DATA ANALYSIS PORTFOLIO

By Ayush Mahanta

Professional Background

Professional Background

Honors & Awards:

- **National Level -** Inline Skating Competitor, KVS National Sports Meet (2017-19)
- **School Captain** - Kendriya Vidyalaya Air Force Station Naliya (Apr 2022)

I am a recent high school graduate from KVS currently pursuing a BECSE at Chandigarh University. My passion for technology and its applications drives my focus on Android app creation using tools like Android Studio and Python Programs. My projects include data analysis and app development. I am eager to apply my theoretical knowledge in real-world scenarios and continuously learn and grow in the tech industry. Additionally, I am interested in expanding my expertise in Cloud Computing, Augmented Reality (AR), And Machine Learning.

Skills:

- Data Analytics
- MySQL
- Python
- Tableau
- PowerBI
- App Development
- Animation
- Ai & Data Science
- Machine Learning

Professional Experience:

Forage

Apr 2024 - May 2024

- Completed *Tata Cybersecurity Security Analyst* Job Simulation.
- Completed *Accenture North America Data Analytics and Visualization* Job Simulation.
- Completed *Quantium Data Analytics* Job Simulation.

Pantech ProEd Pvt Ltd

Apr 2024 - May 2024

- Completed 30-day internship in Data Analytics at *Pantech Prolabs India Pvt Ltd*.

https://www.linkedin.com/in/ayush17mahanta

ayush17mahanta@gmail.com

+91 9685221141

TABLE OF CONTENTS

TABLE OF CONTENTS

01

Data Analytics Process

Something About Our Planning

Yo! So, we are on this crazy adventure exploring India, mouthwatering food, and culture.
India’s like combination of everything like happiness and challenges. And we love challenges!

PLAN

Yo! So, we are on this crazy adventure exploring India, mouthwatering food, and culture.
India has its vast culture beat out major tourist place like South Korea, Egypt, and Australia to be the most visited country in the world.

We will explore its culture, cuisine, bazzars, Spiritual Sites, mode of transportation, etc.

PREPARE

Budget

Making a budget for our road trip is like planning how we will spend our money. It always necessary to make budget for any trip to avoid disappointment and incidents in between trips. We split it into 4 categories: Food, Fuel, Accommodation and Miscellaneous expenses.

Road Trip

In this we have data about road trip like Safety, Tourism Appeal and Weather Preferences. We must need data to have a smooth and pleasant trip. We represent data by graphs.

Destination

By calculating budget and by analysing the data of road trip we put Top 3 Best Destination places infront of our group.
This is contain Top 3 contains high percentage of state safety rating, pleasant weather Preference and Tourism Appeal.

Experience

Past experience is good to enhance upcoming trips and making less mistakes. By learning from past experience we can improve safety measures, expenses, styles and etc. Also we can think what we didn't do in our past trips and what's new should we do in our upcoming trips.

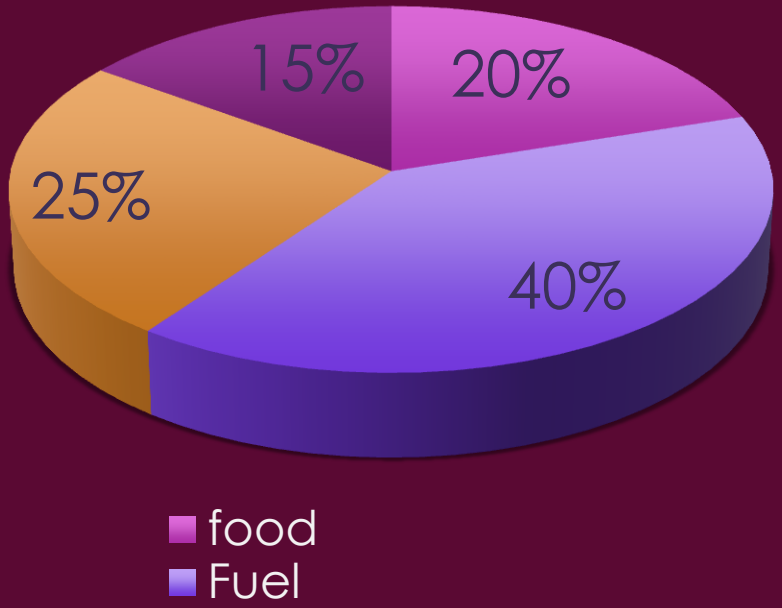
PROCESS

For a perfect trip we have to go through the process which called “WH words”
Like “**WHEN WHERE HOW”**

When: first we have to figure out in which season we will go out for the trip because India has complex seasonal climate, we have to look forward for the weather
Where: In the land of colors we have to find the best destination for our trip
How: Then we have to look around for the mode of transport i.e. train, plan, or car/cab

ANALYZE...

Expenses Category



Food

We allotted 20% expensive for food which necessary in trip. It include water and drinks too.

Fuel

The largest share, 40%, is dedicated to fuel because in fuel is soul of road trip without it we can't get achieve anything.

Accommodation

We Alloted 25% to accommodation guarantees us comfortable stays during the road trip. This category covers hotels, motels, or any overnight arrangements, enhancing the overall travel experience.

Miscellaneous

The remaining 15% is reserved for miscellaneous expenses, including tolls and emergencies. This flexible allocation accommodates unforeseen circumstances and provides a safety net for unexpected road trip events.

CONT....

India Weather Averages

and Climate Information

Months	Temperature (C°)	Temperature (F°)	Rainfall (mm)	Sunshine (hrs)	Tourists (millions)
JAN	15	59	20	10	1.10
FEB	17	63	30	10	1.09
MAR	22	72	20	11	0.97
APR	29	84	40	10	0.77
MAY	33	91	30	8	0.61
JUN	34	93	80	7	0.72
JUL	31	88	170	5	0.81
AUG	30	86	200	4	0.80
SEP	30	86	110	6	0.75
OCT	26	79	10	7	0.90
NOV	21	70	10	9	1.09
DEC	16	61	10	10	1.22

Source : Holidayweather

Northern & Northeast India

Average Daily Temperature from 20-40°C

Southern & Southeast India

Average Daily Temperature : - 32°C

Western India

Average Daily Temperature : 26.4°C

SHARE

WE ARE READY TO TAKE OFF…..

Here we finalize the some of the top destinations and share with our friends and family.

Delhi Agra Jaipur Mumbai Kerala Varanasi

After gathering all the information whether its city or state, which time is suitable or the mode of transportation.

Used some shared photos, information's and stuff to keep everyone on the same page.

ACT….

Now its time to make things happen

- Booking our flights,
- finding accommodations,
- Packed our luggage
- Do all the online booking in advance
- Print all the documents

Get set go……. For **KERALA** because it’s the perfect time to travel the southern part of India as it known for its natural beauty.

Real-life adventurer moves!

CONCLUSION….

- We have decided to go God's Own Country **Kerala** with my family and friends
- We going for 2 weeks with fixed **budget** and make beautiful memories
- **Data Analytics** helps us to make a perfect trip plan
- And it gives as a chance to visit the most beautiful part of the India
- we explore its **stunning backwater, exotic beaches, ayruvedic retreat, delicious cuisine.**

This overall trip will be the best trip for all of us and its ideal destination make this a life time memorable experience.

02

Instagram User Analytics

Instagram User Analytics

Project Description
In this project, as a member of the Instagram product team, I was tasked with analyzing user interactions and engagement to enhance the user experience and support business growth. The analysis focused on answering specific questions posed by the management team, leveraging SQL and MySQL Workbench for data extraction and analysis.

Approach
The approach involved using SQL queries to extract relevant data from the database. Subsequent data analysis aimed to uncover trends, patterns, and insights that could drive business expansion and improve user experiences on Instagram.

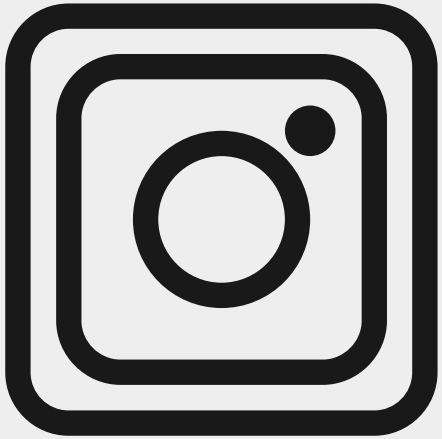
Instagram User Analytics

Tech Stack Used

- MySQL Workbench: Used for database management, data storage, and query execution.
- Microsoft Excel: Utilized for creating visual data representations to simplify the interpretation of insights.

Marketing Analysis

- *Loyal User Reward*: Identified the five longest-tenured users to reward their loyalty.
- *Inactive User Engagement*: Found 26 users who never posted a photo, suggesting targeted promotional emails.
- *Hashtag Research*: Determined the top five most popular hashtags for partner brand usage.
- *Ad Campaign Launch*: Identified Sunday and Thursday as the best days to launch ad campaigns.
- *Contest Winner Declaration*: Found the user with the most likes on a single



Instagram User Analytics

Loyal User Reward
Users with the longest tenure on the platform were identified for loyalty rewards.

Username	Created At
Darby_Herzog	2016-05-06 00:11:21
Emilio_Bernier52	2016-05-06 13:04:30
EJenor88	2016-05-08 01:30:11
Nicole71	2016-05-09 17:30:22
Jordyn.Jacobson2	2016-05-14 07:30:22

Inactive User Engagement
Users who never posted a photo since joining were identified for re-engagement through promotional emails.

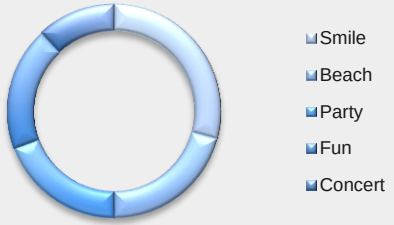
ID	Username	ID	Username
05	aniya_hackett	57	julien_schmid
83	bartholome.bernhard	07	kasandrahomenick
91	bethany20	75	leslie67
80	bardby_herzog	53	linea59
45	david.osinski47	24	maxwell.halvorson
54	duane60	41	mckenna17
90	esmeralda.mraz57	66	mike.auer39
81	esther.zulauf61	49	morgan.kassulke
68	franco_keebler64	71	nia_haag
74	hulda.macejkovic	36	ollie_ledner37
14	jaclyn81	34	pearl7
76	janelle.nikolaus81	21	rocio33
89	jessyca_west	25	tierra.trantow

Instagram User Analytics

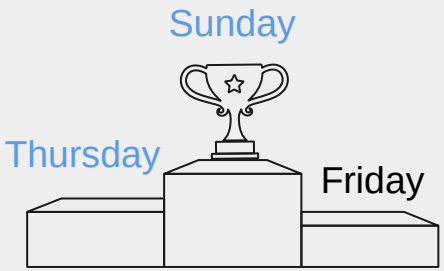
Contest Winner Declaration
The user with the most likes on a single photo was identified.

ID	Username
52	zack_kemmer

Hashtag Research
The top five most commonly used hashtags were identified to enhance partner brand posts.



Ad Campaign Launch
The best days to launch ad campaigns were determined based on user registration patterns, with Sunday and Thursday being the most popular days.



Instagram User Analytics

User Engagement
On average, a user posts about 2.57 times, with a total of 257 photos from 100 users.

Total User

Total Photos

AVG Post Per Users

Bots & Fake Accounts
Identified 13 bots that liked all photos, ensuring data accuracy by removing these accounts.

id	username	likes	id	username	likes
05	aniya_hackett	257	24	maxwell.halvors	257
91	bethany20	257	41	mckenna17	257
54	duane60	257	66	mike.auer39	257
14	jaclyn81	257	71	nia_haag	257
76	janelle.nikolaus	257	36	ollie_ledner37	257
57	julien_schmidt	257	21	rocio33	257
75	leslie67	257			257

Instagram User Analytics

Conclusion

This project allowed me to apply MySQL skills in a real-world context, providing insights into Instagram user behavior and business operations. The removal of bots ensured data accuracy, contributing to the overall success of the analysis. It also enhanced my SQL query writing and execution skills, gained practical experience in data analysis for user behavior insights and improved understanding of business operations and strategic decision-making based on data.

03

Operation Analytics and Investigating Metric Spike

AGENDA

- DESCRIPTION
- APPROACH
- TECH-STACK USED
- ANALYSIS

DESCRIPTION

Operational Analytics is key for improving a company's operations by analyzing data and finding areas for enhancement. As a Data Analyst, you'll collaborate with various teams like operations, support, and marketing to interpret collected data.

You'll focus on investigating sudden changes in metrics, like drops in daily user engagement or sales, by using different datasets. This daily task requires a thorough understanding of the data and how to analyze these spikes effectively.

In your role as Lead Data Analyst at Microsoft, you'll use data to answer questions from different departments, aiming to enhance operations and understand metric changes. This includes syncing data across tools and ensuring everyone has access to the same information for decision-making.

Your approach involves defining metrics, analyzing spikes in metrics like users, events, and email events, and taking actions to improve the company's growth. This project aims to optimize workflows, predict growth, and address any declines in metrics like daily engagement and sales through operational analytics.

APPROACH

Role Overview: Data Analyst Lead at Microsoft

- **Responsibilities:** Perform Operational Analytics, examine entire operations for improvements, collaborate with teams (operations, support, marketing) for insights.
- **Key Focus:** Investigate metric using tables like users', events', and email events'.
- **Approach:**
  - Download all provided data.
  - Create a database with Case Study 1 (Job Data).
  - Upload datasets for Case Study 2 (Investigating metric spikes) to MySQL Workbench.
  - Gather insights, write queries, analyze data, make decisions based on findings.

TECH-STACK USED

MySQL

I used MySQL Workbench to database, store data, and make queries.

Excel

I used Microsoft Excel a lot to create visual representations of the data.

ANALYSIS

Focus on User Engagement for Metric Spike Analysis

- **Key Metric:** User engagement, indicating user satisfaction and product/service quality perception.
- **Case Study Approach:**
  - **Case Study 1 (Job Data):** Created database from provided Excel sheet.
  - **Case Study 2 (Metric Spikes):** Uploaded datasets to MySQL Workbench, used SQL queries to answer questions effectively.
- **Outcome:** Demonstrated SQL's effectiveness in analyzing metric spikes, particularly in user engagement.

Case Study 1

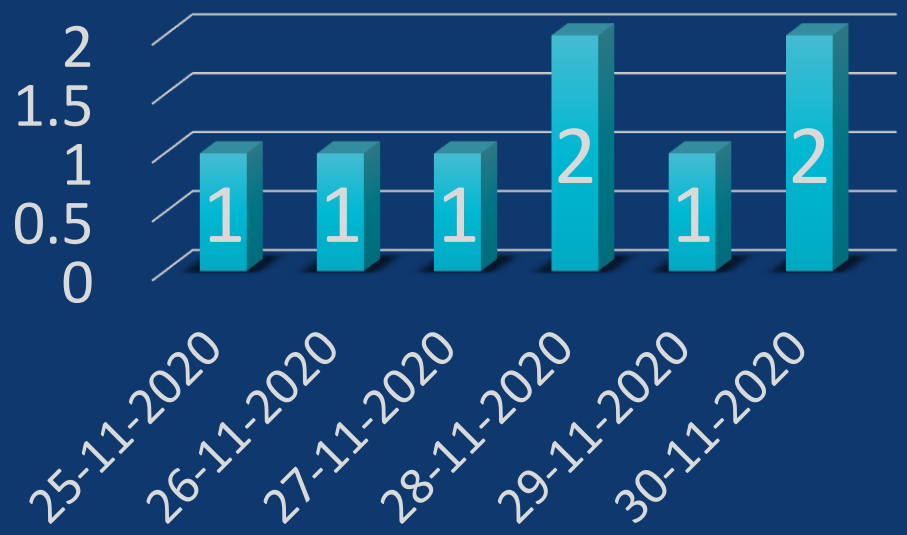
JOB DATA ANALYSIS TOOLS

- Jobs Reviewed Over Time
- Throughput Analysis
- Language Share Analysis
- Duplicate Rows Detection

Jobs Reviewed Over Time

```
SELECT
COUNT(job_id) AS job_reviewed,
DATE(ds) as date
FROM job_data
WHERE DATE(ds) BETWEEN '2020-11-01' AND
'2020-11-30'
GROUP BY date ;
```

job_reviewed	date
2	2020/11/30
1	2020/11/29
2	2020/11/28
1	2020/11/27
1	2020/11/26
1	2020/11/25



This query calculates the number of jobs reviewed per day in November 2020. It counts the number of unique job IDs for each date within the specified date range.

Throughput Analysis

```
SELECT    event,
COUNT(*) AS frequency
FROM    job_data
GROUP BY    event;
```

event	frequency
decision	3
skip	2
transfer	3

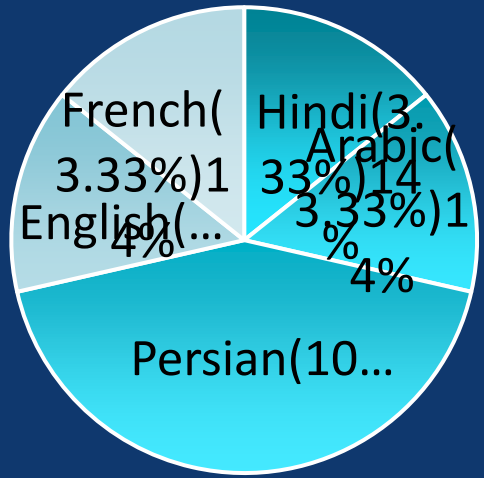
I prefer using the 7-day rolling average for throughput because it smooths out daily fluctuations and provides a more stable metric for analyzing trends over time.

The query returns a result set with columns for date (ds) and avg_throughput, which represents the 7-day rolling average of throughput.

Language Share Analysis

```
SELECT language,
CONCAT(ROUND((COUNT(*) / 30) * 100, 2),
"%") AS percentage_share
FROM    job_data
GROUP BY    language;
```

language	percentage_share
English	3.33%
Arabic	3.33%
Persian	10.00%
Hindi	3.33%
French	3.33%



By Calculating the Persian has highest percentage share (10%) and rest each has percentage share (3.33%) in the last 30 days.

The query returns a result set with columns for language, events_count (the number of events for each language in the last 30 days), and percentage_share (the percentage share of each language in the last 30 days)

Duplicate Rows Detection

```
SELECT * FROM job_data
WHERE (ds, job_id, actor_id, event, language,
time_spent, org)
IN (SELECT ds, job_id, actor_id, event, language,
time_spent,
org FROM job_data
    GROUP BY ds, job_id, actor_id, event,
language, time_spent, org
    HAVING COUNT(*) > 1);
```

Ds	job_id	actor_id	event	language	time_spent	org

No duplicate rows in the data.

The query returns duplicate rows from the job_data table, showing the duplicated records based on the specified columns.

Case Study 2

INVESTIGATING METRIC SPIKE

- Weekly User Engagement
- User Growth Analysis
- Weekly Retention Analysis
- Weekly Engagement Per Device
- Email Engagement Analysis

Weekly User Engagement

SQL Query:

```
SELECT
WEEK(created_at) AS
week_number,
COUNT(DISTINCT
user_id) AS
active_users
FROM users
GROUP BY week_number
ORDER BY week_number;
```

This query calculates the week number based on the created_at date and counts the distinct user_ids for each week.

week_number	active_users	week_number	active_users	week_number	active_users
0	106	**18**	207	**36**	72
1	156	**19**	242	**37**	85
2	157	**20**	215	**38**	90
3	149	**21**	232	**39**	84
4	160	**22**	250	**40**	87
5	181	**23**	246	**41**	73
6	173	**24**	274	**42**	99
7	167	**25**	264	**43**	89
8	163	**26**	257	**44**	96
9	176	**27**	274	**45**	91
10	186	**28**	287	**46**	88
11	161	**29**	288	**47**	102
12	181	**30**	305	**48**	97
13	206	**31**	260	**49**	116
14	197	**32**	316	**50**	124
15	207	**33**	334	**51**	102
16	225	**34**	337	**52**	47
17	219	**35**	81		

Weekly User Engagement

The results show how many users were active each week.

Weekly User Engagement

Interpretation

The results show how many users were active each week.

Insight

We can see if user activity is going up, down, or staying the same over time. This helps us understand how engaging our platform is and if our efforts to attract users are working.

