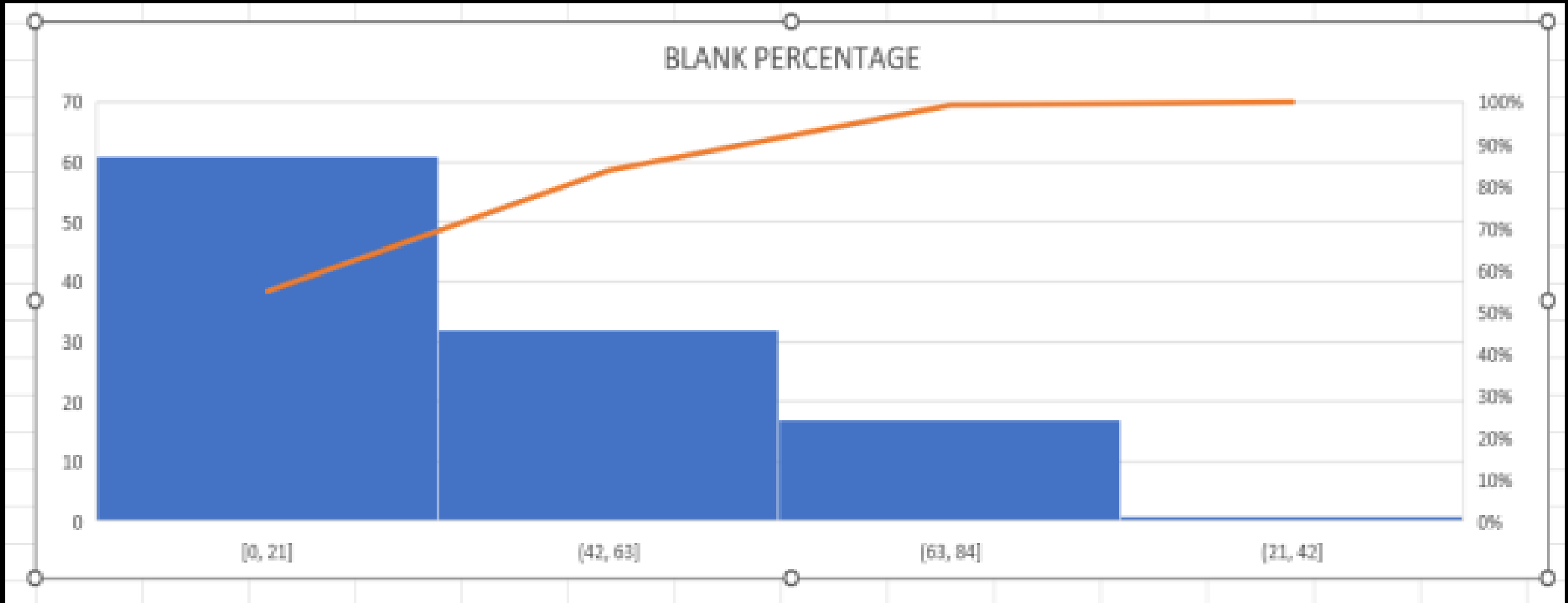
Project 6: Bank Loan Case Study

Project Description

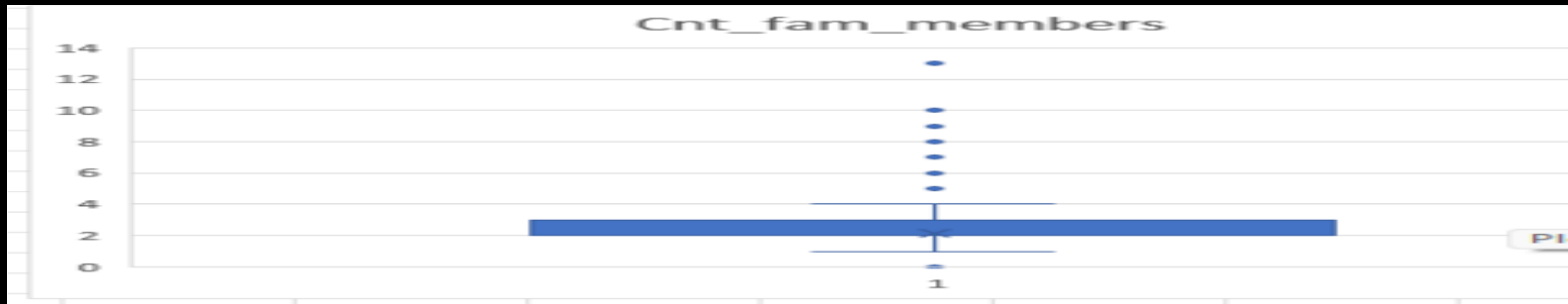
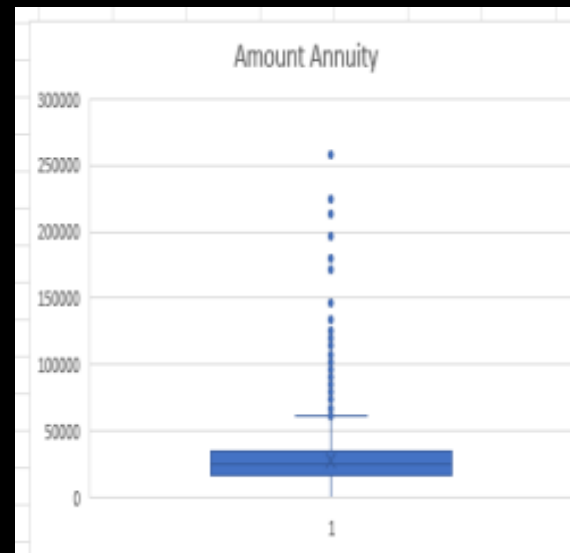
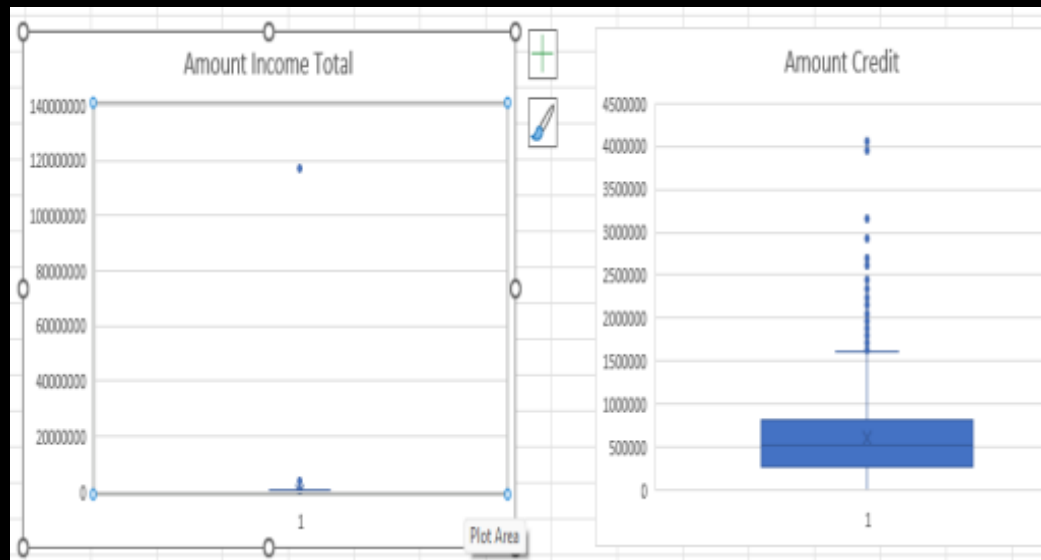
- ◊ Our role is as a data analyst at a finance company that specializes in lending various types of loans to urban customers. our company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.
- ◊ When a customer applies for a loan, your company faces two risks:
- ◊ If the applicant can repay the loan but is not approved, the company loses business.
- ◊ If the applicant cannot repay the loan and is approved, the company faces a financial loss.
- ◊ When a customer applies for a loan, there are four possible outcomes:
- ◊ Approved: The company has approved the loan application.
- ◊ Cancelled: The customer cancelled the application during the approval process.
- ◊ Refused: The company rejected the loan.
- ◊ Unused Offer: The loan was approved but the customer did not use it.
- ◊ Our goal in this project is to use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

