DATA ANALYSIS PORTFOLIO

By

Rashidha

PROFESSIONAL BACKGROUND

I am a versatile and driven professional with a unique blend of creative design acumen and analytical data skills. Possessing a Bachelor of Science in Fashion Design from INIFD and Interior Design from Justice Basheer Ahmed Sayeed College (2018-2021), I have cultivated a strong foundation in design principles, market analysis, and spatial data interpretation.

In the realm of fashion design, I have demonstrated a keen ability to conduct comprehensive market trend analysis, effectively informing the design process and ensuring projects aligned with current consumer demands. Within interior design, I have honed skills in spatial data analysis, leveraging data-driven insights to optimize space utilization and functionality. This analytical approach to design reflects a commitment to both aesthetic appeal and practical application.

Beyond formal education I have proactively developed a robust skillset in social- media marketing Over the past three years, proficiency in Canva has been cultivated, enabling the creation of impactful social media content, website designs, and bespoke invitation cards, enhancing brand visibility and engagement.

Furthermore, I possesses advanced analytical capabilities, particularly in Excel and Power BI. This highlights a strong aptitude for data cleaning, analysis, and the extraction of meaningful insights.

I am a highly adaptable and resourceful individual, capable of seamlessly integrating creative design principles with rigorous data analysis. This unique combination of skills positions me as a valuable asset, capable of contributing to a wide range of projects and initiatives. A commitment to continuous learning and a passion for leveraging data to drive informed decisions are core to my professional ethos.

TABLE OF CONTENTS

PROFESSIONAL BACKGROUND 1

TABLE OF CONTENTS 2-3

DATA ANALYTICS PROCESS 4

- DESCRIPTION 4
- THE PROBLEM 5
- DESIGN 6-8
- CONCLUSION 8

INSTAGRAM USER ANALYTICS 9

- DESCRIPTION 9
- THE PROBLEM 10
- DESIGN 10
- FINDINGS 11-18
- ANALYSIS 18-20
- CONCLUSION 21

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE 23

- DESCRIPTION 23
- THE PROBLEM 23
- DESIGN 25
- FINDINGS 26
- ANALYSIS ..... 49 -51
- CONCLUSION 53

HIRING PROCESS ANALYTICS 54

- DESCRIPTION 54
- THE PROBLEM 55
- DESIGN 56
- FINDINGS 57
- ANALYSIS 58-59
- CONCLUSION 61

TABLE OF CONTENTS (Cont..)

IMDb MOVIE ANALYSIS 62

- DESCRIPTION 62
- THE PROBLEM 63
- DESIGN 63
- FINDINGS 64
- ANALYSIS 70-72
- CONCLUSION 74

BANK LOAN CASE STUDY 74

- DESCRIPTION 74-75
- THE PROBLEM 77
- DESIGN 78-83
- FINDINGS 83-138
- ANALYSIS 138-139
- CONCLUSION 142

ANALYZING THE IMPACT OF CAR FEATURES ON PRICE & PROFITABILITY 144

- DESCRIPTION 144
- THE PROBLEM 145
- DESIGN 146
- FINDIINGS 146
- ANALYSIS 178-179
- CONCLUSION 180

ABC CALL VOLUME TREND ANALYSIS 182

- DESCRIPTION 182
- THE PROBLEM 183
- DESIGN 183
- FINDINGS 184-197
- ANALYSIS 198-199
- CONCLUSION 202

SUMMARY OF LEARNINGS FROM ALL ABOVE PROJECTS 205

APPENDIX 206

DATA ANALYTICS PROCESS

DESCRIPTION

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

THE PROBLEM

Riya, a potential customer, is overwhelmed by the myriad of choices available in the market. She struggles to find a ring that matches her preferences, budget, and expectations. The objective is to understand her purchasing journey, identify pain points, and provide actionable insights to improve the customer experience for future buyers

DESIGN

REAL LIFE CASE SCENARIO - Riya planning to buy a necklace

Online

PLAN (Define):

Riya is to buy a most affordable Infinity Silver necklace

within a specific price range

PREPARE (DATA COLLECTED):

Riya willing to spend around the price range of 299 - 399 for her necklace through searching for the necklace online by cash on delivery

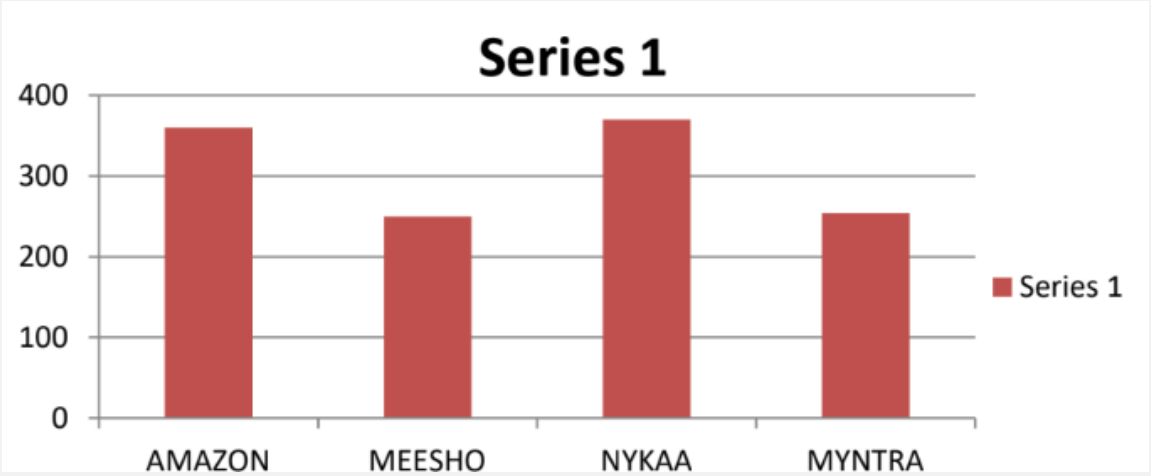
INFINITY NECKLACE

PLATFORMS	COST $
AMAZON	360
NYKAA	379
MEESHO	250
JOKER & WITCH	400
MYNTRA	254
ZAVYA	606

PROCESS (Cleared Data):

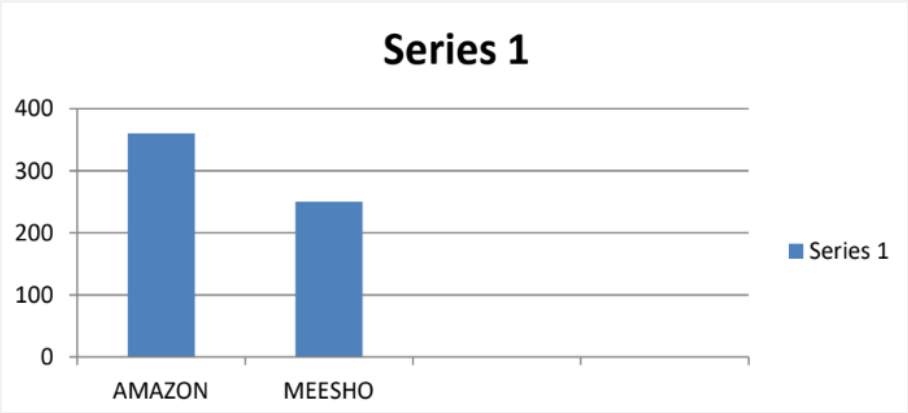
The item around the price range of Rs.299 -399

PLATFORMS	COST $
AMAZON	360
NYKAA	379
MEESHO	250
MYNTRA	254



ANALYZE (To buy):

PLATFORMS	COST $
AMAZON	360
MEESHO	250



Amazon & Meesho offer the most affordable options

SHARE:

Meesho offers the necklace at Rs.250 which is within the

desired price range

Riya finally decided to buy her infinity necklace from Meesho

for the best price

ACT :

Riya bought her necklace.

CONCLUSION

Hence, we have seen how we can use the 6 steps of Data Analytics while making any decision in real life scenarios The 6 steps used to take decisions in real life scenarios are:-

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

INSTAGRAM USER ANALYTICS

DESCRIPTION

User analysis is the process by which we track how users engage and interact with our digital product (software or mobile application) in an attempt to derive business insights for marketing, product & development teams.These insights are then used by teams across the business to launch a new marketing campaign, decide on features to build for an app, track the success of the app by measuring user engagement and improve the experience altogether while helping the business grow. You are working with the product team of Instagram and the product manager has asked you to provide insights on the questions asked by the management team.

THE PROBLEM

Instagram users and content creators often struggle to understand which posts resonate with their audience. With the vast amount of data generated from user interactions, it becomes challenging to identify trends and make data-driven decisions. This project addresses the problem by systematically analyzing user engagement metrics to provide actionable insights.

DESIGN

Steps taken to load the data into the data base

- Using the 'create db' function of MySQL create a database
- Then add tables and column names

Then add the values into them using the 'insert into'

function of MysQL

- By using the 'select' command we can query the desired output

Software used for querying the results

--> MysQL Workbench 8.0 CE

FINDINGS

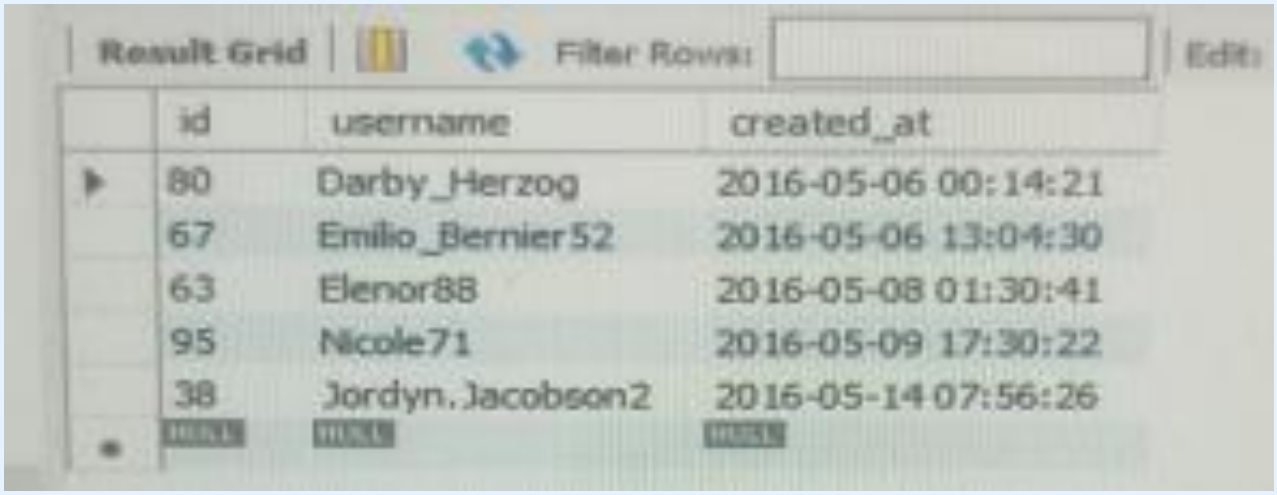
A) Marketing Analysis:

1. Loyal User Reward: Identify the five oldest users on Instagram from the provided database

QUERY:

```
1 •   SELECT *
2     FROM ig_clone.users
3     ORDER BY created_at ASC
4     LIMIT 5;
```

OUTPUT :



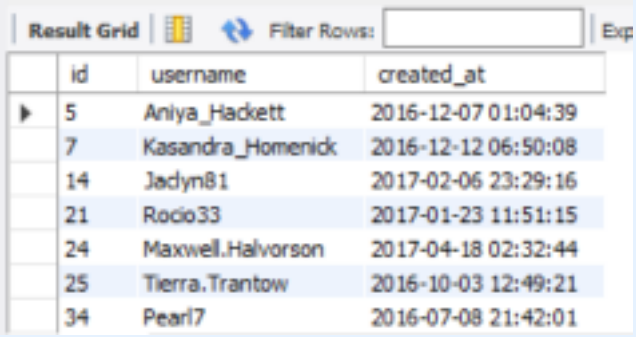
id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26

2. Inactive User Engagement: Identify users who have never posted a single photo on Instagram.

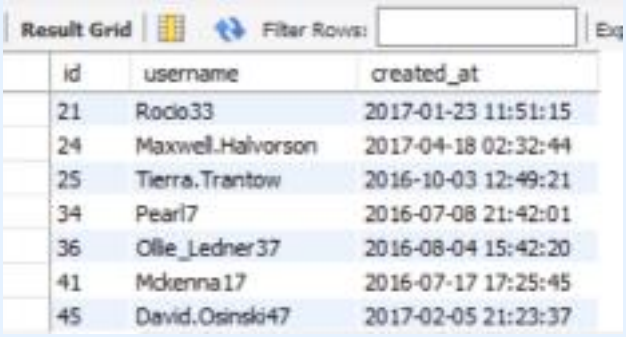
QUERY:

```
SELECT u.*
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p ON u.id
= p.user_id
WHERE p.id IS NULL;
```

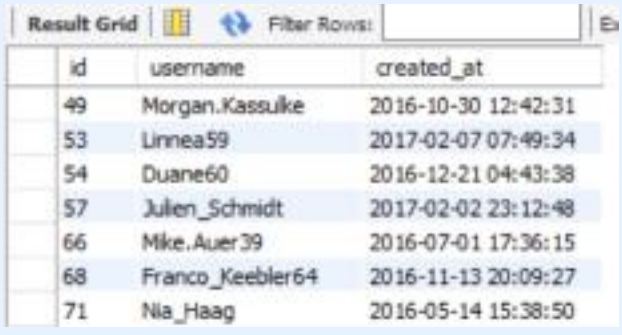
OUTPUT :



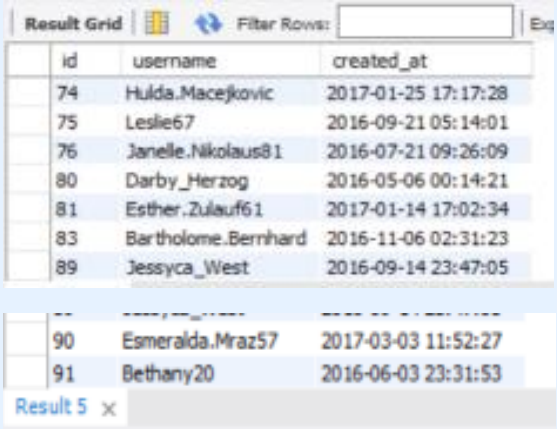
id	username	created_at
5	Aniya_Hackett	2016-12-07 01:04:39
7	Kasandra_Homenick	2016-12-12 06:50:08
14	Jadyn81	2017-02-06 23:29:16
21	Rocio33	2017-01-23 11:51:15
24	Maxwell.Halvorson	2017-04-18 02:32:44
25	Tierra.Trantow	2016-10-03 12:49:21
34	Pearl7	2016-07-08 21:42:01



id	username	created_at
21	Rocio33	2017-01-23 11:51:15
24	Maxwell.Halvorson	2017-04-18 02:32:44
25	Tierra.Trantow	2016-10-03 12:49:21
34	Pearl7	2016-07-08 21:42:01
36	Ollie_Ledner37	2016-08-04 15:42:20
41	Mckenna17	2016-07-17 17:25:45
45	David.Osinski47	2017-02-05 21:23:37



id	username	created_at
49	Morgan.Kassulke	2016-10-30 12:42:31
53	Linnea59	2017-02-07 07:49:34
54	Duane60	2016-12-21 04:43:38
57	Julien_Schmidt	2017-02-02 23:12:48
66	Mike.Auer39	2016-07-01 17:36:15
68	Franco_Keebler64	2016-11-13 20:09:27
71	Nia_Haag	2016-05-14 15:38:50



id	username	created_at
74	Hulda.Macejkovic	2017-01-25 17:17:28
75	Leslie67	2016-09-21 05:14:01
76	Janelle.Nikolaus81	2016-07-21 09:26:09
80	Darby_Herzog	2016-05-06 00:14:21
81	Esther.Zulauf61	2017-01-14 17:02:34
83	Bartholome.Bernhard	2016-11-06 02:31:23
89	Jessyca_West	2016-09-14 23:47:05
90	Esmeralda.Mraz57	2017-03-03 11:52:27
91	Bethany20	2016-06-03 23:31:53

Result 5

3.Contest Winner Declaration: Determine the winner who has most likes on the single photo.

QUERY:

```sql
WITH most_liked_image AS (
    SELECT image_url, COUNT(*) AS like_count
    FROM likes
    JOIN photos ON likes.photo_id = photos.id
    GROUP BY image_url
    ORDER BY like_count DESC
    LIMIT 1
),
most_liked_user AS (
    SELECT p.user_id, u.username, p.image_url
    FROM photos p
    JOIN users u ON p.user_id = u.id
    JOIN most_liked_image mlp ON p.image_url = mlp.image_url
)
```

```sql
SELECT *
FROM most_liked_user
JOIN photos p ON most_liked_user.user_id = p.user_id AND most_liked_user.image_url = p.image_url
JOIN photo_tags pt ON p.id = pt.photo_id
JOIN tags t ON pt.tag_id = t.id;
```

```sql
SELECT *
FROM most_liked_user;
SELECT*
FROM users
WHERE username = 'zack_kemmer93';
    SELECT
     p.image_url,
     c.comment_text,
     COUNT(f.follower_id) AS follows,
     u.username,
     p.created_dat,
     COUNT(l.photo_id) AS likes
FROM
     users as u
JOIN
     photos as p ON u.id = p.user_id
LEFT JOIN
     comments c ON p.id = c.photo_id
LEFT JOIN
     follows f ON follower.id = f.follower_id
LEFT JOIN
     likes l ON p.id = l.photo_id
WHERE
     u.username = 'zack_kemmer93'
GROUP BY
     p.image_url, c.comment_text, p.created_dat, u.username;
```

OUTPUT:

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

user_id	username	image_url
52	Zack_Kemmer93	https://jarret.name

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

image_url	comment_text	follows	username	created_dat	likes
https://jarret.name	dolor sint aut	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	laborum tempora volup...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	tempore accusamus ve...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	ducimus sit maxime	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	labore quas inventore	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	molestias temporibus c...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	odit ad explicabo	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	et omnis ad	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	molestias vel odit	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752

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

image_url	comment_text	follows	username	created_dat	likes
https://jarret.name	est molestiae dolorem	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	vitae occaecati laborios...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	assumenda placeat offi...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	ipsam nobis est	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	ut veritatis reprehenderit	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	perspiciatis et et	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	voluptatem aut ipsa	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	ad impedit ducimus	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	cumque aut omnis	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752

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

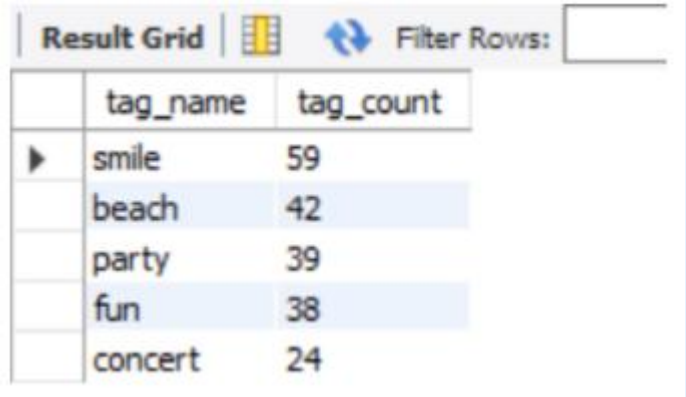
image_url	comment_text	follows	username	created_dat	likes
https://jarret.name	ad id repudiandae	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	incidunt officiis eos	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	eveniet perspiciatis rep...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	rerum aperiam beatae	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	molestiae sapiente est	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	et qui et	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	at mollitia soluta	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752
https://jarret.name	laudantium inventore i...	4752	Zack_Kemmer93	2024-12-13 14:57:03	4752

4.Hashtag Research: Identify and suggest the top five most commonly used hashtags on the Platform.

QUERY:

```
114 •   use ig_clone;
115 •   SELECT  t.tag_name,COUNT(*) AS
116     tag_count
117     FROM ig_clone.tags t
118     JOIN ig_clone.photo_tags pt ON t.id = pt.tag_id
119     GROUP BY t.tag_name
120     ORDER BY tag_count DESC
121     LIMIT 5;
```

OUTPUT:



Result Grid | Filter Rows:

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

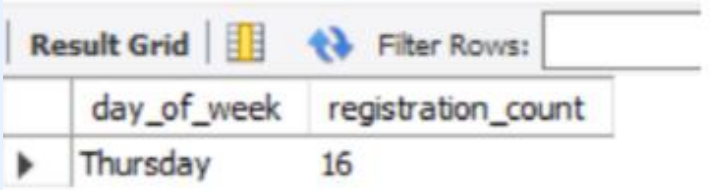
5.Ad Campaign Launch:

Determine the day of the week when most users register on Instagram.Provide insights on when to schedule an ad campaign.

QUERY:

```
use ig_clone;
SELECT DAYNAME(created_at)
AS day_of_week, COUNT(*) AS
registration_count
FROM ig_clone.users
GROUP BY day_of_week
ORDER BY registration_count DESC
LIMIT 1;
```

OUTPUT :



Result Grid | Filter Rows:

day_of_week	registration_count
Thursday	16

B) Investor Metrics:

1.User Engagement: Calculate the average number of posts per user on Instagram. Also, provide the total number of photos on Instagram divided by the total number of users.

QUERY :

```
SELECT
    (SELECT COUNT(*) FROM
ig_clone.photos)/(SELECT
COUNT(*) FROM ig_clone.users) AS
avg_posts_per_user,
    (SELECT COUNT(*) FROM
ig_clone.photos)/(SELECT
COUNT(*) FROM ig_clone.users) AS
avg_photos_per_user_alt
FROM dual;
```

OUTPUT :

Result Grid | Filter Rows: | Expc

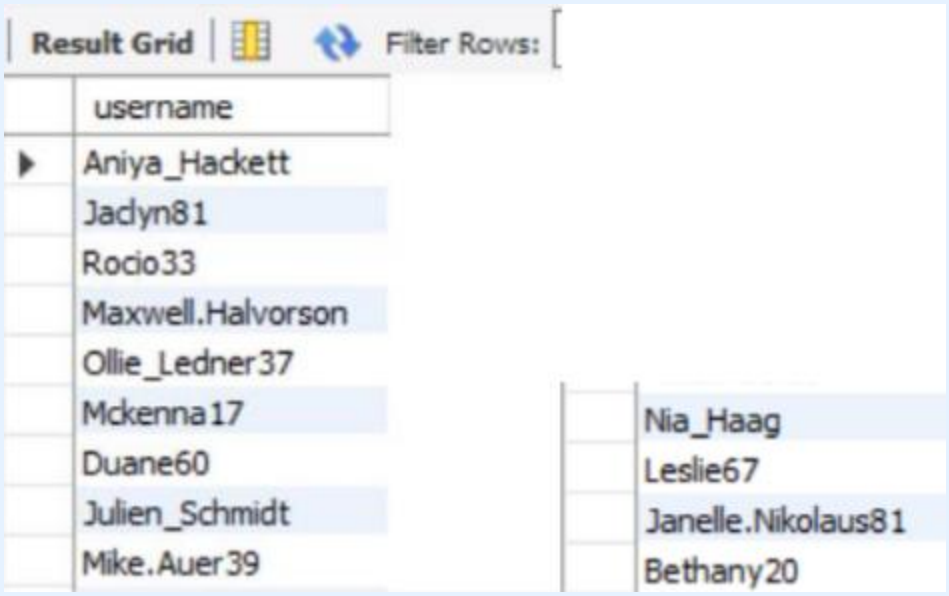
avg_posts_per_user	avg_photos_per_user_alt
2.5700	2.5700

2.Bots & Fake Accounts: Identify users (potential bots) who have liked every single photo on the site, as this is not typically possible for a normal user.

QUERY:

```
SELECT u.username
FROM ig_clone.users u
JOIN ig_clone.likes l ON u.id= l.user_id
GROUP BY u.id
HAVING COUNT(DISTINCT l.photo_id) = (SELECT COUNT(*) FROM ig_clone.photos);
```

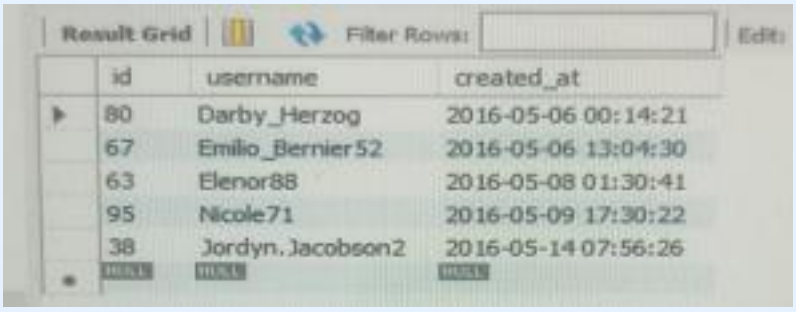
OUTPUT :



ANALYSIS

After performing the analysis I have the following points:-

1. The most loyal users i.e. the top 5 oldest users are:



2.Out of the 100 total users there are 26 users who are inactive and they have never posted any kind of stuff of Instagram may it be any photo,video or any type of text. So, the Marketing team of Instagram needs to re mind such inactive users

3.So, the user named Zack_Kemmer93 with user_id 52 is the winner of the contest cause his photo with photo_id 145 has the highest number of likes i.e. 48

4.The top 5 most commonly used #hashtags along with the total Count are smile(59), beach(42), party(39), fun(38) and concert(24)

5. Most of the users registered on Thursday and Sunday i.e. 16 and hence it would prove beneficial to start AD Campaign on these two days

6.So, there are in total 257 rows i.e. 257 photos in the photos table and 100 rows i.e. 100 ids in the users table which makes the desired output to be 257/100 = 2.57 (avg. users posts on Instagram)

7. Out of the total user id's there are 13 such user id's who have liked each and every post on Instagram (which is not practically possible) and so such user id's are considered as BOTS and Fake Accounts

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

1.Why did the Marketing team wanted to know the most inactive users?

--> So, they can reach out to those users via mail and ask them What's keeping them away from using the Instagram.

2.Why did the Marketing team wanted to know the top 5 #hashtags used?

--> May be the tech team wanted to add some filter features for photos and videos posted using the top 5 mentioned #hashtags

3.Why did the Marketing team wanted to know on which day of the week the platform had the most new users registered?

--> So, that they can run more Ads of various brands during such days and also get profit from it

4.Why did the Investors wanted to know about the average posts per user has on Instagram?

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

5.Why did the Investors wanted to know the count of BOTS an Fake accounts if any?

--> So that the Investors are assured that they are investing into an Asset and not a Future Liability

CONCLUSION

In conclusion, I would like to conclude that not only Instagram but many other social media and commercial firms use such Analysis to find the insights from their

customer data which in turn help the firms to find the customers who will be an Asset to the firm in the future and not some Liability.

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

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE

DESCRIPTION

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they collect. Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operationanalytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc.

Questions like these must be answered daily and for that its very important to investigate metric spike.

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

THE PROBLEM

Organizations often encounter unexpected spikes in operational metrics, such as increased processing times, higher error rates, or surges in customer inquiries. These spikes can disrupt business processes and affect overall performance. The challenge is to pinpoint the root causes of these anomalies and implement effective solutions to mitigate their impact. This project addresses the problem by systematically analyzing the data and identifying key drivers of the metric spike.

DESIGN

Steps taken to load the data into the data base

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

Software used for querying the results

--> MySQL Workbench 8.0 CE

Software used for analyzing using Bar plots

--> Microsoft Excel

FINDINGS

Case Study 1: Job Data Analysis

Tasks:

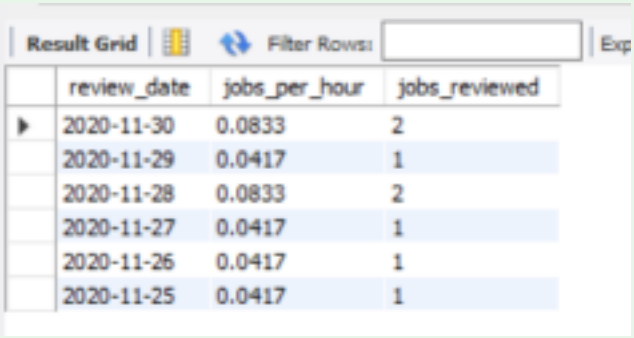
A.Jobs Reviewed Over Time:

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

QUERY:

```
SELECT
STR_TO_DATE(ds, '%m/%d/%Y') AS review_date,
((Count(*) /24) )    As jobs_per_hour,
COUNT(*) AS jobs_reviewed
FROM job_data
GROUP BY
review_date;
```

OUTPUT :



Result Grid | Filter Rows: | Exp

review_date	jobs_per_hour	jobs_reviewed
2020-11-30	0.0833	2
2020-11-29	0.0417	1
2020-11-28	0.0833	2
2020-11-27	0.0417	1
2020-11-26	0.0417	1
2020-11-25	0.0417	1

Insights

Consistent Activity: If the number of jobs reviewed per hour is relatively consistent throughout the day, it suggests a steady workflow and balanced workload distribution.

Peak Hours: Identifying specific hours with higher job review activity can help optimize staffing and resource allocation. For example, if most reviews happen between 10 AM and 2 PM, you might need more staff during these hours.