SQL Query:

```
SELECT AVG(active_users) AS
avg_active_usersFROM (SELECT
WEEK(created_at)
AS week_number,COUNT(DISTINCT
user_id) AS active_users
FROM users GROUP BY week_number)
AS weekly_engagement;
```

avg_active_users
177.0000

User Growth Analysis

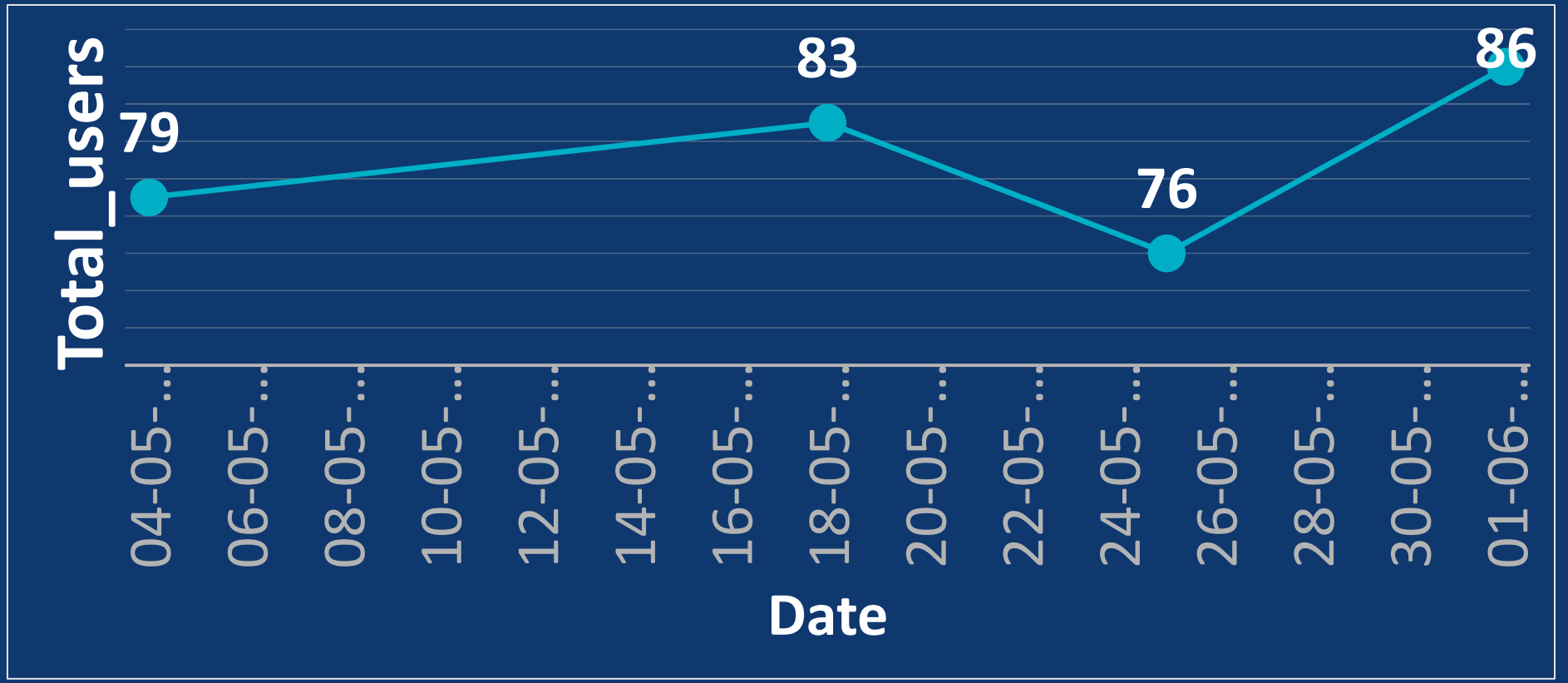
```
SELECT DATE(occurred_at)
AS date,COUNT(DISTINCT user_id)
AS total_users FROM events
GROUP BY DATE(occurred_at)
ORDER BY DATE(occurred_at);
```

This query calculates the total number of distinct users for each date based on the events table.

date	total_users
2014-05-01	293
2014-05-02	358
2014-05-03	145
2014-05-04	79
2014-05-05	257
2014-05-06	310
2014-05-07	323
2014-05-08	312
2014-05-09	352
2014-05-10	135
2014-05-11	92
2014-05-12	259
2014-05-13	313
2014-05-14	300

123 row(s) returned

User Growth Analysis



User Growth Analysis

Interpretation:

- By looking at the results, we can see how the user base is expanding.
- We can identify trends and spikes in user growth, which can help in understanding the impact of marketing campaigns or product updates.
- This analysis can assist in making informed decisions to further accelerate user growth.

Insight

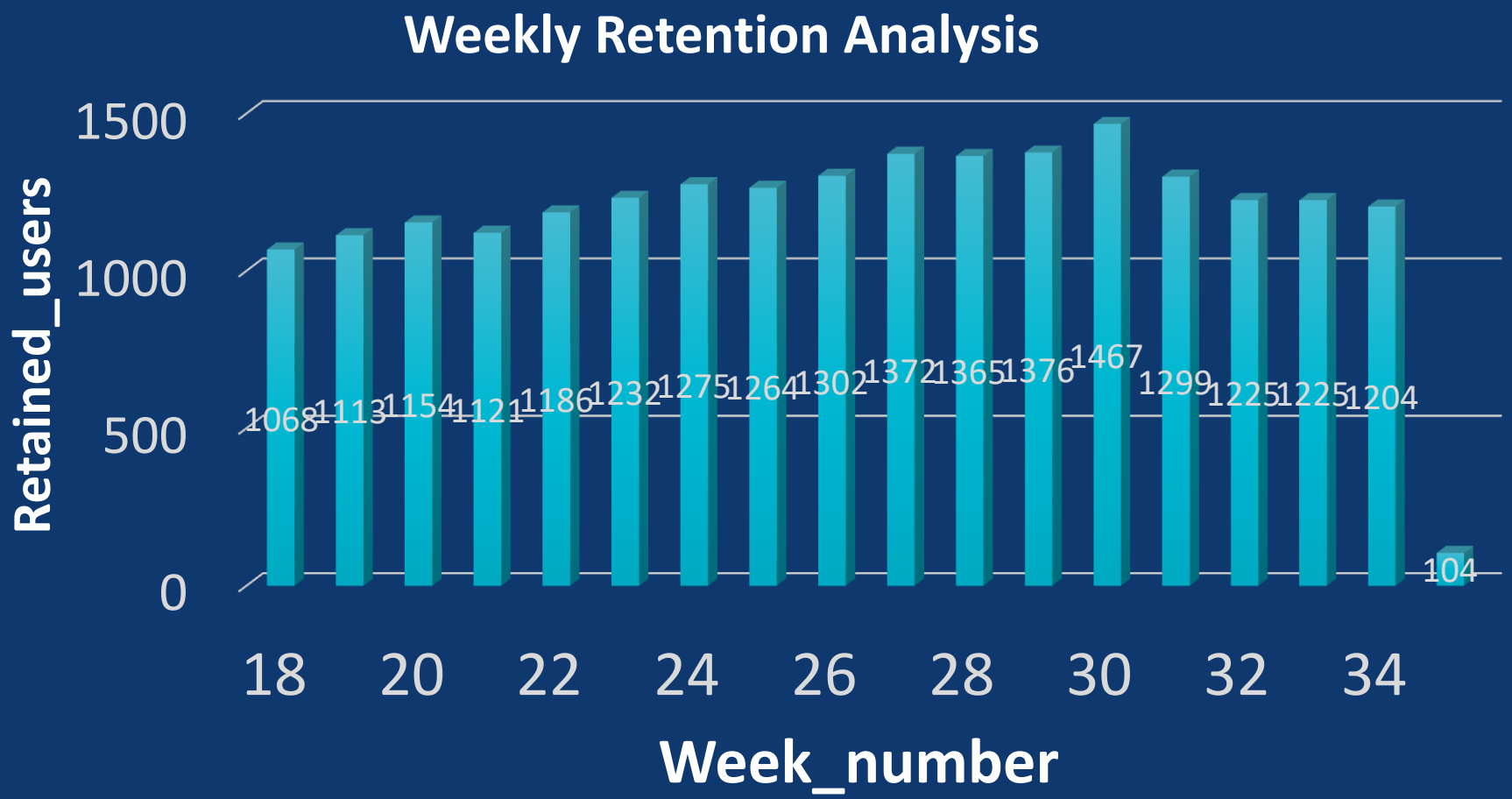
The user count in 2014 shows a gradual increase over time, with the highest number of users recorded on July 18th (455 users) and the lowest on May 25th (76 users). This indicates a growth trend in user numbers, peaking in mid-July and experiencing a gradual decline thereafter.

Weekly Retention Analysis

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

week_number	retained_users	week_number	retained_users
17	663	25	1264
18	1068	26	1302
19	1113	27	1372
20	1154	28	1365
21	1121	29	1376
22	1186	30	1467
23	1232	31	1299
24	1275	32	1225

This query calculates the number of retained users per week based on their sign-up cohort.


Weekly Retention Analysis
Retained_users
1500
1000
500
0
1068
1113
1154
1121
1186
1232
1275
1264
1302
1372
1365
1376
1467
1299
1225
1225
1204
104
18
20
22
24
26
28
30
32
34
Week_number

Weekly Retention Analysis

Interpretation:

The fluctuation in retained users per week suggests varying levels of user engagement or retention strategies. Week 30's peak could indicate successful initiatives or product updates, while the drop in week 17 might indicate a need for improvement in retaining users.

Insight

The number of retained users per week varies, with a peak of 1467 users in week 30 and a low of 663 users in week 17.

Weekly Engagement Per Device

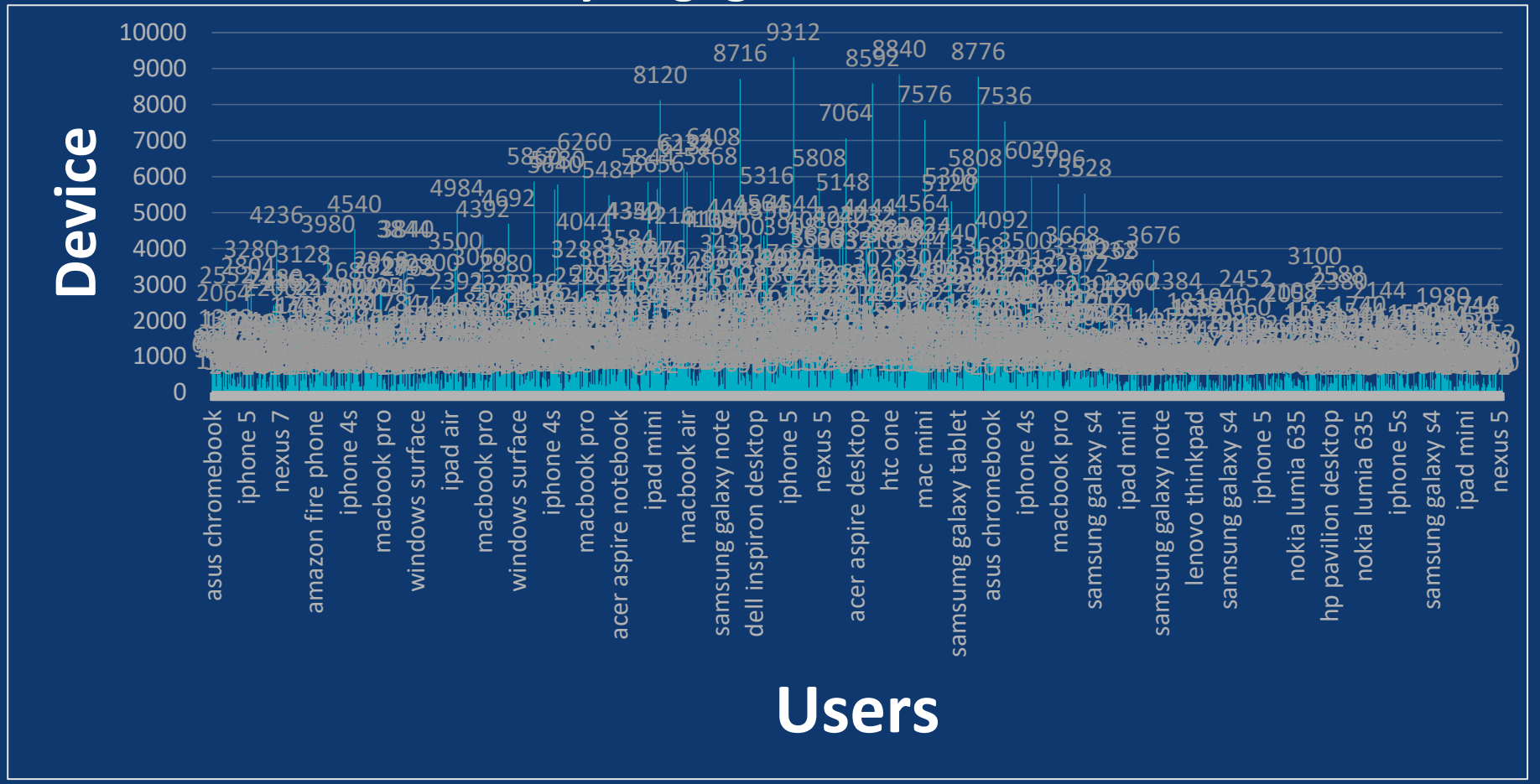
```
select weekofyear(u.created_at)
as weekly , e.device ,
count(u.user_id) as usersfrom
events e
right join users u
on
e.user_id = u.user_idwhere
e.event_type = 'engagement'group
by 1,2order by 1;
```

This query calculates the number of distinct active users per week per device based on the events table.

weekly	device	users
1	acer aspire notebook	220
1	asus chromebook	648
1	dell inspiron desktop	808
1	dell inspiron notebook	704
1	hp pavilion desktop	508
1	htc one	168
1	ipad air	880
1	ipad mini	768
1	iphone 4s	488
1	iphone 5	716
1	iphone 5s	428
1	kindle fire	28
1	lenovo thinkpad	2064
1	mac mini	164

1260 row(s) returned

Weekly Engagement Per Device



Weekly Retention Analysis

Insight

- **Popular Devices**: iPhone 5, MacBook Pro, and iPad Air consistently have high user engagement over the weeks.
- **Varied Engagement:** Devices like Kindle Fire and Nokia Lumia 635 show fluctuating user engagement.
- **Seasonal Trends:** Some devices, like Acer Aspire Notebook, have peaks and dips, possibly due to seasonal factors or product releases.
- **Consistent Growth:** Nexus 5 and Lenovo ThinkPad show a steady increase in user engagement over time.

Interpretation

the data provides insights into user engagement trends with different devices over time. It can help understand user preferences, identify popular devices, and track changes in user behavior. The analysis can inform decision-making regarding product development, marketing strategies, and customer support efforts.

Email Engagement Analysis

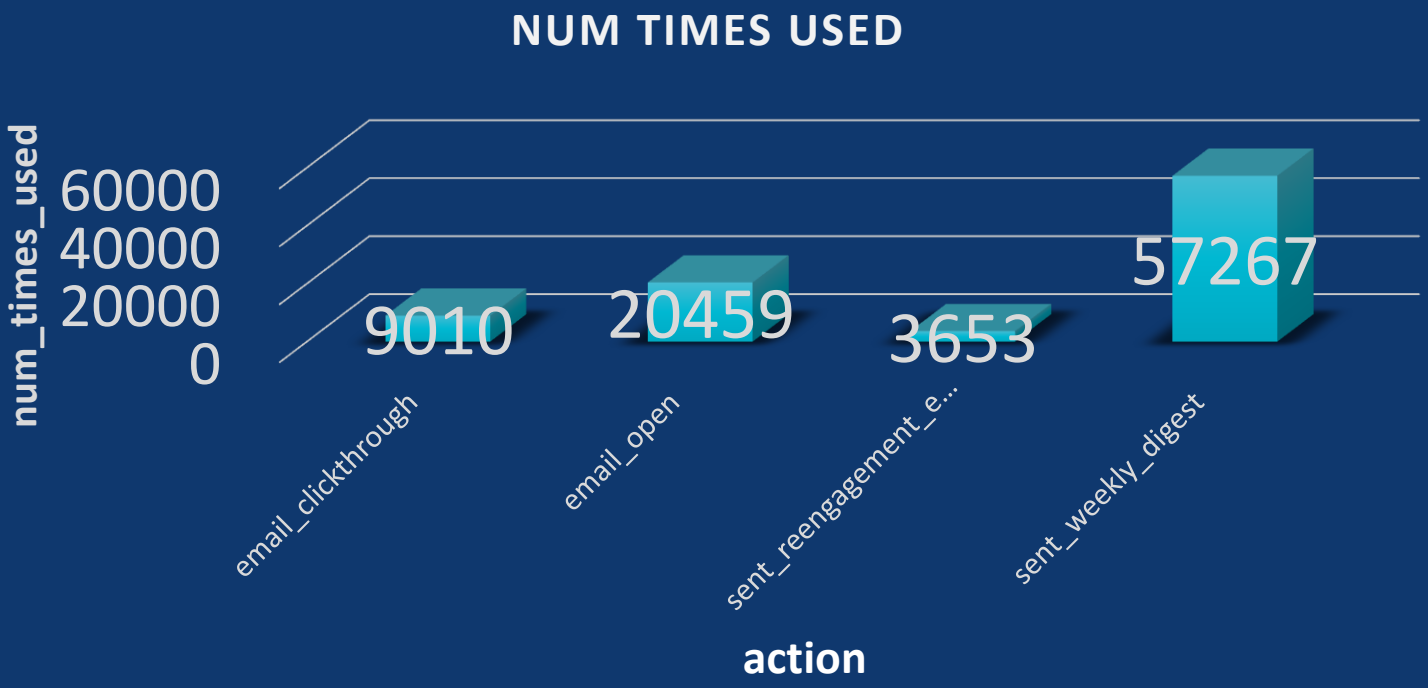
```
SELECT
    u.user_id,
    COUNT(CASE WHEN ee.action = 'sent_weekly_digest' THEN 1 ELSE NULL END) AS
num_sent_weekly_digest,
    COUNT(CASE WHEN ee.action = 'email_open' THEN 1 ELSE NULL END) AS num_email_open,
    COUNT(CASE WHEN ee.action = 'email_clickthrough' THEN 1 ELSE NULL END) AS
num_email_clickthrough,
    COUNT(CASE WHEN ee.action = 'sent_reengagement_email' THEN 1 ELSE NULL END) AS
num_sent_reengagement_email,
    AVG(CASE
            WHEN ee.action = 'sent_weekly_digest' THEN 1
            WHEN ee.action = 'email_open' THEN 1
            WHEN ee.action = 'email_clickthrough' THEN 1
            WHEN ee.action = 'sent_reengagement_email' THEN 1
            ELSE 0
        END) AS avg_email_engagement
FROM
    users u
LEFT JOIN
    email_events ee ON u.user_id = ee.user_id
GROUP BY
    u.user_id;

WITH q1 AS (
    SELECT
        action,
        TIMESTAMPDIFF(WEEK, '2013-01-01 04:40:10', occurred_at) AS wk,
        COUNT(user_id) AS Cnt
    FROM
        email_events
    GROUP BY
        action,
        wk
)
SELECT
    q1.action,
    COUNT(ee.user_id) AS num_times_used,
    ROUND(AVG(q1.Cnt), 2) AS avg_email_engagement
FROM
    q1
LEFT JOIN
    email_events ee ON q1.action = ee.action AND TIMESTAMPDIFF(WEEK, '2013-01-01 04:40:10',
ee.occurred_at) = q1.wk
GROUP BY
    q1.action
ORDER BY
    avg_email_engagement DESC;
```

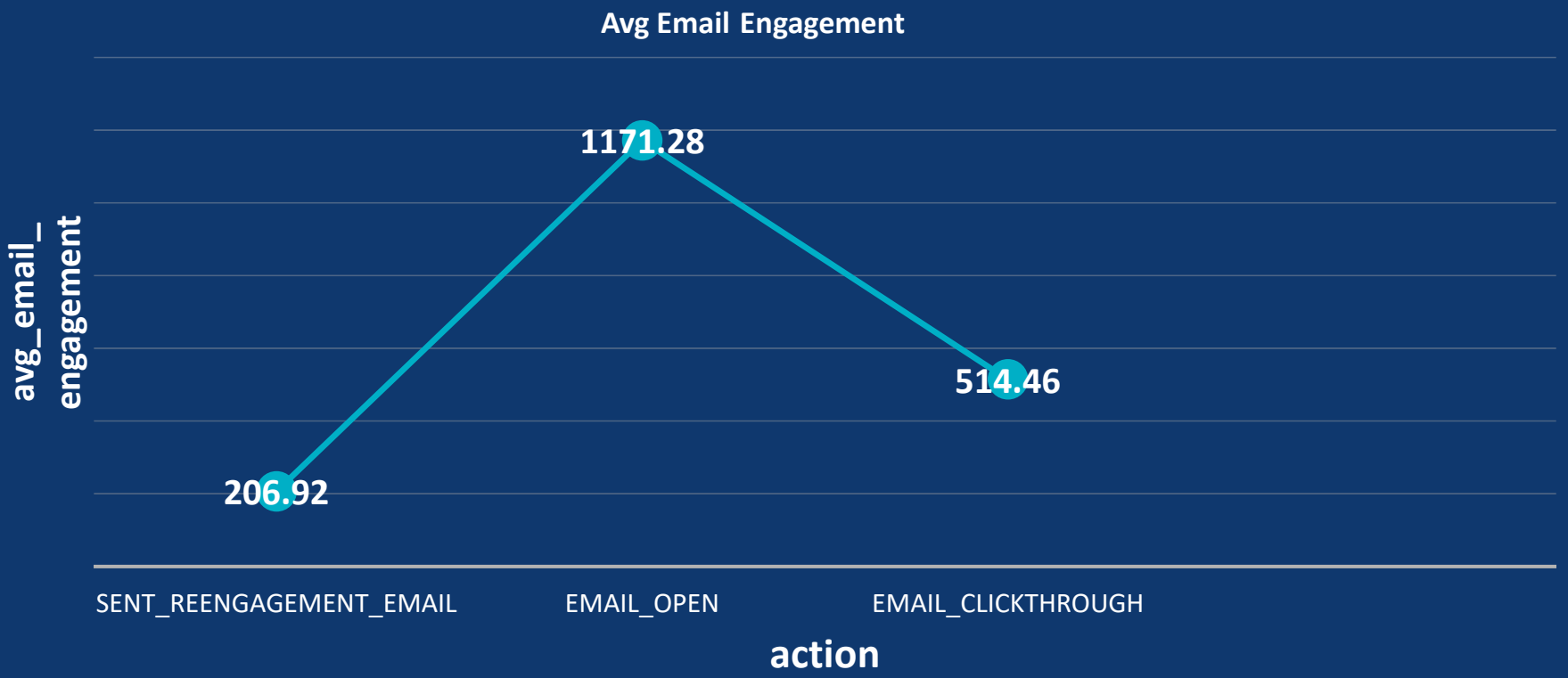
action	num_times_used	avg_email_engagement
email_clickthrough	9010	514.46
email_open	20459	1171.28
sent_reengagement_email	3653	206.92
sent_weekly_digest	57267	3281.11

This query calculates the average number of email engagements per week for each action based on the events in the email_events table.