Task A: Data Analytics Tasks

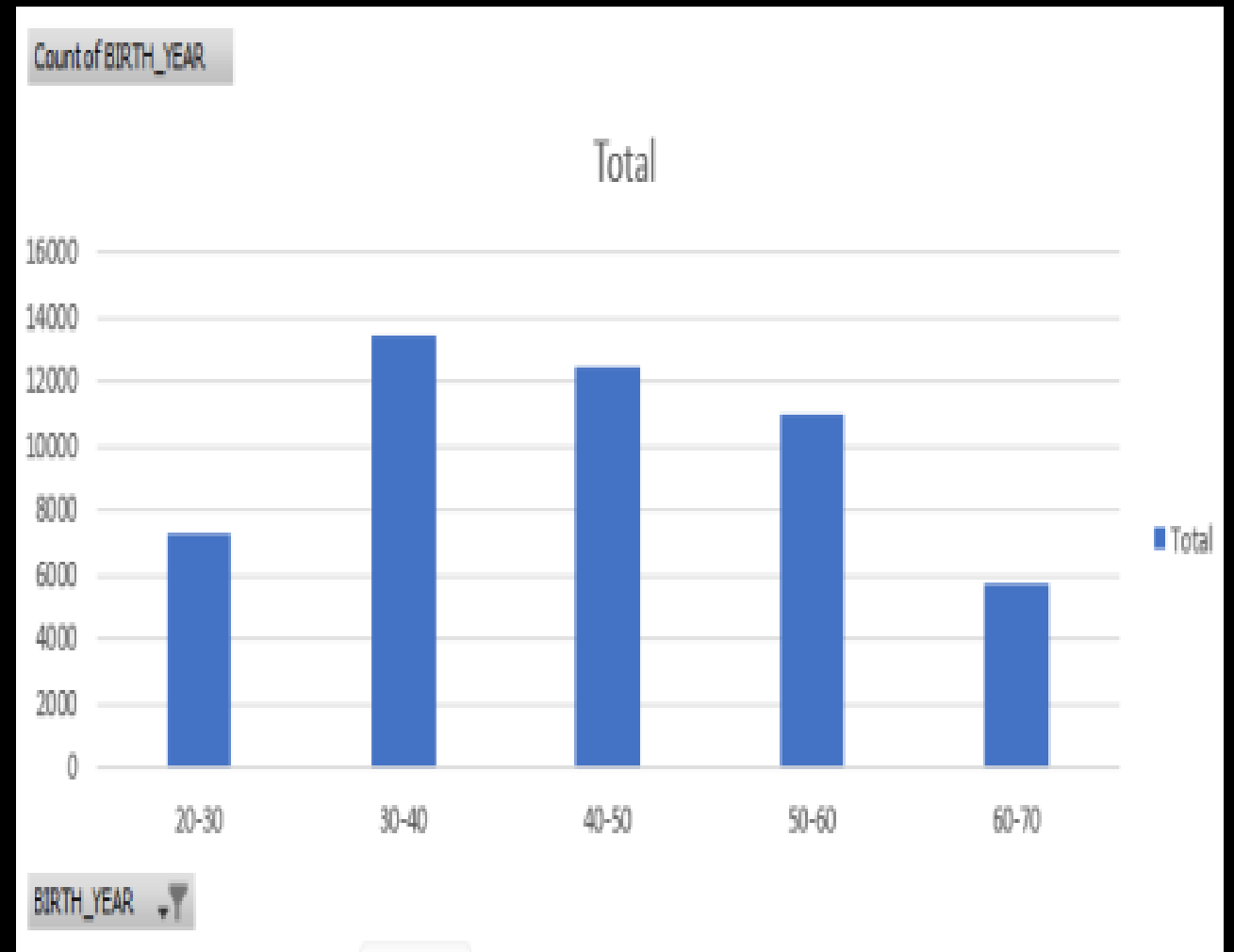
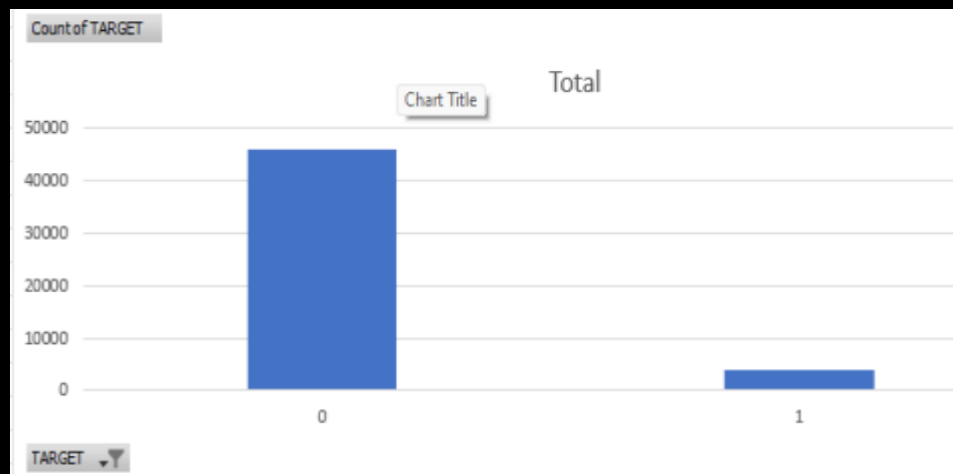
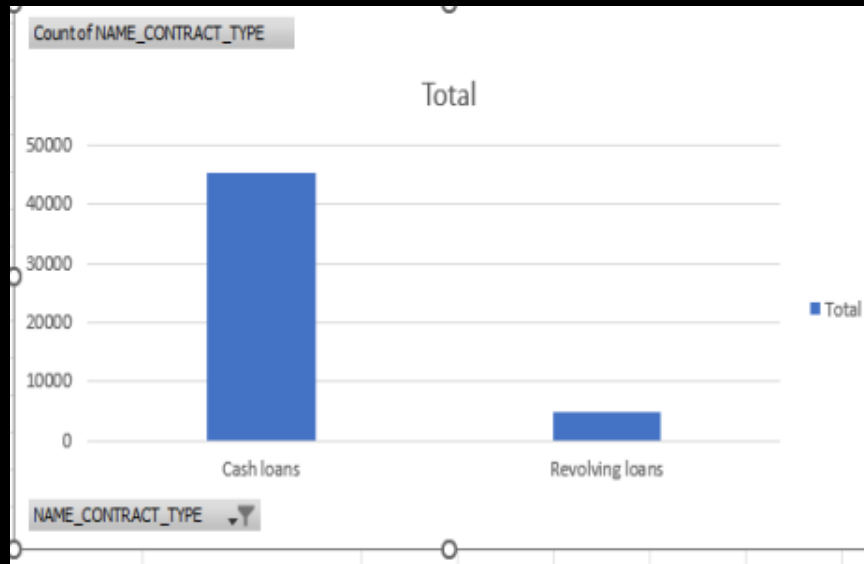
- A. Identify Missing Data and Deal with it Appropriately:



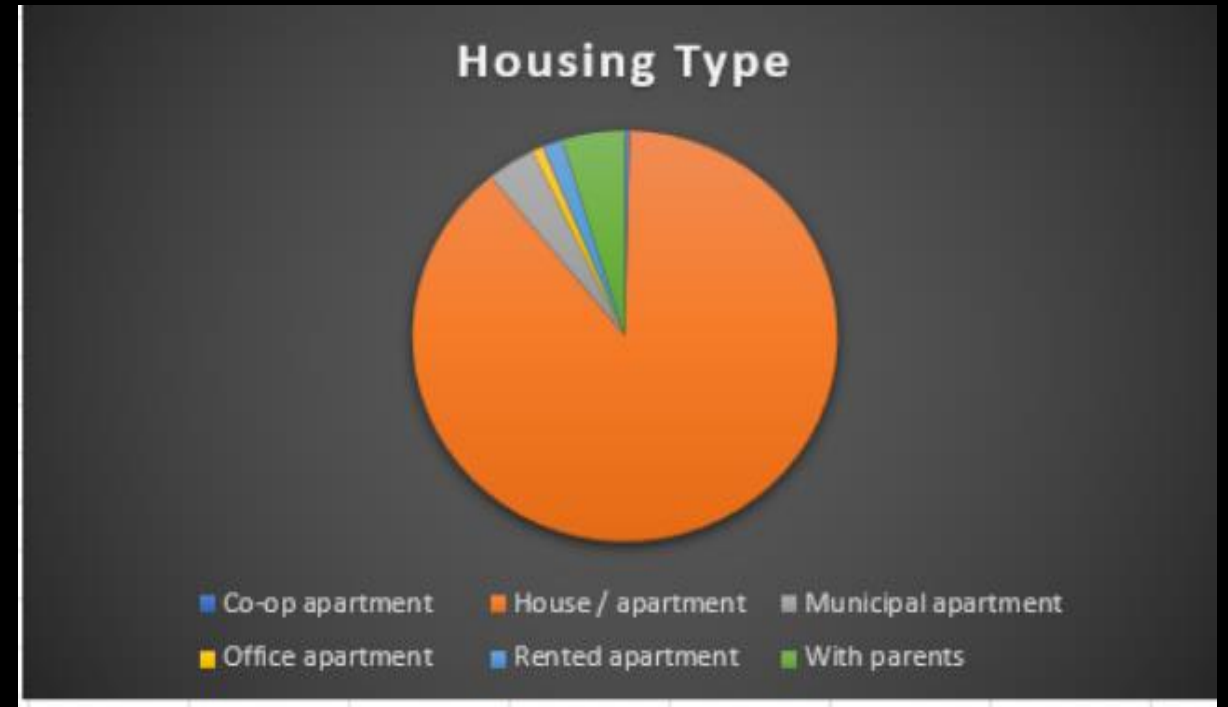
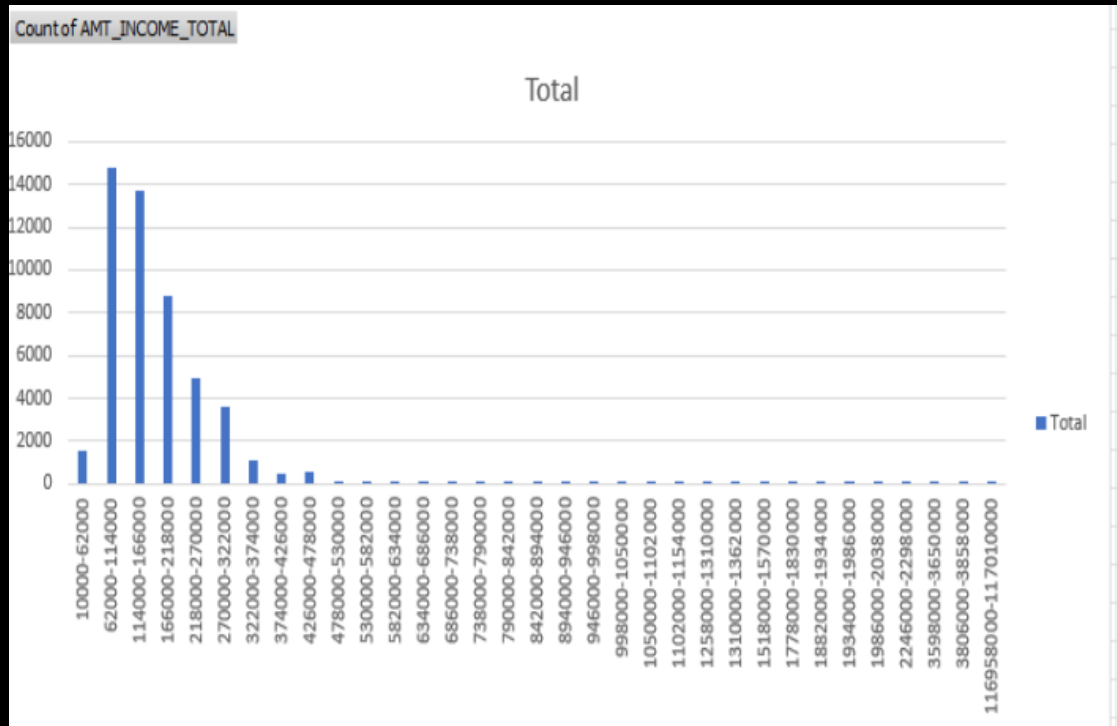
Task B: Identify Outliers in the Datasets.