Daily Variations: If certain days show significantly higher or lower activity, it might indicate patterns related to work schedules, deadlines, or external factors. For instance, higher activity on Mondays might suggest a backlog from the weekend.

Actionable Steps

Optimize Staffing: Use the insights on peak hours to ensure adequate staffing during high activity periods to maintain efficiency and reduce bottlenecks.

Balance Workload: If certain hours or days are consistently busier, consider redistributing tasks to balance the workload more evenly across the week.

Investigate Anomalies: If there are unexpected spikes or drops in activity, investigate the underlying causes. This could involve looking at specific events, system issues, or changes in workflow.

Interpretation

High Activity Hours: If you notice that the number of jobs reviewed peaks during specific hours, it indicates that these are the busiest times. Ensure that you have enough resources available during these period

Low Activity Periods: If there are hours with significantly lower activity, it might be an opportunity to schedule maintenance or training sessions without disrupting the workflow.

Weekly Patterns: If certain days consistently show higher activity, it might be related to weekly cycles, such as preparing for deadlines or handling end-of-week tasks. Understanding these patterns can help in planning and scheduling.

By analyzing these metrics and making data-driven adjustments, you can improve the efficiency and effectiveness of your job review process.

B.Throughput Analysis:

2.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.

QUERY:

```sql
SELECT
    STR_TO_DATE(ds, '%m/%d/%Y') AS date_z,
    COUNT(*) AS events_per_second,
    SUM(time_spent) AS total_time_spent_per_day,
    COUNT(*) / TIMESTAMPDIFF(SECOND, MIN(ds), MAX(ds)) AS events_per_second,
    AVG(COUNT(*)) OVER (ORDER BY DATE(ds) ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_avg_events_per_day,
    AVG(SUM(time_spent)) OVER (ORDER BY DATE(ds) ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS rolling_avg_time_spent
FROM
    job_data
WHERE
    STR_TO_DATE(ds, '%m/%d/%Y') BETWEEN '2020-11-25' AND '2020-11-30'
GROUP BY
    date_z,ds
ORDER BY
    date_z;
```

OUTPUT:

date_z	events_per_second	total_time_spent_per_day
2020-11-25	1	45
2020-11-26	1	56
2020-11-27	1	104
2020-11-28	2	33
2020-11-29	1	20
2020-11-30	2	40

rolling_avg_events_per_day	rolling_avg_time_spent
1.3333	49.6667
1.4000	50.6000
1.5000	49.2500
1.6667	31.0000
1.5000	30.0000
2.0000	40.0000

For many real-world applications, I'd generally prefer using the 7-day rolling average for throughput over the daily metric.

1.Smooths Out Daily Fluctuations: Daily throughput can be heavily influenced by various factors such as time of day, day of the week, and even external events. These fluctuations can make it difficult to identify underlying trends and potential issues. The 7- day rolling average smooths out these daily variations, providing a more stable and consistent view of throughput over time

2.Highlights Long-Term Trends: By averaging throughput over a 7- day window, the 7-day rolling average helps to identify longer- term trends, such as gradual increases or decreases in throughput, seasonal patterns, or the impact of system upgrades or changes.

3. Easier to Spot Sustained Issues: If there is a consistent decline in the 7-day rolling average, it indicates a potential performance issue that requires further investigation. In contrast, daily fluctuations might mask a sustained decline in performance.

4. Improved Decision Making: By providing a more stable and consistent view of throughput, the 7-day rolling average can help with better decision-making related to system capacity planning, resource allocation, and performance optimization. However, it's important to note that the choice between the daily metric and the 7-day rolling average depends on the specific use case and the desired level of detail. If you need to identify and address short-term issues or investigate daily fluctuations, the daily metric might be more appropriate.

Insights
Smoothing Out Fluctuations: The 7-day rolling average helps smooth out daily fluctuations, providing a clearer view of long-term trends. This is useful for identifying consistent patterns in throughput.

Identifying Trends: By looking at the rolling average, you can identify whether throughput is generally increasing, decreasing, or remaining stable over time. This can help you understand the overall performance of your system or process.

Detecting Anomalies: Significant deviations from the rolling average can indicate anomalies. For example, a sudden spike might suggest a surge in demand, while a drop could indicate a system issue or reduced user activity.

Actionable Steps

Investigate Anomalies: If you notice significant spikes or drops, investigate the underlying causes. This could involve looking at system logs, user activity, or external factors that might have influenced throughput.

Optimize Performance: Use the insights from the rolling average to optimize system performance. For example, if you notice consistent periods of high throughput, ensure your system can handle the load during these times.

Plan for Capacity: Understanding long-term trends can help you plan for future capacity needs. If throughput is steadily increasing, you might need to scale up your infrastructure to accommodate growth.

Interpretation

Increasing Trend: If the 7-day rolling average shows a steady increase in throughput, it suggests that your system or service is experiencing growing usage. This is a positive sign, but you should ensure your infrastructure can handle the increased load.

Decreasing Trend: A steady decrease in the rolling average might indicate declining user engagement or potential issues with your system. Investigate the reasons behind the decline and take corrective actions.

Stable Trend: If the rolling average remains relatively stable, it suggests consistent performance. This is generally a good sign, but you should still monitor for any sudden changes.

Spikes and Dips: Sudden spikes might indicate successful marketing campaigns or seasonal demand, while dips could suggest outages or reduced user activity. Understanding these patterns can help you respond more effectively.

By regularly monitoring the 7-day rolling average of throughput and making data-driven adjustments, you can improve system performance and user satisfaction

C.Language share Analysis:

3.Write an SQL query to calculate the percentage share of each language over the last 30 days.

QUERY:

```
SELECT
    language,
    STR_TO_DATE(ds, '%m/%d/%Y') AS date_z,
    ROUND(100.0*COUNT(*)/
SUM(COUNT(*)) OVER (),2)AS
percentage_share
FROM
    job_data
WHERE
   STR_TO_DATE(ds, '%m/%d/%Y') >=
 STR_TO_DATE(ds, '%m/%d/%Y') BETWEEN '2020-11-25' AND '2020-11-30' 'CURDATE(), INTERVAL 30 DAY'
GROUP BY
    language,ds;
```

OUTPUT:

Result Grid | Filter Rows: | Expc

language	date_z	percentage_share
English	2020-11-30	12.50
Arabic	2020-11-30	12.50
Persian	2020-11-29	12.50
Persian	2020-11-28	12.50
Hindi	2020-11-28	12.50
French	2020-11-27	12.50
Persian	2020-11-26	12.50
Italian	2020-11-25	12.50

Insights

Dominant Languages: Identify which languages have the highest percentage share. This indicates the primary languages your users prefer.

Emerging Languages: Look for languages that are increasing in share. This can indicate growing user bases in new regions or demographics.

Declining Languages: Identify languages that are decreasing in share. This might suggest a shift in user preferences or potential issues with content in those languages.

Actionable Steps

Content Localization: For languages with high or growing shares, consider investing in more localized content and features to cater to these users.

User Support: Ensure that customer support is available in the most popular languages to enhance user satisfaction.

Marketing Strategies: Tailor marketing campaigns to target regions or demographics where certain languages are dominant or emerging.

Analyze Declines: Investigate why certain languages are declining in share. This could involve user feedback, content quality, or external factors.

Interpretation

High Percentage Share: If a language like English has a high percentage share, it indicates that a significant portion of your user base prefers English. This might be expected if your primary market is English-speaking countries.

Emerging Languages: If you notice an increase in the share of a language like Spanish, it suggests that your product is gaining traction in Spanish-speaking regions. Consider enhancing your Spanish content and marketing efforts in these areas.

Declining Languages: A decrease in the share of a language like French might indicate that French-speaking users are not as engaged. Investigate potential reasons, such as content relevance or user experience issues.

By regularly monitoring these metrics and making data-driven adjustments, you can better cater to your diverse user base and improve overall engagement.

D.Duplicate Rows Detection:

4. Write an SQL query to display duplicate rows from the job_data table.

QUERY:

```
SELECT
     job_id,
     COUNT(*) AS count
FROM
   job_data
GROUP BY
    job_id
HAVING
     count > 1;
 SELECT
      actor_id,
      COUNT(*) AS count
 FROM
    job_data
 GROUP BY
     actor_id
 HAVING
```

```
        count > 1;
SELECT
        event,
        COUNT(*) AS count
 FROM
    job_data
 GROUP BY
        event
 HAVING
        count > 1;
SELECT
        language,
        COUNT(*) AS count
 FROM
    job_data
```

```
 FROM
    job_data
GROUP BY
     language
HAVING
    count > 1;
SELECT
    org,
     COUNT(*) AS count
FROM
    job_data
GROUP BY
     org
HAVING
    count > 1;
SELECT
    ds,
```

```
     COUNT(*) AS count
FROM
    job_data
GROUP BY
     ds
HAVING
    count > 1;
```

OUTPUT :

Result Grid | Filter R

actor_id	count
1003	2

Result Grid | Filter Ro

event	count
skip	2
transfer	3
decision	3

Result Grid | Fil

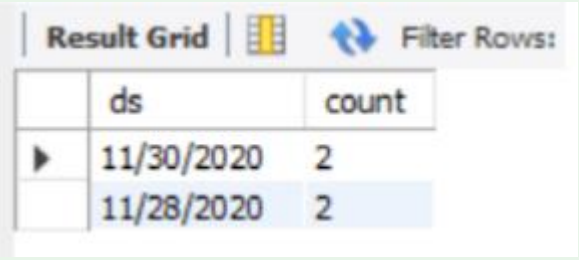
job_id	count
23	3

Result Grid | Filter

language	count
Persian	3

Result Grid | Filter Row

org	count
A	2
B	2
C	2
D	2



Insights

Data Quality Issues: The presence of duplicate rows often indicates data quality issues. This can arise from multiple sources, such as data entry errors, system glitches, or improper data integration processes.

Impact on Analysis: Duplicate rows can skew analysis results, leading to inaccurate insights and decisions. For example, metrics like average salary, job counts, or other aggregations might be inflated or deflated due to duplicates.

System Performance: Duplicate data can increase the size of your database unnecessarily,leading to slower query performance and higher storage costs.

Causes

Data Entry Errors: Manual entry of data can lead to duplicates if the same job information is entered multiple times.

Integration Issues: When merging data from different sources, duplicates can occur if there are no proper deduplication rules in place.

Lack of Unique Constraints: If the table lacks unique constraints or primary keys, it becomes easier for duplicate rows to be inserted.

Actionable Steps

Implement Unique Constraints: Ensure that your table has appropriate unique constraints or primary keys to prevent duplicates from being inserted in the future.

Improve Data Entry Processes: Implement validation checks during data entry to prevent duplicates. This can include using forms with built-in validation or automated scripts that check for existing records before inserting new ones.

Regular Data Audits: Schedule regular audits of your data to identify and address duplicates and other data quality issues promptly.

Interpretation

High Number of Duplicates: If you find a high number of duplicate rows, it suggests significant data quality issues that need immediate attention. Investigate the sources of these duplicates and implement measures to prevent them.

Patterns in Duplicates: Analyze the patterns in duplicate rows. For example, if duplicates are more common for certain job titles or locations, it might indicate specific areas where data entry or integration processes need improvement.

By addressing these issues, you can improve the accuracy and reliability of your data, leading to better analysis and decision-making.

Case Study 2: Investigating Metric Spike

Tasks:

A.Weekly User Engagement:

1. Write an SQL query to calculate the weekly user engagement.

QUERY:

```
127  WITH WeeklyuserEngagement AS (
128      SELECT
129          user_id,
130          company_id,
131          language,
132          DATE_FORMAT('Week',
133  activated_at) AS week_start,
134          SUM(CASE WHEN state = 'active'
135   AND activated_at <= created_at + INTERVAL 7 day THEN 1 ELSE 0 END ) AS WEEKLY_ENGAGEMENT
136     FROM
137         users)
```

```sql
SELECT
    user_id,
    company_id,
    language,
     DATE_FORMAT('Week',
activated_at) AS week_start,
SUM(CASE WHEN state = 'active'
 AND activated_at <= created_at + INTERVAL 7 day THEN 1 ELSE 0 END) AS weekly_engagement
FROM
    users
GROUP BY
    user_id,
    company_id,
    language,
    DATE_FORMAT('Week',
  activated_at)
ORDER BY
   user_id,
   company_id,
   language,
   DATE_FORMAT('Week',
  activated_at);
```

OUTPUT:

	user_id	company_id	language	week_start	weekly_engagement
2	0	5737	english	NULL	0
3	3	2800	german	NULL	0
4	4	5110	indian	NULL	0
5	6	11699	english	NULL	0
6	7	4765	french	NULL	0
7	8	2698	french	NULL	0
8	11	3745	english	NULL	0
9	13	4025	english	NULL	0
10	15	4259	english	NULL	0
11	17	5025	japanese	NULL	0
12	19	326	english	NULL	0
13	20	7	italian	NULL	0
14	21	2606	english	NULL	0
15	22	545	german	NULL	0
16	27	6	japanese	NULL	0
17	30	4148	english	NULL	0
18	31	39	arabic	NULL	0
19	33	10768	english	NULL	0

Full output Data in the below link:

:https://docs.google.com/spreadsheets/d/14KU751E9zu_Ngim7n7PG_VRQ7jTZ2f1o/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Insights

Consistent Engagement: If you see a steady number of weekly active users, it indicates that your product maintains user interest and engagement over time.

Engagement Spikes: Sudden increases in engagement can be linked to specific events such as marketing campaigns, new feature releases, or seasonal trends. Analyzing these spikes can help you understand what drives user engagement.

Retention vs. Churn: High retention rates and low churn rates suggest strong user satisfaction and loyalty. Conversely, high churn rates may indicate issues with user experience or competition.

Actionable Steps

Enhance User Experience: If engagement drops, consider improving the user interface, adding new features, or addressing user feedback to enhance the overall experience.

Targeted Campaigns: Use insights from periods of high engagement to design targeted marketing campaigns that attract and retain users.

User Engagement Strategies: Implement strategies to keep users engaged, such as personalized content, regular updates, and incentives for continued use.

Interpretation

Weekly Active Users (WAU): A high WAU indicates that your product is engaging a significant number of users on a weekly basis. If WAU is low, it might suggest that users are not finding enough value to return regularly.

Engagement Frequency: If users are interacting with your product multiple times a week, it suggests high engagement and dependency on your product. Low frequency might indicate that users are not fully utilizing your product's features.

Retention Rate: High retention rates indicate that users find long-term value in your product. If retention drops after a certain period, it might indicate that users are losing interest or facing issues.

Churn Rate: A high churn rate suggests that users are leaving your product. Investigate the reasons behind this and address any issues to improve retention. By regularly monitoring these metrics and making data-driven adjustments, you can improve user engagement and overall product success.

B.User Growth Analysis:

2. Write an SQL query to calculate the user growth for the product.

QUERY:

```
89  WITH Monthlyusercounts AS (
90      SELECT
91          DATE_FORMAT(occured_at, '%Y-%m-01') AS month,
92          device,
93          user_type,
94          user_id,
95          COUNT(DISTINCT user_id) AS user_count
96      FROM
97          events
98      GROUP BY
99          DATE_FORMAT(occured_at, '%Y-%m-01'), device, user_type,user_id
100 ),
```

```
101  Laggedcounts AS (
102      SELECT
103          month,
104          device,
105          user_type,
106          user_count,
107          user_id,
108          LAG(user_count) OVER (PARTITION BY device, user_type ORDER BY month) AS prev_month_count
109      FROM
110          Monthlyusercounts
111  )
```

```
112  SELECT
113      month,
114      device,
115      user_type,
116      user_id,
117      user_count,
118      prev_month_count,
119      ROUND(((user_count - prev_month_count) / NULLIF(prev_month_count, 0)) * 100, 2) AS growth_percentage
120  FROM
121      LaggedCounts
122    ORDER BY
123        device,
124        user_type,
125        user_id,
126        month;
```

OUTPUT:

	Month	Device	user_type	user_id	user_count	prev_month_count	growth_percentage
2	6/1/2014	acer aspire desktop	1	83	1	1	0
3	5/1/2014	acer aspire desktop	1	227	1		
4	6/1/2014	acer aspire desktop	1	566	1	1	0
5	8/1/2014	acer aspire desktop	1	899	1	1	0
6	5/1/2014	acer aspire desktop	1	1311	1	1	0
7	6/1/2014	acer aspire desktop	1	1311	1	1	0
8	7/1/2014	acer aspire desktop	1	1311	1	1	0
9	8/1/2014	acer aspire desktop	1	1311	1	1	0
10	6/1/2014	acer aspire desktop	1	1543	1	1	0
11	7/1/2014	acer aspire desktop	1	1543	1	1	0
12	8/1/2014	acer aspire desktop	1	1543	1	1	0
13	5/1/2014	acer aspire desktop	1	1554	1	1	0
14	6/1/2014	acer aspire desktop	1	1554	1	1	0
15	8/1/2014	acer aspire desktop	1	1646	1	1	0
16	5/1/2014	acer aspire desktop	1	1979	1	1	0

Full output data in the below link :

https://docs.google.com/spreadsheets/d/19Ohy3lHbmS56QNUy6hlljV_ksvVo-8bd/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Insights

Consistent Growth: If you see a steady increase in new and active users, it indicates that your marketing and user acquisition strategies are effective. It also suggests that your product is meeting user needs and expectations.

Spikes in Growth: Sudden spikes in user growth can be linked to specific events such as marketing campaigns, product launches, or seasonal trends. Analyzing these spikes can help you understand what drives user acquisition.

Retention vs. Churn: High retention rates and low churn rates indicate strong user satisfaction and loyalty. Conversely, high churn rates may suggest issues with user experience, product value, or competition.

Actionable Steps

Enhance Marketing Strategies: If certain campaigns or channels are driving significant growth, consider investing more in those areas.

Improve Onboarding: If you notice a drop in user retention, focus on improving the onboarding process to ensure new users understand the value of your product.

Engage Users: Implement strategies to keep users engaged, such as personalized content, regular updates, and incentives for continued use.

Analyze Feedback: Collect and analyze user feedback to identify pain points and areas for improvement.

Interpretation

New User Growth: If you see a high number of new users in a particular month, investigate what marketing activities or product changes occurred during that time. This can help you replicate successful strategies.

Active User Growth: A steady increase in active users suggests that your product is retaining users well. If growth is stagnant, consider enhancing features or user engagement strategies.

Retention Rate: High retention rates indicate that users find long-term value in your product. If retention drops after a certain period, it might indicate that users are losing interest or facing issues.

By regularly monitoring these metrics and making data-driven adjustments, you can improve user growth and overall product success.

C.Weekly Retention Analysis:

3. Write an SQL query to calculate the weekly retention of users based on their sign-up cohort.

QUERY:

```
304 •    SELECT DISTINCT user_id
305      FROM users
306      WHERE activated_at < NOW() -INTERVAL 7 day
307      AND user_id IN (
308      SELECT DISTINCT user_id
309      FROM users
310      WHERE activated_at IS NOT NULL
311      );
312 •    SELECT
313      (SELECT COUNT(DISTINCT user_id)
314      FROM users
315      WHERE activated_at >= NOW() - INTERVAL 7 DAY)*1.0/(SELECT COUNT(*) FROM users)AS active_user_ratio;
```

```sql
    SELECT
      DATE(created_at) AS sign_up_date,
      COUNT(DISTINCT user_id) AS total_users,
      COUNT(DISTINCT CASE
          WHEN activated_at >= NOW() - INTERVAL 7 DAY THEN user_id
      END) AS active_users,
      COUNT(DISTINCT CASE
          WHEN activated_at >= NOW() - INTERVAL 7 DAY THEN user_id
      END) * 1.0 / COUNT(DISTINCT user_id) AS active_user_ratio
  FROM
      users
  GROUP BY
      DATE(created_at)
  ORDER BY
      sign_up_date;
```

OUTPUT:

user_id
0
3
4
6
7
8
11
13
15

Result Grid | Filter Rows: | Export:

sign_up_date	total_users	active_users	active_user_ratio
2013-01-01	7	0	0.00000
2013-01-02	7	0	0.00000
2013-01-03	6	0	0.00000
2013-01-04	1	0	0.00000
2013-01-05	2	0	0.00000
2013-01-06	3	0	0.00000
2013-01-07	4	0	0.00000

Full output data in the below link :

https://docs.google.com/spreadsheets/d/1nLI1O8RNwLBBR0ELs9JEwjm3VO_887jx/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Insights

Retention Trends: Look for patterns in retention rates over time. For example, if retention rates drop significantly after the first week, it might indicate issues with user onboarding or initial engagement.

Cohort Performance: Compare different cohorts to see if certain groups of users (e.g., those who signed up during a specific marketing campaign) have higher or lower retention rates. This can help you identify successful strategies or areas for improvement.

Long-Term Retention: Identify how long users typically stay active. If retention rates stabilize after a certain period, it might indicate a core group of loyal users.

Actionable Steps

Improve Onboarding: If you notice a significant drop in retention after the first week, consider enhancing your onboarding process to better engage new users.

Targeted Campaigns: Use insights from high-performing cohorts to design targeted marketing campaigns that attract and retain similar users.

User Engagement: Implement strategies to keep users engaged over time, such as personalized content, regular updates, and incentives for continued use.

Interpretation

Week 1 Retention: If you see a high retention rate in the first week, it suggests that your initial engagement strategies are effective. However, if there's a sharp drop-off in subsequent weeks, it might indicate that users are not finding long-term value in your product.

Cohort Differences: If a particular cohort (e.g., users who signed up during a specific promotion) shows higher retention rates, analyze what was different about that cohort's experience. This could provide insights into what drives user engagement and retention.

Stabilization Point: If retention rates stabilize after a certain number of weeks, it indicates that you've identified a core group of loyal users. Focus on understanding their behavior and preferences to further enhance their experience.

By regularly monitoring these metrics and making data-driven adjustments, you can improve user retention and overall engagement.

D.Weekly Engagement Per Device:

4. Write an SQL query to calculate the weekly engagement per device

QUERY:

```
331 •   SELECT
332        EXTRACT(year FROM occured_at)
333     AS year,
334        EXTRACT(WEEK FROM occured_at)
335     AS week,
336       device,
337       COUNT(DISTINCT user_id) AS active_users
338   FROM
339       events
340   WHERE
341       event_type = 'engagement'
342   GROUP BY
343         year,week,device
344   ORDER BY
345         year,week,device;
```

OUTPUT:

Result Grid | Filter Rows: | Exp

year	week	device	active_users
2014	17	acer aspire desktop	9
2014	17	acer aspire notebook	20
2014	17	amazon fire phone	4
2014	17	asus chromebook	21
2014	17	dell inspiron desktop	18
2014	17	dell inspiron notebook	46
2014	17	hp pavilion desktop	14
2014	17	htc one	16

Full output data in the below link :

https://docs.google.com/spreadsheets/d/1WBg4y_MJOd0avvy-XQjsOEghqyuFlpvR/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Insights

High Engagement on Specific Devices: If you see consistently high engagement on certain devices (e.g., mobile phones), it indicates that users prefer using your service on those devices. This could guide you to optimize your service for those devices.

Weekly Trends:

If engagement spikes during certain weeks, investigate what might have caused these spikes. It could be due to marketing campaigns, new feature releases, or seasonal trends.

Low Engagement on Certain Devices: If some devices show low engagement, it might indicate issues with the user experience on those devices. Consider investigating and improving the user interface or performance on those devices.

Actionable Steps

Optimize for Popular Devices: Focus development and marketing efforts on the devices with the highest engagement.

Investigate Spikes and Dips: Analyze the reasons behind any significant changes in engagement to replicate successful strategies or address issues.

Improve Underperforming Devices: Enhance the user experience on devices with lower engagement to boost overall user activity.

E.Email Engagement Analysis:

5. Write an SQL query to calculate the email engagement metrics.

QUERY:

```
SELECT
    COUNT(CASE WHEN action =
    'email_open' THEN 1 END )* 100.0 /
    COUNT(CASE WHEN action =
    'sent_weekly_digest' THEN 1 END) AS
    open_rate
    FROM
        emailevents;
SELECT
    COUNT(CASE WHEN action =
    'email_clickthrough' THEN 1 END )* 100.0/COUNT(CASE WHEN action = 'email_open' THEN 1 END ) AS ctr
    FROM
       emailevents;
SELECT
   COUNT(CASE WHEN action =
   'email_conversion' THEN 1 END )* 100.0 / COUNT(CASE WHEN action = 'email_clickthrough' THEN 1 END ) AS conversion_rate
   FROM
      emailevents;
```

OUTPUT:

open_rate
35.72564

ctr
44.03930

conversion_rate
0.00000

Open Rate

Insight

High Open Rate: Indicates that your subject lines are compelling and your audience is interested in your emails.

Low Open Rate: Suggests that your subject lines may not be engaging enough or your emails are not reaching the intended audience.

Actionable Steps

Test different subject lines (A/B testing).

Ensure your email list is up-to-date and targeted.

Personalize subject lines to increase relevance.

Click-Through Rate (CTR)

Insight

High CTR: Indicates that your email content is engaging and your call-to-action (CTA) is effective.

Low CTR: Suggests that your email content or CTA may not be compelling enough.

Actionable Steps

Improve the clarity and placement of your CTA.

Ensure your email content is relevant and valuable to the reader.

Use engaging visuals and concise, persuasive text.

Conversion Rate

Insight

High Conversion Rate: Indicates that your email campaign is effective in driving the desired actions (e.g., purchases, sign-ups).

Low Conversion Rate: Suggests that there may be issues with the landing page or the offer itself.

Actionable Steps

Optimize your landing page for conversions (clear CTA, user-friendly design).

Ensure the offer is compelling and matches the expectations set in the email.

Test different offers and landing page designs.

Overall Interpretation

By analyzing these metrics, you can identify strengths and weaknesses in your email campaigns. High open rates but low CTRs might indicate good subject lines but poor email content. Conversely, high CTRs but low conversion rates might suggest issues with your landing page or offer.

Regularly monitoring these metrics and making data-driven adjustments will help you improve the effectiveness of your email marketing efforts.

ANALYSIS

From the tables I have infer the following:-

1. number of distinct job reviewed per day is 0.0083

2. number of non-distinct jobs reviewed per day is 0.0111

3. 7 day rolling average throughput for 25, 26, 27, 28, 29 and 30 Nov 2020 are 1, 1, 1,1.25, 1.2 and 1.3333 respectively (for both distinct and non-distinct)

Percentage Share of each language i.e. Arabic, English, French, Hindi, Italian and Persian are 12.5, 12.5, 12.5, 12.5, 12.5 and 37.5 respectively (for both distinct and non-distinct)

There are 2 duplicates values/rows having job_id = 23 and language = Persian in both the rows

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

Why there is a difference of values between the number of distinct jobs reviewed per day and number of non-distinct jobs reviewed per day?

----> May be due to repeated values in two or more rows or the dataset consisted of duplicate rows

Why one shall use 7 day rolling average for calculating throughput and not daily metric average?

----> For calculating the throughput we will be using the 7-day rolling because 7-day rolling gives us the average for all the days right from day 1 to day 7 Whereas daily metric gives us average for only that particular day itself.

Why is it that percentage share of all other languages is 12.5% but that of language = Persian' is 37.5?

-----> In such cases there are two chances i.e. either there were duplicate rows having language as 'Persian' or there were really two or more unique people who were speaking in Persian language

Why do we need to look for duplicate rows in an dataset?

----> Duplicates have a direct influence of the Analysis going wrong and may led to wrong Business Decision leading to loss to the company or any entity; so to avoid these one must look for duplicates and remove them where necessary

From the tables I have infer the following:-

- The weekly user engagement is the highest for week 31 i.e. 1685
- There are in total 9381 active users from 1st week of 2013 to the 35th week of 2014
- The email_opening_rate is 33.5833 and email_clicking_rate is 14.78988

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

Why is the weekly user engagement so less in the beginning and then got increased?

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

Why is weekly retention so important?

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

Why is weekly engagement per device plays an important role?

----> Based on the reviews from users weekly engagement per device helps the firms on which devices they must focus more and which devices need more improvements so they also get a good review in users weekly engagement per device

Why is Email Engagement plays an important role?

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

CONCLUSION

In Conclusion, I would like to conclude that Operation Analytics and Investigating Metric Spike are very necessary and they must be done on daily, weekly, Monthly, Quarterly

or Yearly basis based on the Business needs of the firm. Also, any firm/entity must focus on the Email Engagement with the customers; the firm must use catchy headings along

with reasonable discounts and coupons so as to increase their existing customer base

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

HIRING PROCESS ANALYTICS

DESCRIPTION

Hiring process is the fundamental and the most important function of a company.
Here,the MNCS get to know about the major underlying trends about the hiring process.
Trends such as- number of rejections, number of interviews, types of jobs,
vacancies etc. are important for a company to analyse before hiring freshers or any other individual. Thus, making an opportunity for a Data Analyst job here too!
Being a Data Analyst, your job is to go through these trends and draw insights out of it for hiring department to work upon.
You are working for a MNC such as Google as a lead Data Analyst and the company has provided with the data records of their previous hirings and have asked you to answer certain questions making sense out of that data.

THE PROBLEM

Organizations often face challenges in their hiring processes, including lengthy recruitment cycles, high turnover rates, and difficulties in identifying the best candidates. These issues can lead to increased costs, reduced productivity, and a negative impact on company culture.