Email Engagement Analysis



Email Engagement Analysis



Thank You

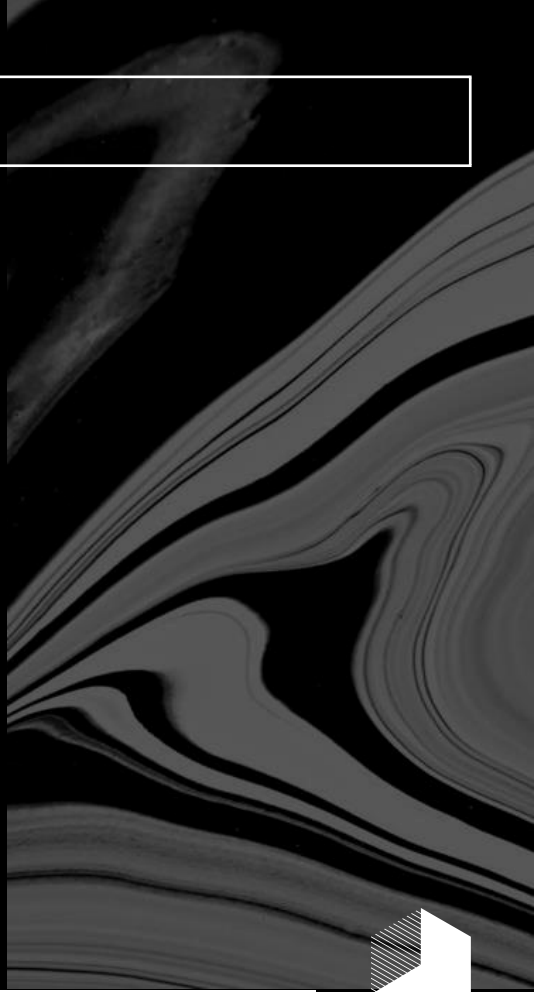
By Ayush Mahanta

04

Hiring Process Analytics

AGENDA

- Description
- Approach
- Tech - Stack Used
- Insights
- Result



PROJECT DESCRIPTION

The project involves analyzing a multinational company's hiring process data to gain insights and improve the hiring process. Key tasks include handling missing data, clubbing columns, detecting and handling outliers, and summarizing findings using statistical measures and visualizations.

APPROACH

Data Cleaning

Check for missing values and decide on a strategy to handle them. Combine columns with multiple categories if possible to simplify analysis.

Outlier Detection

Identify outliers and decide whether to remove, replace, or leave them as is.

Data Summary

Calculate averages, medians, and other statistical measures. Create visualizations to understand the data better.

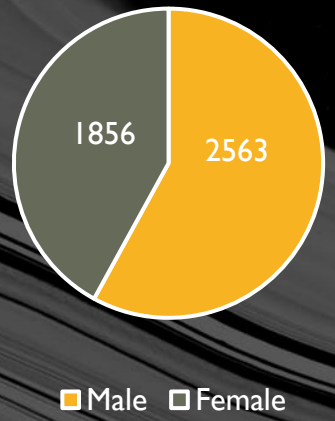
TECH-STACK USED

A. HIRING ANALYSIS

=COUNTIFS(D:D,“Male",C:C,"Hired")	=COUNTIFS(D:D,"Female",C:C,"Hired")
Total no. of Male hired 2563	Total no. of Female hired1856

INSIGHT: THIS WILL GIVE YOU THE COUNT OF MALES AND FEMALES HIRED. THE COMPANY HIRED MORE MALES(2563) COMPARED TO FEMALE EMPLOYEES(1856).

Male : Female proportion of hired employees



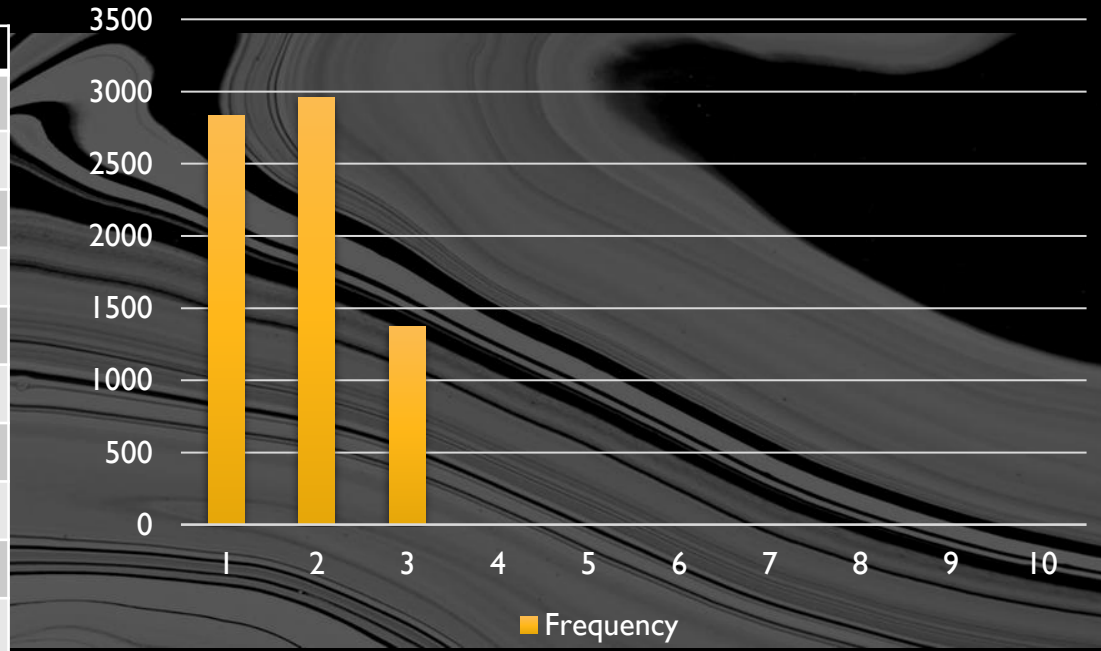
B. SALARY ANALYSIS

=AVERAGE(G:G)	49976.05594

INSIGHT: THE AVERAGE SALARY OFFERED IN THE COMPANY IS **49,976.06**

C. SALARY DISTRIBUTION

S.No.	Class Intervals	Frequency
1.	0 - 40,000	2831
2.	40,001 - 80,000	2963
3.	80,001 - 120,000	1370
4.	120,001 - 160,000	0
5.	160,001 - 200,000	1
6.	200,001 - 240,000	0
7.	240,001 - 280,000	0
8.	280,001 - 320,000	1
9.	320,001 - 360,000	0
10.	360,001 - 400,000	2

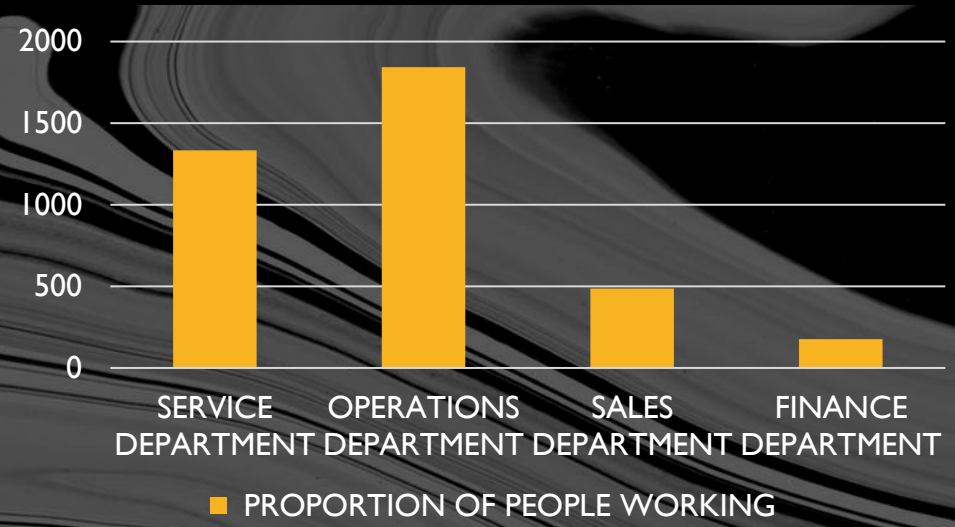


=COUNTIFS(G:G, ">="&0, G:G, "<="&40000)	2831

Insight: The majority of employees fall within the salary range 40,000-80,000.

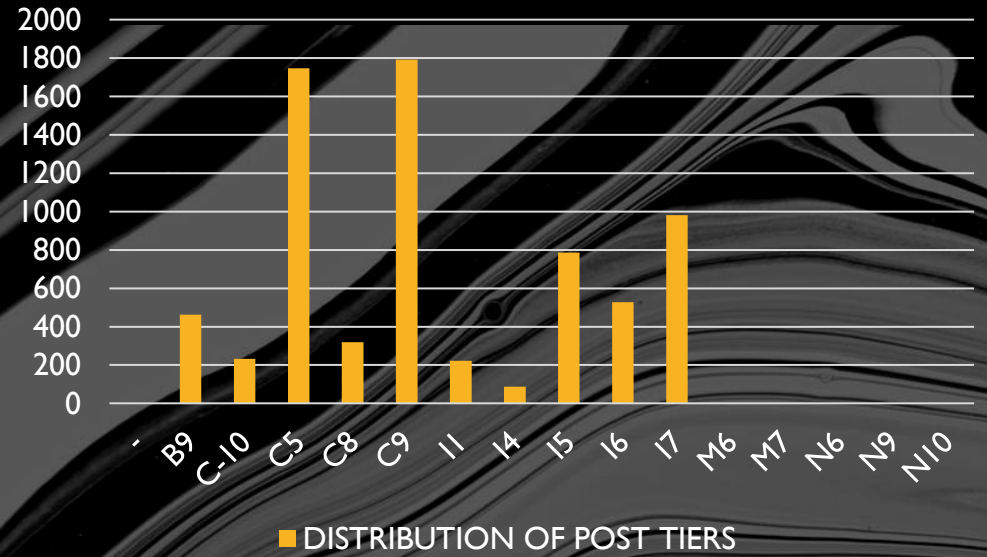
D. DEPARTMENTAL ANALYSIS

DEPARTMENTS	PROPORTION OF PEOPLE WORKING
SERVICE DEPARTMENT	1332
OPERATIONS DEPARTMENT	1843
SALES DEPARTMENT	485
FINANCE DEPARTMENT	176
PURCHASE DEPARTMENT	230
PRODUCTION DEPARTMENT	246
MARKETING DEPARTMENT	202
HUMAN RESOURCE DEPARTMENT	70
=COUNTIFS(E:E,"Service Department",C:C,"Hired"	1332



Insight: The majority of employees work in the 'Operation Department', followed by the 'Service Department'.

E. POSITION TIER ANALYSIS



=COUNTIFS(F:F,"b9")	463

Insight: The highest proportion of employees are in the C9 post tier, followed by C5 and I7.

POST NAME	DISTRIBUTION OF POST TIERS
-	1
B9	463
C-10	232
C5	1747
C8	320
C9	1792
I1	222
I4	88
I5	787
I6	527
I7	982
M6	3
M7	1
N6	1
N9	1
N10	1

CONCLUSION

- This project showed me how valuable data analytics is for a company's hiring process.
- It helps to understand things like how many people were rejected, why they were rejected, who is applying for jobs, and how many positions are open.
- This information helps the hiring team make better decisions based on data.

EXCEL LINK:

https://docs.google.com/spreadsheets/d/11I2BQnp-gdEVmIMPtDiSNQaSEYLPIb8p/edit?usp=drive_link&ouid=108396890359637253084&rtpof=true&sd=true

05

IMDB Movie Analysis

INTRODUCTION

In this project, I analyzed a dataset of movies from IMDB. I looked at different aspects of the movies like genres, duration, language, directors, and budget to see how they relate to the movie's IMDB rating and financial success.

Approach

DATA CLEANING

In this step, I preprocess the data to make it ready for analysis. This involves fixing any missing values, removing duplicate entries, changing data types if needed, and possibly creating new features.

DATA ANALYSIS

Here, I explore the data to understand how different variables are related. I look at things like how movie ratings are influenced by factors such as genre, director, and budget. I also consider the year the movie was released, the actors in it, and other relevant factors.

Report Generation

Finally, I created a report that summarizes my findings. This report includes what I learned from the data analysis and is supported by visualizations and statistics.

Tech-Stack Used

Insight

TASK

Movie Genre Analysis

Movie Duration Analysis

Language Analysis

Director Analysis

Budget Analysis

Movie Genre Analysis

This analysis provides valuable insights for filmmakers and producers regarding audience preferences and industry trends

- **Genre Popularity:** Comedy is the most represented genre, while Family, Musical, Romance, Thriller, and Western are less common.
- **Rating Trends:** Documentaries tend to have the highest average ratings, while Thrillers have the lowest.
- **Score Variability:** Westerns have the most varied ratings, while Biographies are more consistent.
- **Quality and Consistency:** Biographies consistently score well, while Comedies have a more mixed reception.
- **Genre Appeal:** Action, Adventure, Drama, and Crime genres have broad appeal, while Family, Musical, Romance, and Thriller genres are more niche..

Analysis

Row Labels	Count of imdb_score	Average of imdb_score	Max of imdb_score	Min of imdb_score	Var of imdb_score	StdDev of imdb_score
Action	935	6.285989305	9	2.1	1.078186788	1.038357736
Adventure	367	6.561307902	8.6	2.3	1.264345826	1.124431334
Animation	46	6.763043478	8	4.5	0.945937198	0.972593028
cc	206	7.151941748	8.9	4.5	0.489337675	0.69952675
Comedy	1026	6.164424951	8.8	1.9	1.074567328	1.036613393
Crime	252	6.945238095	9.3	3.3	0.75475811	0.868768157
Documentary	43	6.951162791	8.5	1.6	2.005415282	1.41612686
Drama	676	6.821745562	8.8	2.1	0.82072643	0.905939529
Family	3	6.5	7.9	5.7	1.48	1.216552506
Fantasy	35	6.234285714	7.9	4.3	0.799378151	0.894079499
Horror	156	5.813461538	8.5	2.3	1.015624069	1.007781757
Musical	2	6.75	7.2	6.3	0.405	0.636396103
Mystery	23	6.586956522	8.5	3.3	1.23027668	1.109178381
Romance	2	6.65	7.1	6.2	0.405	0.636396103
Sci-Fi	8	6.5875	8.2	5	1.064107143	1.031555691
Thriller	3	5.3	6.3	4.8	0.75	0.866025404
Western	3	6.766666667	8.9	4.1	5.973333333	2.444040371
(blank)						
Grand Total	**3786**	**6.462466984**	**9.3**	**1.6**	**1.118324058**	**1.05750842**

Top 5 Genres

Action

I found the most common genres in the dataset and studied how they affect IMDB scores. I discovered that Action, Biography, Crime, Comedy, and Drama are the most common genres.

Movie Duration Analysis

1.Movie Duration:

- **Mean Duration**: The average duration of movies is approximately 109.81 minutes.
- **Variability**: The duration varies widely, with a range from 34 to 330 minutes, and a standard deviation of 22.76, indicating significant variability in movie lengths.
- **Distribution**: The distribution of movie durations is positively skewed (skewness = 2.35), meaning that there are more shorter-duration movies than longer-duration ones. The kurtosis value of 12.41 indicates a very peaked distribution.

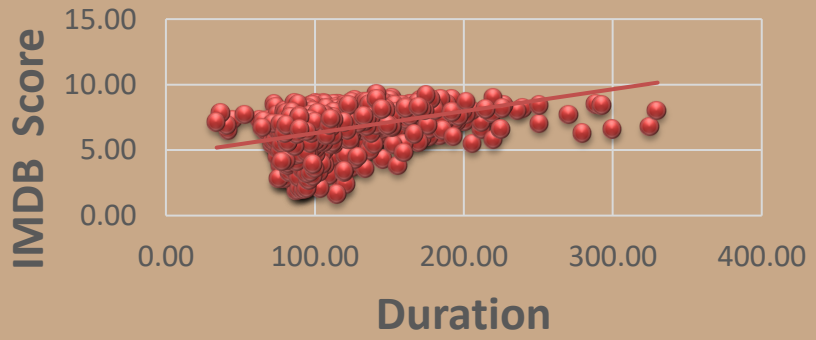
2.IMDb Scores:

- **Average IMDb Score**: The average IMDb score is 6.46 out of 10.
- **Variability**: The IMDb scores range from 1.6 to 9.3, with a standard deviation of 1.06, indicating some variability in the ratings.
- **Distribution**: The distribution of IMDb scores is negatively skewed (skewness = -0.73), meaning that there are more higher-rated movies than lower-rated ones. The kurtosis value of 1.13 indicates a moderately peaked distribution.

Analysis

DURATION		IMDB_SCORE	
Mean	109.808505	Mean	6.462466984
Standard Error	0.369949997	Standard Error	0.017186741
Median	105	Median	6.6
Mode	101	Mode	6.7
Standard Deviation	22.763201	Standard Deviation	1.05750842
Sample Variance	518.16332	Sample Variance	1.118324058
Kurtosis	12.40512587	Kurtosis	1.133330499
Skewness	2.347508256	Skewness	-0.725634682
Range	296	Range	7.7
Minimum	34	Minimum	1.6
Maximum	330	Maximum	9.3
Sum	415735	Sum	24466.9
Count	3786	Count	3786
Largest(1)	330	Largest(1)	9.3
Smallest(1)	34	Smallest(1)	1.6
Confidence Level(95.0%)	0.725320612	Confidence Level(95.0%)	0.033696168

Duration Vs IMDB score



Language Analysis

1.Language Impact on IMDb Scores:

- **Highest Average IMDb Scores**: Languages with the highest average IMDb scores include Portuguese (7.76), German (7.69), and Japanese (7.63), indicating generally well-received movies in these languages.
- **Lowest Average IMDb Scores**: Bosnian (4.3) and Kazakh (6.0) have the lowest average IMDb scores, suggesting lower reception or fewer movies in these languages.

2.Variability in IMDb Scores:

- **Most Consistent Scores**: Arabic, Aramaic, Dzongkha, and Zulu have only one movie each, but with identical IMDb scores, indicating consistent reception for movies in these languages.
- **Most Varied Scores**: Hindi movies have a wide range of IMDb scores, from 4.8 to 8.0, suggesting varied reception among viewers.

3.Impact of Sample Size:

- Languages with larger sample sizes, such as English (3606 movies), may have more stable average IMDb scores compared to languages with smaller sample sizes, like Icelandic (1 movie).

4.Overall Trends:

- The average IMDb score across all languages is 6.46, with a standard deviation of 1.06, indicating a moderate level of variability in movie ratings across languages.
- English movies, with the largest sample size, have an average IMDb score of 6.42, suggesting a relatively consistent reception for English-language movies.