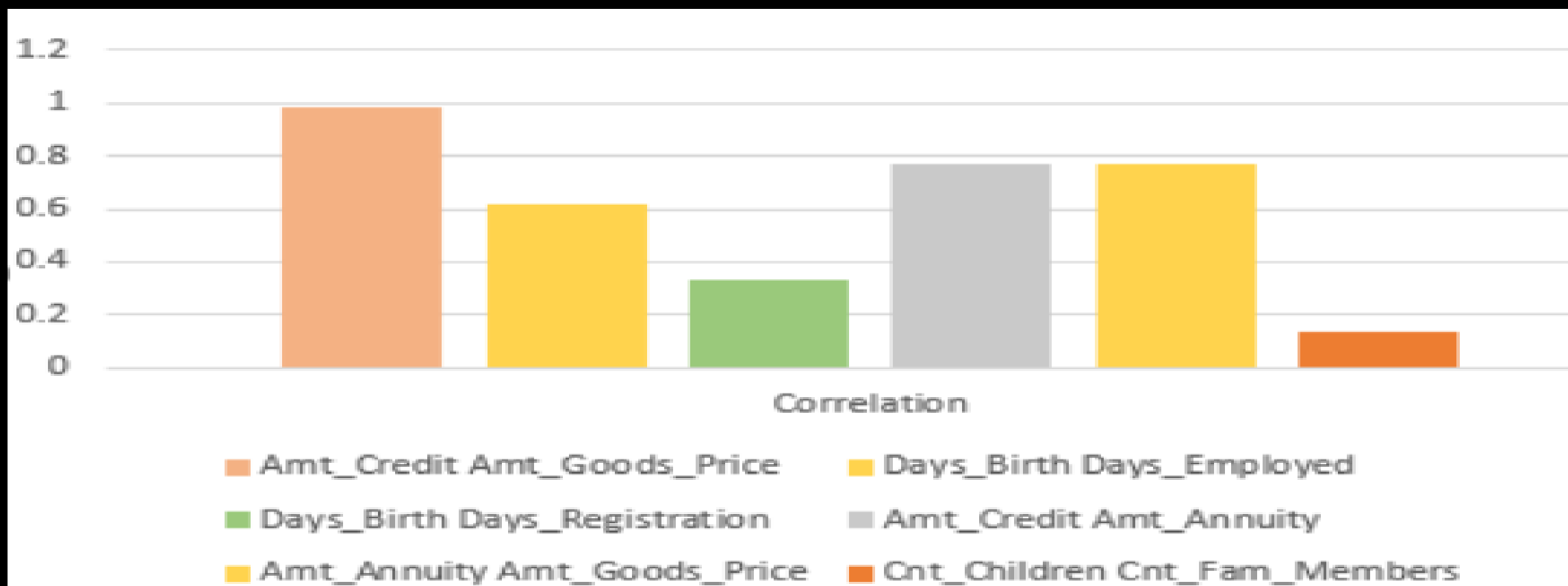
TASK C: Analyze Data Imbalance



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



TASK E: Identify Top Correlations for Different Scenarios



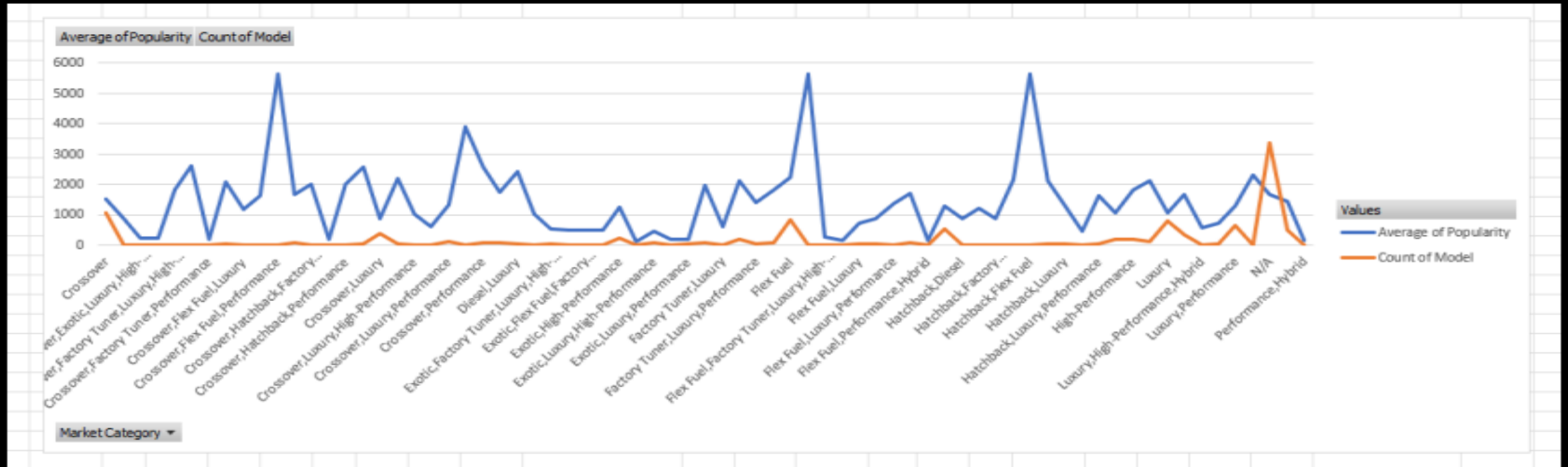
Project 7: Analyzing the Impact of Car Features on Price and Profitability

Project Description

- The automotive industry has been rapidly evolving over the past few decades, 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. In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.
- For the given dataset, as a Data Analyst, the client has asked How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?
- This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer

Task 1: 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.

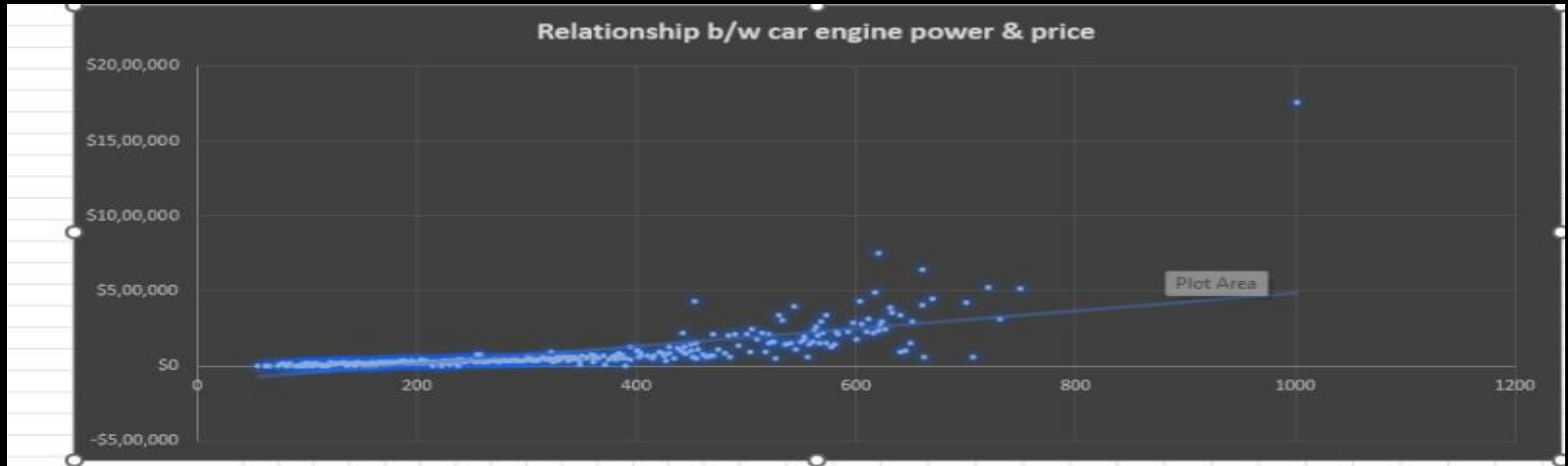


Insights:

From the chart we can clearly observe that crossover, Flex Fuel, Diesel, Hatch Back are most popular in the market and crossover has the highest selling cars.