This project addresses the problem by systematically analyzing hiring data to uncover trends, bottlenecks, and areas for improvement

Organizations face several challenges in their hiring processes:

- Inefficient recruitment workflows causing delays
- Lack of visibility into hiring bottlenecks
- Suboptimal candidate experience
- Extended time-to-hire periods
- Need for data-driven hiring decisions
- Diversity and inclusion challenges in recruitment

DESIGN

Before starting the actual analysis I have:-

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

Software used for doing the overall Analysis:-

----> Microsoft Excel

FINDINGS

A.Hiring Analysis:

1. Determine the gender distribution of hires. How many males and females have been hired by the

company?

OUTPUT :

Male Hired :

=COUNTIF(D2:D7168,"Male")

= 4084

Female Hired :

=COUNTIF(D2:D7168,"Female")

= 2675

B.Salary Analysis:

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

OUTPUT :

=AVERAGE(G2:G7168)

= 49983.02902

C.Salary Distribution:

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

OUTPUT :

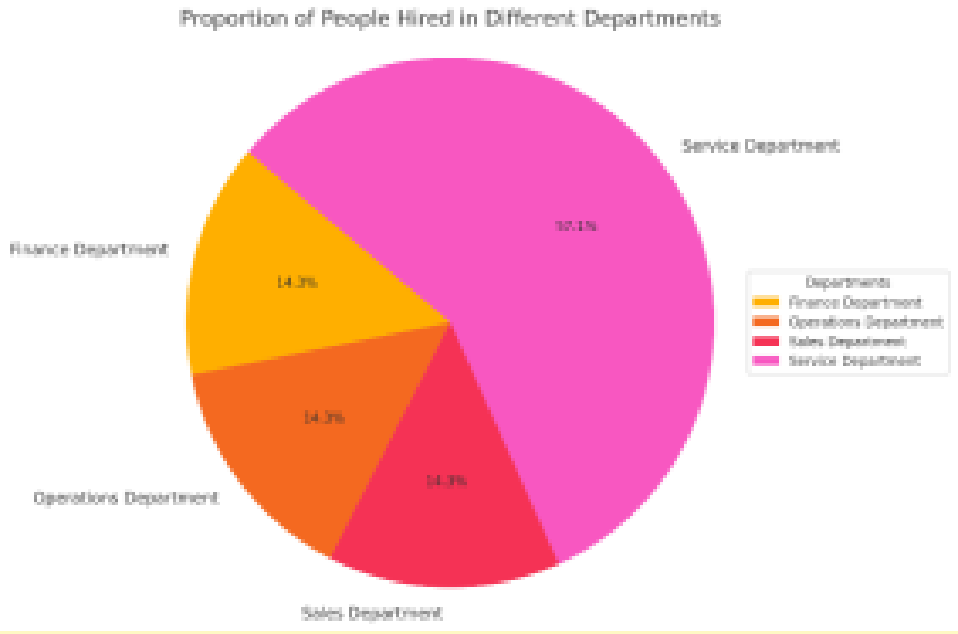
Minimum value : 100, class width : 19,995

Offered Salary	class intervals
56553	100-20,094
22075	20,095-40089
70069	40090-60084
3207	60085-80079
29668	80080-100074
85914	100075-120069
69904	120070-140064
11758	140065-160059
15156	160060-180054
49515	180055-200049

D.Departmental Analysis:

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

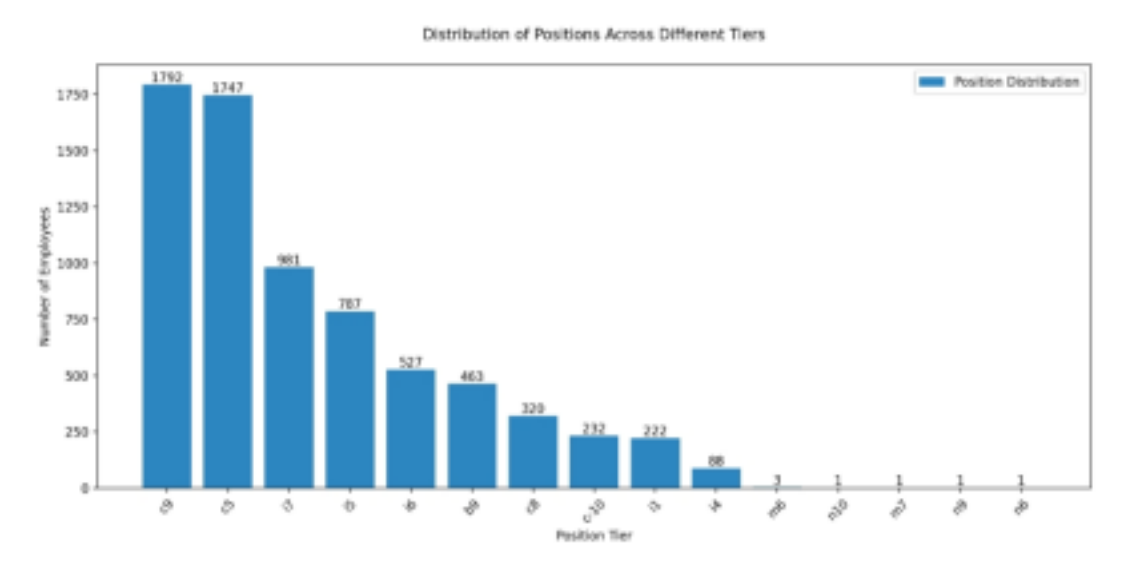
OUTPUT :



E.Position Tier Analysis:

5. 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.

OUTPUT :



ANALYSIS

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

Why is there so much difference in the total number of Males and Females hired?

---> Since, the Company is an MNC and people from all around the world work here; such difference exists due to the fact that the men-women equality has not yet reached to each and every part of the world. Some regions in the Gulf countries and in African continents along with some Asian countries face this problem

Why is it that there are less number of people who have salaries more than 85000 and there are more number of people who have salaries 35000 to 60000?

----> It is a fact that there are some positions in company who require a specialist person with years of experience in that particular field of work and hence company looks for such people and offer them higher salary packages also such people regularly prove themselves an asset to the company.For any company there are more people having the salary in the range 35000 to 60000; such people have spent 3-4 years in the company and their salary and increments are decided based on their monthly, quarterly and yearly performance.

Why is that the Operations department has the highest number of people working?

----> Operations Department works like a central hub for all other departments, all the execution tasks are carried out by this department. Operations department has the highest work load when compared to all other departments

CONCLUSION

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

Hiring Process Analytics is done on monthly, quarterly or yearly basis as per the needs and policies of the companies For any company the Operations Department has the highest number of workforce due to the workload on this department as this department acts as a central hub for all the executive tasks carried out

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

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

IMDb MOVIE ANALYSIS

DESCRIPTION

You are provided with dataset having various columns of different IMDB Movies.

You are required to Frame the problem.

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

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

THE PROBLEM

The film industry is highly competitive, and understanding what makes a movie successful can be challenging. Filmmakers and producers often struggle to predict audience preferences, leading to high-risk investments. Additionally, movie enthusiasts may find it difficult to discover films that match their interests. This project addresses the problem by systematically analyzing IMDb data to uncover trends, patterns, and correlations that can inform decision-making in the film industry.
The movie industry faces several challenges in understanding:
- Factors contributing to movie success
- Impact of genres on audience reception
- Relationship between budget and revenue
- Influence of directors and actors on ratings
- Role of movie duration in audience satisfaction
- Language and regional impact on global success
- Evolution of audience preferences over time

DESIGN

1. Firstly I made a copy of the raw data where I can perform the
Analysis so that what ever changes I made it will not affect the original data
2. Then dropping the columns which have no use for the analysis that we will be doing
3. Columns like 'Color', 'director_facebook_likes, 'actor_3_facebook_likes, 'actor_2_name','actor_1_facebook_likes''cast_total_facebook_likes, 'actor_3_name', 'facenumber_in_posts','plot_keywords','movie_imdb_link','content_rating','actor_2_facebook_likes','aspect_ratio', 'movie_facebook_likes' are the columns
containing irrelevant data for the analysis tasks provided. So, these columns needs to be dropped.
4. After dropping the irrelevant columns now we need to remove The rows from the dataset having anyone of its column value As blank/NULL
5. Then we need to get rid off the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab

FINDINGS

A.Movie Genre Analysis:

1. 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.

OUTPUT :

	Genre	Count	Mean	Median	Mode	Range	Variance	Std Dev	IMDb Scores
2	Drama	1893	6.79	6.9	6.7	7.2	0.8	0.9	[6.7, 7.2, 7.7, 7.3,
3	Comedy	1461	6.19	6.3	6.7	6.9	1.07	1.04	[7.8, 6.8, 7.3, 6.3,
4	Thriller	1117	6.38	6.4	6.5	6.3	0.94	0.97	[6.8, 8.5, 5.9, 7.0,
5	Action	959	6.29	6.3	6.1	6.9	1.08	1.04	[7.9, 7.1, 6.8, 8.5,
6	Romance	859	6.44	6.5	6.5	6.4	0.91	0.95	[6.2, 7.8, 7.2, 7.7,
7	Adventure	781	6.45	6.6	6.6	6.6	1.24	1.11	[7.9, 7.1, 6.8, 6.6,

The link below contains the rest of the output :

https://docs.google.com/spreadsheets/d/1p4Er_zbFiW1dzKn3M0wg_Hk49hpPOf3F/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

B. Movie Duration Analysis:

2. Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

OUTPUT :

		DURATION
2	COUNT	3756
3	MEAN	110.2579872
4	STD	22.64671656
5	MIN	37
6	25%	96
7	50%	106
8	75%	120
9	MAX	330

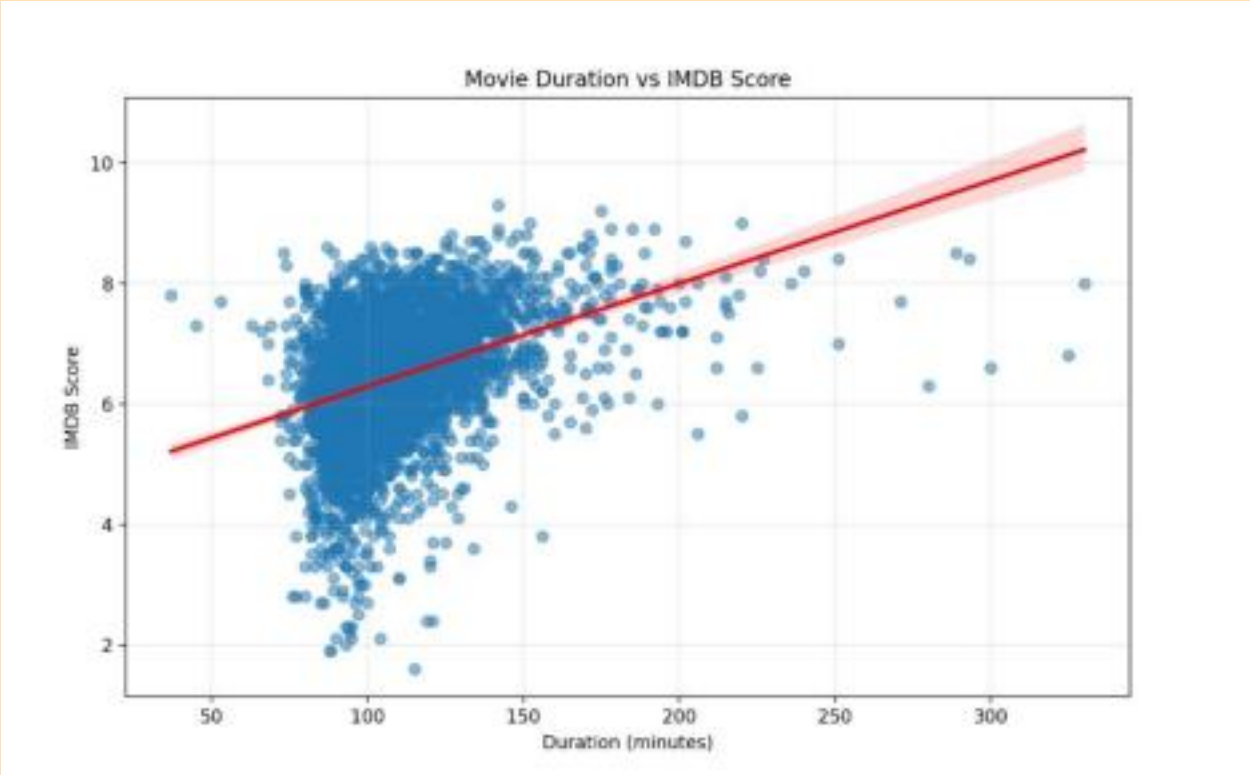
Min: The shortest movie duration is 37 minutes

25%(1st quartile): 25% of movies have a duration of 96 minutes or less

50%(Median): The median movie duration is 106 minutes

75%(3rd quartile): 75% of movies have a duration of 120 minutes or less

Max: The longest movie duration is 330 minutes.



Correlation coefficient between duration and IMDB Score: 0.366221017357080 07

C. Language Analysis

3. Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics

OUTPUT

	A	B	C	D	E	F	G
1	language	Count	Mean IMDB	Median IMDB	Std Dev	Min IMDB	Max IMDB
2	German	10	7.77	7.8	0.71	6.1	8.5
3	Japanese	10	7.66	8	0.99	6	8.7
4	French	34	7.36	7.3	0.52	5.8	8.4
5	Mandarin	15	7.08	7.4	0.77	5.6	7.9
6	Spanish	23	7.08	7.2	0.86	5.2	8.2
7	English	3598	6.43	6.5	1.05	1.6	9.3

Based on the analysis of the IMDB Movies dataset the key findings about languages and their impact on the IMDB scores :

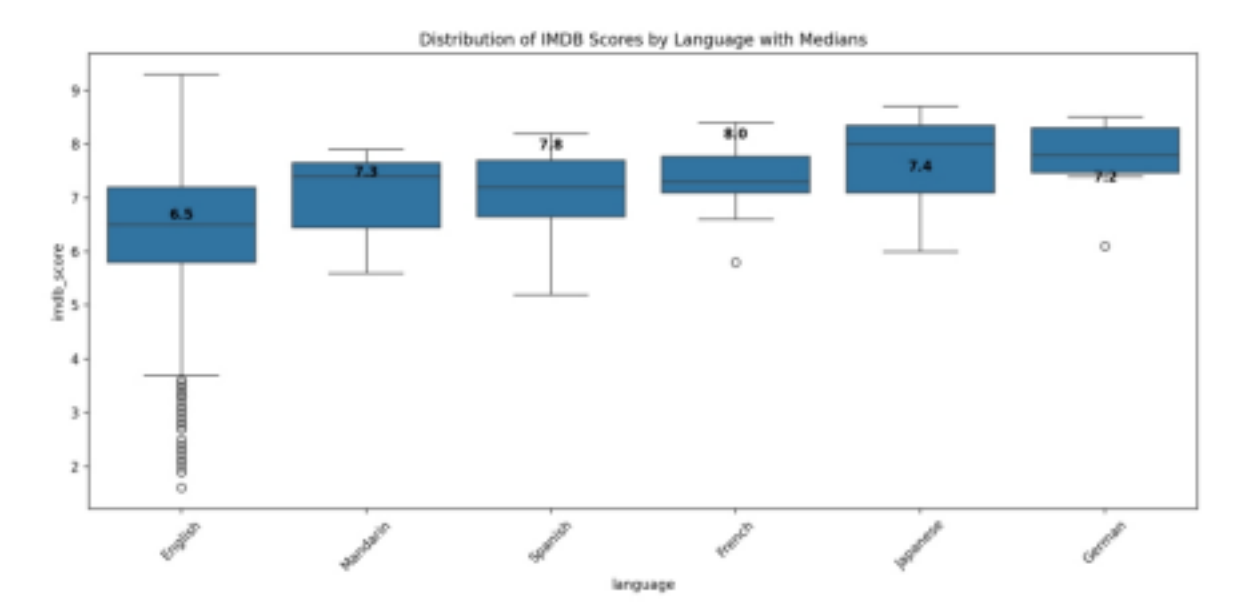
1.Most common Languages:

- English dominates the dataset with 3,598 movies
- French is second with 34 movies
- Spanish follows with 23 movies
- Mandarin has 15 movies
- Japanese has 10 movies
- German has 10 movies

2. IMDB Score statistics by language:

- Japanese films have the highest average rating (7.66) and median(8.0)
- French films follow with an average of 7.36
- Spanish and mandarin films have similar averages around 7.08

- English films show the widest spread of scores,indicating more variability in quality
- English films have the lowest average rating (6.43) among the top languages
- Japanese films have the highest median scores but also show considerable variation
- French films show more consistent ratings with a smaller spread
- Non – English films generally tend to have higher ratings,which might be due to selection bias
- The german language movie has mean IMDB score is 7.77 with a standard deviation of 0.71 and the median score is 7.8



This analysis suggests that while English-language films dominate the database non-english films that make it into the IMDB database tend to receive higher ratings on average however this could be influenced by selection bias,as often only the most notable non-english films gain international recognition and entries in IMDB.

D. Director Analysis:

4. Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

OUTPUT

The analysis of top directors based on IMDB scores

The overall statistics for context :

Overall IMDB score statistics :

Mean score : 6.47

Median score : 6.60

Standard deviation : 1.06

25th percentile :5.90

75th percentile : 7.20

The top directors

1.First place:

Director: sergio leone

Number of movies: 3

Average score : 8.43

Score standard deviation: 0.45

Percentile rank : 98.7th percentile

2.Director: Christopher Nolan

Number of movies : 8

Average score: 8.43

Score standard deviation: 0.54

Percentile rank : 98.7th percentile

3.Animation Masters:

Director: Pete doctor

Number of movies: 3

Average score : 8.23

Score standard deviation: 0.12

Percentile Rank : 97.7th percentile

Director: Hayao Miyazaki

Number of Movies : 4

Average score : 8.22

Score standard deviation: 0.39

Percentile rank: 97.7th percentile

Insights:

Average score of top directors: 8.04

Average number of movies per top director : 4.4

Score difference from overall mean : 1.58

The analysis shows that the top directors consistently perform well above the overall mean score of 6.47 Both sergio leone and Christopher Nolan lead with identical average scores of 8.43,placing them in the 98.7th percentile.nolan maintained this high average across 8 films,while leone achieved it with 3 classics.

The data also reveals strong showings from animation directors (pete docter,hayao Miyazaki) and auteur filmmakers (quentin tarantino),all scoring above the 97th percentile,demonstrating consistent excellence across multiple films.

E. Budget Analysis:

5. Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

OUTPUT :

The analysis of movie budgets and gross earnings:

Correlation analysis:

Correlation coefficient between budget and gross earnings : 0.0995 indicating a weak positive correlation

This relatively low correlation coefficient suggests that there isn't a strong linear relationship between movie budgets and box office earnings.

Summary statistics (in millions):

Average Budget : $46.24M

Average Gross : $52.61M

Average Profit : $6.38M

Number of movies analyzed : 3756

The movies with the highest profit margins are

: 1.James Cameron -

Gross: $760,505,847,

Budget: $237,000,000,

Profit margin : $523,505,847

2. Colin Trevorrow -

Gross: $652,177,271,

Budget: $150,000,000,

Profit Margin: $502,177,271

3.James Cameron -

Gross: $658,672,302,

Budget: $200,000,000,

Profit Margin : $458,672,302

4.George Lucas -

Gross: $460,935,665,

Budget: $11,000,000,

Profit Margin : $449,935,665

5.Steven Spielberg -

Gross: $434,949,459,

Budget: $10,500,000,

Profit Margin: $424,449,459

These movies have the highest profit margins in the dataset.

ANALYSIS

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

Why is it that the Most rated IMDB movie and the highest profit movie not the same?

-----> Maybe, due to fact that during the IMDB rating only recognized and people who know how to vote on IMDB have the access to the IMDB portal. On the other hand the profit is calculated on the basis of the tickets sold in theatres worldwide.

Why there are more number of votes during the decade 2001-2010?

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

Why is it that only movies having language as 'English' are the top 5 ranked movies on the basis of IMDB?

-----> Movies having language as English were having country of origin as USA; Also it is a well known fact that USA economy was robust during those days. So the social media investors looked for directors made movies so as to gain some financial gains

Why is it that only Drama and Comedy had the highest popularity?

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

Why is it that there were more number of votes for the decade 2001-2010 than compared to 2011-2020, though there was advancement in graphics and animation during 2011-2020?

----> It is a fact that there was a great and immense growth of technology not only in the graphics and animation sector but in all aspects of life; Also it was during this interval VPN was introduced; VPN led to piracy (illegal distribution of film) due to which most of people avoided going to theatres.

CONCLUSION

In Conclusion, I would like to conclude that IMDB Movie Analysis or any such analysis is done not only by Movie makers before movie production, but it is also done by various investors, stake-holders, theatre outlet owners. Normal people would not mind to do such analysis but such analysis plays an crucial part during the pre-production phase of the movies and also during the post-production phase

Also, it is not necessary that the movie with the highest IMDB rating will have the highest profit.

Profit is calculated truly on the basis on the number of tickets sold by theatres all over the world

Most of the people are tired with their daily lives and they prefer movies with Comedy/ Drama genre or both, and they would not go for movies with Action/Horror genre So, directors and production team must keep in mind the above

points and shall do the pre-production analysis before the commencement of filming

BANK LOAN CASE STUDY

DESCRIPTION

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

Two types of risks are associated with the bank's decision:

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

When a client applies for a loan, there are four types of decisions that

could be taken by the client/company:

1.Approved: The company has approved loan application

2.Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

4.Unused Offer: Loan has been cancelled by the client but on different stages of the process.

THE PROBLEM

Financial institutions face significant challenges in making informed loan approval decisions while maintaining a balance between risk management and customer satisfaction. The current loan approval process may be inefficient and potentially inconsistent, leading to increased default risks and missed opportunities for viable customers.

The bank lacks a data-driven understanding of the key factors influencing loan approval outcomes, which impacts their ability to:

1. Make consistent and objective loan approval decisions
2. Identify high-risk applications early in the process
3. Optimize the approval process for different customer segments
4. Balance risk management with business growth

Specific Challenges

- Difficulty in identifying the most significant variables that predict loan approval success
- Lack of understanding of the relationships between different applicant characteristics
- Presence of data quality issues, including missing values and outliers
- Significant data imbalance in loan approval outcomes
- Need for better segmentation and analysis of different customer profiles

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

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

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

DESIGN

Firstly create a copy of the raw data

Then the percentage of null values needs to be analyzed and those columns that have more than 50% of the null data have to be dropped And those columns with less than 50% of the null data have to be replaced with mean or median or the highest occurring categorical variables.

The following columns needs o be dropped as they have more than 50% of the NULL values

- OWN_CAR_AGE
- EXT_SOURCE_1
- E APARTMENTS_AVG
- BASEMENTAREA_AVG
- YEARS_BUILD_AVG
- COMMON_AREA_AVG
- ELEVATORS_AVG
- ENTRANCES_AVG
- FLOORSMAX_AVG
- FLOORSMIN_AVG
- LANDAREA_AVG

- LIVINGAPARTMENTS_AVG
- LIVINGAREA_AVG
- NONLIVINGAPARTMENTS_AVG
- NONLIVINGAREA_AVG
- APARTMENTS_MODE
- BASEMENTAREA_MODE
- YEARS_BUILD_MODE
- COMMON_AREA_MODE
- ELEVATORS_MODE
- ENTRANCES_MODE
- FLOORSMAX_MODE
- FLOORSMIN_MODE
- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MODE
- NONLIVINGAPARTMENTS_MODE
- NONLIVINGAREA_MODE
- APARTMENTS_MEDIAN
- BASEMENTAREA_MEDIAN
- YEARS_BUILD_MEDIAN
- COMMON_AREA_MEDIAN
- ELEVATORS_MEDIAN
- ENTRANCES_MEDIAN
- FLOORSMAX_MEDIAN
- FLOORSMIN_MEDIAN
- LANDAREA_MEDIAN

- FLOORSMIN_MEDIAN
- LANDAREA_MEDIAN
- LIVINGAPARTMENTS_MEDIAN
- LIVINGAREA_MEDIAN
- NONLIVINGAPARTMENTS_MEDIAN
- NONLIVINGAREA_MEDIAN
- FONDKAPREMONT_MODE
- HOUSETYPE_MODE
- WALLSMATERIAL_MODE

Then drop those columns which are irrelevant for doing the Data Analysis. The following columns needs to be dropped:-

- FLAG_MOBILE
- FLAG_EMPLOY_PHONE
- FLAG_WORK_PHONE
- FLAG_CONT_MOBILE
- FLAG_PHONE
- FLAG_EMAIL
- CNT_FAMILY_MEMBERS
- REGION_RATING_CLENT
- REGION_RATING_CLENT_W_CITY
- EXT_SOURCE_3
- YEAR_BEGINEXPLUATATION_AVG
- YEAR_BEGINEXPLUATATION_MODE
- YEAR_BEGINEXPLUATATION_MEDIAN

- TOTAL_AREA_MODE
- EMERGENCYSTATE_MODE
- DAYS_LAST_PHONE_CHANGE
- FLAG DOC 2
- FLAG DOC 3
- FLAG DOC 4
- FLAG DOC 5
- FLAG DOC 6
- FLAG DOC 7
- FLAG DOC 8
- FLAG DOC 9
- FLAG DOC 10
- FLAG DOC 11
- FLAG DOC 12
- FLAG DOC 13
- FLAG DOC 14
- FLAG DOC 15
- FLAG DOC 16
- FLAG DOC 17
- FLAG DOC 18
- FLAG DOC 19
- FLAG DOC 20
- FLAG DOC 21

Replacing Blanks in Occupation_Type column of the Application

Dataset with the highest occurring categorical variable

--> Highest occurring categorical variable is 'Laborers'

Replacing Blanks in AMT_ANNUTIY column of the Application Dataset

with the median of the AMT_ANNUITY as there exists outliers in the AMT ANNUITY column

--> Median of AMT_ANNUITY = 24903

Replacing Blanks in AMT_GOODS PRICE column of the Application

Dataset with the median of the AMT_GOODS PRICE as there exists outliers in the AMT_GOODS PRICE column

--> Median of AMT_GOODS PRICE = 450000

Replacing Blanks in Name_Type_Suite column of the Application

Dataset with the highest occurring categorical variable

--> Highest occurring categorical variable is 'Unaccompanied'

Replacing Blanks in Organization_type column of the Application

Dataset with the highest occurring categorical variable

--> Highest occurring categorical variable is 'Business Entity Type 3'

The following columns of the previous application datasets need to be dropped as they are irrelevant for doing the data analysis

- HOUR_APPR_PROCESS_START
- WEEKDAY_APPR_PROCESS_START_PREV
- FLAG_LAST_APPL_PER_CONTRACT

- NFLAG_LAST_APPL_IN_DAY
- SK_ID_CURR
- WEEKDAY_APPR_PROCESS_START

Removing the rows with the values 'XNA' &'XAP' for the column:

NAME_TYPE_SUITE

----> Replace Blanks with Unaccompained

AMT_ANNUITY :- Replace Blanks with 21340(median)

FINDINGS

A.Identify Missing Data and Deal with it Appropriately:

1.Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features. Create a bar chart or column chart to visualize the proportion of missing values for each variable.