Row Labels	Count of Imdb_Score	Average of Imdb_Score	Max of Imdb_Score	Min of Imdb_Score	Varp of Imdb_Score	StdDevp of Imdb_Score
Aboriginal	2	6.95	7.5	6.4	0.3025	0.55
Arabic	1	7.2	7.2	7.2	0	0
Aramaic	1	7.1	7.1	7.1	0	0
Bosnian	1	4.3	4.3	4.3	0	0
Cantonese	8	7.2375	7.8	6.5	0.16984375	0.412121038
Czech	1	7.4	7.4	7.4	0	0
Danish	3	7.9	8.3	7.3	0.186666667	0.43204938
Dari	2	7.5	7.6	7.4	0.01	0.1
Dutch	3	7.566666667	7.8	7.1	0.108888889	0.329983165
Dzongkha	1	7.5	7.5	7.5	0	0
English	3606	6.421436495	9.3	1.6	1.107446744	1.052352956
Filipino	1	6.7	6.7	6.7	0	0
French	37	7.286486486	8.4	5.8	0.306574142	0.553691378
German	13	7.692307692	8.5	6.1	0.379171598	0.615769111
Hebrew	3	7.5	8	7.2	0.126666667	0.355902608
Hindi	10	6.76	8	4.8	1.1124	1.05470375
Hungarian	1	7.1	7.1	7.1	0	0
Icelandic	1	6.9	6.9	6.9	0	0
Indonesian	2	7.9	8.2	7.6	0.09	0.3
Italian	7	7.185714286	8.9	5.3	1.144081633	1.069617517
Japanese	12	7.625	8.7	6	0.741875	0.861321659
Kazakh	1	6	6	6	0	0
Korean	5	7.7	8.4	7	0.26	0.509901951
Mandarin	14	7.021428571	7.9	5.6	0.544540816	0.737930089
Maya	1	7.8	7.8	7.8	0	0
Mongolian	1	7.3	7.3	7.3	0	0
None	1	8.5	8.5	8.5	0	0
Norwegian	4	7.15	7.6	6.4	0.2475	0.497493719
Persian	3	8.133333333	8.5	7.5	0.202222222	0.449691252
Portuguese	5	7.76	8.7	6.1	0.7664	0.875442745
Romanian	1	7.9	7.9	7.9	0	0
Russian	1	6.5	6.5	6.5	0	0
Spanish	26	7.05	8.2	5.2	0.656346154	0.810151933
Swedish	1	7.6	7.6	7.6	0	0
Telugu	1	8.4	8.4	8.4	0	0
Thai	3	6.633333333	7.1	6.2	0.135555556	0.368178701
Vietnamese	1	7.4	7.4	7.4	0	0
Zulu	1	7.3	7.3	7.3	0	0
(blank)						
Grand Total	3786	6.462466984	9.3	1.6	1.118028674	1.05736875

Analysis

Director Analysis

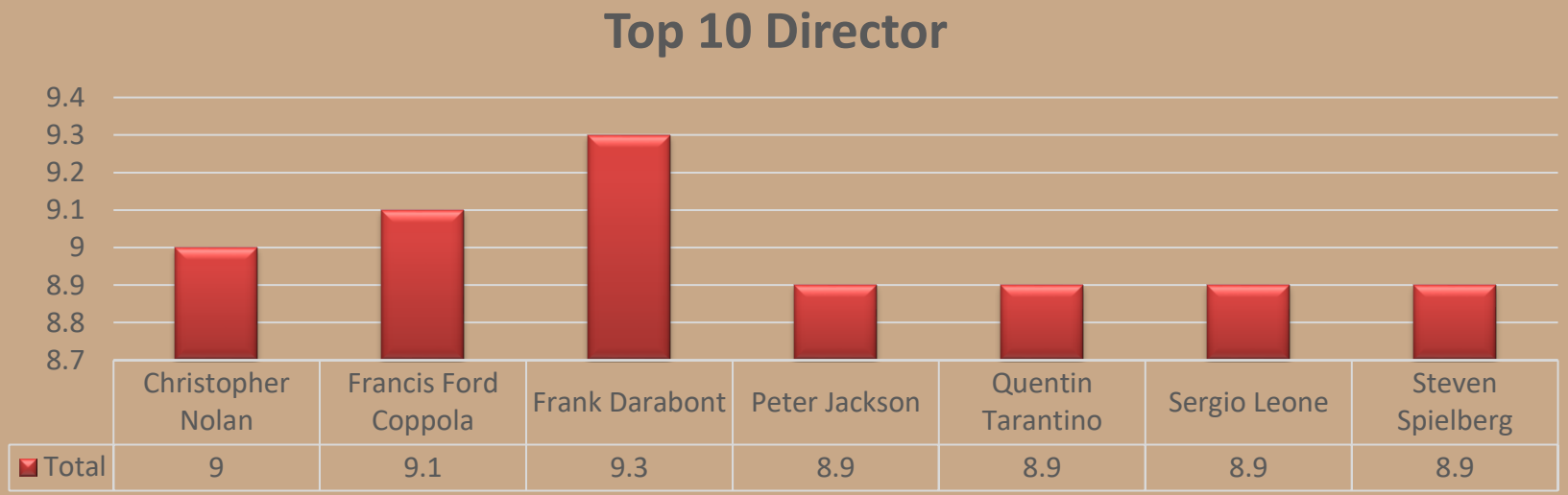
1.Highly Acclaimed Directors: Directors like Frank Darabont, Francis Ford Coppola, and Christopher Nolan have exceptionally high average IMDb scores of 9.0 or above. This indicates a consistent track record of directing highly acclaimed and well-received movies.

2.Consistency in High Ratings: Directors such as Peter Jackson, Quentin Tarantino, Sergio Leone, and Steven Spielberg also have high average IMDb scores, all around 8.9. This suggests a consistent ability to deliver movies that are highly appreciated by audiences and critics alike.

3.Overall Average IMDb Score: The overall average IMDb score across all directors is 9.0125, indicating that the movies directed by these directors are generally considered of high quality and are well-received.

Analysis

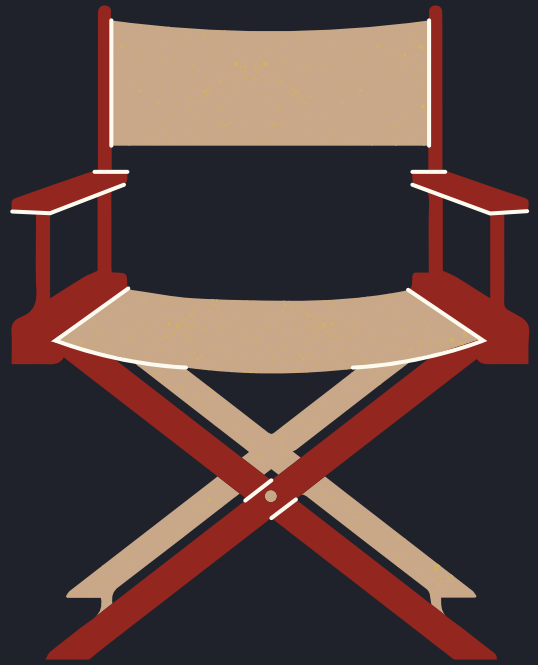
Row Labels	Average of imdb_score
Christopher Nolan	8.9
David Fincher	8.8
Francis Ford Coppola	9.1
Frank Darabont	9.3
Irvin Kershner	8.8
Peter Jackson	8.85
Quentin Tarantino	8.9
Robert Zemeckis	8.8
Sergio Leone	8.9
Steven Spielberg	8.9
Grand Total	8.930769231



Budget Analysis

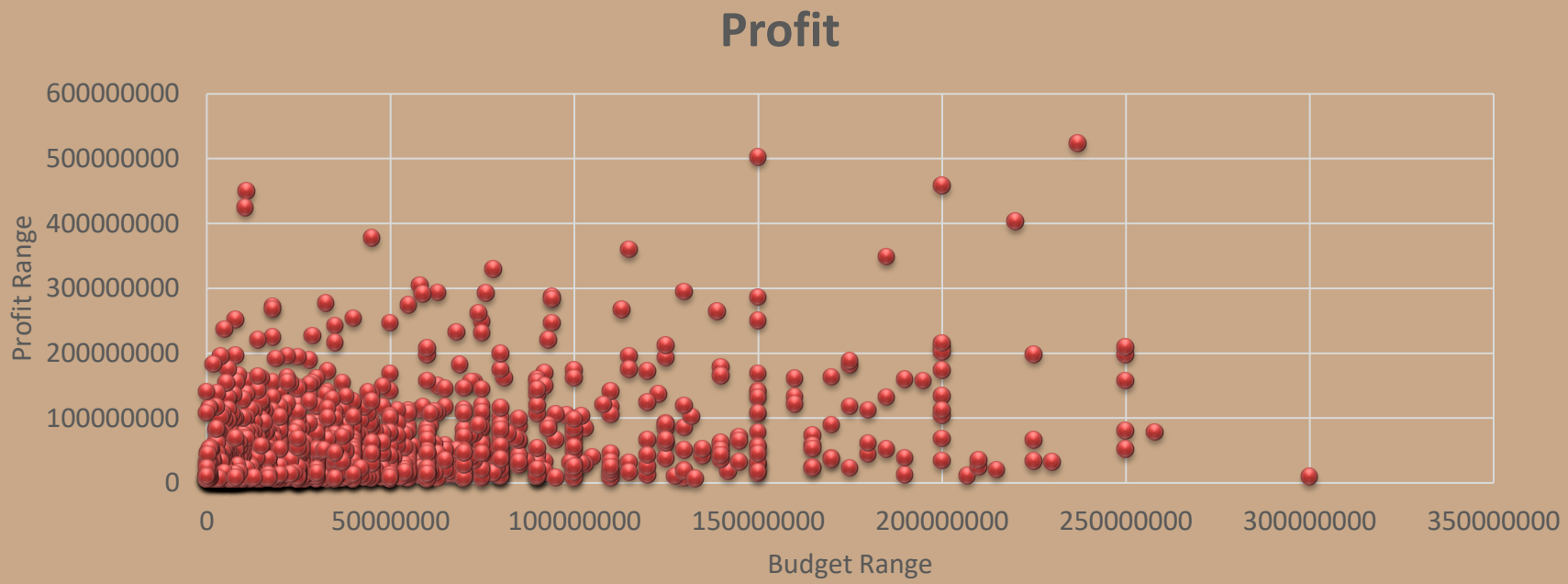
I found out how movie budgets relate to financial success by studying the correlation between budgets and gross earnings. I found a positive correlation between the two variables, but the strength of this relationship depended on factors like genre and director.

If we keep the budget over 200,000,000, there's a high chance of earning a greater profit margin.



MAX	523505847
MIn	-12213298588

Analysis



CONCLUSION

This project helped me learn about what makes movies popular, successful and financial performance. By analyzing genres, durations, languages, directors, and budgets, I was able to identify trends and patterns that shed light on audience preferences and industry dynamics. My findings provide actionable insights for filmmakers, producers, and studios seeking to optimize their movie production strategies and maximize audience appeal.

Excel Drive link :

https://docs.google.com/spreadsheets/d/1rr8cE81SRHX86jjkUBGrdPKZyv9wHLSz/edit?usp=drive_l

ink&ouid=108396890359637253084&rtpof=true&sd=true

06

Bank Loan Case Study

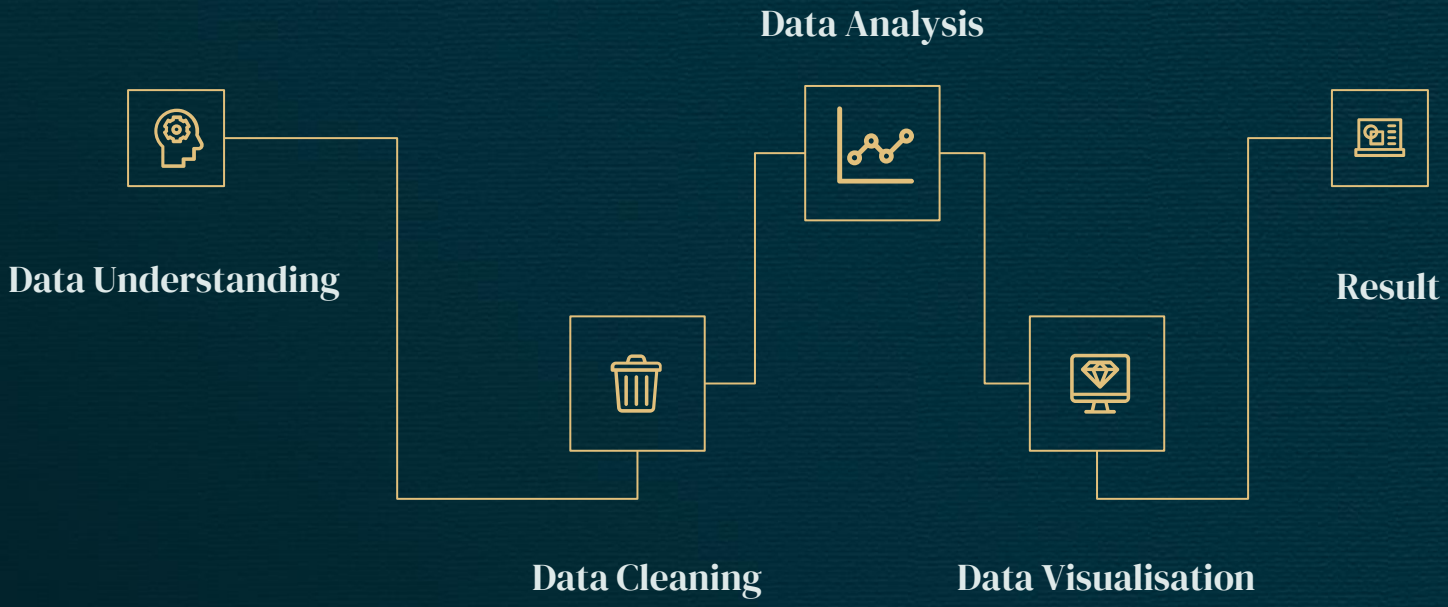
Bank Loan Case Study

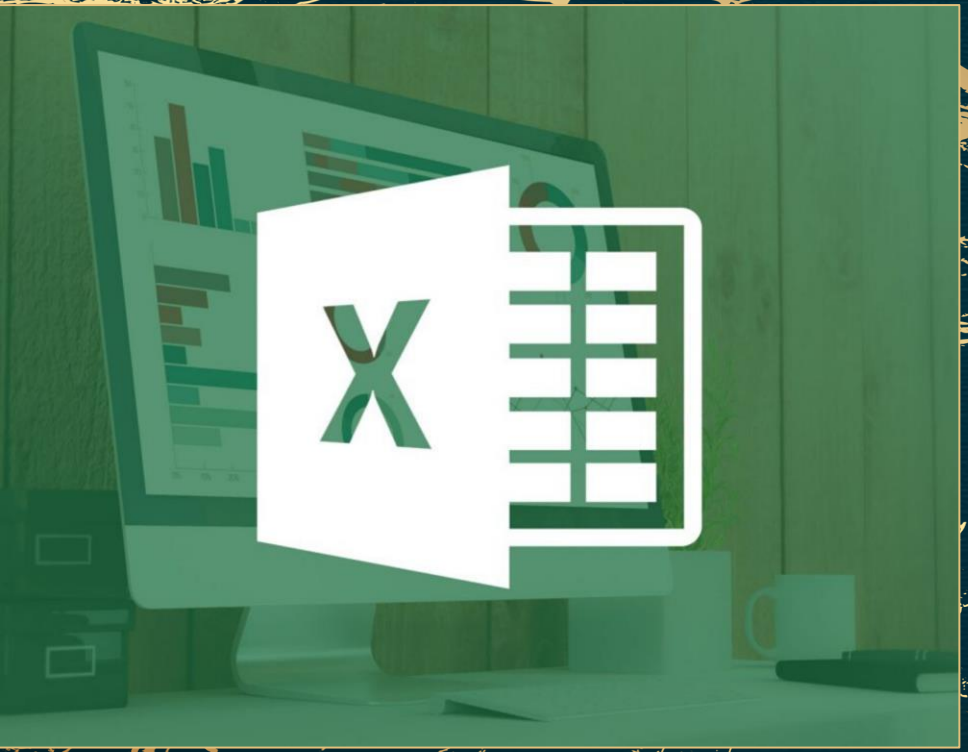
By Ayush Mahanta

PROJECT DESCRIPTION

In this project, I will figure out signs that show if a client might struggle to pay back their installments. I will use this information to decide whether to deny a loan, give less money, or offer a loan with higher interest rates to risky applicants. This helps ensure that people who can pay back the loan don't get turned down. I will study the data using Exploratory Data Analysis (EDA), which helps summarize and understand the characteristics of the dataset by looking for patterns and relationships to help with decision-making and analysis.

APPROACH


Data Analysis
Data Understanding
Result
Data Cleaning
Data Visualisation



Tech Stack

All the analysis has been performed in excel. This tool is also used to create graphical representation of the results and to understand the result set better.

Excel Link

:https://docs.google.com/spreadsheets/d/1HmMMXhynS3GySzA2xJIFXFksXyN6_SiU/edit?usp=drive_link&ouid=108396890359637253084&rtpof=true&sd=true

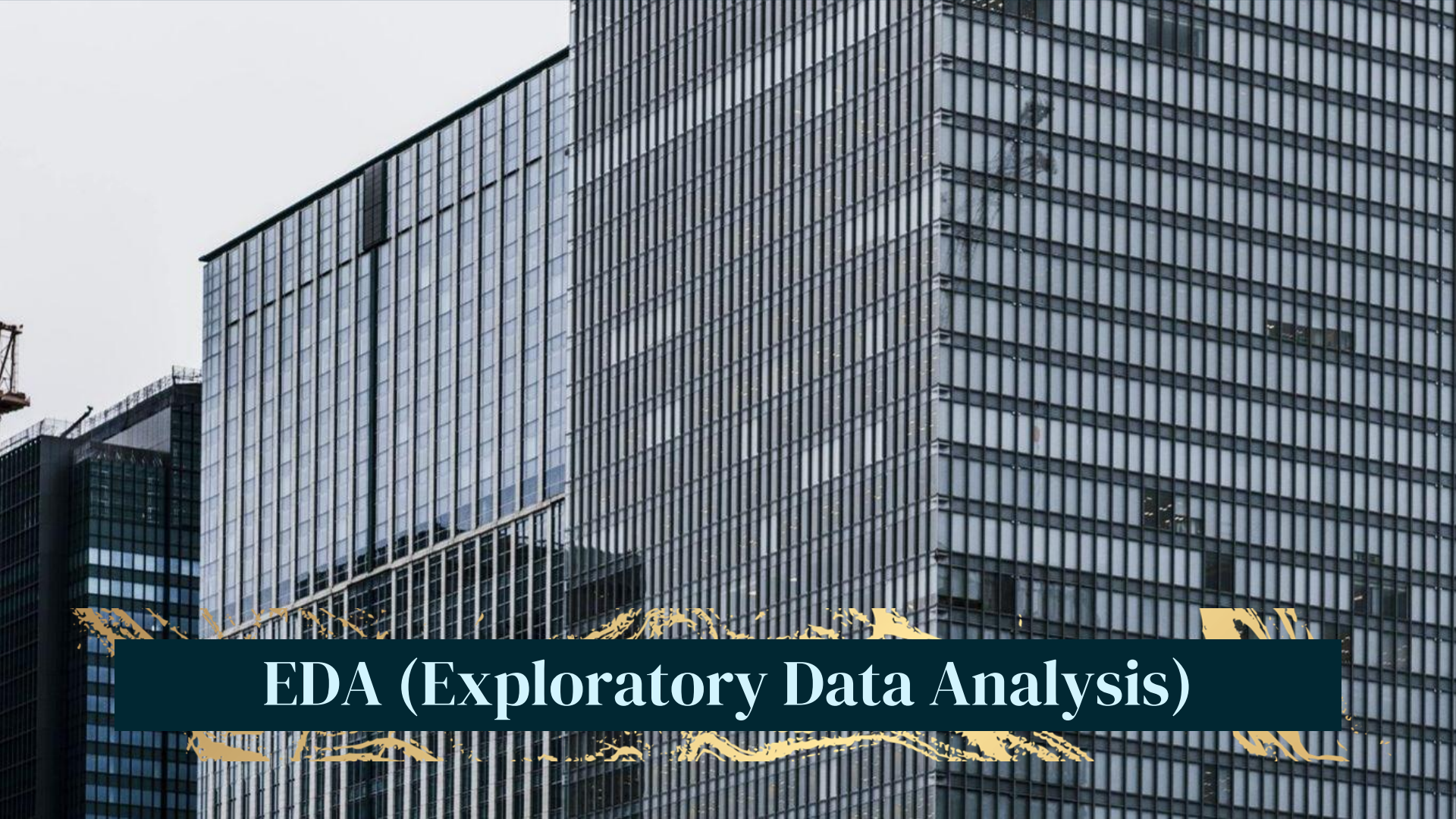
01

Data Cleaning

Columns to Drop

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

We see there are 50 columns with missing values greater than 30% we will drop those columns. These are the columns which mainly contain the residential details of the client



EDA (Exploratory Data Analysis)

Representation of columns having >30% NULL values.