Task 2: What is the relationship between a car's engine power and its price?

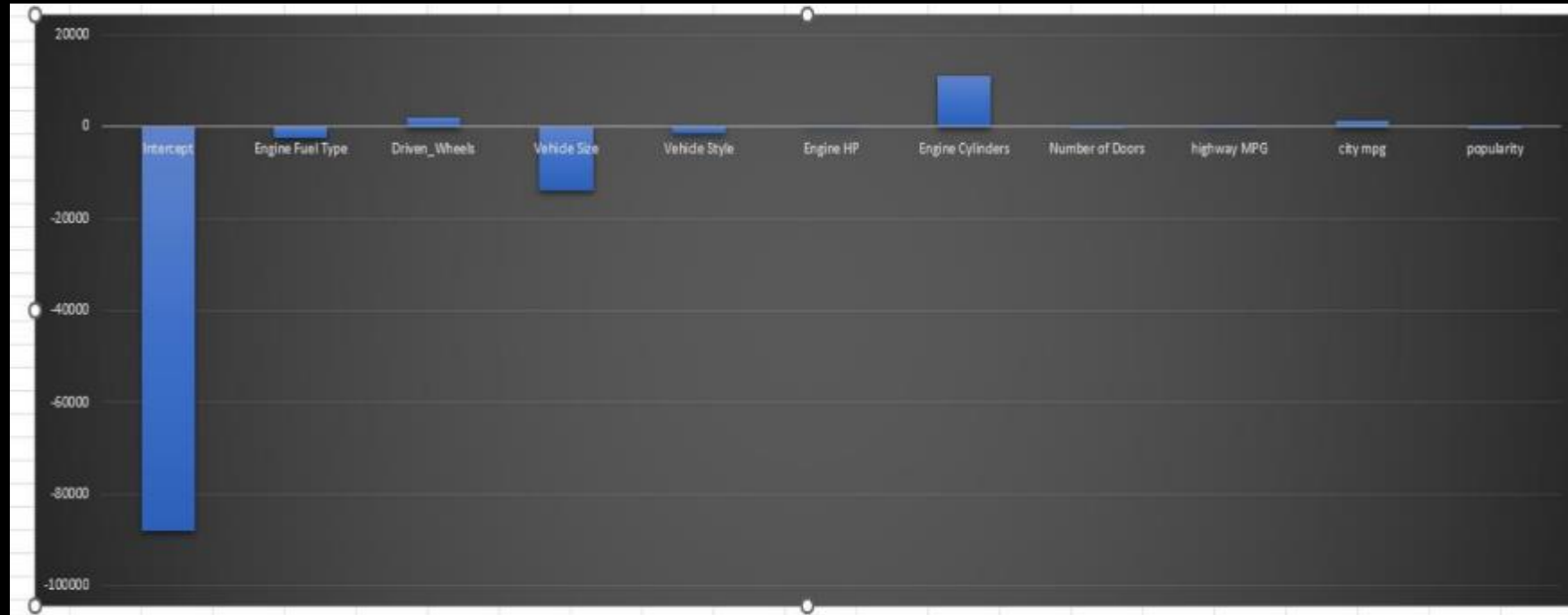
- Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables



- Insights:
- Engine power and MSRP have a positive relation.
- Thus Increasing the price with the increase of car engine power.

Task 3: 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.

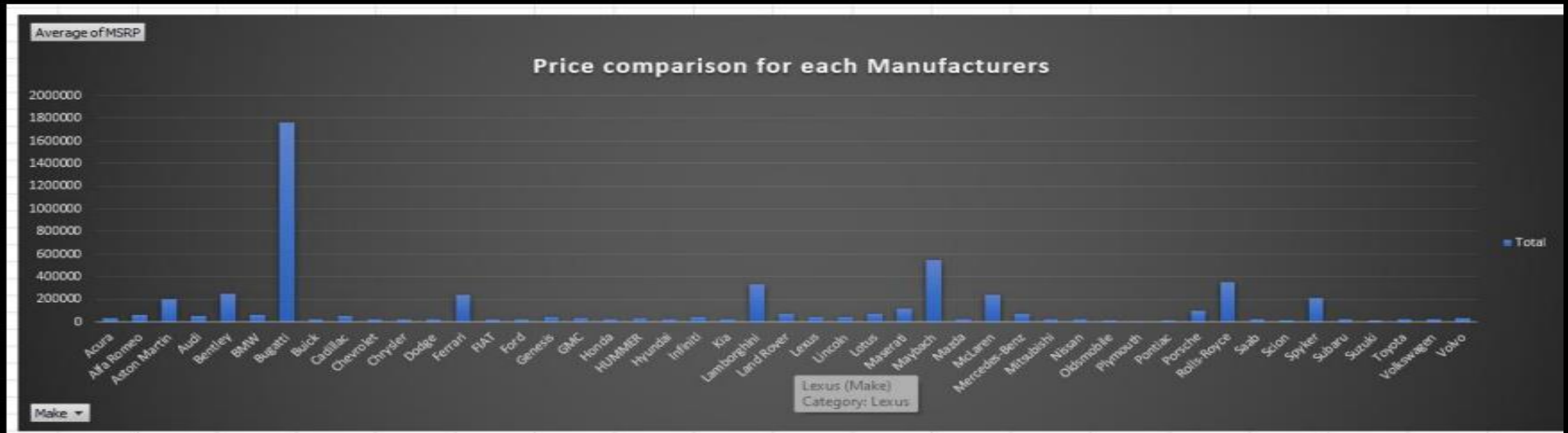


Insights:

Engine cylinder are most important in determining the car price.