OUTPUT :

	A	B	C	D	E	F
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
2	100002	1	Cash loans	M	N	Y
3	100003	0	Cash loans	F	N	N
4	100004	0	Revolving loans	M	Y	Y
5	100006	0	Cash loans	F	N	Y
6	100007	0	Cash loans	M	N	Y
7	100008	0	Cash loans	M	N	Y
8	100009	0	Cash loans	F	Y	Y
9	100010	0	Cash loans	M	Y	Y
10	100011	0	Cash loans	F	N	Y
11	100012	0	Revolving loans	M	N	Y
12	100014	0	Cash loans	F	N	Y
13	100015	0	Cash loans	F	N	Y
14	100016	0	Cash loans	F	N	Y
15	100017	0	Cash loans	M	Y	N
16	100018	0	Cash loans	F	N	Y
17	100019	0	Cash loans	M	Y	Y

The Rest of the columns which I have used to find missing values is given in the below link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

TO FIND BLANKS :

Used conditional formatting – new Rule – Format only cells that contain – format only that contain

section –blanks –format -ok

To Find empty missing cells :

COUNT BLANK used to count the empty missing cells

COUNT BLANK FUNCTION :

=COUNTBLANK(A:A)

To find the total no of missing cells

COUNT A is used to count the total no of missing cells

=COUNTA (A:A)

So I have found the total no of missing values and the count of missing cells

As the the percentages of missing values is found to quantify the extent of missing data relative to

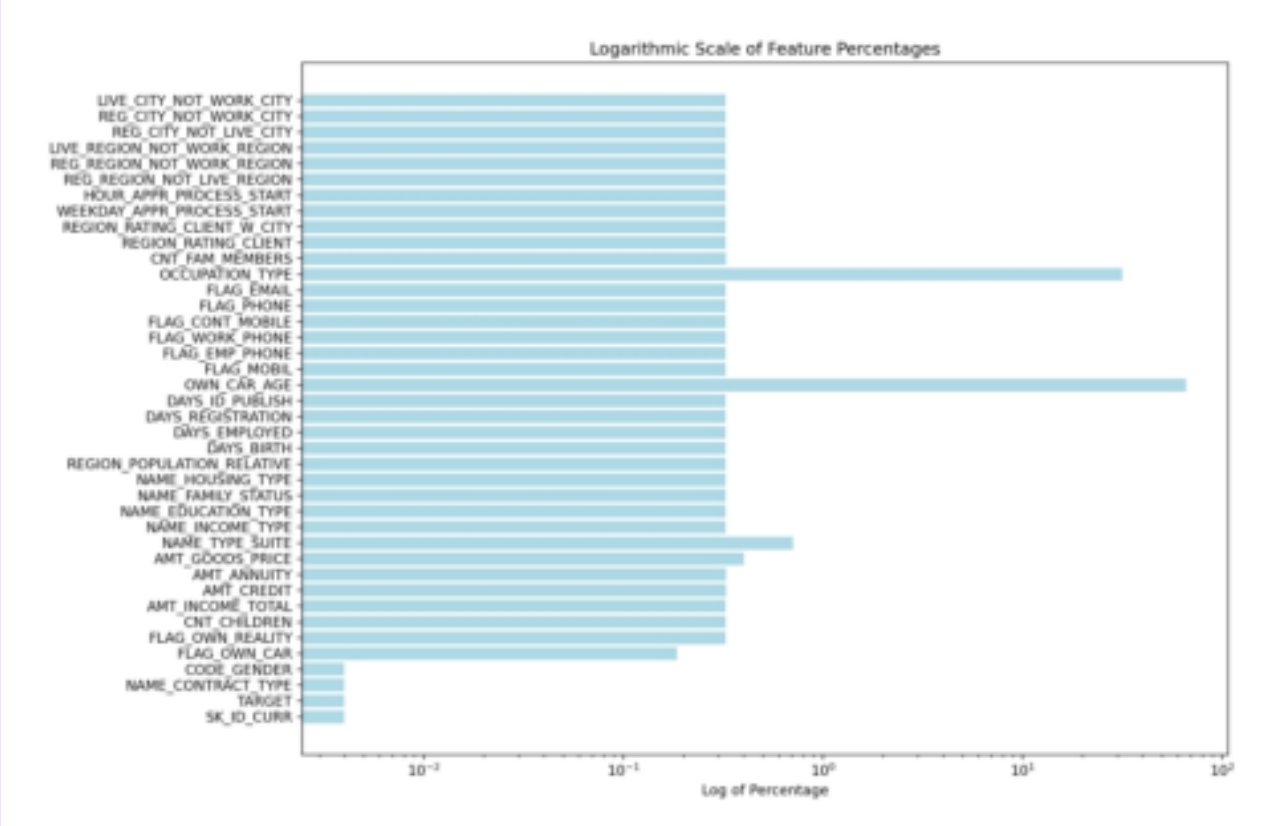
the total no of entries for each variable

I have found percentages of all values

MISSING VALUES & PERCENTAGES

	A	B	C
1		MISSING VALUES	PERCENTAGE
2	SK_ID_CURR	2	0.003987082
3	TARGET	2	0.003987082
4	NAME_CONTRACT_TYPE	2	0.003987082
5	CODE_GENDER	2	0.003987082
6	FLAG_OWN_CAR	93	0.185399306
7	FLAG_OWN_REALITY	163	0.324947171
8	CNT_CHILDREN	163	0.324947171
9	AMT_INCOME_TOTAL	163	0.324947171
10	AMT_CREDIT	163	0.326940717
11	AMT_ANNUITY	164	0.326940712
12	AMT_GOODS_PRICE	201	0.400701726
13	NAME_TYPE_SUITE	355	0.707707029
14	NAME_INCOME_TYPE	163	0.324947171
15	NAME_EDUCATION_TYPE	163	0.324947171
16	NAME_FAMILY_STATUS	163	0.324947171
17	NAME_HOUSING_TYPE	163	0.324947171
18	REGION_POPULATION_RELATIVE	163	0.324947171
19	DAYS_BIRTH	163	0.324947171
20	DAYS_EMPLOYED	163	0.324947171
21	DAYS_REGISTRATION	163	0.324947171
22	DAYS_ID_PUBLISH	163	0.324947171
23	OWN_CAR_AGE	33113	66.01212073
24	FLAG_MOBIL	163	0.324947171
25	FLAG_EMP_PHONE	163	0.324947171
26	FLAG_WORK_PHONE	163	0.324947171
27	FLAG_CONT_MOBILE	163	0.324947171
28	FLAG_PHONE	163	0.324947171
29	FLAG_EMAIL	163	0.324947171
30	OCCUPATION_TYPE	15817	31.53183685
31	CNT_FAM_MEMBERS	164	0.326940712
32	REGION_RATING_CLIENT	163	0.324947171
33	REGION_RATING_CLIENT_W_CITY	163	0.324947171
34	WEEKDAY_APPR_PROCESS_START	163	0.324947171
35	HOUR_APPR_PROCESS_START	163	0.324947171
36	REG_REGION_NOT_LIVE_REGION	163	0.324947171
37	REG_REGION_NOT_WORK_REGIO	163	0.324947171
38	LIVE_REGION_NOT_WORK_REGIO	163	0.324947171
39	REG_CITY_NOT_LIVE_CITY	163	0.324947171
40	REG_CITY_NOT_WORK_CITY	163	0.324947171
41	LIVE_CITY_NOT_WORK_CITY	163	0.324947171
42	ORGANIZATION_TYPE	163	0.324947171
43	EXT_SOURCE_1	28335	56.48696218
44	EXT_SOURCE_2	289	0.576133328
45	EXT_SORCE_3	10107	20.14871815
46	APARTMENTS_AVG	25548	50.93098361
47	BASEMENTAREA_AVG	29362	58.53434871
48	YEARS_BEGINEXPLUATATION_AVG	24557	48.95538455
49	YEARS_BUILD_AVG	33402	66.58825406
50	COMMONAREA_AVG	35123	70.01913799
51	ELEVATORS_AVG	26814	53.45480643

So As by finding percentage of missing values I have found the proportion of the missing values And chart used to visualize the proportion of missing values for each variable is given below



Logarithmic scale is used for the percentage differences in the chart to be seen more clearly

B. Identify Outliers in the Dataset:

2. Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables. Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers

OUTPUT :

	A	B	C	D	E	F
1	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY
2	100002	1	Cash loans	M	N	Y
3	100003	0	Cash loans	F	N	N
4	100004	0	Revolving loans	M	Y	Y
5	100006	0	Cash loans	F	N	Y
6	100007	0	Cash loans	M	N	Y
7	100008	0	Cash loans	M	N	Y
8	100009	0	Cash loans	F	Y	Y
9	100010	0	Cash loans	M	Y	Y
10	100011	0	Cash loans	F	N	Y
11	100012	0	Revolving loans	M	N	Y
12	100014	0	Cash loans	F	N	Y
13	100015	0	Cash loans	F	N	Y
14	100016	0	Cash loans	F	N	Y
15	100017	0	Cash loans	M	Y	N
16	100018	0	Cash loans	F	N	Y

The Rest of the columns which I have used to find outliers is given in the below link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have found outliers using Z-score method & to calculate z score I have found the mean and standard deviation

To calculate Z-score I have subtracted mean from value and divide by standard deviation

To calculate Mean :

=AVERAGE(A2: A100)

To calculate standard deviation :

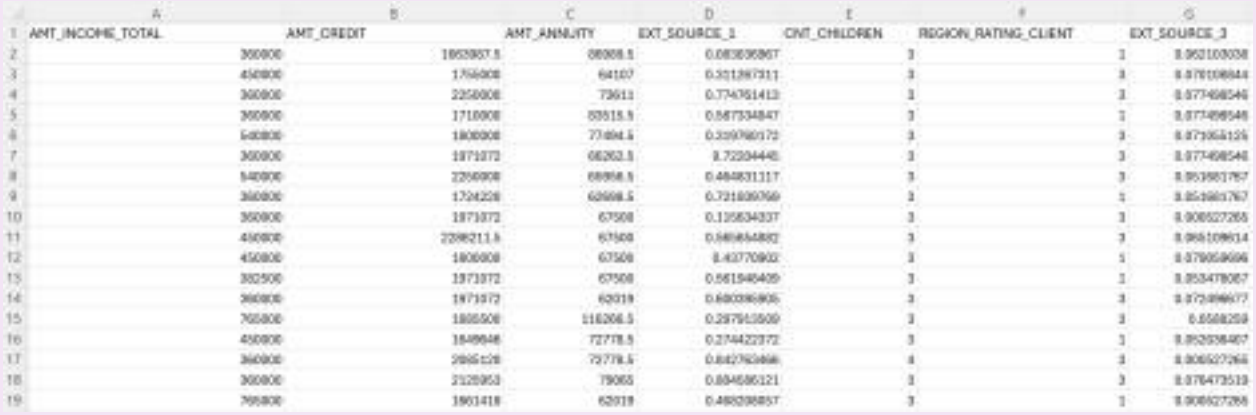
=STDEV.P (A2:A100)

z- score formula :

=(A2 - B1) / B2

I have found outliers using Z-score where values less than -3 and values greater than -3 are considered outliers

TheAMT_INCOME_TOTAL,AMT_CREDIT,AMT_ANNUITY,EXT_SOURCE_1,CNT_CHILDREN,REGION_RATING_CLIENT,EXT_SOURCE_3 are the columns from the dataset where outliers were found.



	A	B	C	D	E	F	G
1	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	EXT_SOURCE_1	CNT_CHILDREN	REGION_RATING_CLIENT	EXT_SOURCE_3
2	360000	1863857.5	86886.5	0.083836867	3	1	0.962108038
3	450000	1756908	64107	0.501287311	3	3	0.878108844
4	360000	2250000	73611	0.774751413	3	3	0.877498546
5	360000	1710000	83515.5	0.587334847	3	1	0.877498546
6	540000	1800000	77494.5	0.339760172	3	3	0.873664126
7	360000	1971072	68263.5	0.7220444	3	3	0.877498546
8	540000	2250000	68958.5	0.484831117	3	3	0.953981767
9	360000	1724220	62668.5	0.721830766	3	1	0.953981767
10	360000	1971072	67508	0.135834337	3	3	0.908527266
11	450000	2298211.5	67508	0.685654882	3	3	0.966109614
12	450000	1800000	67508	0.42770902	3	1	0.879059696
13	382500	1971072	67508	0.561945409	3	1	0.953478067
14	360000	1971072	62019	0.680086805	3	3	0.972496677
15	765800	1800508	116206.5	0.287513509	3	3	0.8568259
16	450000	1849648	72778.5	0.27442272	3	1	0.952038467
17	360000	2665128	72778.5	0.642762666	4	3	0.908527266
18	360000	2128953	79065	0.684586121	3	3	0.878473519
19	765800	1901418	62019	0.469238857	3	1	0.908527266

The rest of the outliers output is in the link below :

https://docs.google.com/spreadsheets/d/1b9cbpOIEyhBRMTom_q0YRpNZjvd5Zu1p/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

To find if all the datapoints are valid or not I have used these below functions

FOR AMT_INCOME_TOTAL

=IF(OR(A208<10000,A208>1000000),"INVALID","VALID")

FOR AMT_CREDIT

=IF(OR(B2<5000,B2>500000),"INVALID","VALID")

IN AMT_CREDIT column all the values are invalid

FOR AMT_ANNUITY

=IF(OR(C2<1000,C2>50000),"INVALID","VALID")

In AMT_ANNUITY column all the values are invalid

FOR EXT_SORCE_1

=IF(OR(D2<0,D2>1),"INVALID","VALID")

FOR CNT_CHILDREN

=IF(OR(E2<0,E2>10),"INVALID","VALID")

FOR REGION_RATING_CLIENT

=IF(OR(F2<1,F2>3),"INVALID","VALID")

FOR EXT_SOURCE_3

=IF(OR(G2<0,G2>1),"INVALID","VALID")

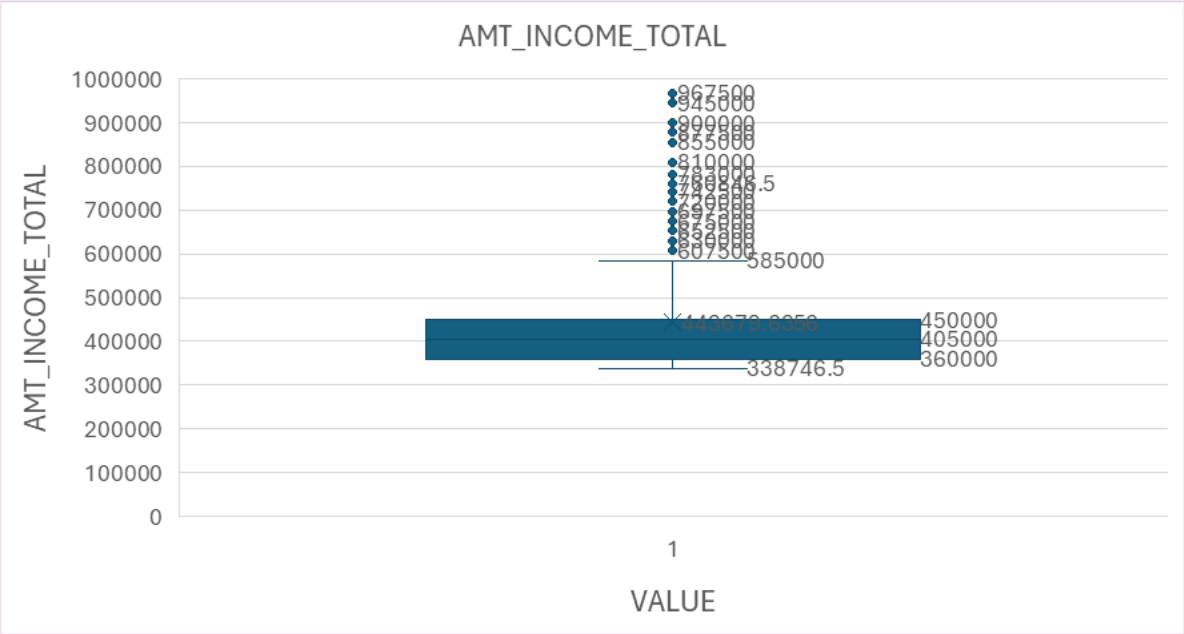
K	L	M	N	O	P	Q
AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	EXT_SOURCE_1	CNT_CHIDREN	REGION_RATING_CL	EXT_SOURCE_3
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID
VALID	INVALID	INVALID	VALID	VALID	VALID	VALID

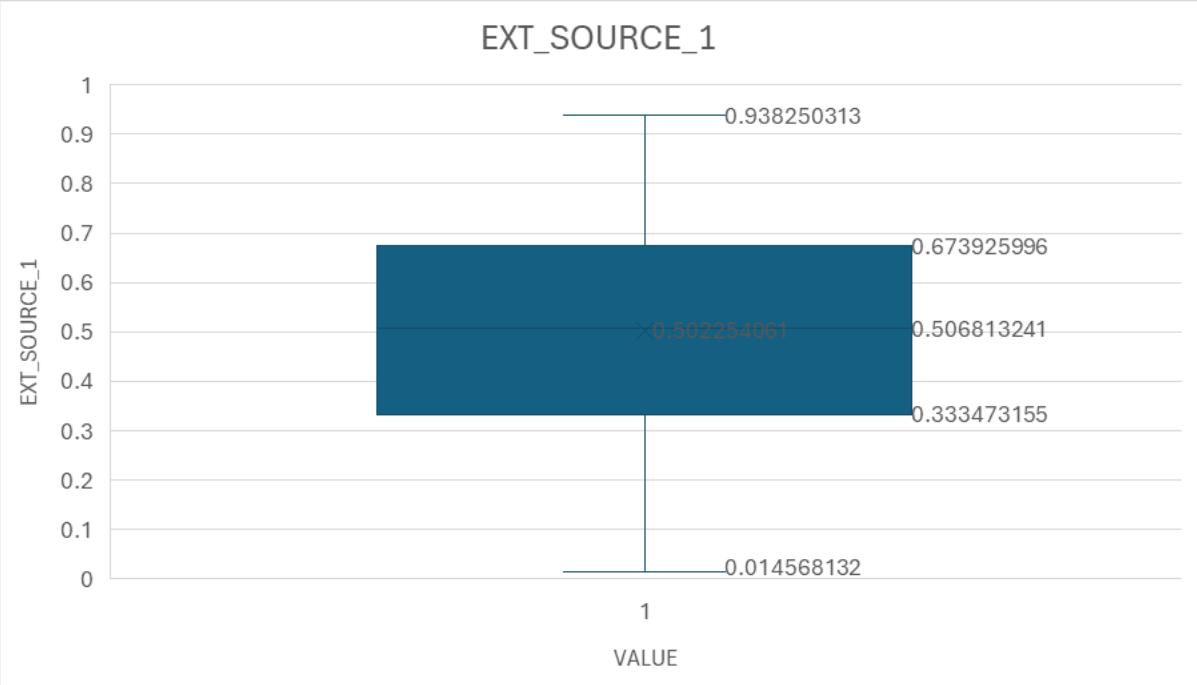
The valid/invalid data points of the data are there in file link below

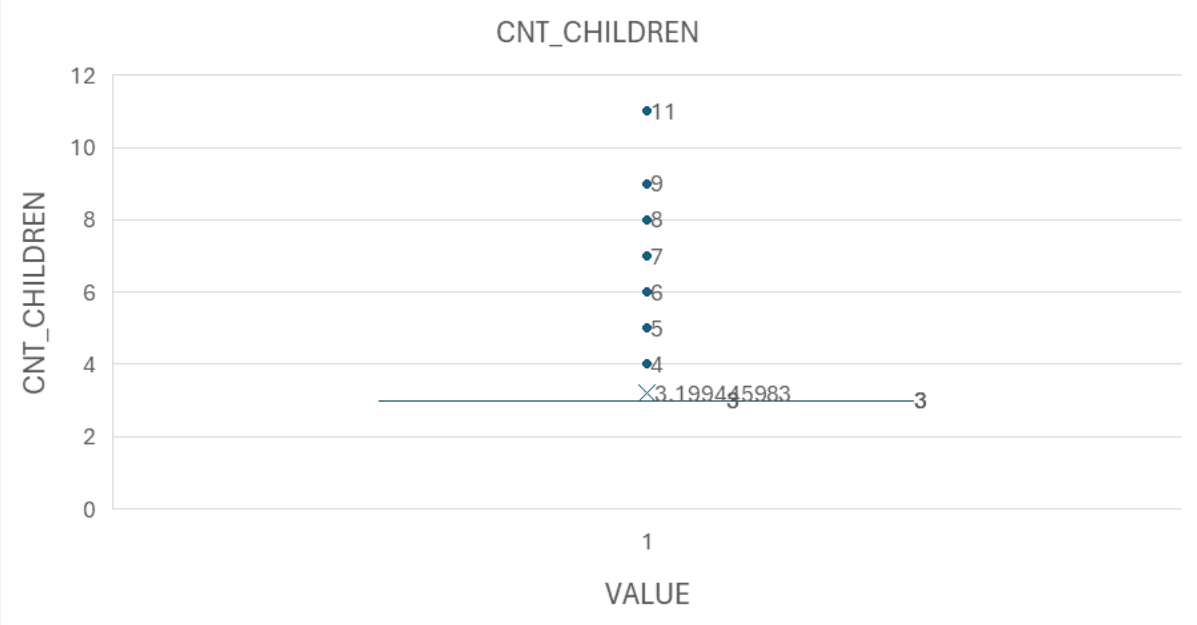
https://docs.google.com/spreadsheets/d/1TrCWbTN1DJhCJflxaRfpLZ9-ypGvzIi3/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

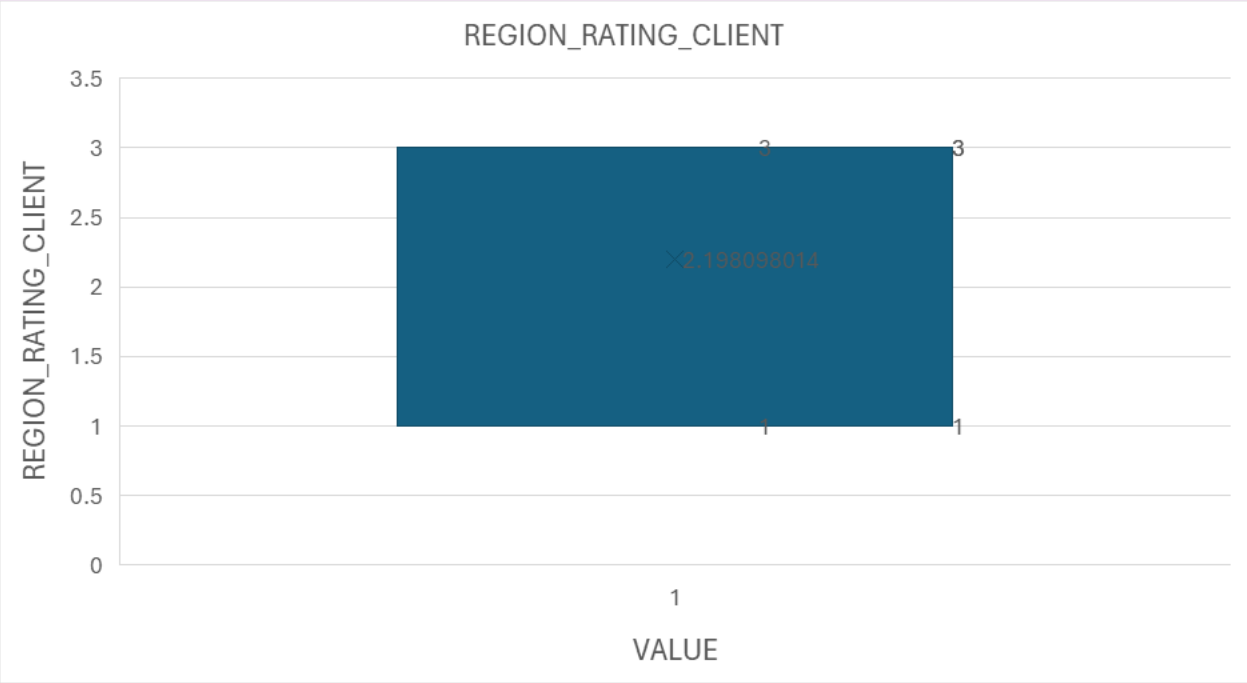
The Data points are used to plot these below box plots

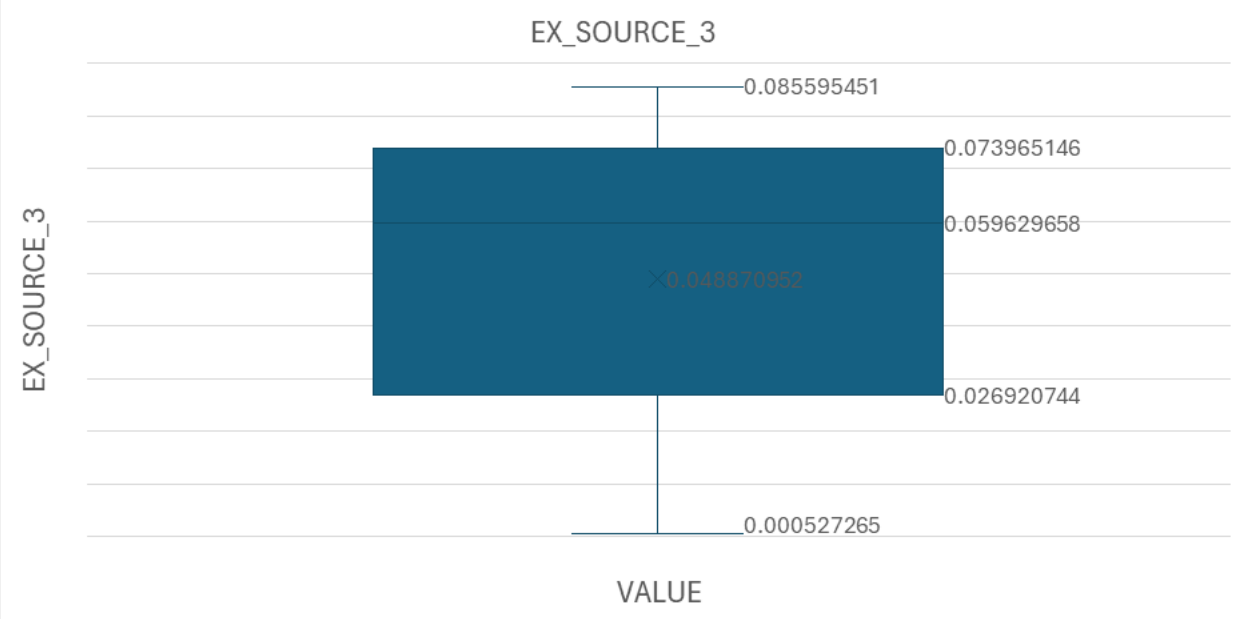
BOX PLOTS :




EXT_SOURCE_1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
EXT_SOURCE_1
0.938250313
0.673925996
0.502254061
0.506813241
0.333473155
0.014568132
1
VALUE


CNT_CHILDREN
12
10
8
6
4
2
0
CNT_CHILDREN
11
9
8
7
6
5
4
3.199445983
3
3
1
VALUE


REGION_RATING_CLIENT
REGION_RATING_CLIENT
3.5
3
2.5
2
1.5
1
0.5
0
3
3
2.198098014
1
1
1
VALUE


EX_SOURCE_3
EX_SOURCE_3
0.085595451
0.073965146
0.059629658
0.048670952
0.026920744
0.000527265
VALUE

C. Analyze Data Imbalance:

3.Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

OUTPUT :

The columns taken are :

code_gender,name_type_suite,name_income_type,name_education_type,name_family_status,name_housing_type,occupation_type,organization_type,housetype_mode, wallsmateal_mode,emergency state_mode

	A	B	C	D	E	F	G	H	I	J
1	CODE_GENDER	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	ORGANIZATION_	HOUSETYPE_MODE	WALLSMATERIAL_MODE	EMERGENCYSTATE_MODE
2	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Business Entity T	block of flats	Stone, brick	No
3	F	Family	State servant	Higher education	Married	House / apartment	School	block of flats	Block	No
4	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Government	block of flats	Panel	No
5	F	Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment	Business Entity T	block of flats	Panel	No
6	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Religion	block of flats	Stone, brick	No
7	M	Spouse, partner	State servant	Secondary / secondary special	Married	House / apartment	Other	block of flats	Stone, brick	No
8	F	Unaccompanied	Commercial associate	Higher education	Married	House / apartment	Business Entity T	block of flats	Panel	No
9	M	Unaccompanied	State servant	Higher education	Married	House / apartment	Other	block of flats	Mixed	No
10	F	Children	Pensioner	Secondary / secondary special	Married	House / apartment	XNA	block of flats	Panel	No
11	M	Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	Electricity	block of flats	Stone, brick	No
12	F	Unaccompanied	Working	Higher education	Married	House / apartment	Medicine	block of flats	Wooden	No
13	F	Children	Pensioner	Secondary / secondary special	Married	House / apartment	XNA	block of flats	Panel	Yes
14	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Business Entity T	block of flats	Others	No
15	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Self-employed	block of flats	Block	No
16	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Transport: type 2	block of flats	Panel	No
17	M	Family	Working	Secondary / secondary special	Single / not married	Rented apartment	Business Entity T	block of flats	Stone, brick	No
18	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Government	block of flats	Stone, brick	No
19	F	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Construction	block of flats	Stone, brick	No
20	F	Other_A	Working	Secondary / secondary special	Widow	House / apartment	Housing	block of flats	Stone, brick	No
21	F	Unaccompanied	State servant	Higher education	Single / not married	House / apartment	Kindergarten	block of flats	Stone, brick	No
22	M	Unaccompanied	Working	Secondary / secondary special	Married	House / apartment	Self-employed	block of flats	Panel	No
23	F	Unaccompanied	Commercial associate	Secondary / secondary special	Married	House / apartment	Trade: type 7	block of flats	Panel	No
24	F	Unaccompanied	Working	Secondary / secondary special	Married	Rented apartment	Self-employed	block of flats	Panel	No

The columns taken to find data imbalance are given in the link below :

https://docs.google.com/spreadsheets/d/1oVteOqRRD-49Eq0_ChBve4iZUBNufCLV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=COUNTIFS(A2:A50000,"1",B2:B50000,"M")

I have used countif function to calculate count of each category comparing to Target column

	A	B	C	D	E	F	G	H	I	J
1		CODE_GENDER			NAME_TYPE_SUITE					
2		F	M	XNA	Children	Family	Group of people	Other_A	Other_B	Spouse, partner
3	TARGET									
4	0	30559	15412	2	502	6037	36	129	246	1704
5	1	2264	1762	0	40	512	0	8	13	145

The rest of the output of counts where I have calculated the count of each category comparing to Target column is in the link below

https://docs.google.com/spreadsheets/d/10QDDrkD32XXK70ZV8w6Vbq-emor5eCMB/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have calculated the sum of target 0 and target 1 rows using sum function

=SUM(A1:A10)

I have calculated total of (sum of target 0 and target 1 rows) using sum function

=SUM(A1:A10,B1:B10)

I have calculated proportions of target variable using proportion formula

=B6/B8

I have calculated the proportion of target 0 and target 1

6	SUM(0)	422931
7	SUM (1)	36980
8	TOTAL SUM	459911
9	PROPORTION(0)	0.919593139
10	PROPORTION(1)	0.080406861

I have calculated the ratio of data imbalance through the formula

Ratio of data imbalance = proportion of Majority value / proportion of Minority value

Majority value(target variable 0) = 0.919593139

Minority value (target variable 1) = 0.080406861

Ratio of data imbalance = 0.919593139 / 0.080406861 = 11.44

Then I have calculated the percentage of both target 0 & target 1 proportions

	TARGET	
12		
13	**0**	91.96%
14	**1**	8.04%

I have inserted a pie chart to display the distribution of target variable and highlighted the class imbalance

DISTRIBUTION OF TARGET VARIABLE

NOTE: The chart indicates a ratio imbalance of 11.44,highlighting a significant disparity among the target variable

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

4. Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features. Create histograms, bar charts, or box plots to visualize the distributions of variables.

Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables an scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

OUTPUT :

1) UNIVARIATE ANALYSIS :

For univariate analysis the columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, CNT_CHILDREN, DAYS_BIRTH, DAYS_EMPLOYED, AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT,

AMT_DOWN_PAYMENT, AMT_GOODS_PRICE, NAME_CONTRACT_TYPE, NAME_CONTRACT_STATUS are taken

	A	B	C	D	E	F	G
1	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	CNT_CHILDREN	DAYS_BIRTH	DAYS_EMPLOYED	AMT_ANNUITY
2	202500	406597.5	24700.5	0	-9461	-637	1730.43
3	270000	1293502.5	35698.5	0	-16765	-1188	25188.615
4	67500	135000	6750	0	-19046	-225	15060.735
5	135000	312682.5	29686.5	0	-19005	-3039	47041.335
6	121500	513000	21865.5	0	-19932	-3038	31924.395
7	99000	490495.5	27517.5	0	-16941	-1588	23703.93
8	171000	1560726	41301	1	-13778	-3130	11368.62
9	360000	1530000	42075	0	-18850	-449	13832.775

The rest of the columns chosen to perform univariate analysis are in the link below :

https://docs.google.com/spreadsheets/d/1N8ho6bKiiMOMwOGUt60p_Hj9SGPq88_L/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance,quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

=COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	CNT_CHILDREN	DAYS_BIRTH	DAYS_EMPLOYED	AMT_ANNUITY.1
count	49999	49999	49998	49999	49999	49999	39408
mean	170767.5905	599700.5815	27107.37736	0.419848397	-16022.04208	63219.42449	15482.20397
std	531819.0951	402415.4339	14562.94444	0.724038548	4361.40027	140794.6057	14530.99679
min	25650	45000	2052	0	-25184	-17531	0
max	117000000	4050000	258025.5	11	-7680	365243	234478.395
Median	145800	514777.5	24939	0	-15731	-1221	10879.8075
Range	116974350	4005000	255973.5	11	17504	382774	234478.395
mode	135000	450000	9000	0	-13429	365243	2250
percentage	12.315789	43.2505127	1.95494978	41.98	-1.155125	4.55939615	0.88006075
variance	2.82832E+11	1.61938E+11	212079350.6	0.524231818	19021812.32	19823120985	211149867.6
quantiles(0.25)	112500	270000	16456.5	0	-19644	-2786	6122.835
quantiles(0.5)	145800	514777.5	24939	0	-15731	-1221	10879.8075
quantiles(0.75)	202500	808650	34596	1	-12378.5	-292	19668.915
Percentile(1%)	45000	76410	6174	0	-24426.02	-10942.18	2117.33505
Percentile(5%)	67500	135000	9000	0	-23197	-6824.2	2666.58075
percentile(10%)	81000	180000	11002.5	0	-22180.2	-4942.4	3708.3105
percentile(25%)	112500	270000	16456.5	0	-19644	-2786	6122.835
percentile(50%)	145800	514777.5	24939	0	-15731	-1221	10879.8075
percentile(75%)	202500	808650	34596	1	-12378.5	-292	19668.915
percentile(90%)	270000	1132573.5	45954	2	-10291.8	365243	33643.845
percentile(95%)	337500	1350000	53248.5	2	-9381.9	365243	45000
percentile(99%)	477045	1847703.6	70007.31	3	-8244	365243	69105.61395

COUNTS CALCULATED FOR CATEGORICAL COLUMN :

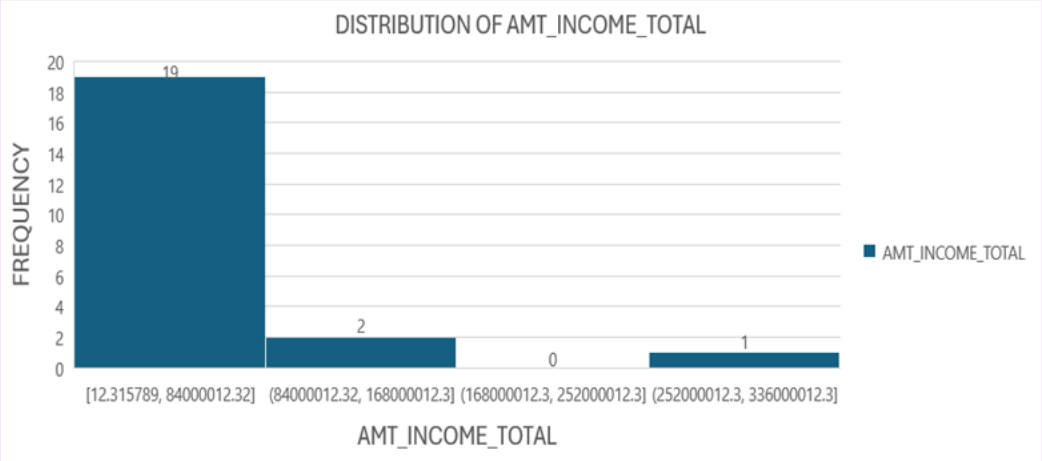
	NAME_CONTRACT_TYPE COUNTS	NAME_CONTRACT_STATUS COUNTS
Approved	0	31885
canceled	1	8594
cash loans	20855	0
consumer loans	23510	0
refused	0	8660
revolving loans	5625	0
unused offer	0	859
XNA	8	0

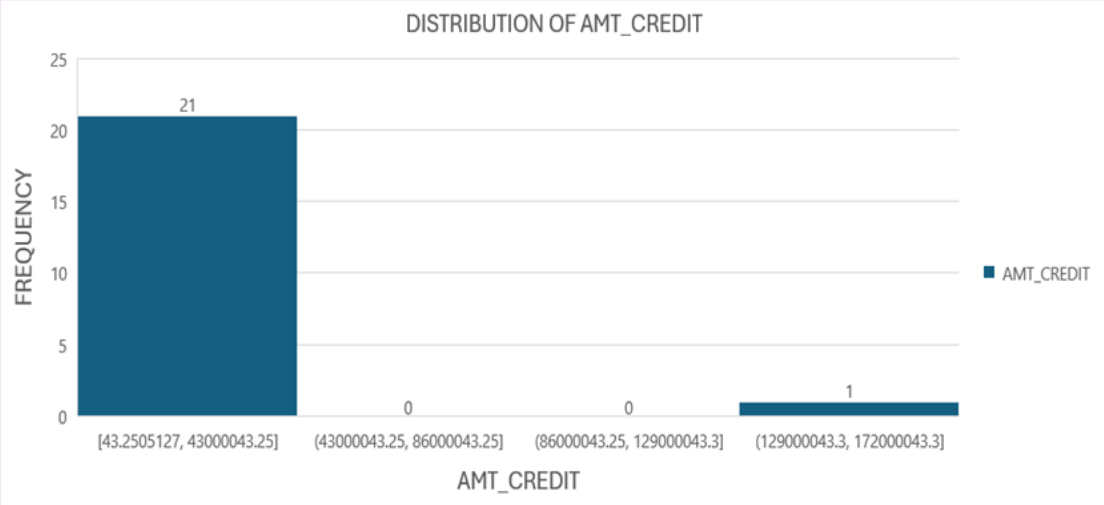
The rest of the descriptive analysis columns output link is given below :

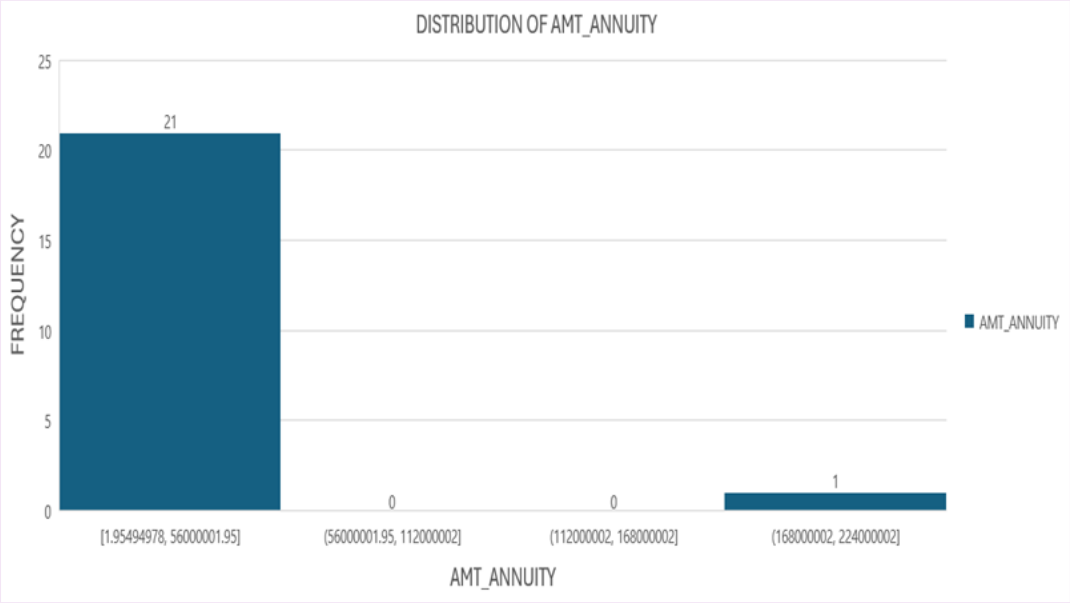
https://docs.google.com/spreadsheets/d/1-cJcCWAY_A7eAVPrI7gKQY7S3k-6ho1m/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

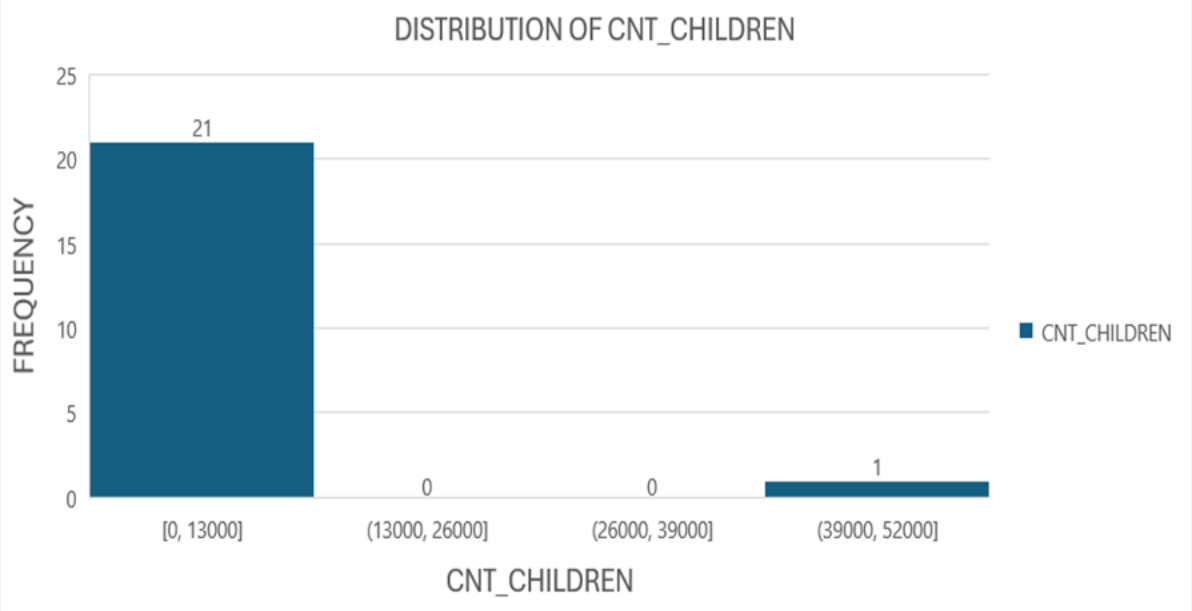
I have used histograms and bar chart to show distribution of individual variable below

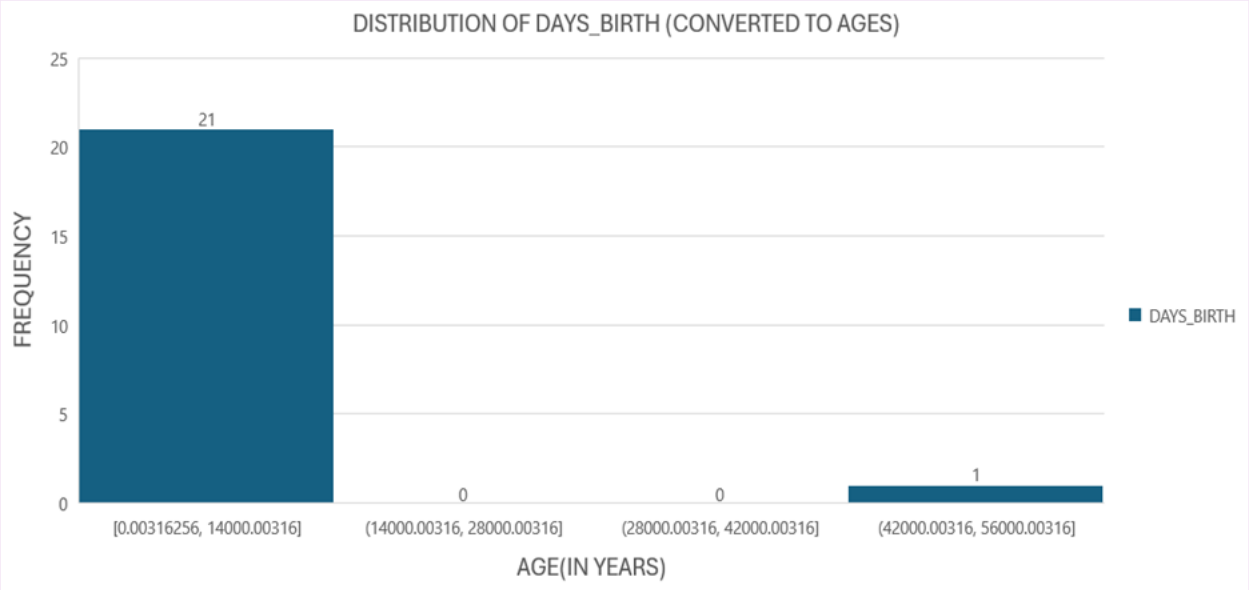
HISTOGRAMS :







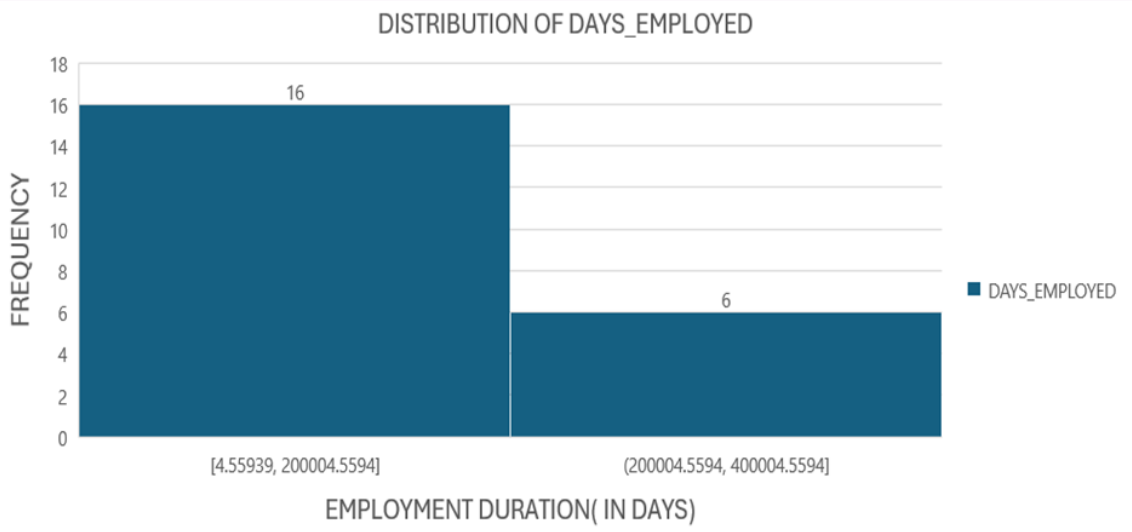


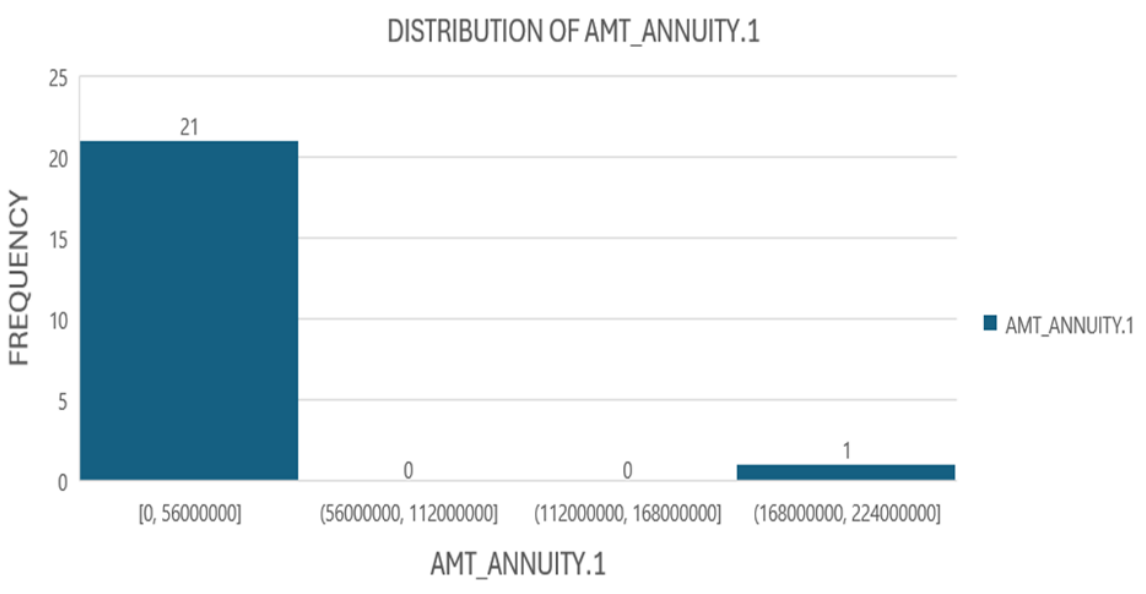


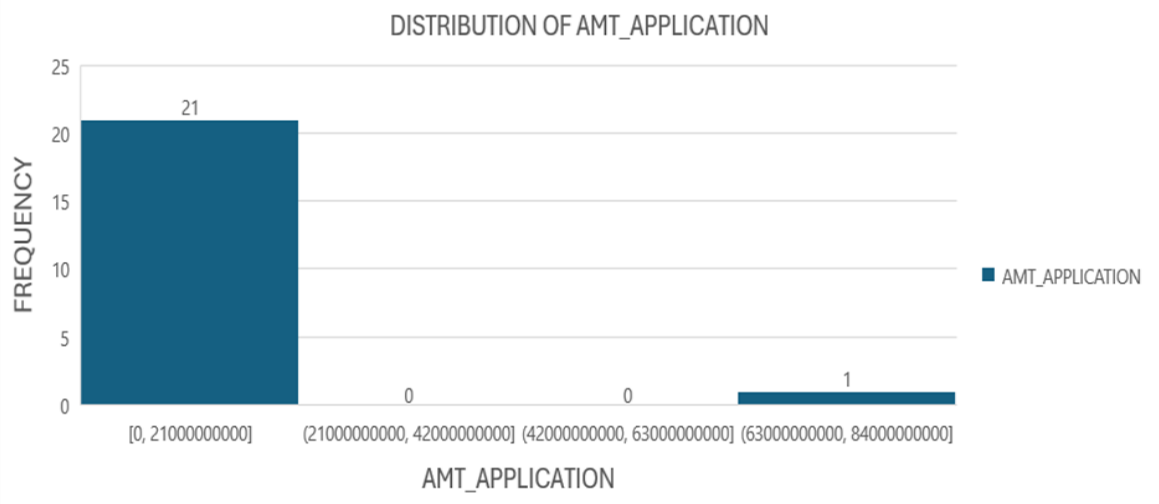
DAYS_BIRTH column is converted to years for positive values on X –Axis

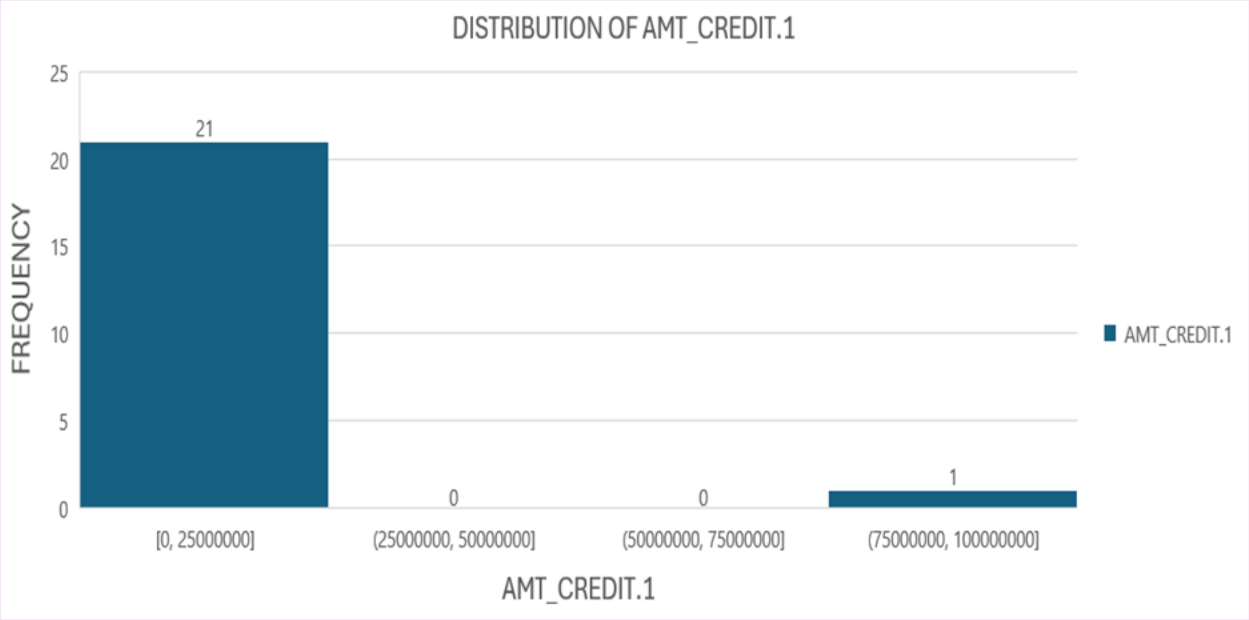
The formula used to convert age in days into age in years :

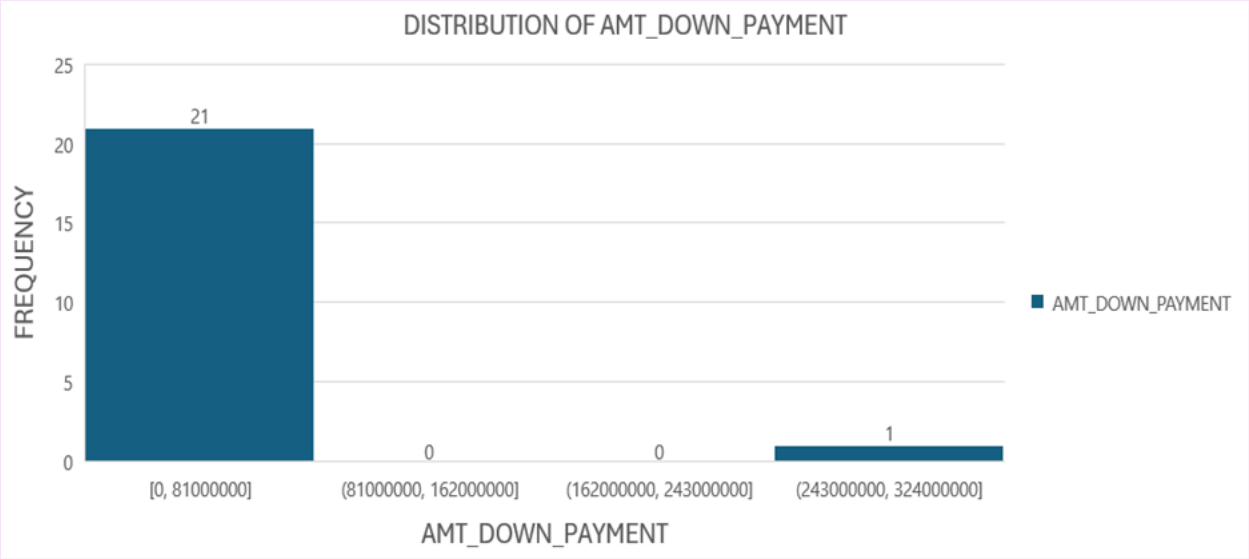
= the age value / 365.25

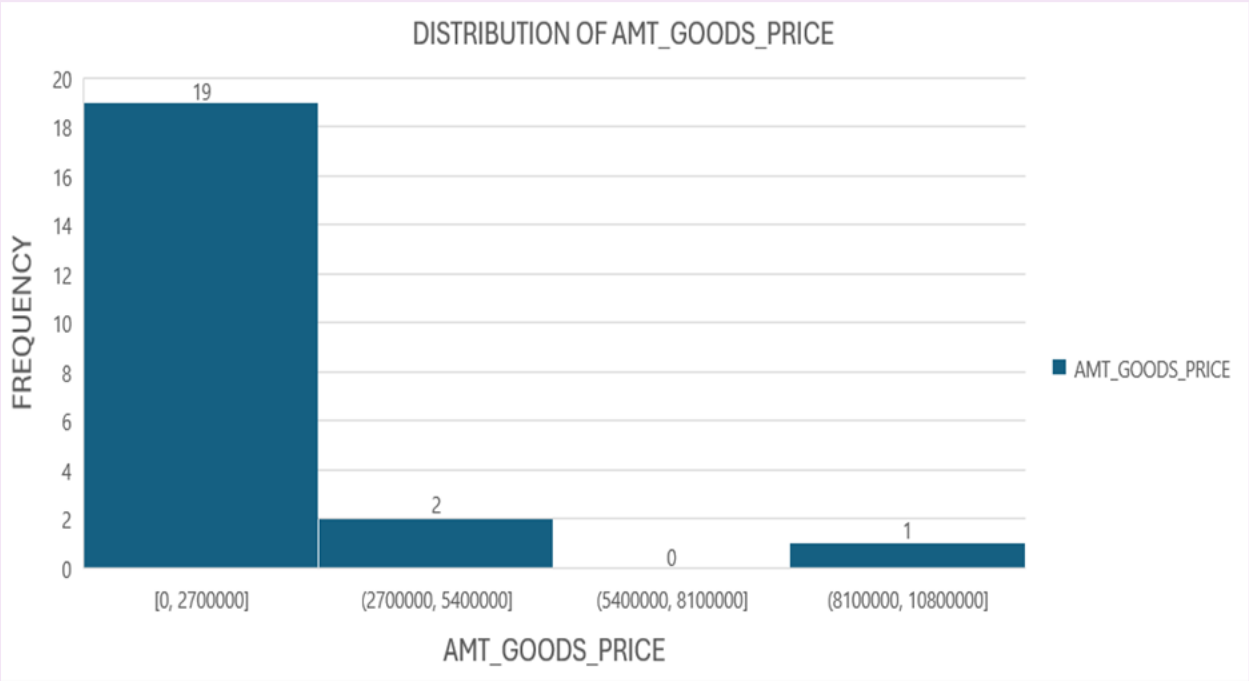

DISTRIBUTION OF DAYS_EMPLOYED
FREQUENCY
18
16
14
12
10
8
6
4
2
0
16
6
[4.55939, 200004.5594]
(200004.5594, 400004.5594]
EMPLOYMENT DURATION(IN DAYS)
DAYS_EMPLOYED


DISTRIBUTION OF AMT_ANNUITY.1
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 56000000]
(56000000, 112000000]
(112000000, 168000000]
(168000000, 224000000]
AMT_ANNUITY.1
AMT_ANNUITY.1