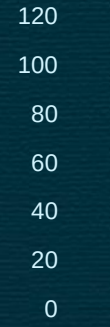
Comparison

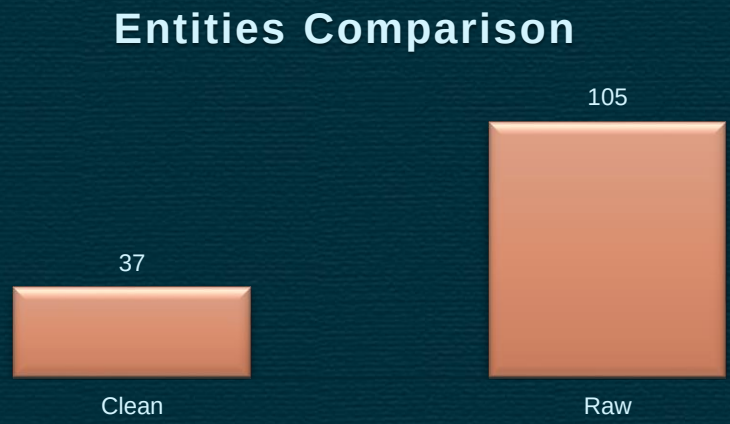
Updated Data

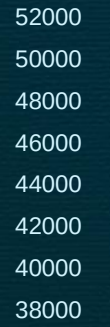
Clean	Clean Records
37	42857

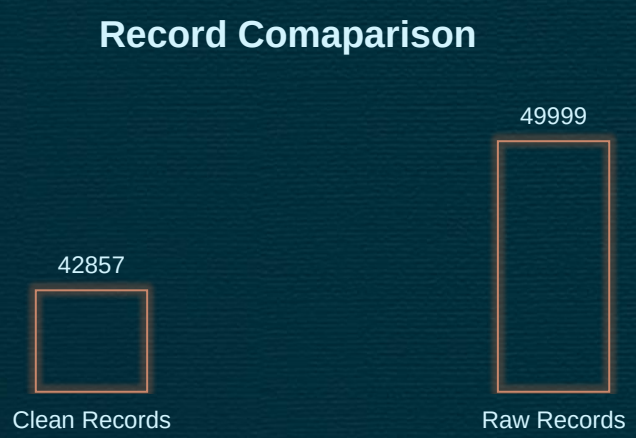
Raw Data

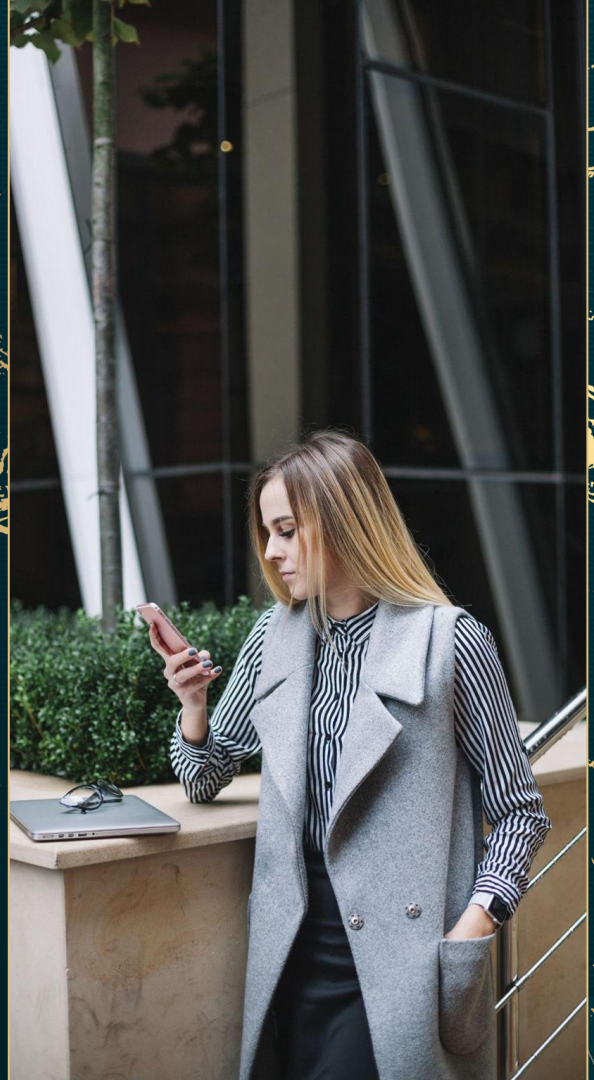
Raw	Raw Records
105	49999





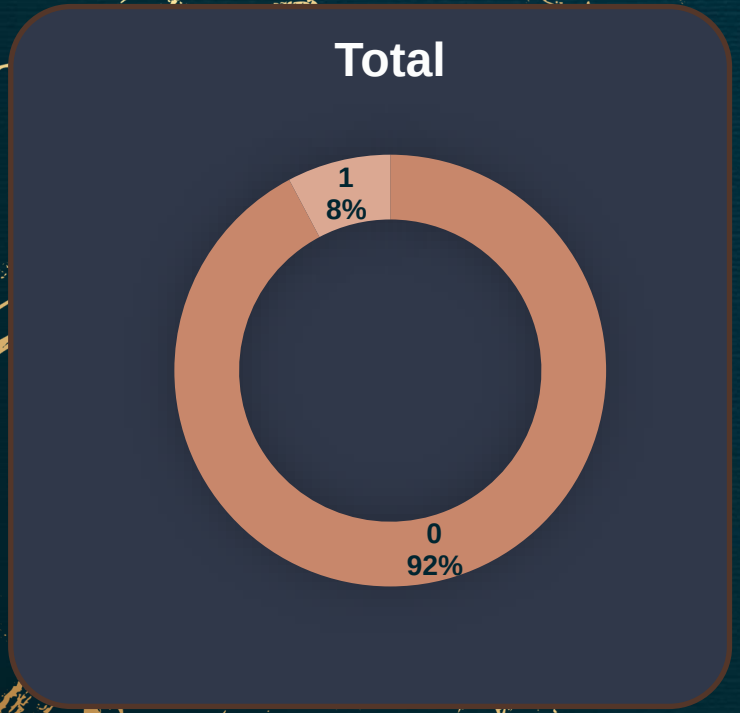






02

Data Imbalance



Data Imbalance

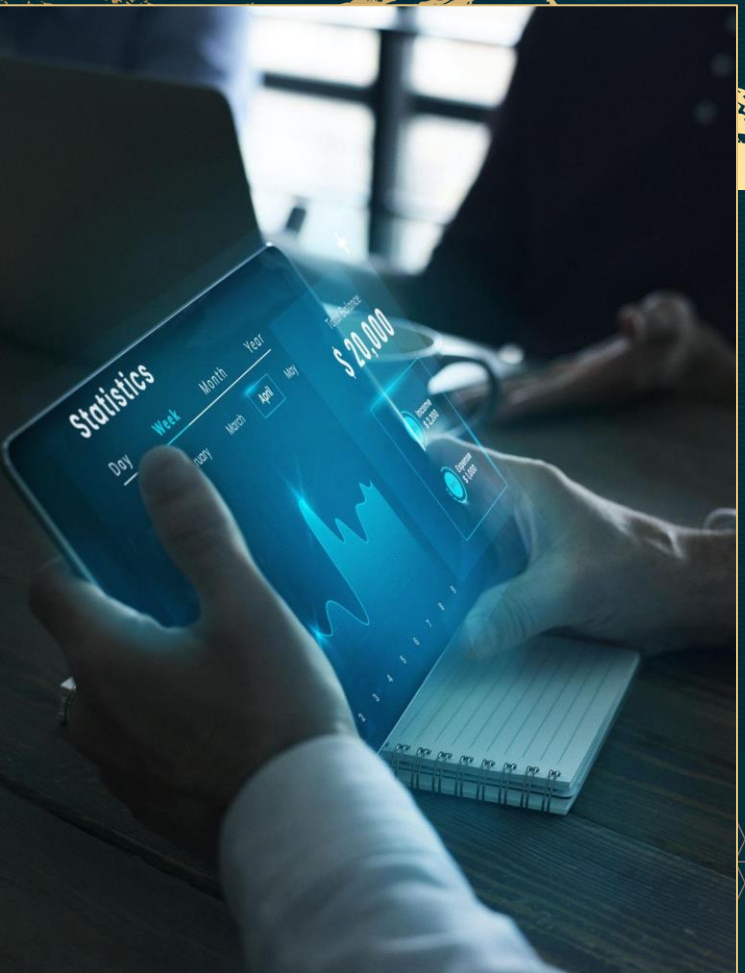
We see that we have :

- 92% as loan re-payers
- 8% as De faulters

Which gives us a clear indication that the data is highly imbalanced

03

Outliers



IDENTIFYING OUTLIERS

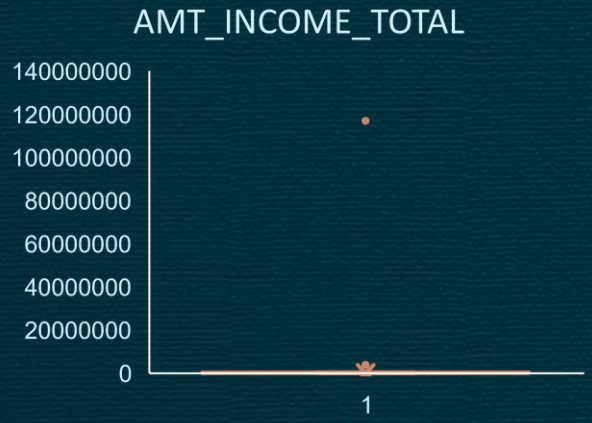
Steps to find Outliers using Tukey's method:

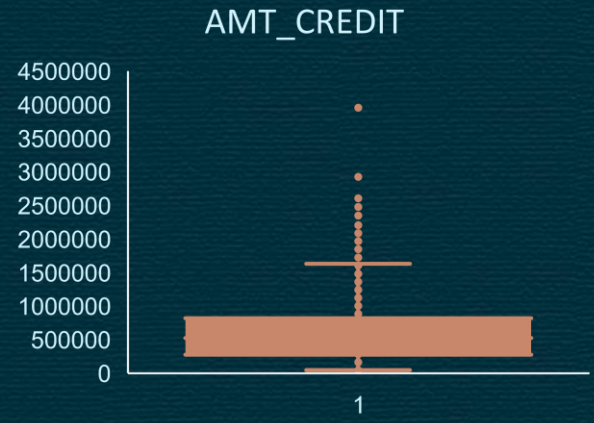
1. Finding 1st Quartile Q1 and 3rd Quartile Q3
2. Finding Inner Quarter Range(IQR)
3. Finding Upper Bound(Q3+(1.5*IQR)
4. Finding Lower Bound(Q1-(1.5*IQR)
5. Now, any data point above Upper Bound or Below Lower Bound Considered as Outlier

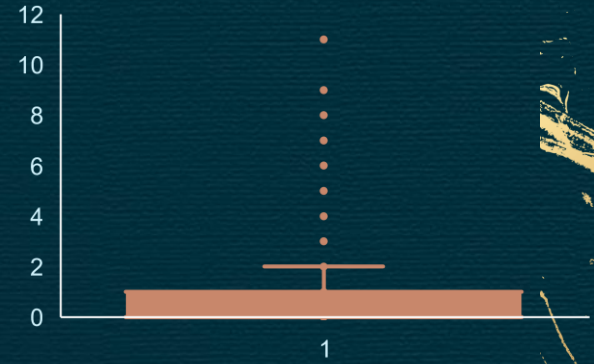
Below , I plotted some Box-Whisker Plots to find there are any outliers present in the data.

Outliers

- The first quartile almost missing for CNT_CHILDREN that means most of the data are present in the first quartile.
- There is single high value data point as outlier present in AMT_INCOME_TOTAL and DAYS_EMPLOYED. Removal this point will drastically impact the box plot for further analysis.
- The first quartiles is slim compare to third quartile for AMT_CREDIT, AMT_ANNUITY, DAYS_REGISTRATION. This mean data are skewed towards first quartile.







Quartile

AMT INCOME TOTAL

Quartile - 1
112500
Quartile - 3
202500
Inter Quartile Range
90000
UPPER LIMIT
337500
Lower Limit
-22500

CNT CHILDREN

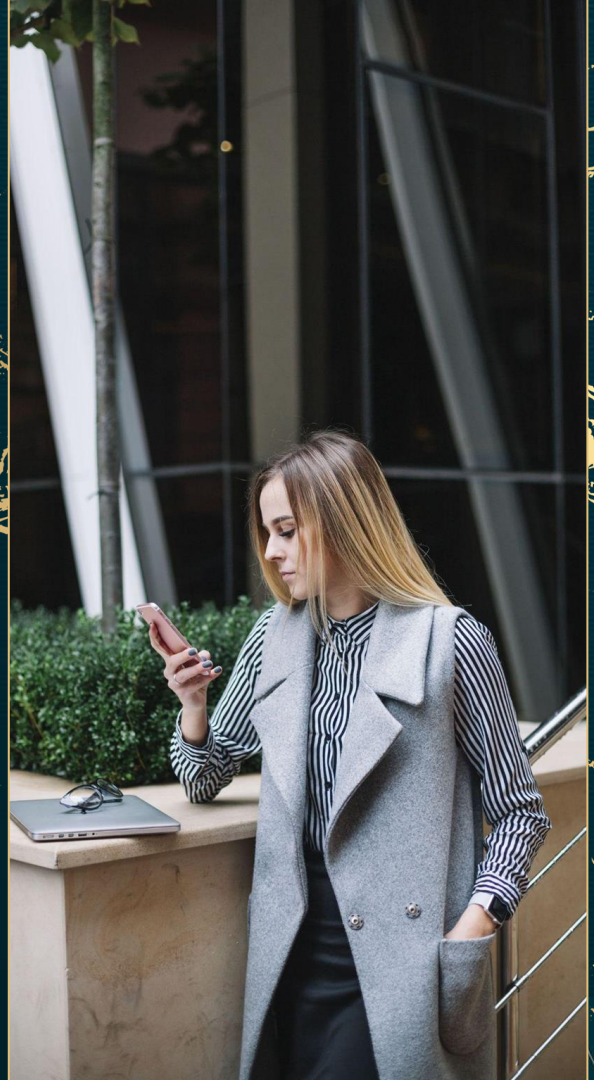
Quartile - 1
0
Quartile - 3
1
Inter Quartile Range
1
UPPER LIMIT
2.5
Lower Limit
-1.5

Days Employed (YRS)

Quartile - 1
2.652054795
Quartile - 3
15.89863014
Inter Quartile Range
13.24657534
UPPER LIMIT
35.76849315
Lower Limit
-17.21780822

AMT CREDIT

Quartile - 1
273636
Quartile - 3
816660
Inter Quartile Range
543024
UPPER LIMIT
1631196
Lower Limit
-540900



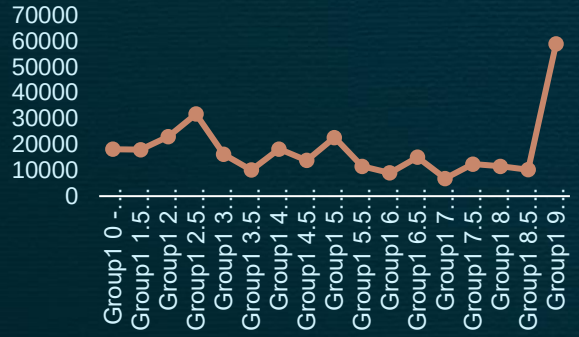
04

Univariate

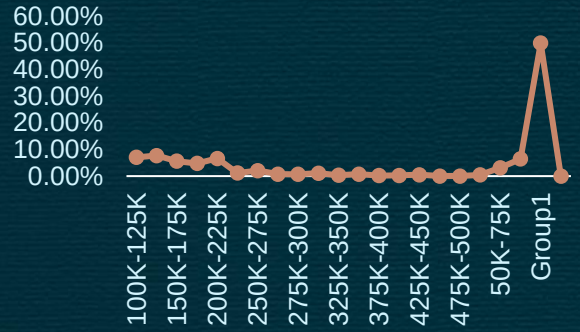
UNIVARIATE ANALYSIS

- Univariate analysis is statistical method to analyze the data with one variable.
- It involves the examining the distribution of single variable and deriving insights from it.
- Univariate analysis of categorical variables involves summarizing and examining the frequency or proportion of each category to gain better understanding of distribution and relationship between variables

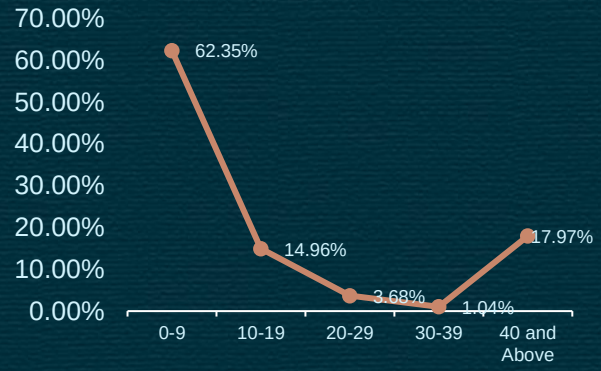
Segmented Applicants Per Credit Bins



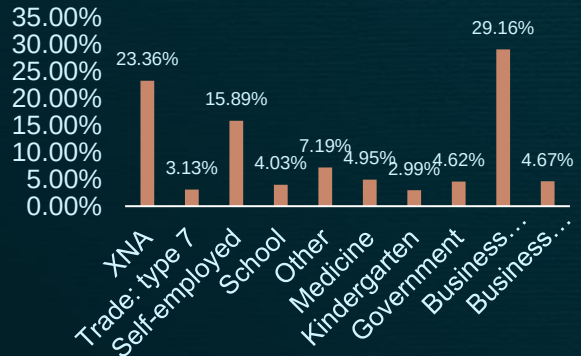
Segmented Target Applicants Per Income Bins



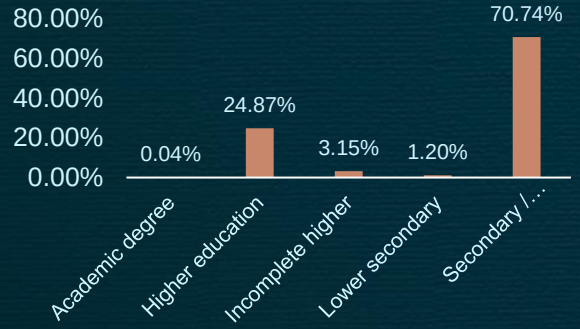
Segmented employment years



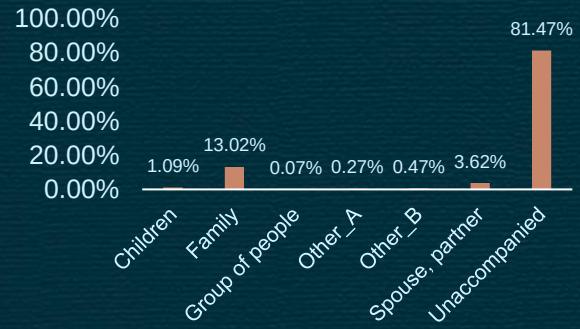
Segmented organization type



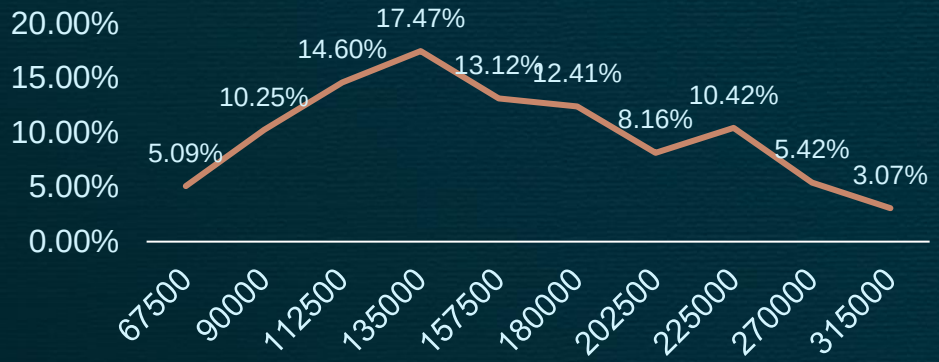
Segmented Education Type distribution



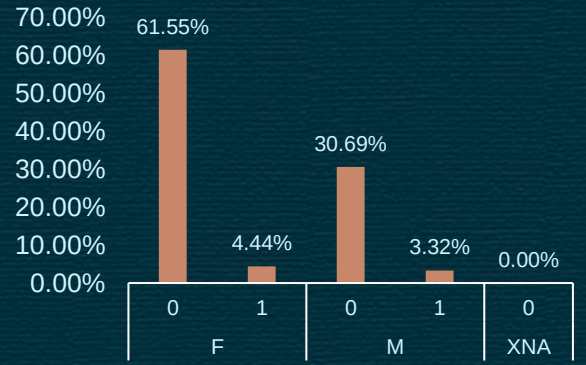
Segmented NAME_TYPE_SUITE



Amount Income



Applicants Per Credit Bins



Amount Credit

05

Bivariate Analysis

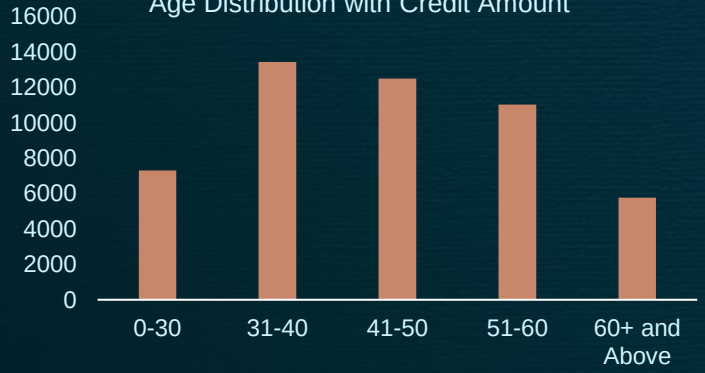
Bivariate Analysis

Bivariate analysis is a statistical method used to analyze data involving two variables. It's about looking at how two variables relate to each other. When we do bivariate analysis with categorical variables, we are interested in understanding the relationship between two categories or groups. This could mean comparing the frequency or proportion of each category across the two variables to see if there's any connection or pattern between them.