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

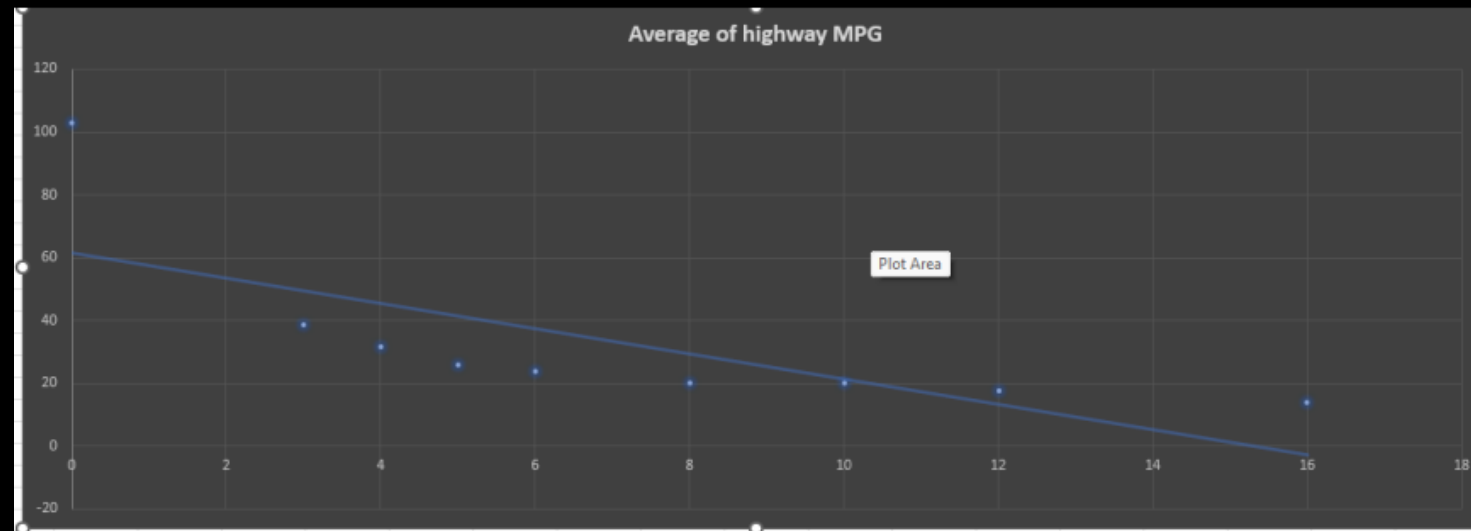


Insights:

From the chart we can observe that Bugatti have the highest average car price.

Task 5: 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.



Insights:

It is quite visible that as the number of cylinders increases the efficiency decreases.

And the correlation coefficient value is -0.614703148

ABC CALL VOLUME TREND ANALYSIS

Project Description

- We'll be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. You'll be provided with a dataset that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).
- A Customer Experience (CX) team plays a crucial role in a company. They analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, among others.

Task 1

- Calculate the average duration of calls received for each time bucket

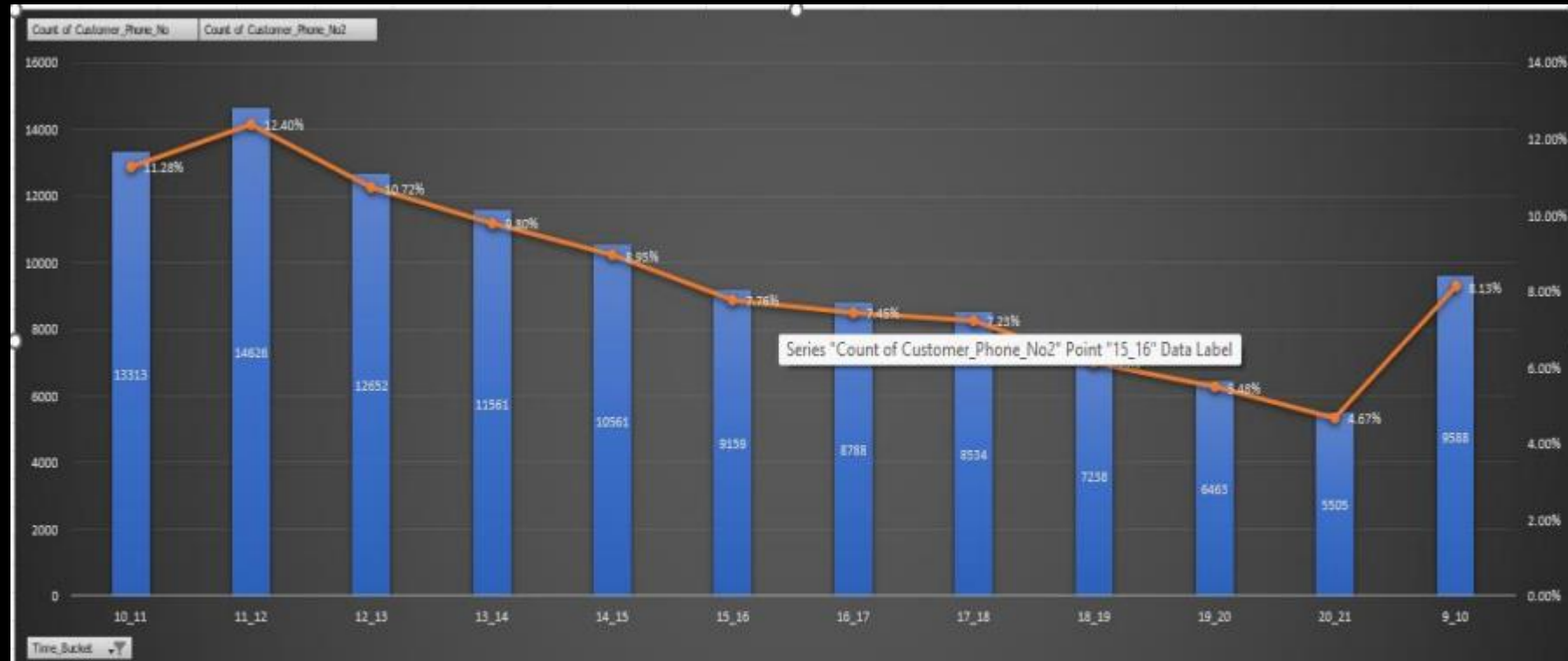


Insights:

The highest average call duration is observed between 4pm to 9 pm and leasr between 12 pm to 2pm

Task 2: Call Volume Analysis:

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

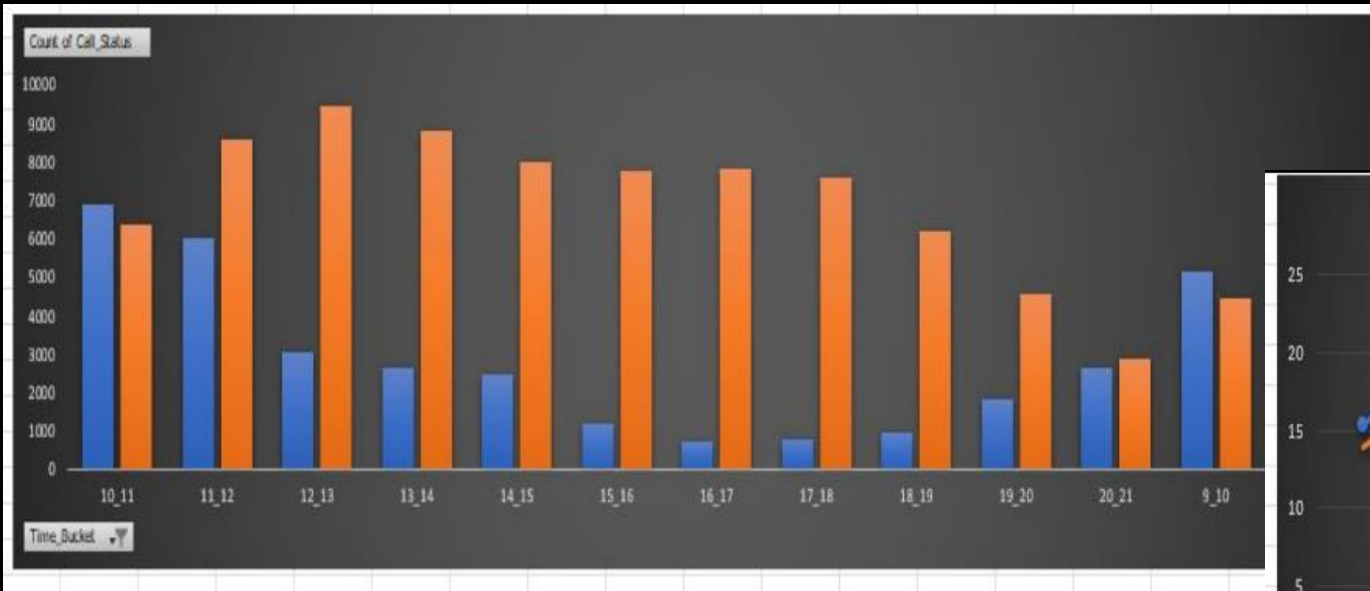


Insights:

From the chart we can observe that between 10 am to 2pm the highest number of calls are received from the customer.

Task 3: Manpower Planning:

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



Insights:

From the chart we can see that from 11 am to 2pm more number of calls are answered.

From the chart we can see that the abandon rate is highest at the beginning and at the end of the day shift.



Insights:

Agent needed are in orange color , Agent working are represented using blue color.

During 8 pm to 9 pm less number of agent were working also requirement was also low.

Task 4: Night Shift Manpower Planning

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



Learnings:

- How to deal with large and complex data and perform the given task using excel, Statistics and Sql and how to extract meaningful insights from raw data , how to transform and clean data
- We have learned how to create interactive dashboards , we have learned and used advanced level SQL and excel for solving the given tasks.

Project Links:

- Project 1 Link:
• <https://drive.google.com/file/d/1ldMWffK7nAjsMh5T39JhXGXWQatN4bts/view?usp=drivesdk>
- Project 2 Link:
• <mailto:https://drive.google.com/file/d/1ggBs9jxCeiM2n7E8IzWHJAP0tGwnPpd0/view?usp=drivesdk>
- Project 3 Link:
• <mailto:https://drive.google.com/file/d/1-Bj9swlWsc3Y5MlnjBnRgV5PksyPsUFc/view?usp=drivesdk>
- Project 4 Link:
• mailto:https://drive.google.com/file/d/1Rq7tyOESo1jfLiRTIJ_6PW_rSz3OY06e/view?usp=drivesdk

Project Links:

- Project 5 Link:
- <mailto:https://drive.google.com/file/d/1S8VQiMJXacSMuE9aFSd84o3d4WHtNnoK/view?usp=drivesdk>
- Project 6 Link:
- <mailto:https://drive.google.com/file/d/1SCvn7LJinlVTBtTGsBzK0d3SQrliH6cn/view?usp=drivesdk>
- Project 7 Link:
- <https://drive.google.com/file/d/1SL9AAA68JNqZUbsSrJhbcH-5-Prm3eE/view?usp=drivesdk>
- Project 8 Link:
- https://drive.google.com/file/d/1SS6nc8fq5kHjADn_lXnScc4LiN8HSmUW/view?usp=drivesdk