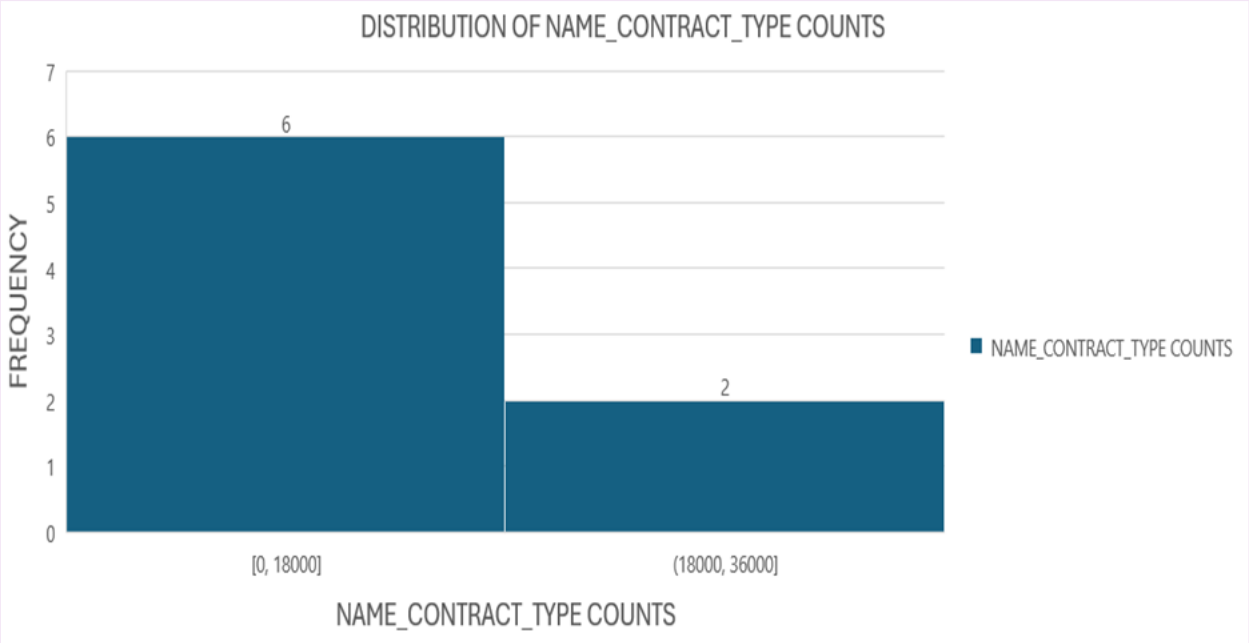

DISTRIBUTION OF AMT_APPLICATION
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 21000000000]
(21000000000, 42000000000]
(42000000000, 63000000000]
(63000000000, 84000000000]
AMT_APPLICATION
AMT_APPLICATION

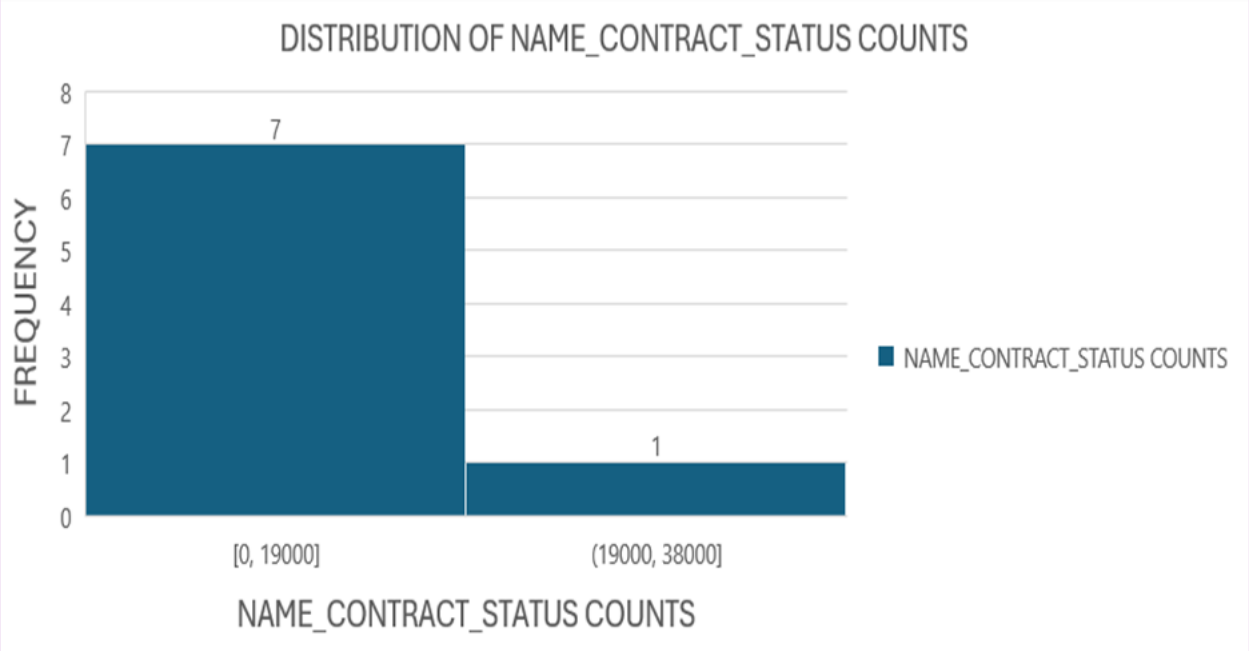

DISTRIBUTION OF AMT_CREDIT.1
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 25000000]
(25000000, 50000000]
(50000000, 75000000]
(75000000, 100000000]
AMT_CREDIT.1
AMT_CREDIT.1


DISTRIBUTION OF AMT_DOWN_PAYMENT
FREQUENCY
25
20
15
10
5
0
21
0
0
1
[0, 81000000]
(81000000, 162000000]
(162000000, 243000000]
(243000000, 324000000]
AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT

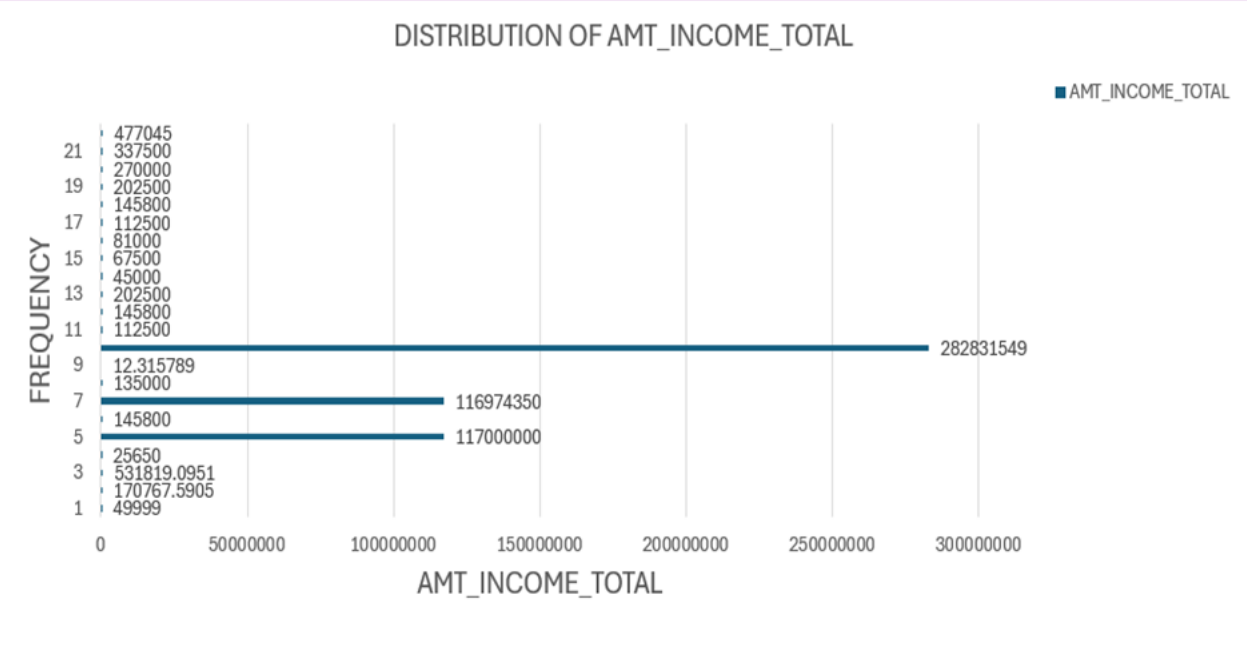

DISTRIBUTION OF AMT_GOODS_PRICE
FREQUENCY
20
18
16
14
12
10
8
6
4
2
0
19
2
0
1
[0, 2700000]
(2700000, 5400000]
(5400000, 8100000]
(8100000, 10800000]
AMT_GOODS_PRICE
AMT_GOODS_PRICE

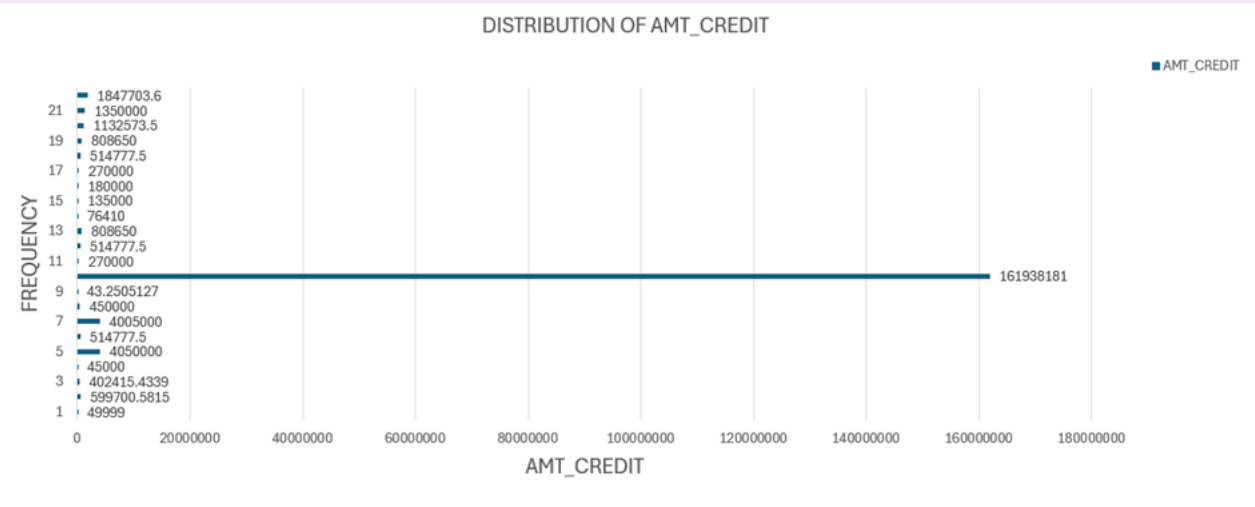
DISTRIBUTION OF VARIABLE COUNTS

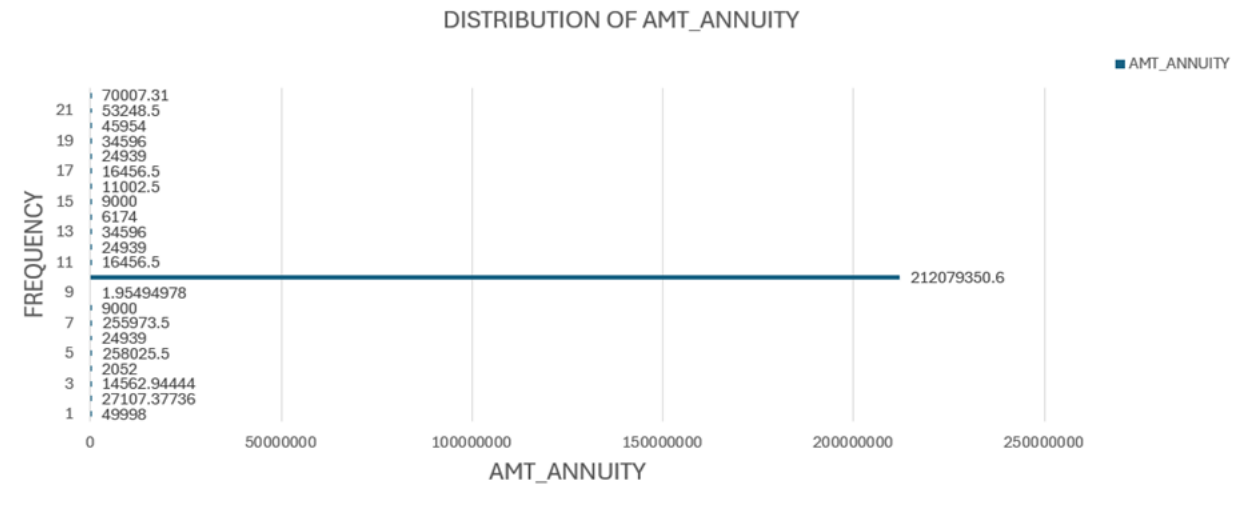


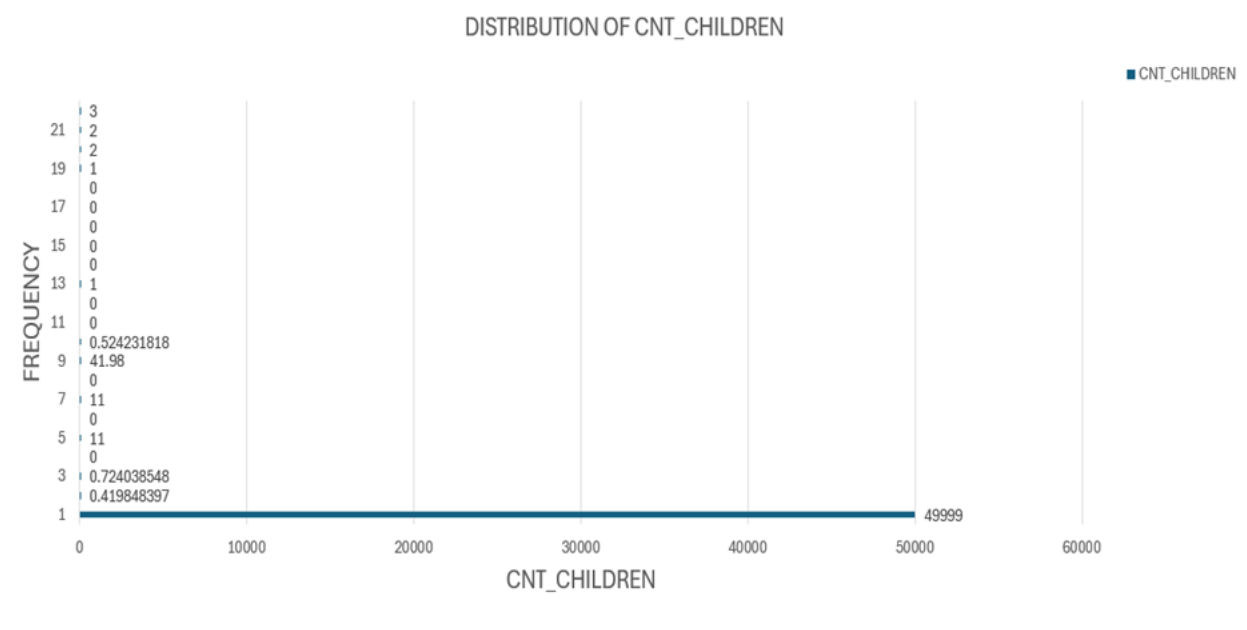


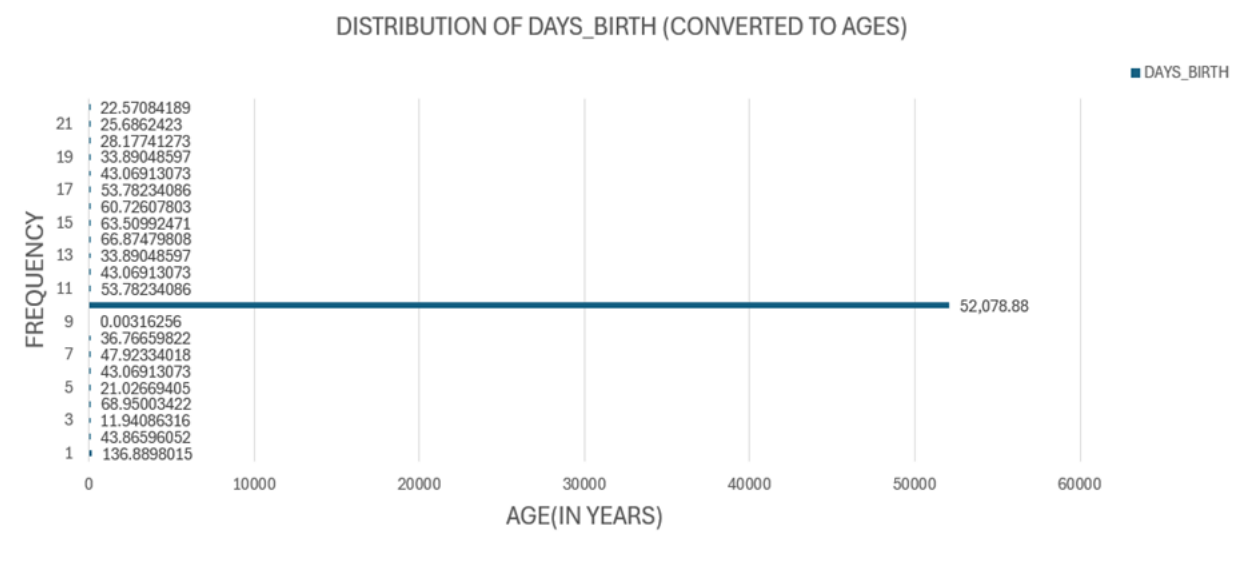
BAR CHARTS :


DISTRIBUTION OF AMT_INCOME_TOTAL
AMT_INCOME_TOTAL
FREQUENCY
AMT_INCOME_TOTAL
477045
337500
270000
202500
145800
112500
81000
67500
45000
202500
145800
112500
282831549
12.315789
135000
116974350
145800
117000000
25650
531819.0951
170767.5905
49999
0
50000000
100000000
150000000
200000000
250000000
300000000
1
3
5
7
9
11
13
15
17
19
21


DISTRIBUTION OF AMT_CREDIT
AMT_CREDIT
FREQUENCY
AMT_CREDIT
1847703.6
1350000
1132573.5
808650
514777.5
270000
180000
135000
76410
808650
514777.5
270000
161938181
43.2505127
450000
4005000
514777.5
4050000
45000
402415.4339
599700.5815
49999
0
20000000
40000000
60000000
80000000
100000000
120000000
140000000
160000000
180000000
1
3
5
7
9
11
13
15
17
19
21


DISTRIBUTION OF AMT_ANNUITY
AMT_ANNUITY
FREQUENCY
AMT_ANNUITY
70007.31
53248.5
45954
34596
24939
16456.5
11002.5
9000
6174
34596
24939
16456.5
212079350.6
1.95494978
9000
255973.5
24939
258025.5
2052
14562.94444
27107.37736
49998
0
50000000
100000000
150000000
200000000
250000000
1
3
5
7
9
11
13
15
17
19
21

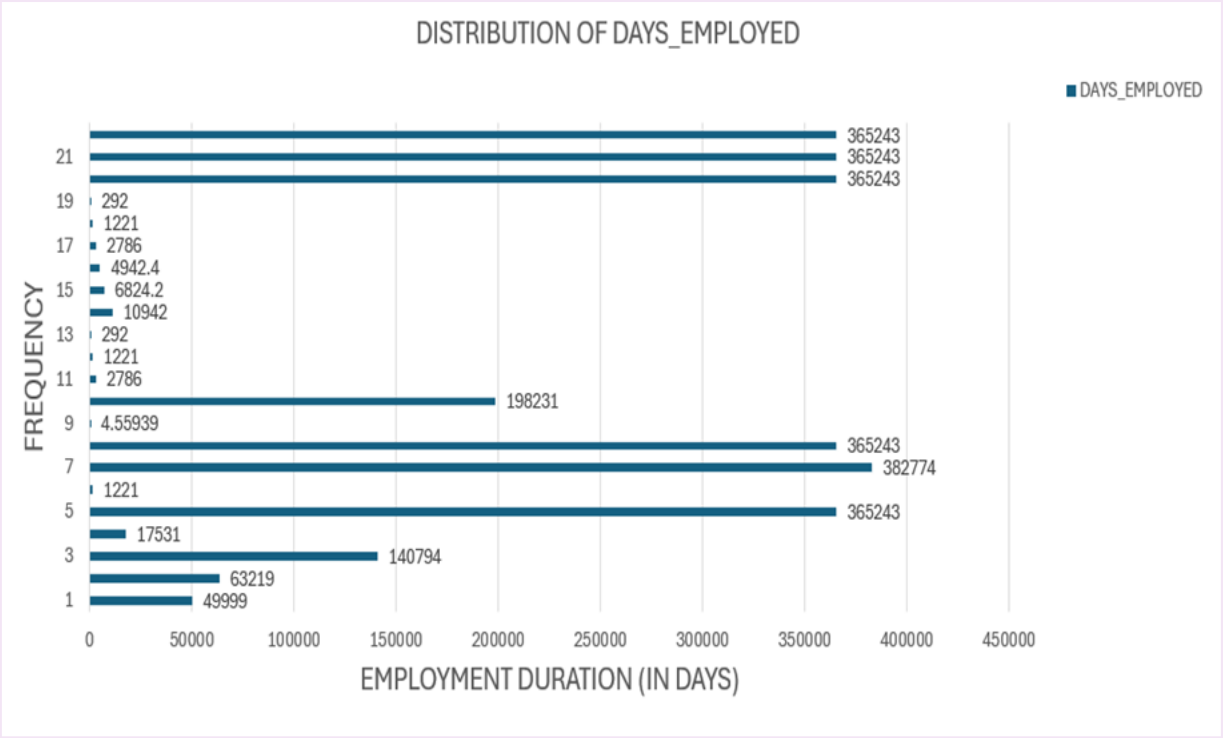


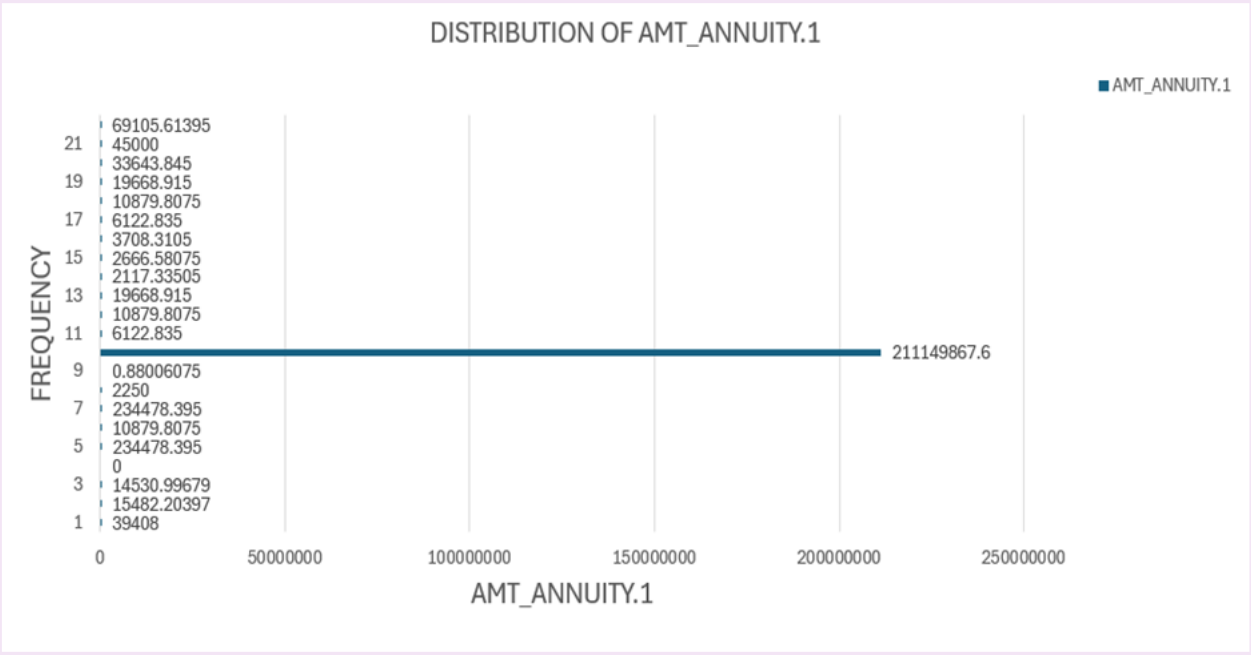


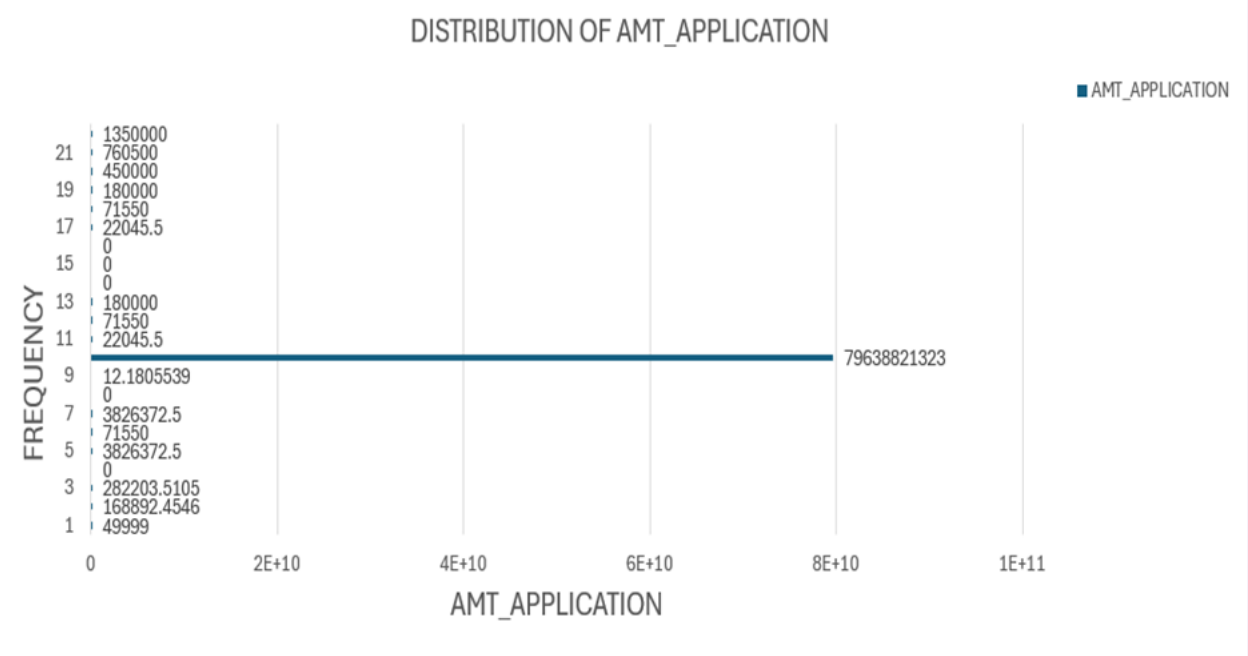
DAYS_BIRTH column is converted to years for positive values on X – Axis

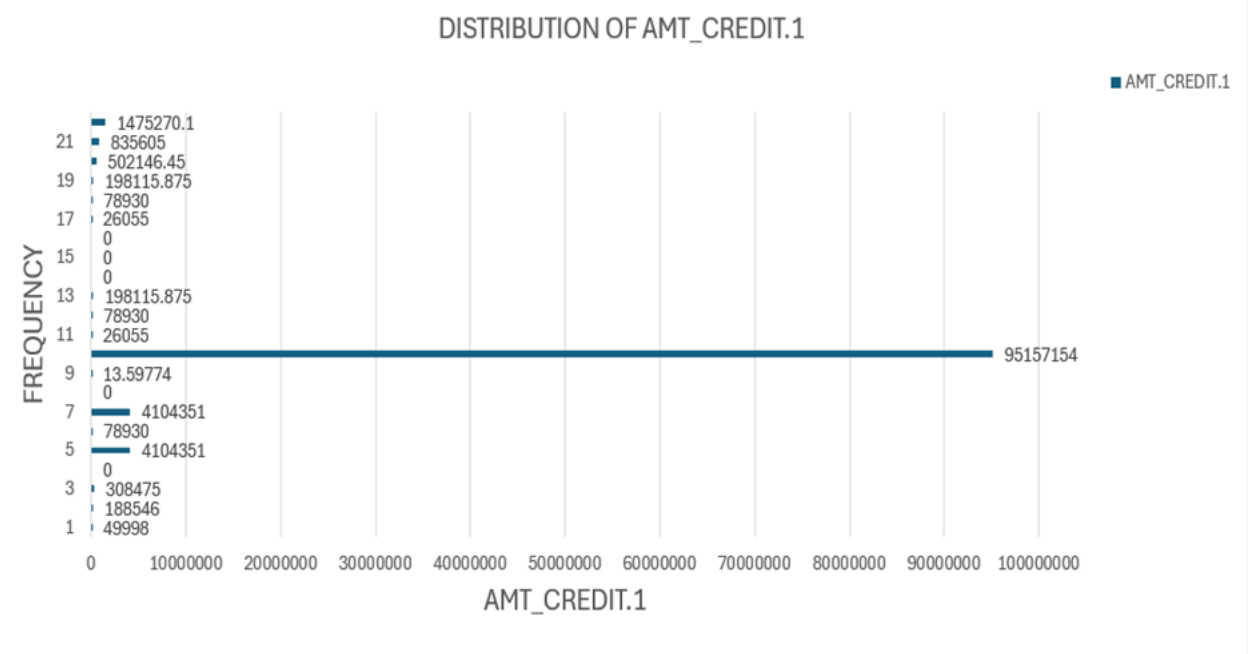
The formula used to convert age in days into age in years :