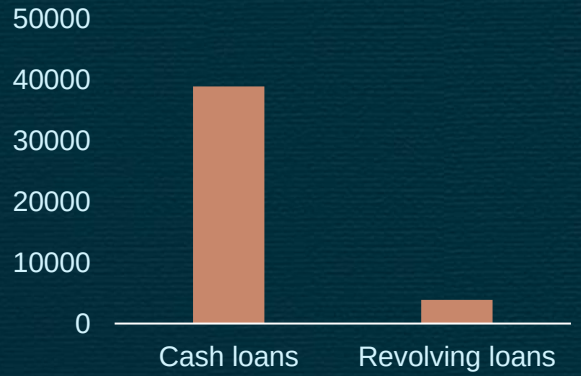
Average Credit Amount Per Credits Bins

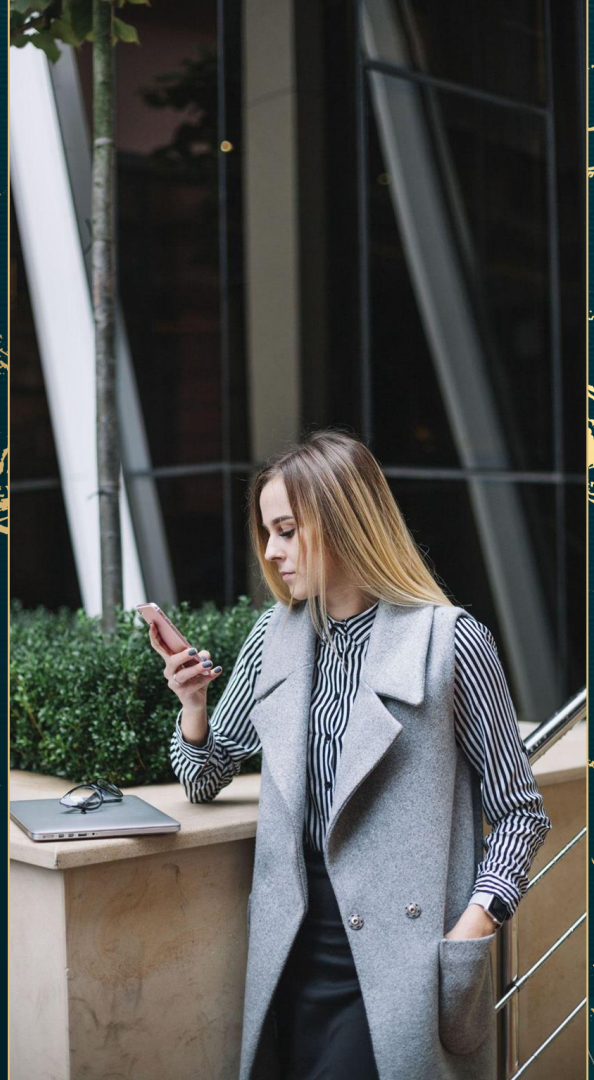
Average Amount Credited Per Name Income Type Segmented

Age Distribution with Credit Amount



NAME_CONTRACT_TYPE





06

Correlation

Correlation For Timely Payments

TARGET 0								
CNT Of Children	1	0.009300239	0.004117726	-0.025043214	-9.7283E-07	0.002368016	0.003949962	0.026733198
AMT_Income_Total	0.009300239	1	0.063511987	0.026543706	-0.000502928	-0.003429977	0.003573112	-0.035127419
AMT_Credit	0.004117726	0.063511987	1	0.098803701	-0.004437401	-0.000222036	-0.004349401	-0.102897278
Region Population Relative	-0.025043214	0.026543706	0.098803701	1	-0.003303866	-0.006023773	0.000844092	-0.527253308
Days_Birth(Yrs)	-9.7283E-07	-0.000502928	-0.004437401	-0.003303866	1	0.523388736	0.50850239	0.003170646
Days_Employed(YRS)	0.002368016	-0.003429977	-0.000222036	-0.006023773	0.523388736	1	0.296604311	0.001543287
Days_ID_Publish(YRS)	0.003949962	0.003573112	-0.004349401	0.000844092	0.50850239	0.296604311	1	0.003088812
Region_Rating_Client	0.026733198	-0.035127419	-0.102897278	-0.527253308	0.003170646	0.001543287	0.003088812	1
	CNT Of Chindren	AMT_Income_Total	AMT_Credit	Region_Population_Relative	Days_Birth(YRS)	Days_Employed(YRS)	Days_ID_Publish(YRS)	Region_Rating_Client

TARGET 1								
CNT Of Children	1	0.009300239	0.004117726	-0.025043214	-9.7283E-07	0.002368016	0.003949962	0.026733198
AMT_Income_Total	0.009300239	1	0.063511987	0.026543706	-0.000502928	-0.003429977	0.003573112	-0.035127419
AMT_Credit	0.004117726	0.063511987	1	0.098803701	-0.004437401	-0.000222036	-0.004349401	-0.102897278
Region Population Relative	-0.025043214	0.026543706	0.098803701	1	-0.003303866	-0.006023773	0.000844092	-0.527253308
Days_Birth(Yrs)	-9.7283E-07	-0.000502928	-0.004437401	-0.003303866	1	0.523388736	0.50850239	0.003170646
Days_Employed(YRS)	0.002368016	-0.003429977	-0.000222036	-0.006023773	0.523388736	1	0.296604311	0.001543287
Days_ID_Publish(YRS)	0.003949962	0.003573112	-0.004349401	0.000844092	0.50850239	0.296604311	1	0.003088812
Region_Rating_Client	0.026733198	-0.035127419	-0.102897278	-0.527253308	0.003170646	0.001543287	0.003088812	1
	CNT_Of_Children	AMT_Income_Total	AMT_Credit	Region_Population_Relative	Days_Birth(YRS)	Days_Employed(YRS)	Days_ID_Publish(YRS)	Region_Rating_Client

Result

- As people get older and gain more experience, they're less likely to miss loan payments. So, banks should give more attention to older and experienced clients.
- Clients with higher education tend to miss fewer payments compared to those with lower education levels like high school or lower secondary.
- Men tend to miss loan payments more often than women.
- Corporate clients are safer bets compared to labor class clients.
- People from Region Rating 3 have the highest percentage of defaulters, so banks could make stricter loan policies for clients from this region. Clients from Region 1 are the safest bet.
- As clients get older, they tend to take larger loan amounts, and since older clients have lower default rates, they are less risky and more profitable for the bank.
- Banks should focus more on clients with contract types like 'Student,' 'Pensioner,' and 'Businessman' who have housing types other than 'Co-op apartment' for successful payments.
- Banks should be cautious with clients whose income type is 'Working' as they have the highest number of missed payments.
- For loans for the purpose of 'Repairs,' although there are more rejections, there are also difficulties in paying on time.
- There are some areas where loan payments are significantly delayed. Banks should be cautious when giving loans for these purposes.
- Banks should avoid giving loans for co-op apartments as they have difficulties in payment.
- Banks can focus more on housing types like 'with parents,' 'House/apartment,' and 'municipal apartment' for successful payments.

Conclusion

This project helps in handling the large datasets. How exploratory data analysis can be applied to large datasets. When dealing with the large datasets it is also important to select only those columns which are extremely useful to our analysis. Finding correlations columns can become very convenient while dealing with large datasets as it saves time selecting which columns should be considered for analysis. The project also helps in understanding the various terminologies used in the banking domain.

07

Analyzing the Impact of Car Features on Price and Profitability

2

PROJECT DESCRIPTION

The car industry is changing fast, with a focus on fuel efficiency and new tech. People want electric and hybrid cars, but gasoline cars are still popular. Companies need to understand what customers want to make more money.

This project aims to study how different features of cars affect their prices and profitability in the automotive industry.
By analyzing data about over 11,000 car models, companies can see which features and types of cars sell best. This helps them decide what new cars to make and how much to charge.

3

APPROACH

01 Data Collection and Familiarization

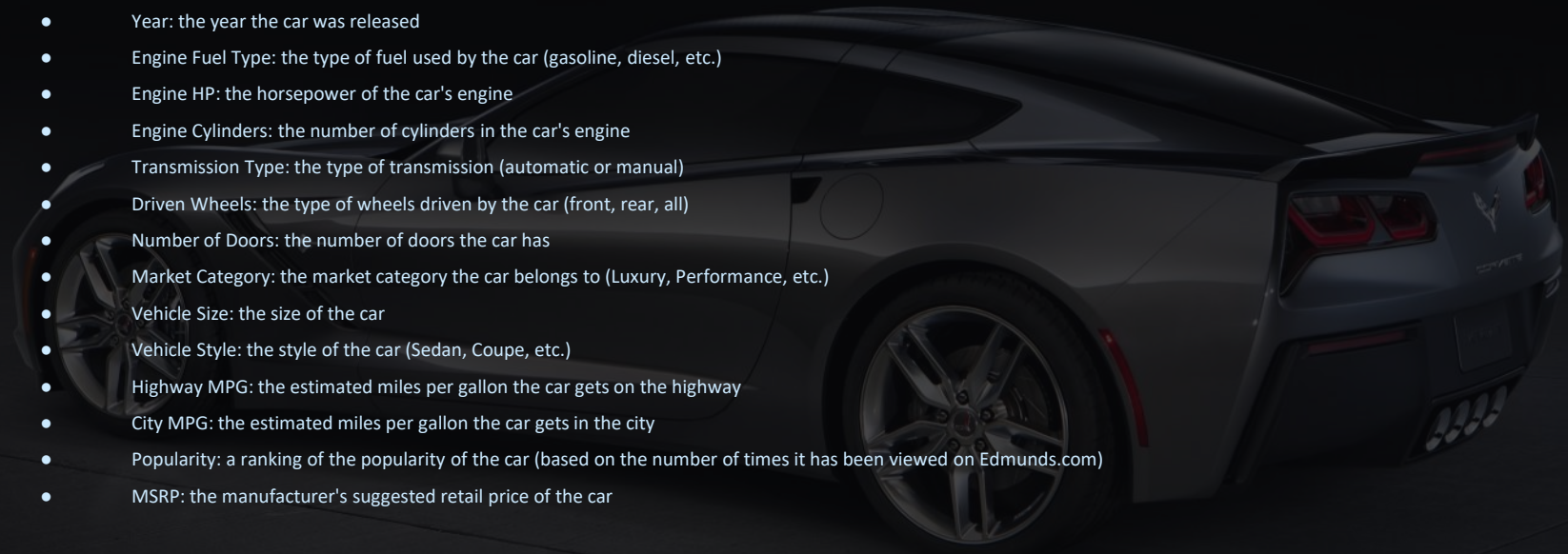
02 Data Cleaning and Preparation

03 Data Analysis

04 Building the Interactive Dashboard

Dataset Description

- Make: the make or brand of the car
- Model: the specific model of the car
- Year: the year the car was released
- Engine Fuel Type: the type of fuel used by the car (gasoline, diesel, etc.)
- Engine HP: the horsepower of the car's engine
- Engine Cylinders: the number of cylinders in the car's engine
- Transmission Type: the type of transmission (automatic or manual)
- Driven Wheels: the type of wheels driven by the car (front, rear, all)
- Number of Doors: the number of doors the car has
- Market Category: the market category the car belongs to (Luxury, Performance, etc.)
- Vehicle Size: the size of the car
- Vehicle Style: the style of the car (Sedan, Coupe, etc.)
- Highway MPG: the estimated miles per gallon the car gets on the highway
- City MPG: the estimated miles per gallon the car gets in the city
- Popularity: a ranking of the popularity of the car (based on the number of times it has been viewed on Edmunds.com)
- MSRP: the manufacturer's suggested retail price of the car



5

TECH-STACK

Microsoft Excel:

For Cleaning Data,

Analyzing Data And Visualization

Microsoft Power Point:

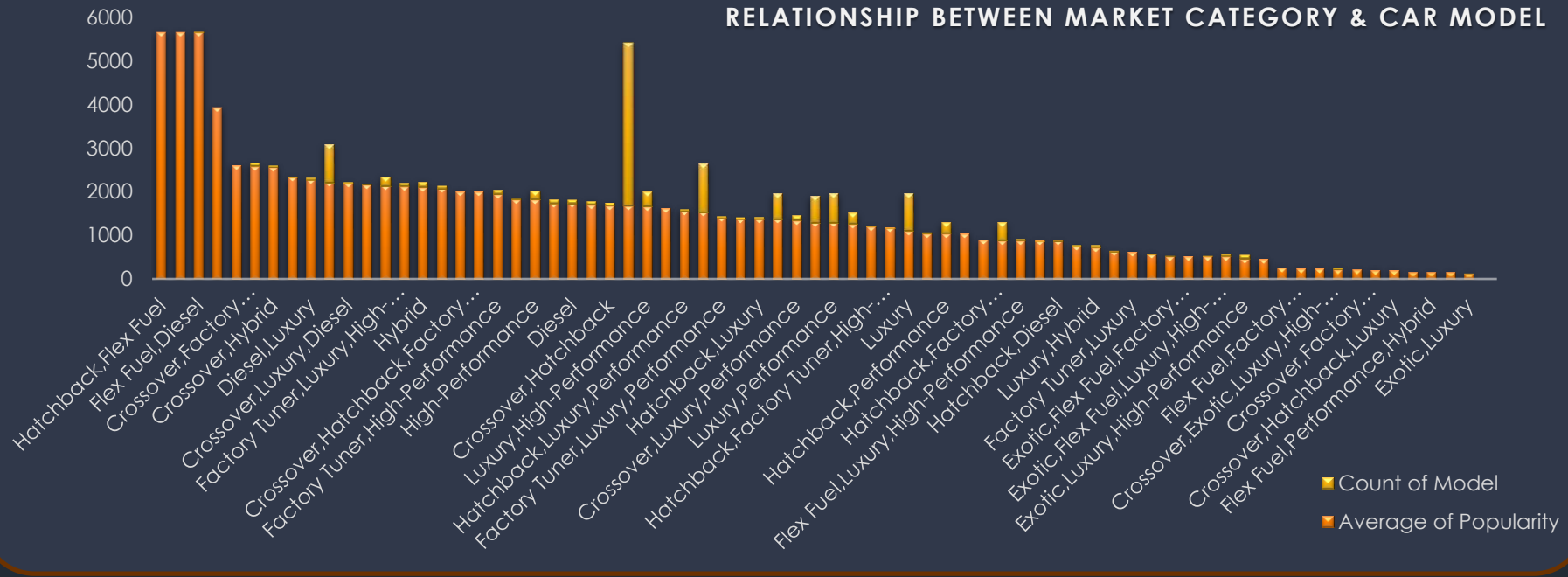
For Creating The Report

Excel Link: CLICK HERE!


/INSIGHTS

7

Popularity

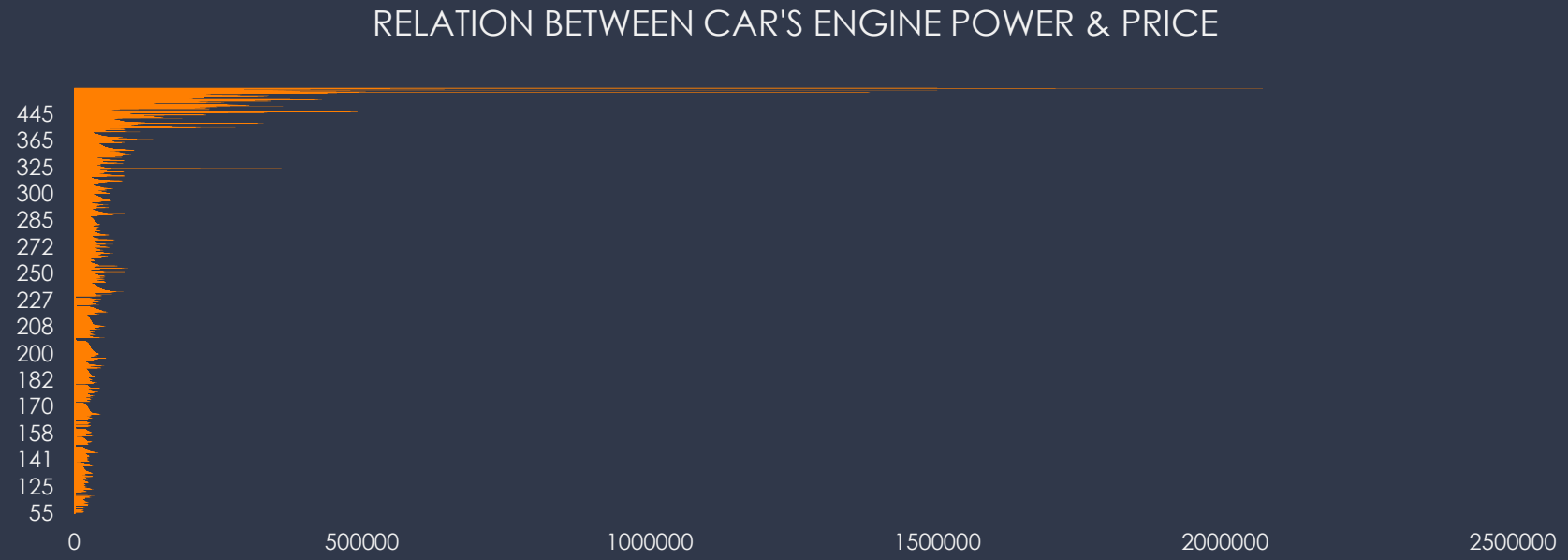


The quantity of car models in different market categories,

along with their respective popularity scores.

8

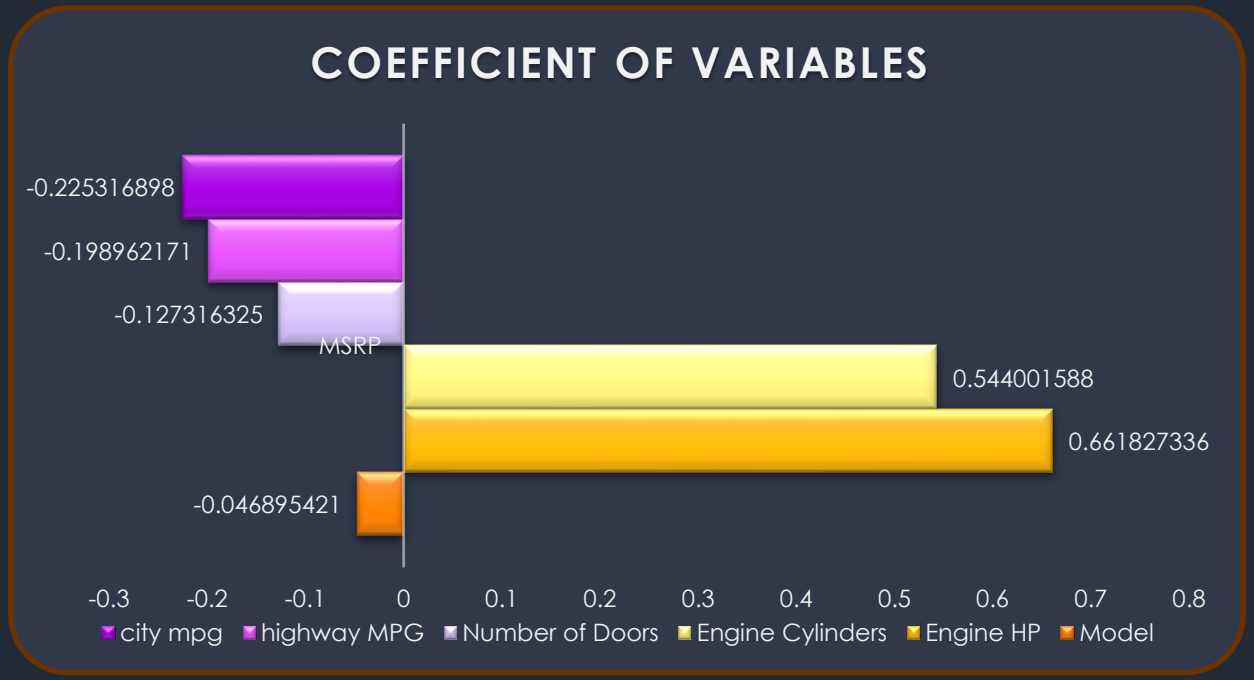
Regression Analysis



The graph exhibits a positive trendline slope, suggesting a direct correlation between a car's engine power and its price. This implies that vehicles equipped with more powerful engines generally have higher price tags.

9

Car's engine power vs price



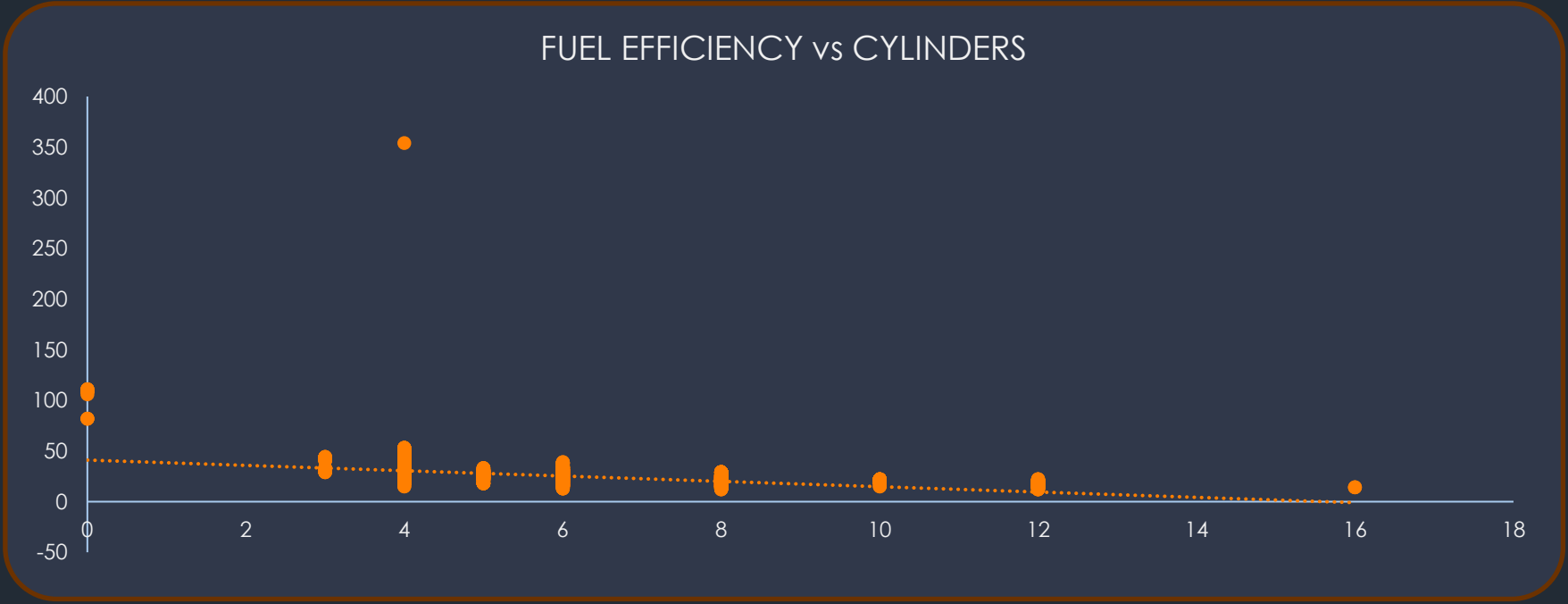
	Model	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg
MSRP	-0.046895421	0.661827336	0.544001588	-0.127316325	-0.198962171	-0.225316898

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.681376129
R Square	0.464273429
Adjusted R Square	0.464003474
Standard Error	44006.94687
Observations	11914

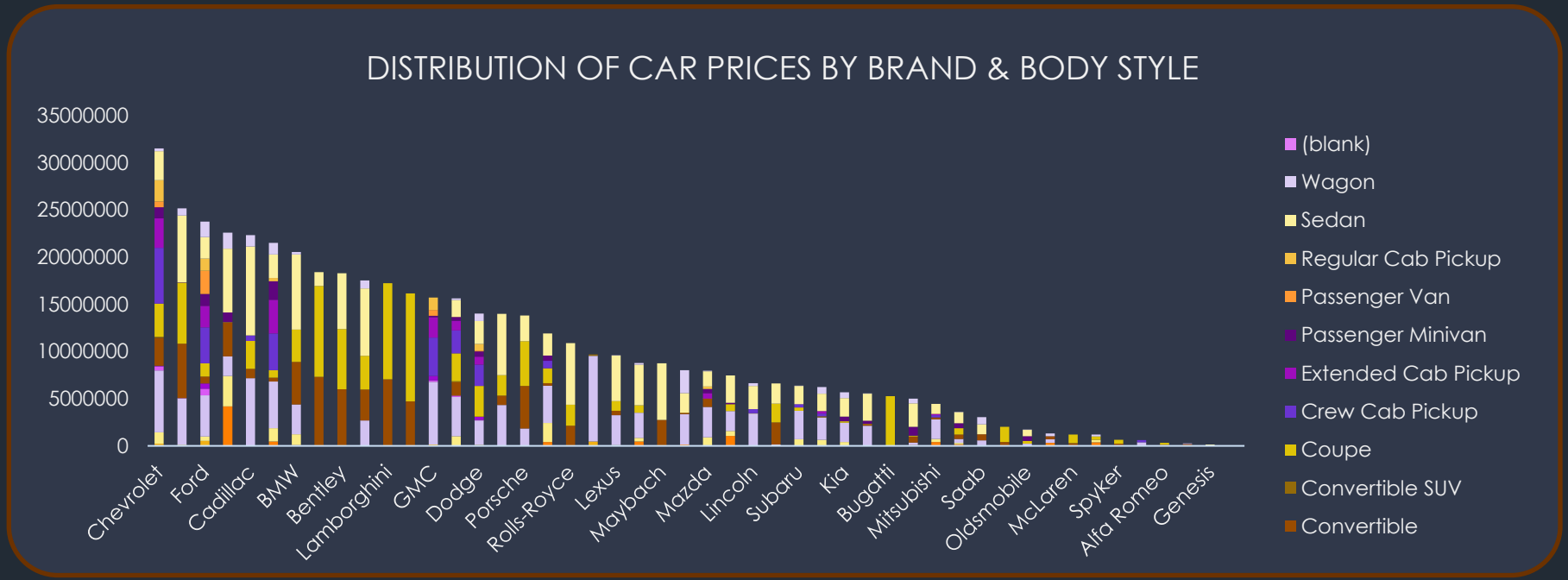
From the bar chart we can conclude that the strongest relationship with price is of Engine Cylinders and the negative relationship is with Number of Doors, which means that as the number of doors in a vehicle increases, the price tends to decrease, and vice versa

Fuel efficiency Vs Cylinders



Upon conducting the analysis, it was observed that there exists a negative correlation between the number of cylinders in a car's engine and its highway miles per gallon (MPG). This implies that as the number of cylinders increases, the fuel efficiency of the vehicle tends to decrease. The graph visually depicts this relationship, with the trendline exhibiting a negative slope, indicating a decline in highway MPG as the number of cylinders in the engine increases.

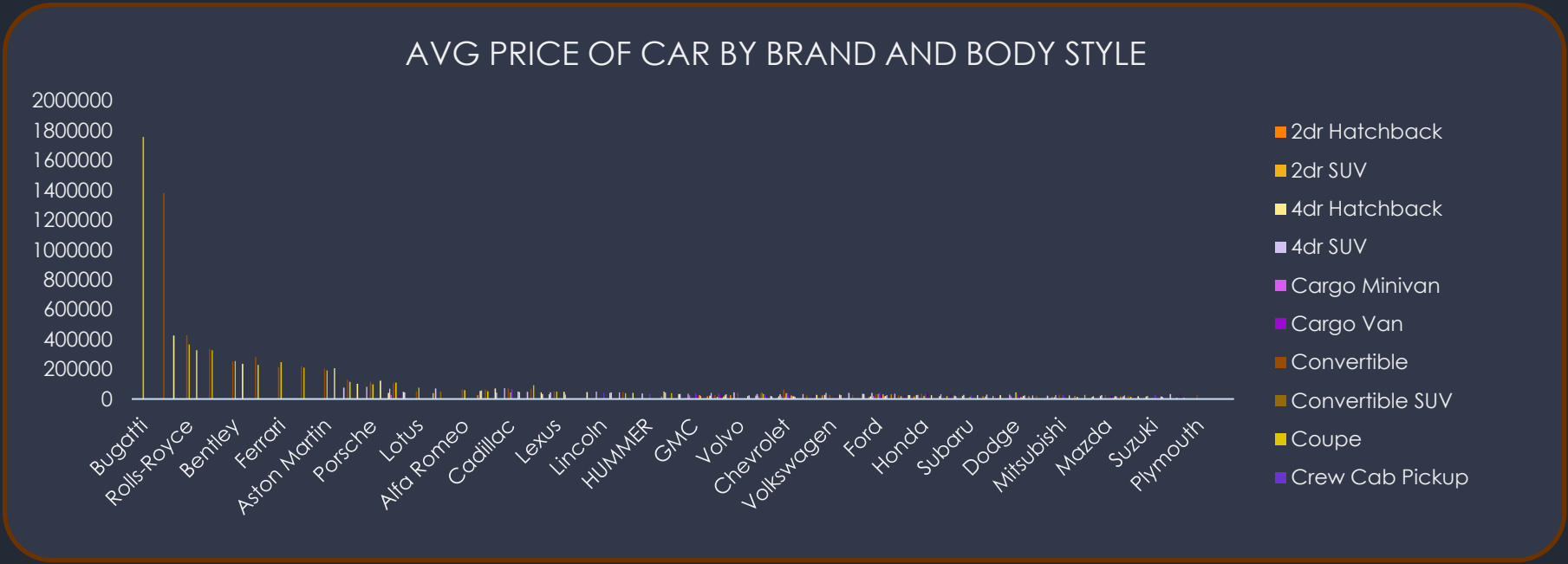
Car Price Distribution



This analysis offers valuable insights into the variations in car prices based on brand and body style. Such insights can prove beneficial for manufacturers in optimizing their pricing strategies and enhancing profitability. Additionally, the utilization of slicers enables a deeper exploration of the data, allowing for a more detailed examination of specific details and patterns. The following brand Chevrolet, Mercedes-Benz and Ford have the highest sum MSRP ,similarly the vehicle styles sedan, coupe and passenger minivan have the highest sum MSRP

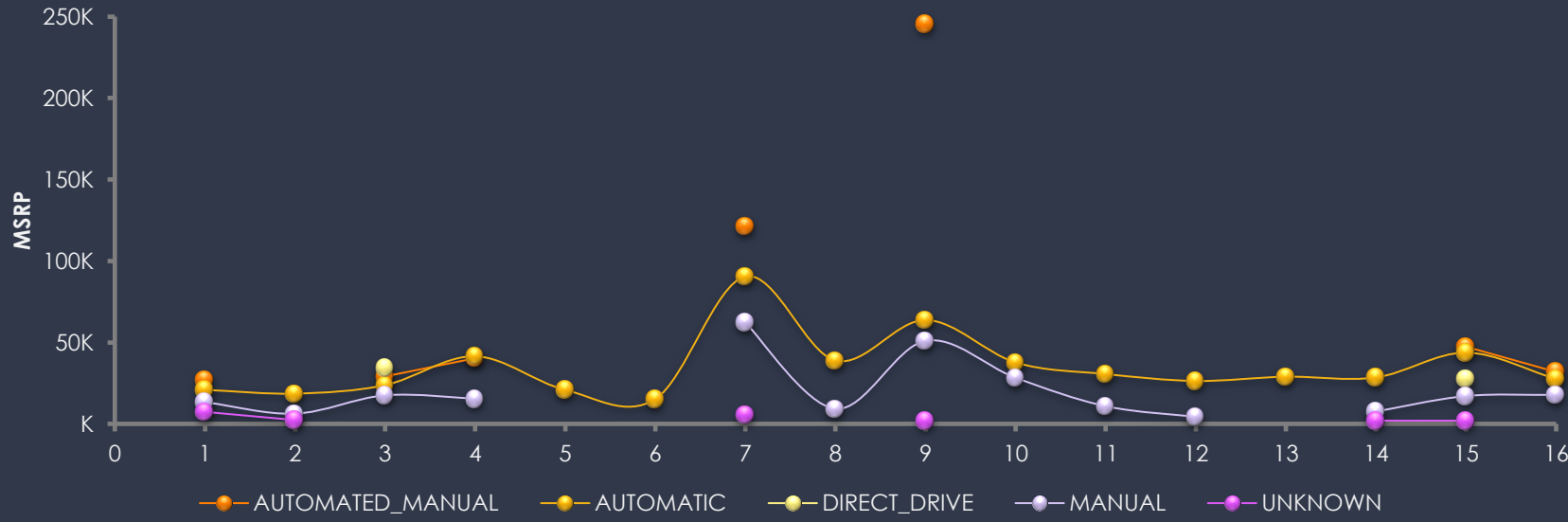
12

Avg Price of Car(body & style)



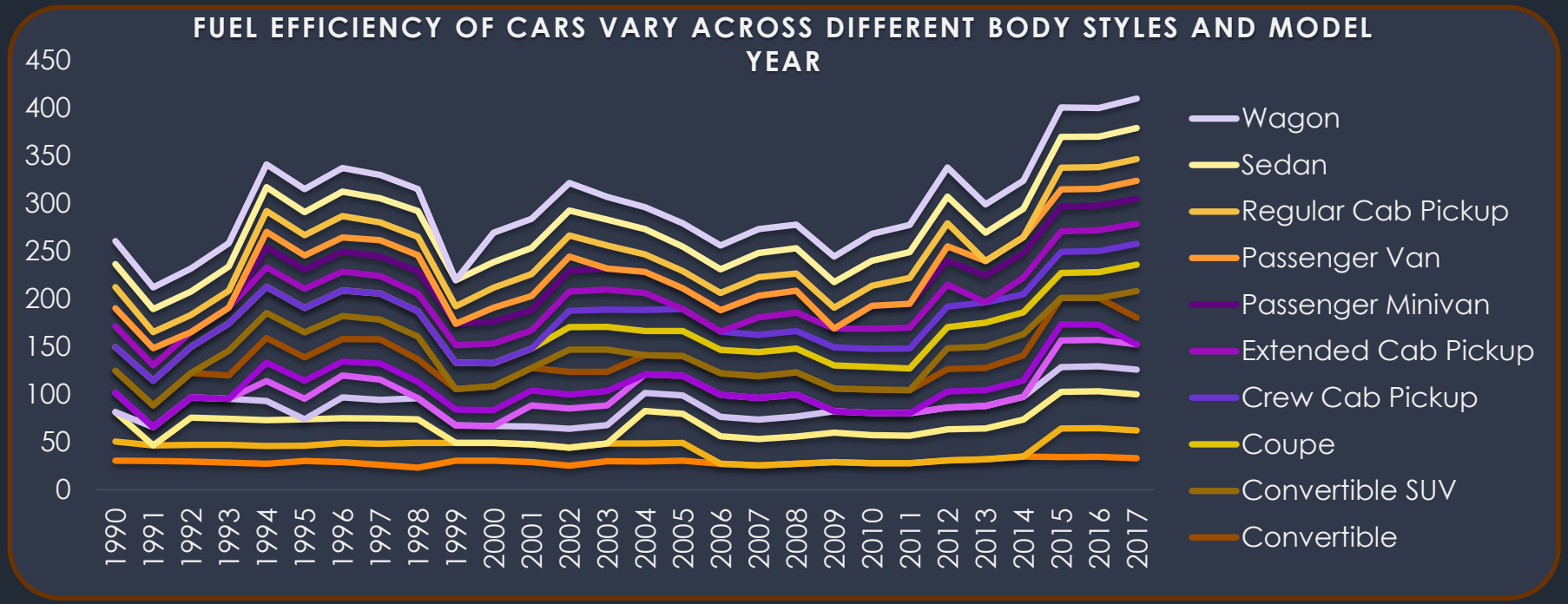
This analysis can offer insights into how prices differ among vehicles produced by each manufacturer, and whether certain manufacturers generally produce cars that are more expensive or less expensive overall.

Transmission Type

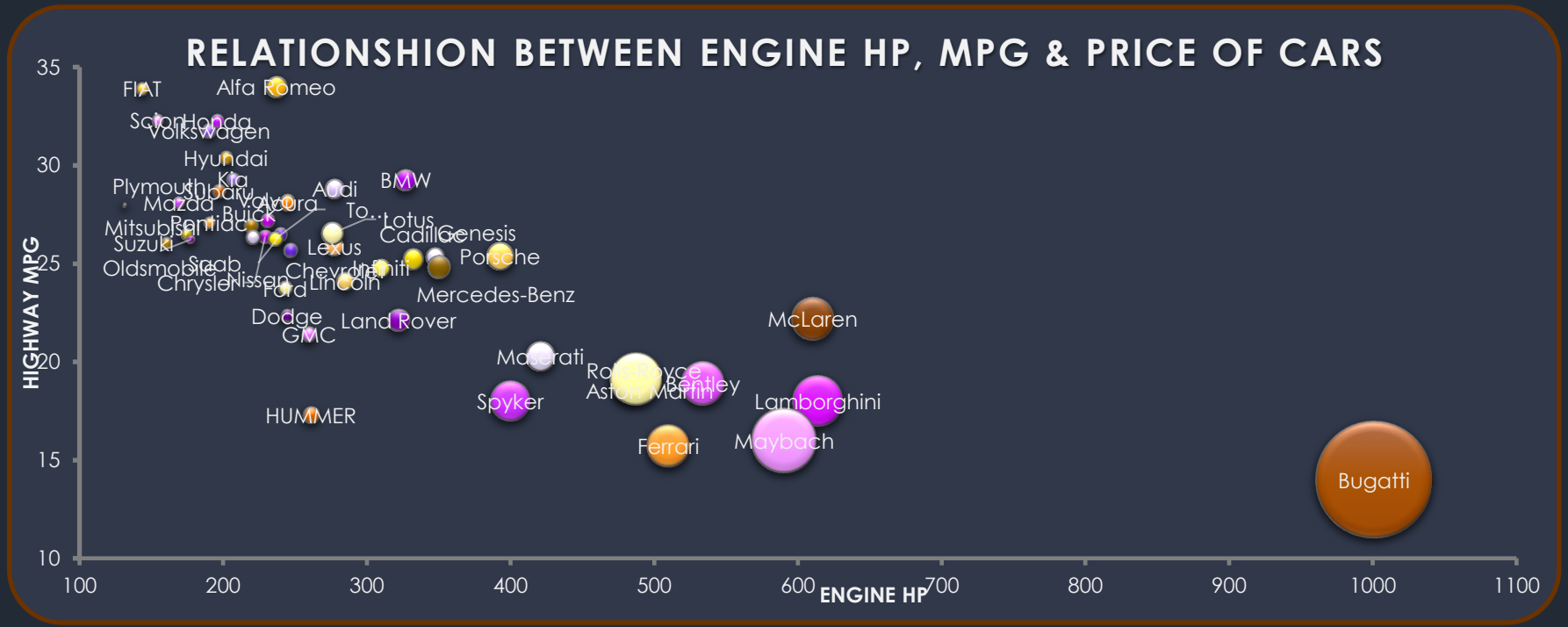


During the analysis, it was observed that body styles such as 4Dr hatchback, 2Dr hatchback, and sedan tend to exhibit higher overall fuel efficiency. This suggests that these particular body styles have shown a greater focus on fuel efficiency improvements compared to other styles.

Fuel Efficiency Over Time



Relation HP, MPG, MSRP



The analysis reveals that certain brands exhibit higher average horsepower, while others have lower average horsepower. This trend holds true for average miles per gallon (MPG) and manufacturer's suggested retail price (MSRP) as well. In other words, there are variations in average horsepower, MPG, and MSRP across different brands, indicating disparities in performance, fuel efficiency, and pricing within the automotive market.

Dashboard

Brand	Transmission Type	Driven Wheels	Number of Doors	Vehicle Size	Vehicle Style
Chevrolet	Automatic	Front Wheel Drive	4	Automatic	Sedan
1115	8231	4732	8263	4708	3013

Filter Panel

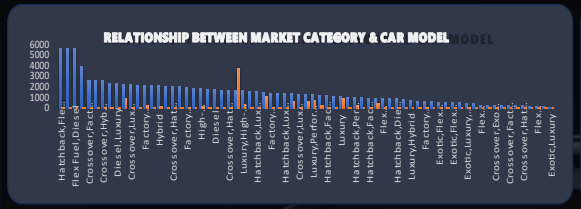
Make
- Chevrolet
- Chrysler
- Dodge

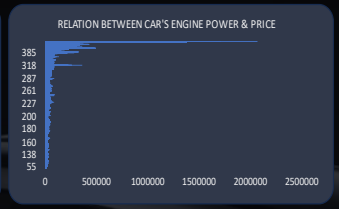
Model
- 62
- 80
- 86

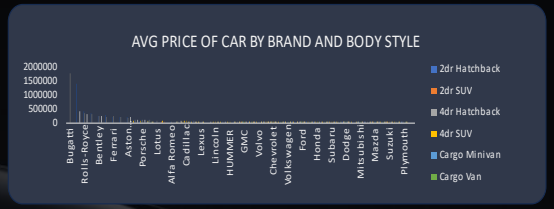
Market Category
- Crossover
- Crossover,Diesel
- Crossover,Exotic,Lu...

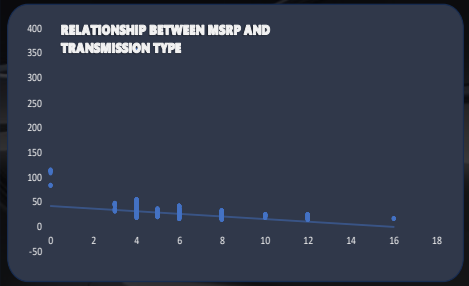
Transmission Type
- AUTOMATED_MAN...
- AUTOMATIC
- DIRECT_DRIVE

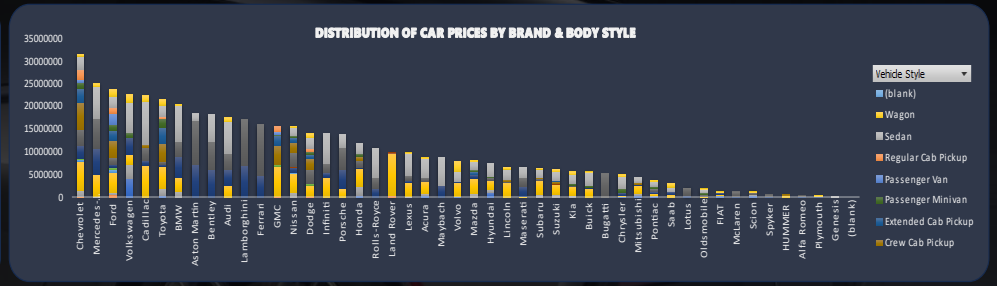
Vehicle Style
- 2dr Hatchback
- 2dr SUV
- 4dr Hatchback

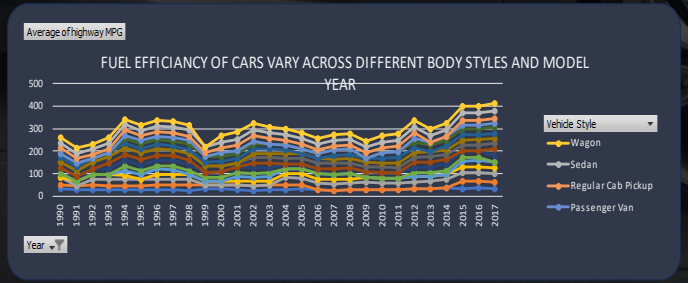


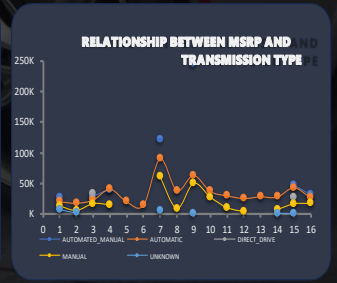


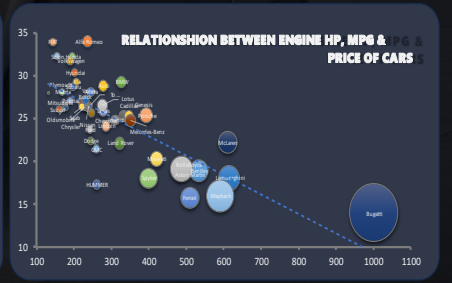












Result

Visuals:

I used visualizations such as pivot tables, scatter graphs, and bar charts to show the report. These visualizations help people understand results better and make decisions easier.