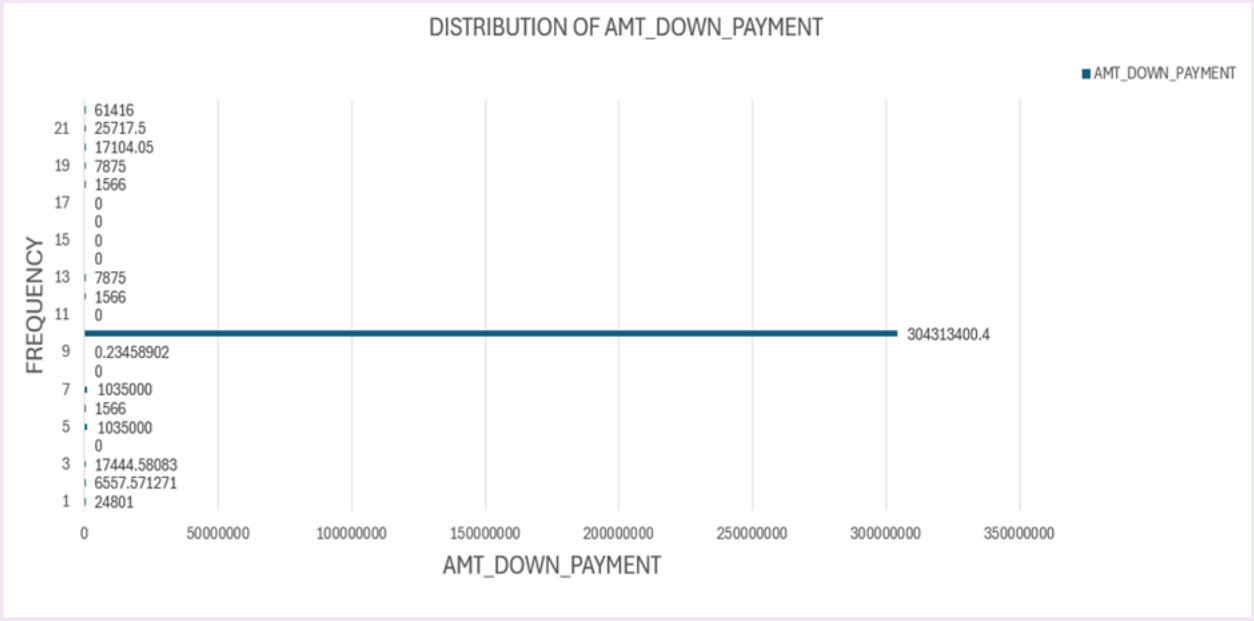
= the age value / 365.25

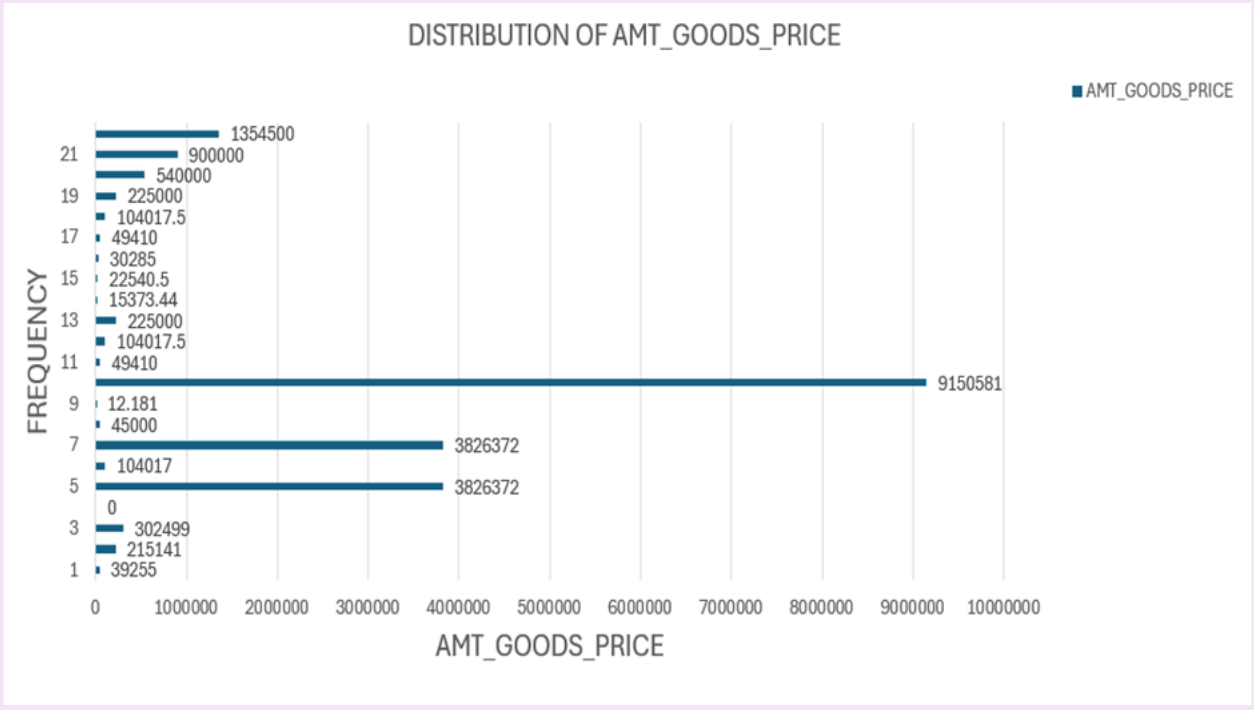

DISTRIBUTION OF DAYS_EMPLOYED
DAYS_EMPLOYED
FREQUENCY
EMPLOYMENT DURATION (IN DAYS)
365243
365243
365243
292
1221
2786
4942.4
6824.2
10942
292
1221
2786
198231
4.55939
365243
382774
1221
365243
17531
140794
63219
49999
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
1
3
5
7
9
11
13
15
17
19
21


DISTRIBUTION OF AMT_ANNUITY.1
AMT_ANNUITY.1
FREQUENCY
AMT_ANNUITY.1
69105.61395
45000
33643.845
19668.915
10879.8075
6122.835
3708.3105
2666.58075
2117.33505
19668.915
10879.8075
6122.835
211149867.6
0.88006075
2250
234478.395
10879.8075
234478.395
0
14530.99679
15482.20397
39408
0
50000000
100000000
150000000
200000000
250000000
1
3
5
7
9
11
13
15
17
19
21

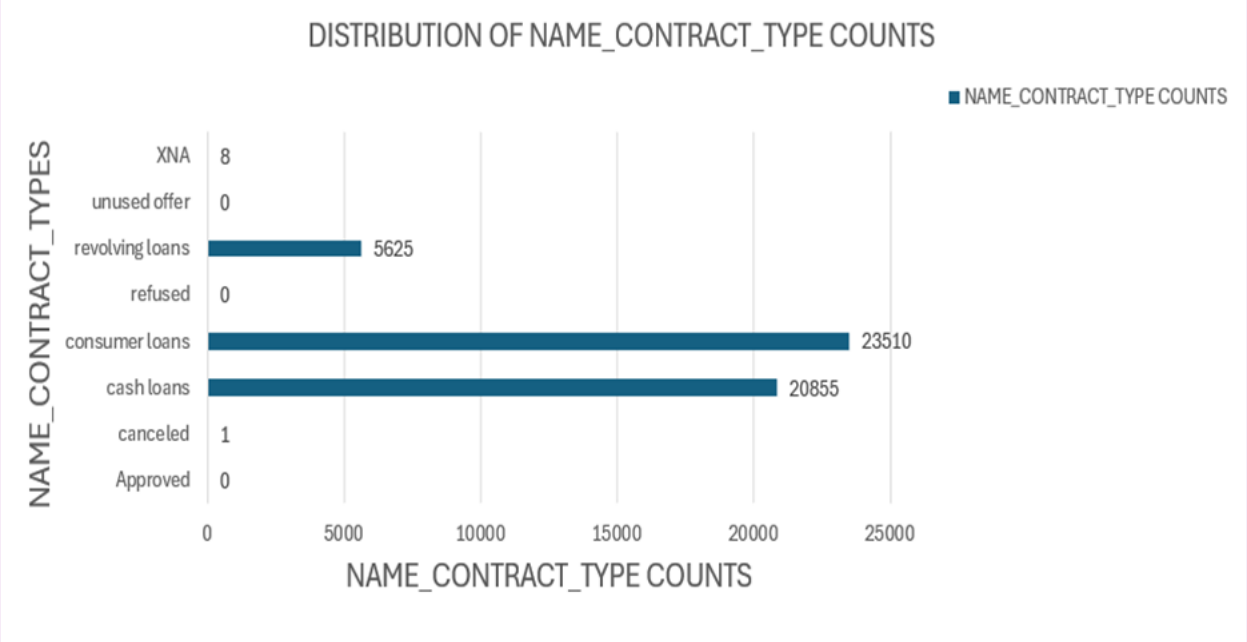

DISTRIBUTION OF AMT_APPLICATION
AMT_APPLICATION
FREQUENCY
1350000
760500
450000
180000
71550
22045.5
0
0
0
180000
71550
22045.5
79638821323
12.1805539
0
3826372.5
71550
3826372.5
0
282203.5105
168892.4546
49999
0
2E+10
4E+10
6E+10
8E+10
1E+11
AMT_APPLICATION

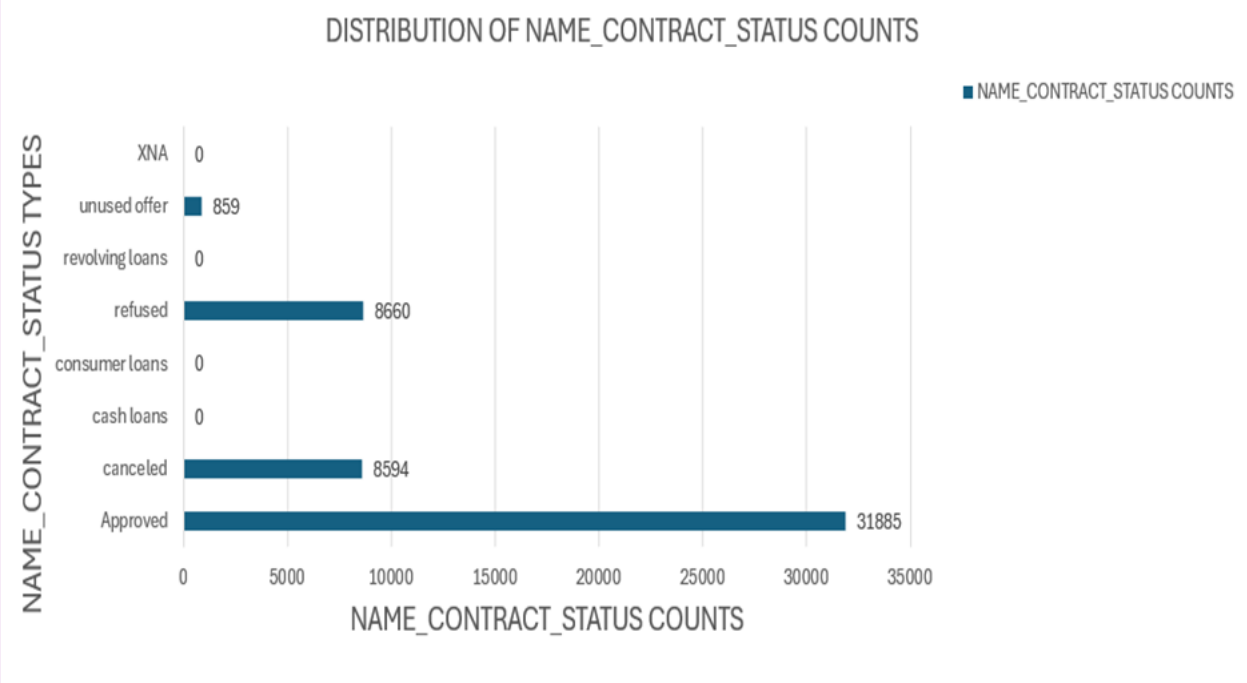

DISTRIBUTION OF AMT_CREDIT.1
AMT_CREDIT.1
FREQUENCY
1475270.1
835605
502146.45
198115.875
78930
26055
0
0
0
198115.875
78930
26055
95157154
13.59774
0
4104351
78930
4104351
0
308475
188546
49998
0
10000000
20000000
30000000
40000000
50000000
60000000
70000000
80000000
90000000
100000000
AMT_CREDIT.1


DISTRIBUTION OF AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
FREQUENCY
AMT_DOWN_PAYMENT
61416
25717.5
17104.05
7875
1566
0
0
0
0
7875
1566
0
304313400.4
0.23458902
0
1035000
1566
1035000
0
17444.58083
6557.571271
24801
0
50000000
100000000
150000000
200000000
250000000
300000000
350000000
1
3
5
7
9
11
13
15
17
19
21


DISTRIBUTION OF AMT_GOODS_PRICE
AMT_GOODS_PRICE
FREQUENCY
AMT_GOODS_PRICE
1354500
900000
540000
225000
104017.5
49410
30285
22540.5
15373.44
225000
104017.5
49410
9150581
12.181
45000
3826372
104017
3826372
0
302499
215141
39255
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000
1
3
5
7
9
11
13
15
17
19
21

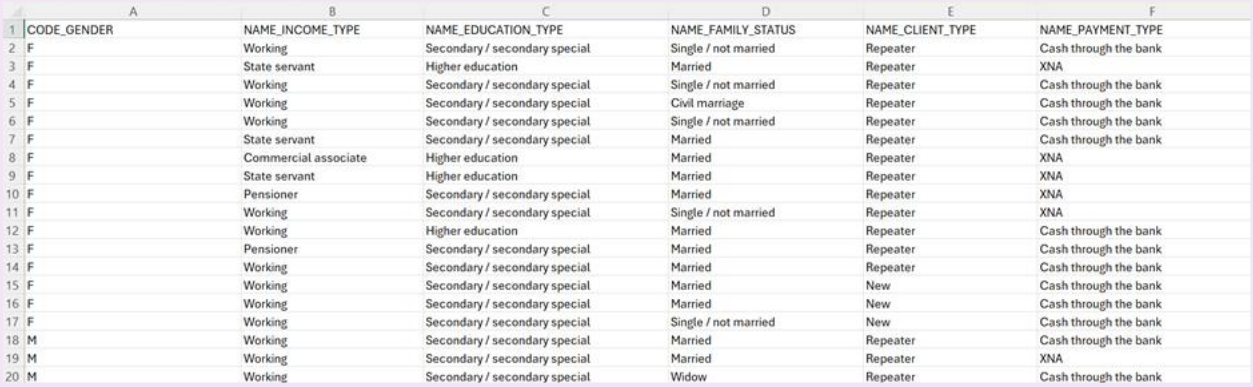
DISTRIBUTION OF VARIABLE COUNTS





2)SEGMENTED UNIVARIATE ANALYSIS:

For Segmented univariate Analysis the columns CODE_GENDER, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_CLIENT_TYPE, NAME_PAYMENT_TYPE, CHANNEL_TYPE, NAME_CONTRACT_TYPE, NAME_CONTRACT_TYPE.1, OWN_CAR_AGE are taken



	A	B	C	D	E	F
1	CODE_GENDER	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_CLIENT_TYPE	NAME_PAYMENT_TYPE
2	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
3	F	State servant	Higher education	Married	Repeater	XNA
4	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
5	F	Working	Secondary / secondary special	Civil marriage	Repeater	Cash through the bank
6	F	Working	Secondary / secondary special	Single / not married	Repeater	Cash through the bank
7	F	State servant	Secondary / secondary special	Married	Repeater	Cash through the bank
8	F	Commercial associate	Higher education	Married	Repeater	XNA
9	F	State servant	Higher education	Married	Repeater	XNA
10	F	Pensioner	Secondary / secondary special	Married	Repeater	XNA
11	F	Working	Secondary / secondary special	Single / not married	Repeater	XNA
12	F	Working	Higher education	Married	Repeater	Cash through the bank
13	F	Pensioner	Secondary / secondary special	Married	Repeater	Cash through the bank
14	F	Working	Secondary / secondary special	Married	Repeater	Cash through the bank
15	F	Working	Secondary / secondary special	Married	New	Cash through the bank
16	F	Working	Secondary / secondary special	Married	New	Cash through the bank
17	F	Working	Secondary / secondary special	Single / not married	New	Cash through the bank
18	M	Working	Secondary / secondary special	Married	Repeater	Cash through the bank
19	M	Working	Secondary / secondary special	Married	Repeater	XNA
20	M	Working	Secondary / secondary special	Widow	Repeater	Cash through the bank

The rest of the columns chosen to perform segmented univariate analysis are in the given link below :

https://docs.google.com/spreadsheets/d/16sDbOypfawu5Guv0hyRxp8crEOJFisYP/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Segments chosen are given below :

1) CODE_GENDER segmented by NAME_INCOME_TYPE

2) NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE

3) NAME_EDUCATION_ TYPE segmented by NAME_FAMILY_STATUS

4) NAME_FAMILY_STATUS segmented by NAME_CLIENT_TYPE

5) NAME_CLIENT_TYPE segmented by NAME_PAYMENT_TYPE

6) NAME_PAYMENT_TYPE segmented by CHANNEL_TYPE

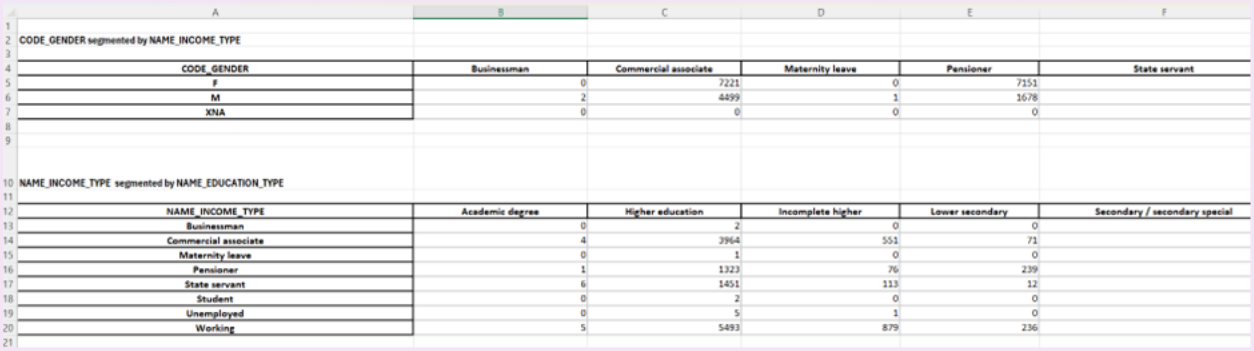
7) NAME_CONTRACT_TYPE segmented by CHANNEL_TYPE

8) NAME_CONTRACT_TYPE segmented by CODE_GENDER

9) NAME_CONTRACT_TYPE.1 segmented by OWN_CAR_AGE

10) OWN_CAR_AGE segmented by CODE_GENDER

These segments are done using pivot table



CODE_GENDER segmented by NAME_INCOME_TYPE

CODE_GENDER	Businessman	Commercial associate	Maternity leave	Pensioner	State servant
F	0	7221	0	7151	
M	2	4499	1	1678	
XNA	0	0	0	0	

NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE

NAME_INCOME_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
Businessman	0	2	0	0	
Commercial associate	4	3964	551	71	
Maternity leave	0	1	0	0	
Pensioner	1	1323	76	239	
State servant	6	1451	113	12	
Student	0	2	0	0	
Unemployed	0	5	1	0	
Working	5	5493	879	236	

The Rest of the segments done using pivot tables are in the given link below :

https://docs.google.com/spreadsheets/d/19_q79GtofixKYSdwJPeJdAQ4VDfVO1jn/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Descriptive Analysis is done to each of the segment's output found using pivot table

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance, quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

   =COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

CODE_GENDER segmented by NAME_INCOME_TYPE

CODE_GENDER	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student
F	0	7221	0	7151	2519	1
M	2	4499	1	1678	976	3
XNA	0	0	0	0	0	0
count	3	3	3	3	3	3
mean	675.9090909	2722.818182	1845.636364	763.7727273	2440.0625	346.375
std	1517.649684	5076.551357	4904.395342	1803.59432	5105.476073	803.4144878
min	0	0	0	0	0	0
max	6303	22586	22909	7151	19307	2311
Median	4.5	562.5	94.5	11	1.5	14.5
Range	6303	22586	22909	7151	19307	2311
mode	0	0.0,2.0	0	0	0	1
percentage	37.5	37.5	37.5	37.5	37.5	37.5
variance	2303260.563	25771373.68	24053093.67	3252952.47	26065885.93	645474.8393
quantiles(0.25)	0	5.75	0.25	0.25	0	1
quantiles(0.5)	4.5	562.5	94.5	11	1.5	14.5
quantiles(0.75)	659	3613.25	1750.5	238.25	2064.5	139.75
Percentile(1%)	0	0	0	0	0	0.07
Percentile(5%)	0	0.05	0	0	0	0.35
percentile(10%)	0	1.1	0	0	0	0.7
percentile(25%)	0	5.75	0.25	0.25	0	1
percentile(50%)	4.5	562.5	94.5	11	1.5	14.5
percentile(75%)	659	3613.25	1750.5	238.25	2064.5	139.75
percentile(90%)	1385.4	7048.2	3340.8	2077.6	7160	948.1

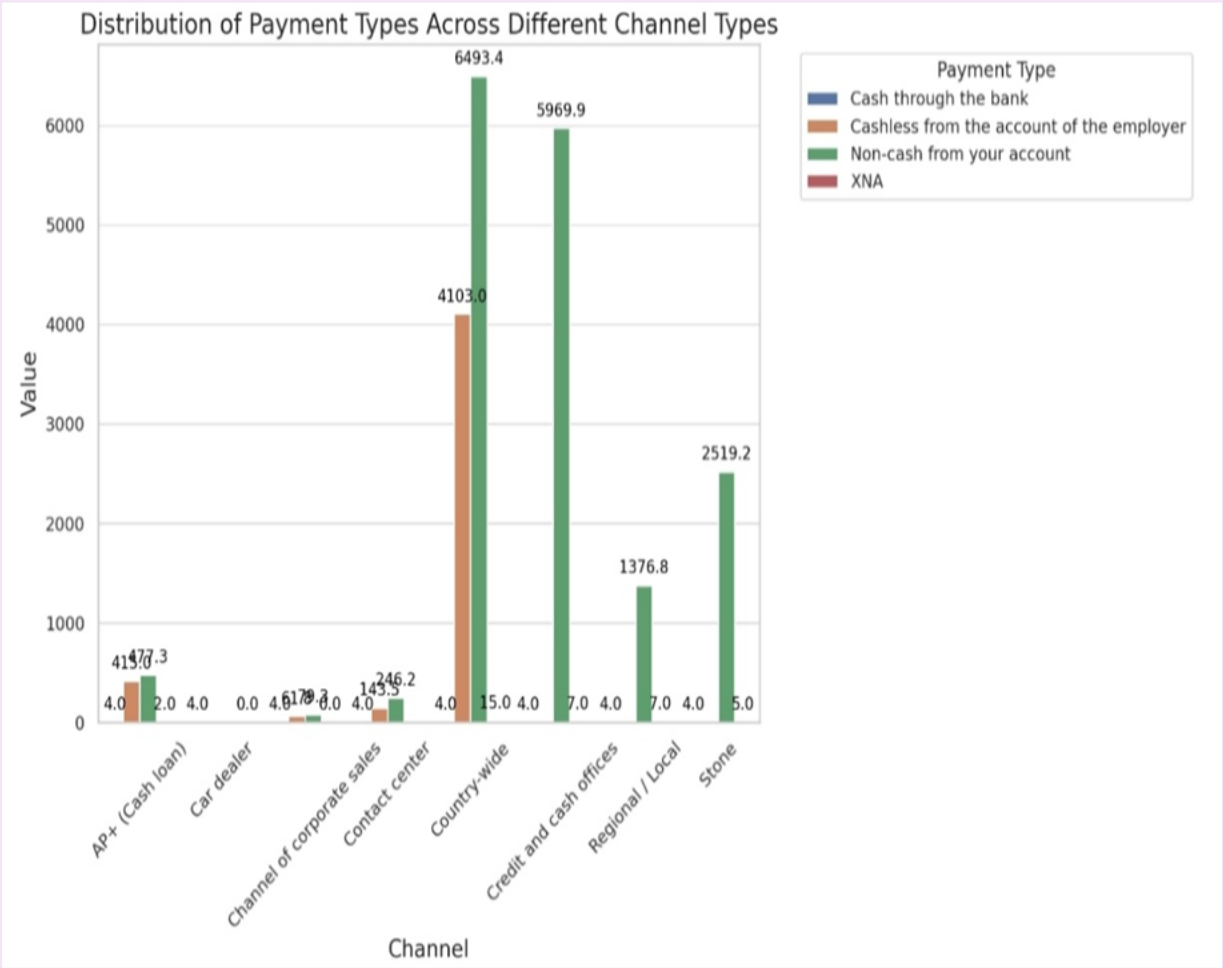
The Rest of the segment's descriptive analysis output is given in the below link :

https://docs.google.com/spreadsheets/d/1ERFXED5g68gr3Vd_fwofd2Rog3rlfiHE/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

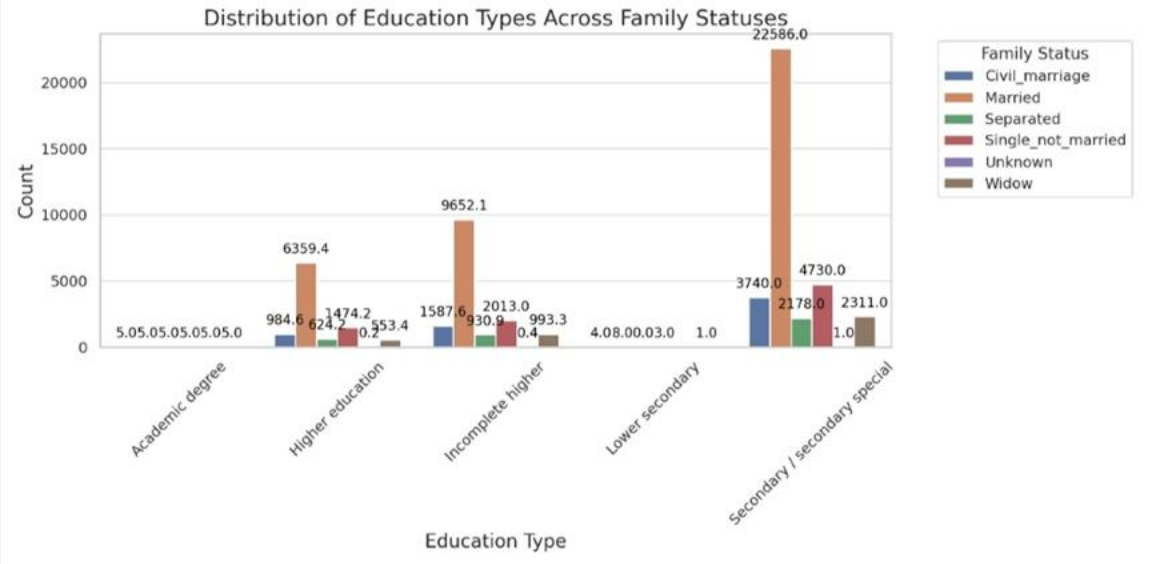
The bar charts plotted to compare variable distributions across different scenarios are given below

BAR CHARTS:

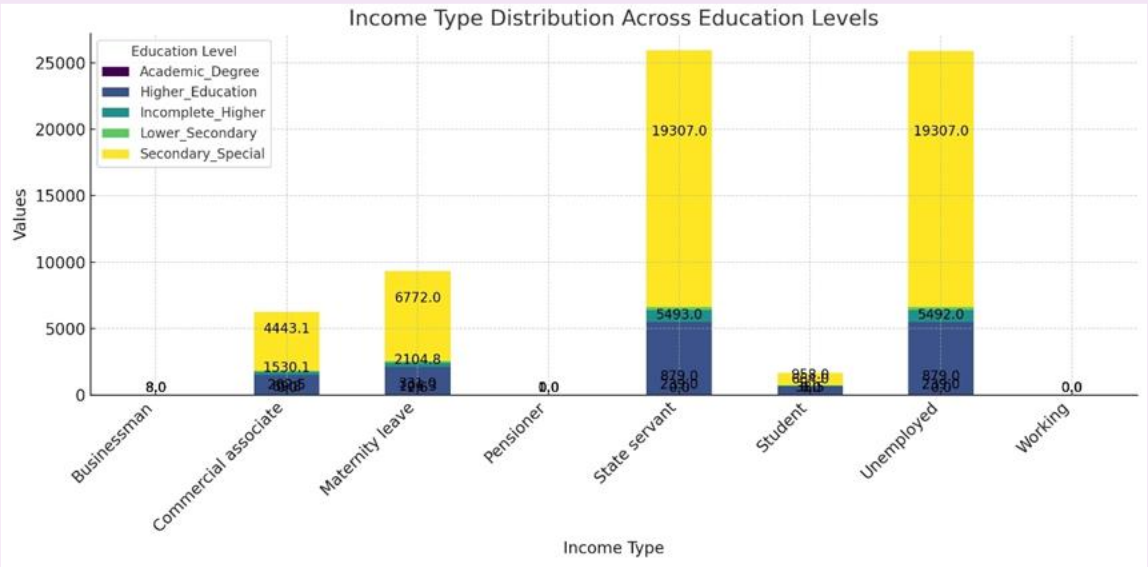
NAME_PAYMENT_TYPE segmented by CHANNEL_TYPE