Discussion:

I found in my study is really important for car makers. It tells them what people like in cars, how to set prices, and what's happening in the market. Knowing this helps them make better decisions to make more money and compete better.

Future Plans:

In the future, I can explore more about how prices change, use up-to-date market info, and use advanced computer techniques to predict what might happen. Also, it's important to keep an eye on what people like and what's happening in the market to stay competitive in the car industry.

Conclusion

During the project, I learned and used different ways to analyze data in Microsoft Excel. I got really good at using pivot tables, which helped me understand big sets of data quickly by spotting trends and unusual things. I also learned how to do regression analysis, which is a fancy way of saying I figured out how different things are connected and used that to make predictions. I got comfortable with making charts and graphs that make data easy to understand. I even made cool interactive dashboards that let people explore the data themselves. All of this helped me find important information and make smart decisions based on it. Overall, this project taught me a lot about Excel and how to use it to understand data better.

08

ABC Call Volume Trend Analysis

Project Description

The project looks at information from an insurance company's Customer Care team over 23 days. It includes details about the staff, how long calls last, and what happened during the calls. They're using smart computer tools like IVR, RPA, Predictive Analytics, and Intelligent Routing to make customers happier. Different people in the team help customers using phones and other ways. They're trying to bring in more customers and make them happy to help the business grow. Ads are super important for getting people interested and buying stuff. Smart thinking is used to figure out the best ways to advertise without spending too much money. The project is all about studying customer care info, checking out cool computer tools, learning about job opportunities, and seeing how ads help a business grow.

APPROACH

01

Data Collection

02

Data Preparation

03

Data Analysis

04

Visualisation

05

Result

06

Conclusion

Tech Stack

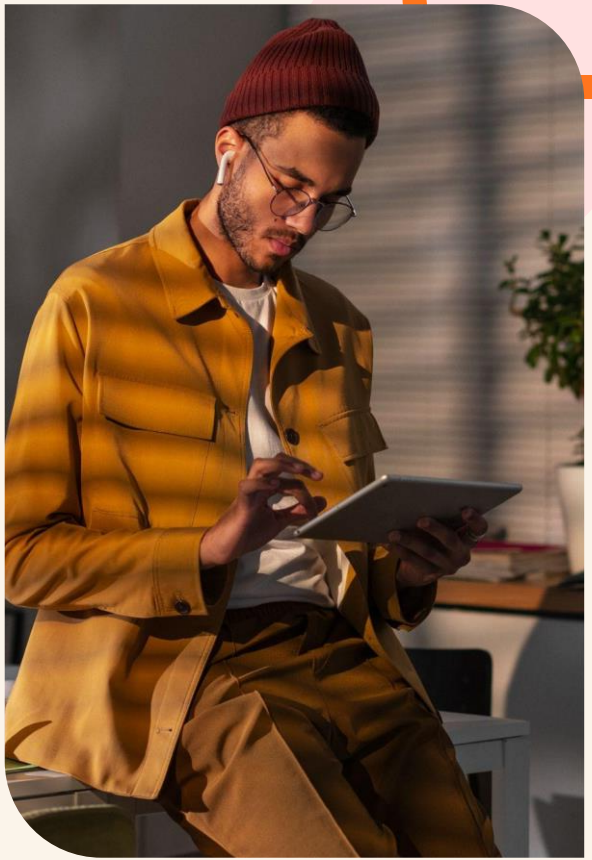
Microsoft Excel:
For Preparing Data, Analyzing Data
And Visualization

Excel Link: CLICK HERE!

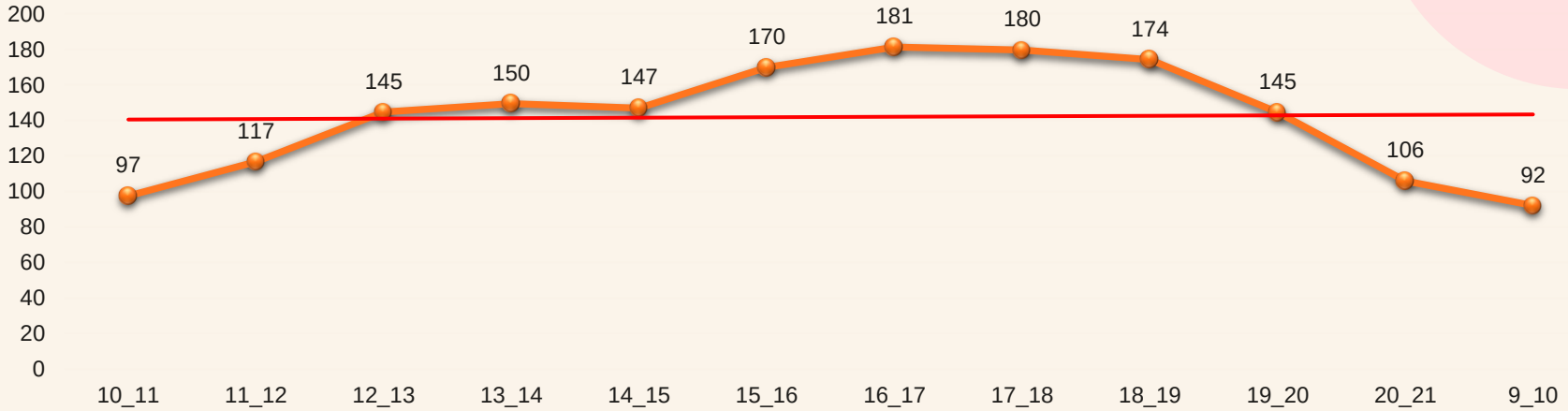
Data Analytics Tasks

- Average Call Duration
- Call Volume Analysis
- Manpower Planning
- Night Shift Manpower Planning

INSIGHTS

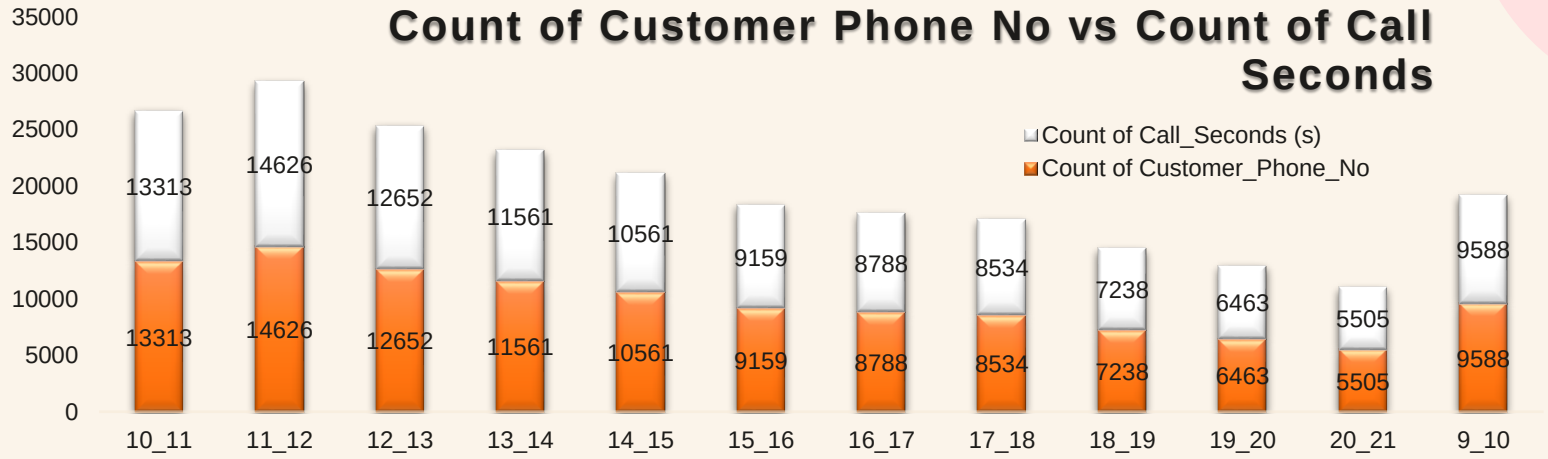


Average Call Duration



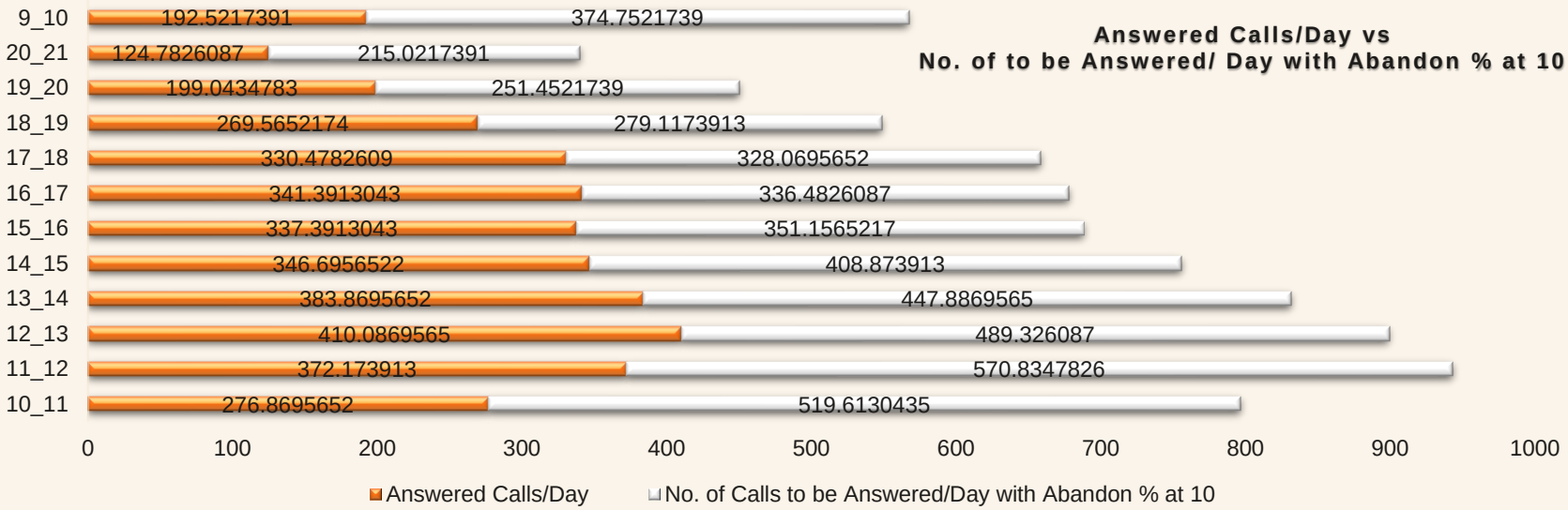
We looked at how long calls last during different times of the day. We found that, on average, calls answered by agents last about 198.6 seconds. We also saw that the longest calls happen between 10 am to 11 am and from 7 pm to 8 pm, while the shortest ones occur between 12 noon to 1 pm. This helps us know when it's busiest and quietest for handling calls during the day.

Call Volume Analysis



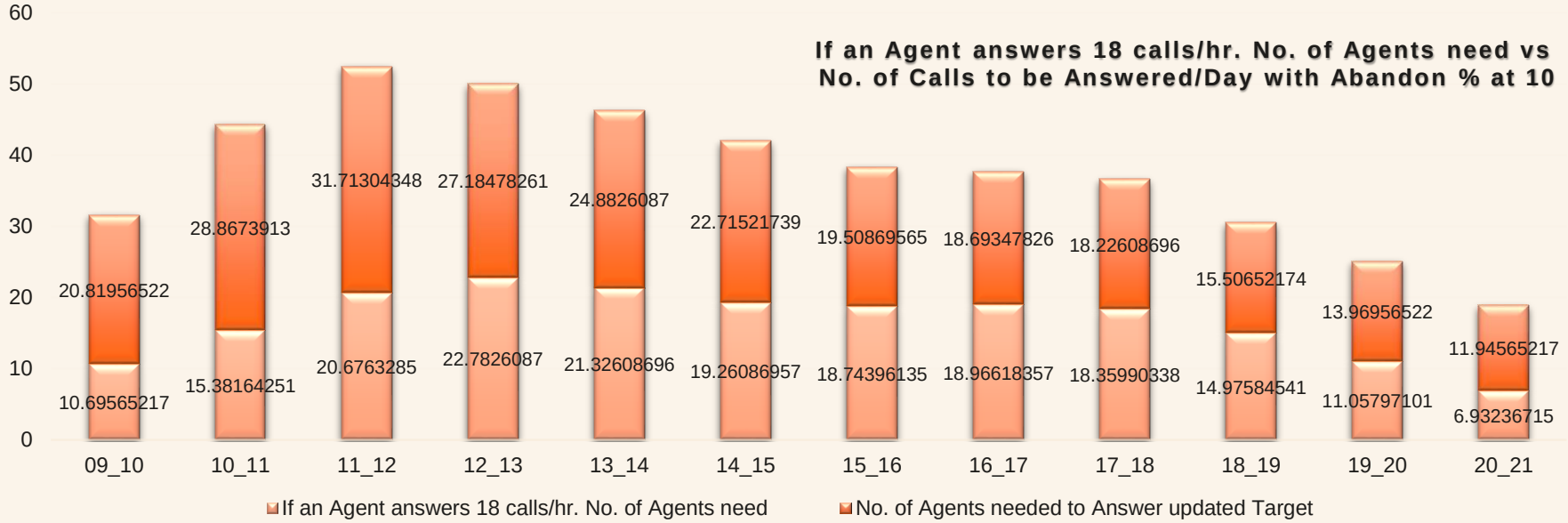
We checked when customers call throughout the day. Most calls happen between 11 am and 12 noon when customers are most engaged. But between 8 pm and 9 pm, there are fewer calls, maybe because it's dinner time or fewer staff are available. This helps companies know when to have more staff ready for calls to give better customer service.

Manpower Planning



We can see the difference between the current number of calls and the projected number of calls per day based on the updated Abandon Rate. From this comparison, we determine the required number of agents for each time period

Manpower Planning



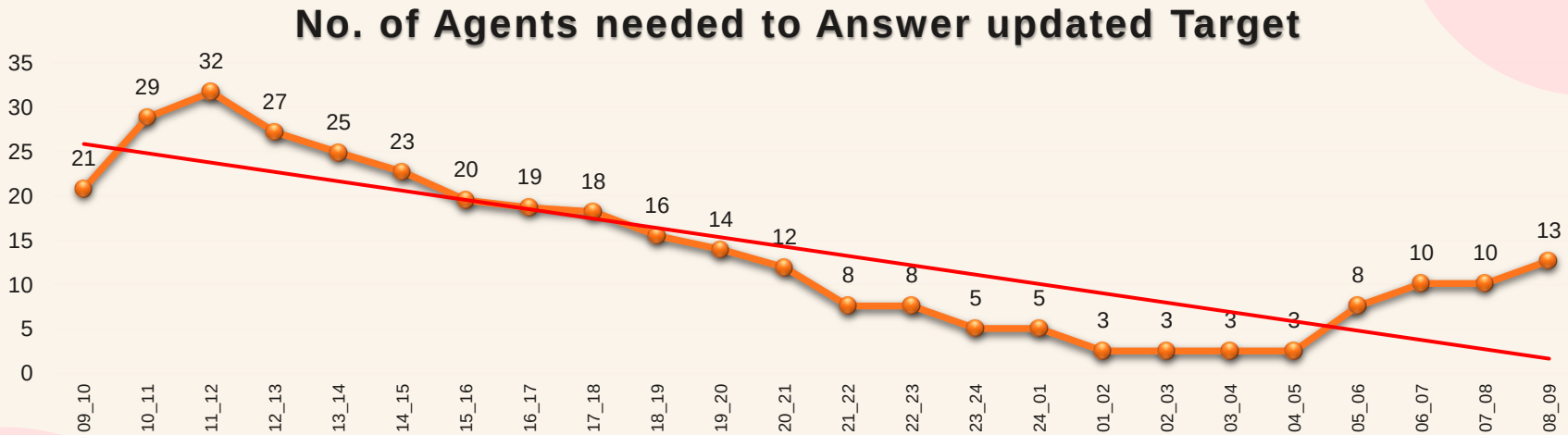
The bar chart shows the relationship between the number of calls to be answered per day and the number of agents needed to answer those calls, at different abandonment rates. The abandonment rate is the percentage of calls that are abandoned by the caller before being answered by an agent.

Night Shift Manpower Planning

There appears to be a need for more staff between 9pm and midnight, with a steady requirement of 12 employees. Then, the staffing needs decrease throughout the night, reaching a low of 4 employees between 4 am and 5 am. There is a slight bump back up to 8 employees by 6 am and then down to 2 by 7 am.

Overall, the graph suggests that staffing needs are highest early in the night shift, then taper off as the night progresses.

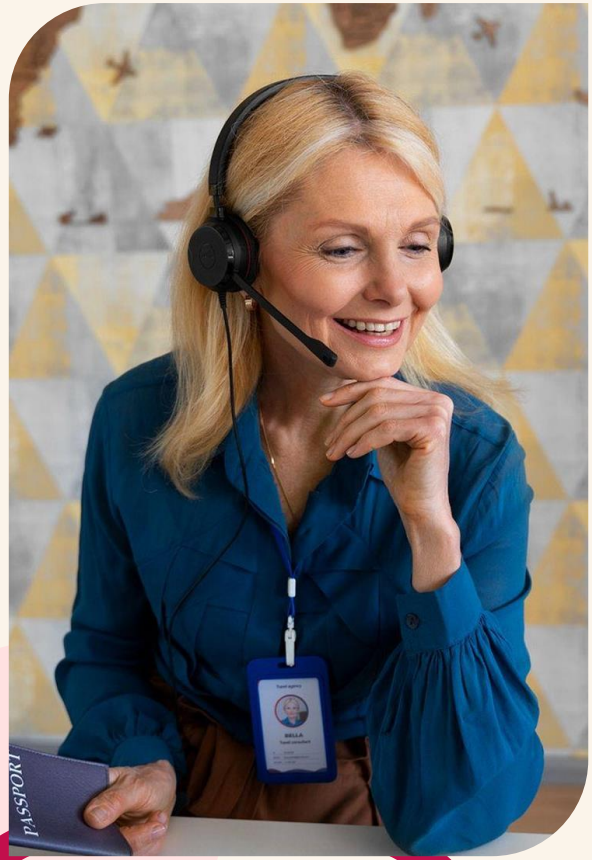
Night Shift Manpower Planning



The graph shows a general downward trend in the number of agents needed. This could be due to a number of factors, such as increased efficiency of the agents, or a decrease in the number of calls or interactions requiring an agent response.

Results

1. Fewer calls come in during the evening, so the company can reduce the number of agents then.
2. They could hire 15 agents specifically for the night shift or move some daytime workers to nights.
3. Adjusting shifts, like having some agents work early and others late, can ensure enough staff during busy times.
4. Splitting the workforce into three shifts means there's always someone available to help customers.
5. Unusual data points were found during analysis, which might change the results if removed.
6. These insights help the company use staff more efficiently, improve customer service, and be available when customers need help.



Conclusion

This project taught me some important things. First, forecasting, which means predicting what might happen in the future based on what's happening now. Then, problem-solving, which is finding solutions when things don't go as planned. Next, manpower management, which is about organizing and using staff efficiently to get things done. It was a tough project, but it helped me get better at using Excel and analyzing data. Now, I feel more comfortable working with Excel and looking at data to figure things out. Overall, it was challenging, but it made me learn and improve my skills.

Appendix

Project S.No.	Project Title	PDFs	Excels
Project 1	Data Analytics Process	Click Here	N/A
Project 2	Instagram User Analytics	Click Here	N/A
Project 3	Operation Analytics and Investigating Metric Spike	Click Here	N/A
Project 4	Hiring Process Analytics	Click Here	Click Here
Project 5	IMDB Movie Analysis	Click Here	Click Here
Project 6	Bank Loan Case Study	Click Here	Click Here
Project 7	Impact of Car Features	Click Here	Click Here
Project 8	ABC Call Volume Trend	Click Here	Click Here