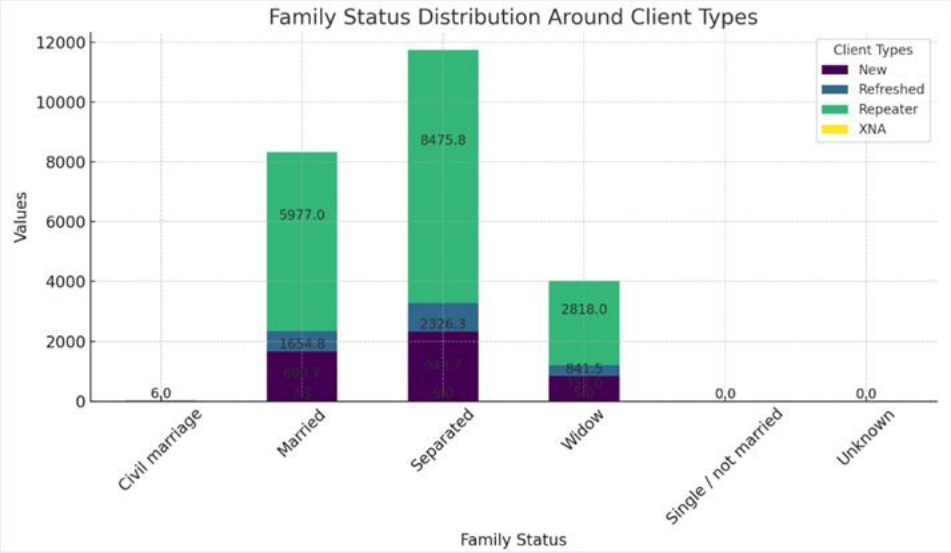
NAME_EDUCATION_ TYPE segmented by NAME_FAMILY_STATUS



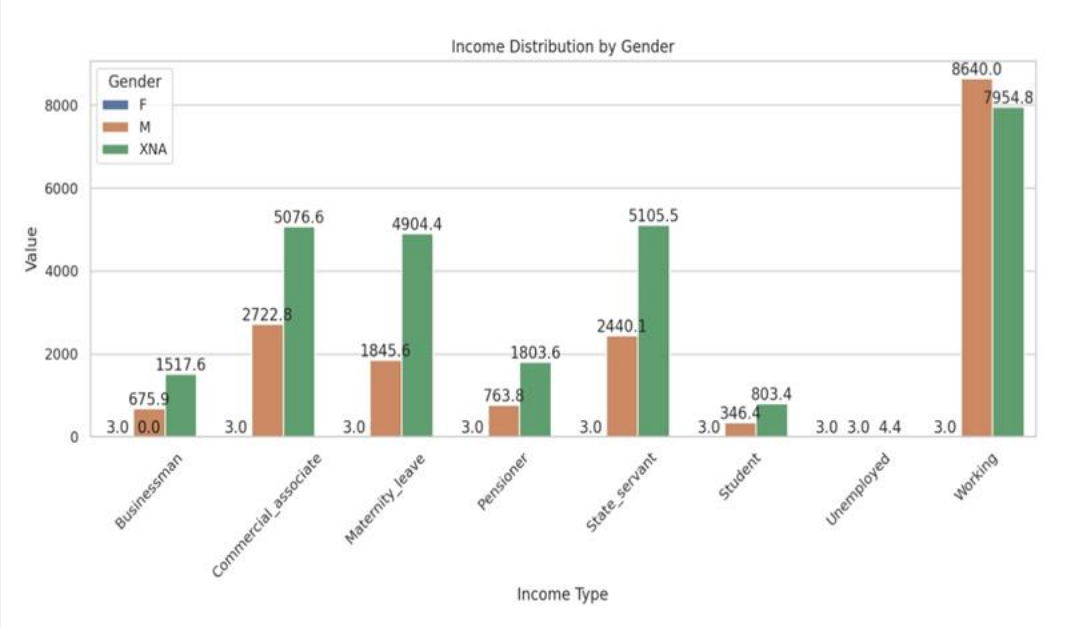
NAME_INCOME_TYPE segmented by NAME_EDUCATION_TYPE



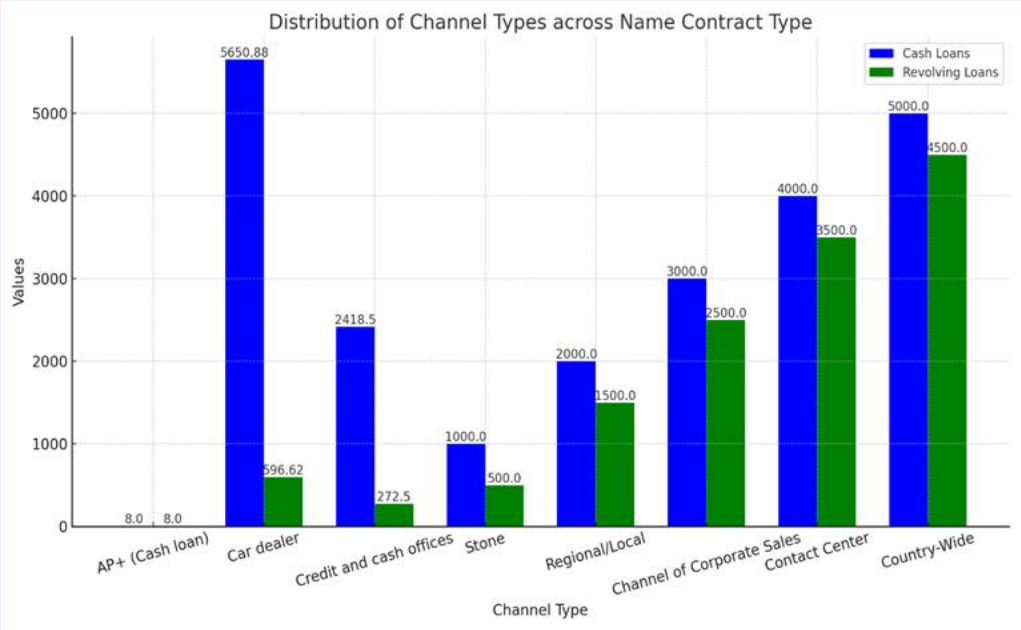
NAME_FAMILY_STATUS segmented by NAME_CLIENT_TYPE



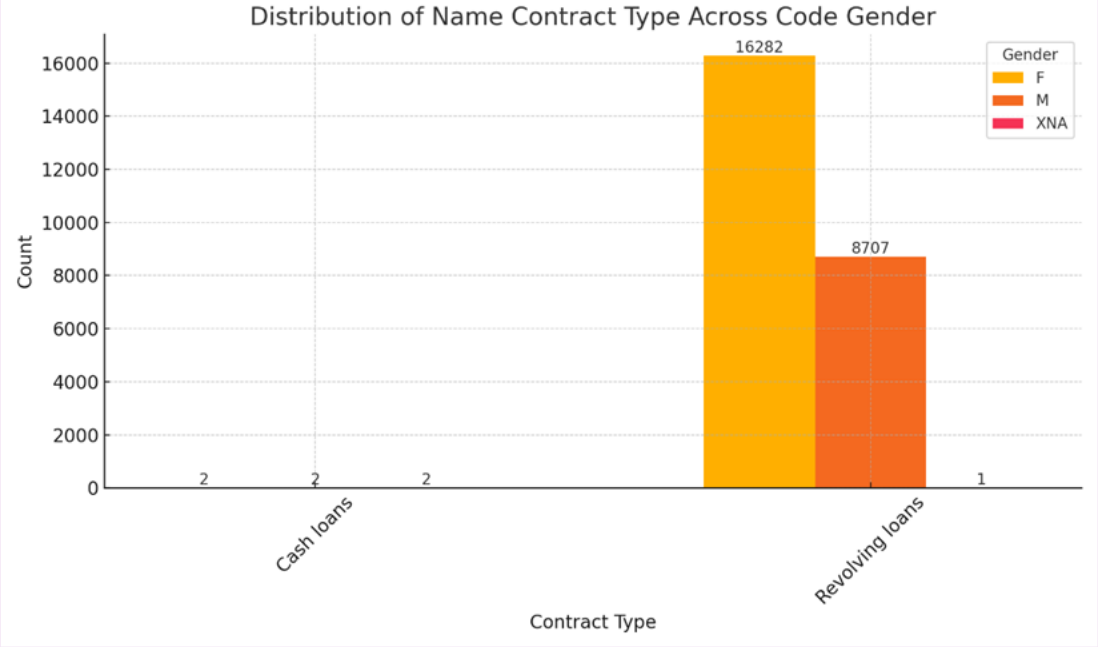
CODE_GENDER segmented by NAME_INCOME_TYPE



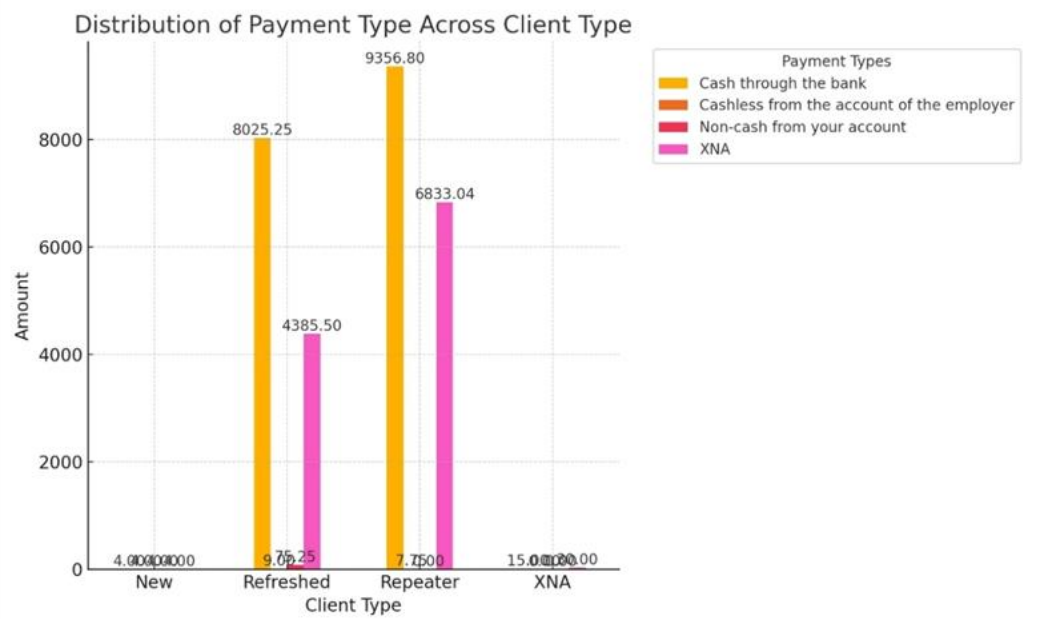
CHANNEL _TYPE segmented by NAME_CONTRACT_TYPE



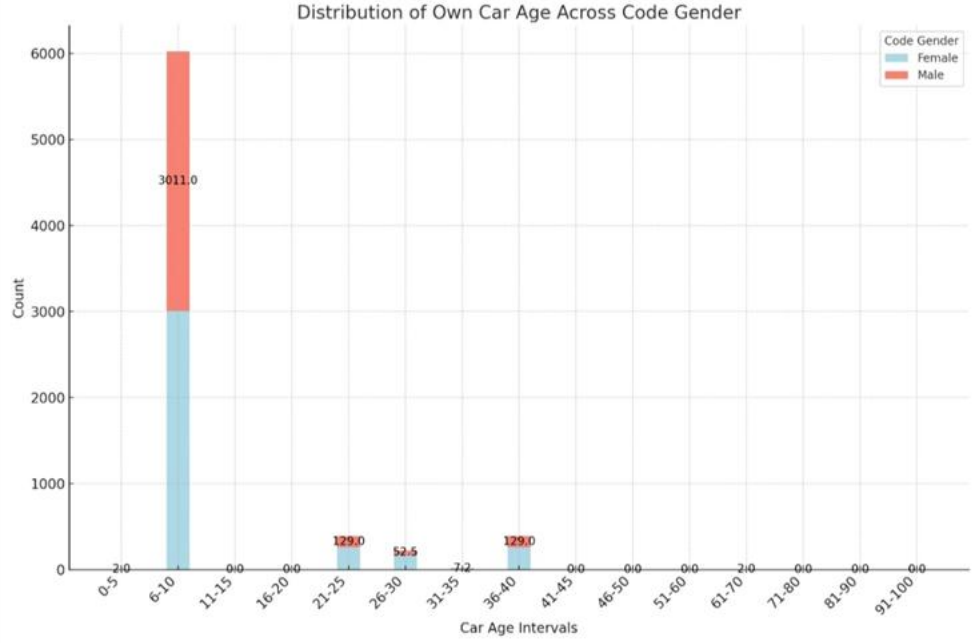
NAME_CONTRACT_TYPE segmented by CODE_GENDER



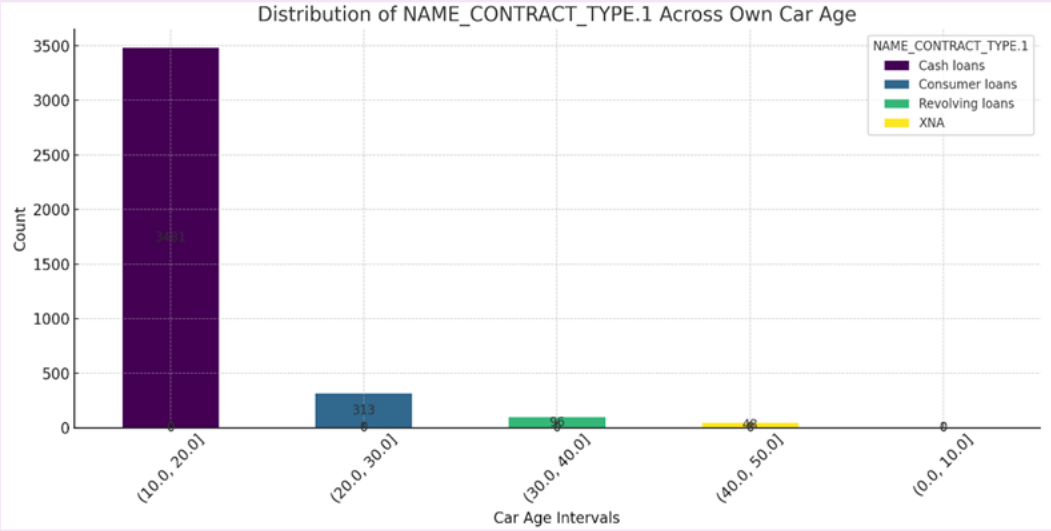
NAME_CLIENT_TYPE segmented by NAME_PAYMENT_TYPE



OWN_CAR_AGE segmented by CODE_GENDER



NAME_CONTRACT_TYPE.1 segmented by OWN_CAR_AGE



3. bivariate analysis

For Bivariate analysis the columns AMT_CREDIT , NAME_CONTRACT_TYPE .1, AMT_APPLICATION, AMT_CREDIT 1, AMT_INCOME_TOTAL, TARGET, AMT_DOWN_PAYMENT are taken

	A	B	C	D	E	F	G
1	AMT_CREDIT	NAME_CONTRACT_TYPE .1	AMT_APPLICATION	AMT_CREDIT 1	AMT_INCOME_TOTAL	TARGET	AMT_DOWN_PAYMENT
2	17145	Consumer loans	17145	17145	202500	1	0
3	679671	Cash loans	607500	679671	270000	0	0
4	136444.5	Cash loans	112500	136444.5	67500	0	0
5	470790	Cash loans	450000	470790	135000	0	12649.5
6	404055	Cash loans	337500	404055	121500	0	1350
7	340573.5	Cash loans	315000	340573.5	99000	0	0
8	0	Cash loans	0	0	171000	0	9000
9	0	Cash loans	0	0	360000	0	0
10	0	Cash loans	0	0	112500	0	0
11	0	Cash loans	0	0	135000	0	0
12	335754	Cash loans	270000	335754	112500	0	0
13	246397.5	Cash loans	211500	246397.5	38419.155	0	13500
14	174361.5	Cash loans	148500	174361.5	67500	0	0
15	57564	Consumer loans	53779.5	57564	225000	0	4500
16	27252	Consumer loans	26550	27252	189000	0	0
17	119853	Consumer loans	126490.5	119853	157500	0	0
18	27297	Consumer loans	26955	27297	108000	0	7573.5
19	180000	Revolving loans	180000	180000	81000	0	9000

The rest of the Full Data of the columns chosen to perform bivariate analysis are in the given link below :

https://docs.google.com/spreadsheets/d/1cTCFZvgNtDc8gfoOWMKdS_ehlnsNCWKh/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Segments chosen are given below :

1) AMT_CREDIT segmented by AMT_CREDIT_1

2) APPLICATION_AMT segmented by TARGET

3) INCOME_TOTAL segmented by DOWN_PAYMENT

4) INCOME_TOTAL segmented by TARGET

5) CREDIT_AMT segmented by INCOME_TOTAL

6) DOWN_PAYMENT segmented by APPLICATION_AMT

7) CREDIT_AMT_1 segmented by DOWN_PAYMENT

8) CONTRACT_TYPE_1 segmented by TARGET

These segments are done using pivot table

	A	B
1		
2	amt_credit segmented by amt_credit_1	
3		
4		AMT_CREDIT
5	0	9435
6	6948	1
7	7879.5	1
8	8281.08	1
9	8649	1
10		
11	Application_amt segmented by target	
12		
13		AMT_APPLICATION
14	0	45973
15	1	4026

The Rest of the segments done using pivot tables are in the given link below :

https://docs.google.com/spreadsheets/d/1F0Fem3IY7msm2a3nPVK7lsPB7VBMjRG4/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Analysis is done to each of the segment's output found using pivot table

To find Correlation Coefficient I have used the CORREL() Function :

= CORREL(A5 : A9 , B5 : B9)

AMT_CREDIT SEGMENTED BY AMT_CREDIT_1

2	amt_credit segmented by amt_credit_1	
3		
4		AMT_CREDIT
5	0	9435
6	6948	1
7	7879.5	1
8	8281.08	1
9	8649	1

The Correlation between amt_credit_1 & amt_credit : -0.984

Indicating a negative correlation between the two variables suggesting that as the amt_credit_1 increases the amt_credit tends to decrease

APPLICATION_AMT SEGMENTED BY TARGET

11	Application_amt segmented by target	
12		
13		AMT_APPLICATION
14	0	45973
15	1	4026

The Correlation between Application_amt & target : -0.2955118903717489

Indicating a negative correlation between the two variables suggesting that as the amt of application increases the target variable tends to decrease

INCOME_TOTAL SEGMENTED BY DOWN_PAYMENT

17	Income_Total segmented by Down_payment	
18		
19		AMT_INCOME_TOTAL
20	0	11912
21	0.045	1
22	0.09	5
23	0.135	1
24	0.18	2

The Correlation between AMT_Income_Total & Down_payment : -0.707

Indicating a strong negative correlation between the two variables suggesting that as the down payment increases the total income tends to decrease

INCOME_TOTAL SEGMENTED BY TARGET

26	INCOME_TOTAL segmented by TARGET	
27		
28		AMT_INCOME_TOTAL
29	0	45973
30	1	4026

The Correlation between Income_Total & Target : - 0.9999999999999999

Indicating a negative correlation between the two variables suggesting that as the income_total decreases, the target value increases or vice versa

CREDIT_AMT SEGMENTED BY INCOME_TOTAL

32	CREDIT_AMT segmented by INCOME_TOTAL	
33		
34		AMT_CREDIT
35	25650	2
36	27000	9
37	28350	1
38	28575	1
39	28800	1

The Correlation between Credit_amt & Income_Total : -0.3992454827590575

Indicating a negative correlation between the two variables suggesting that as the income increases the credit amount tends to decrease

DOWN_PAYMENT SEGMENTED BY APPLICATION_AMT

41	DOWN_PAYMENT segmented by APPLICATION_AMT	
42		
43		AMT_DOWN_PAYMENT
44	0	5364
45	6120	0
46	6916.5	1
47	8281.08	0
48	9450	1

The Correlation between Down_payment & Application_Amt : -0.9376

Indicating a negative correlation between the two variables suggesting that as the application_amt increases the amt_down_payment tends to decrease and vice versa

CREDIT_AMT_1 SEGMENTED BY DOWN_PAYMENT

50	CREDIT_AMT_1 segmented by DOWN_PAYMENT	
51		
52		AMT_CREDIT 1
53	0	11912
54	0.045	1
55	0.09	5
56	0.135	1
57	0.18	2

The Correlation between Down_payment & Credit_amt_1 is -0.7071

Indicating a negative correlation between the two variables suggesting that as the Down_payment increases the credit_amt_1 tends to decrease and vice versa

CONTRACT_TYPE_1 SEGMENTED BY TARGET

50	CREDIT_AMT_1 segmented by DOWN_PAYMENT	
51		
52		AMT_CREDIT 1
53	0	11912
54	0.045	1
55	0.09	5
56	0.135	1
57	0.18	2

The Correlation between Contract_type_1 & Target is -1.0000

Indicating a negative correlation between the two variables suggesting that as the target increases, the contract_type_1 value decreases

The Descriptive Analysis is done to each one of columns taken

For Descriptive analysis -

TofindCount,Mean,Std,Minimum,Maximum,Median,Range,Mode,Percentage,Variance, quartile,percentile values These following functions are used :

1. COUNT FUNCTION:

=COUNT(A1:A50000)

2.MEAN FUNCTION:

=AVERAGE(A1: A50000)

3.STANDARD DEVIATION FUNCTION :

=STDEV.S(A1:A50000)

4.MINIMUM FUNCTION :

=MIN(A1:A50000)

5.MAXIMUM FUNCTION :

=MAX(A1:A50000)

6.RANGE FUNCTION :

=MAX(A1:A50000) – MIN(A1:A50000)

7.MODE FUNCTION :

=MODE(A1:A50000)

8.VARIANCE FUNCTION:

=VAR.S(A1: A50000)

9.QUARTILE FUNCTION:

Q1 = QUARTILE.INC(A1:A50000,1)

Q2 = QUARTILE.INC(A1:A50000,2)

Q3 = QUARTILE.INC(A1:A50000,3)

10.PERCENTILE FUNCTION:

1st Percentile:

=PERCENTILE.INC(A1:A50000, 1/100)

5th Percentile :

=PERCENTILE.INC(A1:A50000, 5/100)

10th Percentile:

=PERCENTILE.INC(A1:A50000, 10/100)

25th Percentile:

=PERCENTILE.INC(A1:A50000, 25/100)

50th Percentile :

=PERCENTILE.INC(A1:A50000, 50/100)

75th Percentile:

=PERCENTILE.INC(A1:A50000, 75/100)

90th Percentile :

=PERCENTILE.INC(A1:A50000, 90/100)

95th Percentile:

=PERCENTILE.INC(A1:A50000, 95/100)

99th Percentile:

=PERCENTILE.INC(A1:A50000, 99/100)

Percentage was also found for all columns

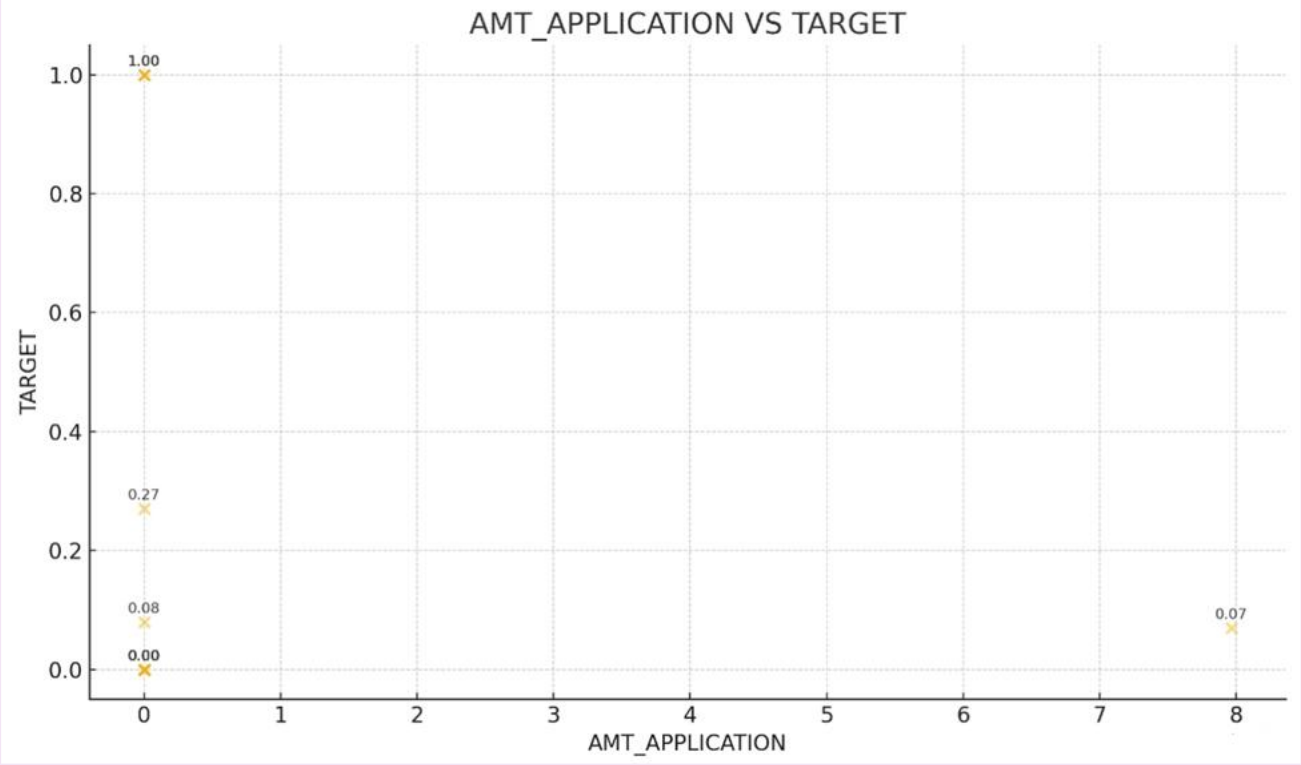
	A	B	C	D	E	F	G
1		AMT_CREDIT	AMT_APPLICATION	AMT_CREDIT 1	AMT_INCOME_TOTAL	TARGET	AMT_DOWN_PAYMENT
2	**count**	49,999	49,999	49,999	49,999	49,999	24,801
3	**mean**	188,542.89	168,892.45	188,542.89	170,767.59	0.08	6,557.57
4	**std**	308,470.52	282,200.69	308,473.60	531,819.10	0.27	17,444.58
5	**min**	0	0	0	25,650.00	0	0
6	**max**	4,104,351.00	3,826,372.50	4,104,351.00	117,000,000.00	1.00	1,035,000.00
7	**Median**	78,907.50	71,550	78,907.50	145,800.00	0.00	1,566.00
8	**Range**	4,104,351.00	3,826,372.50	4,104,351.00	116,974,350.00	1.00	1,035,000.00
9	**mode**	0.00	0.00	0.00	135,000.00	0.00	0.00
10	**percentage**	18.87%	78.12%	81.13%	100.00%	8.05%	25.78%
11	**variance**	95,154,059,586.69	79,637,228,515.08	95,155,962,744.02	282,831,549,942.21	0.07	304,313,400.40
12	**quantiles(0.25)**	26,055.00	22,045.50	26,055.00	112,500.00	0.00	0.00
13	**quantiles(0.5)**	78,907.50	71,550.00	78,907.50	145,800.00	0.00	1,566.00
14	**quantiles(0.75)**	198,126.00	180,000.00	198,105.75	202,500.00	0.00	7,875.00
15	**Percentile(1%)**	0.00	0.00	0.00	45,000.00	0.00	0.00
16	**Percentile(5%)**	0.00	0.00	0.00	67,500.00	0.00	0.00
17	**percentile(10%)**	0.00	0.00	0.00	81,000.00	0.00	0.00
18	***percentile(25%)***	26,055.00	22,045.50	26,055.00	112,500.00	0.00	0.00
19	**percentile(50%)**	78,907.50	71,550.00	78,907.50	145,800.00	0.00	1,566.00

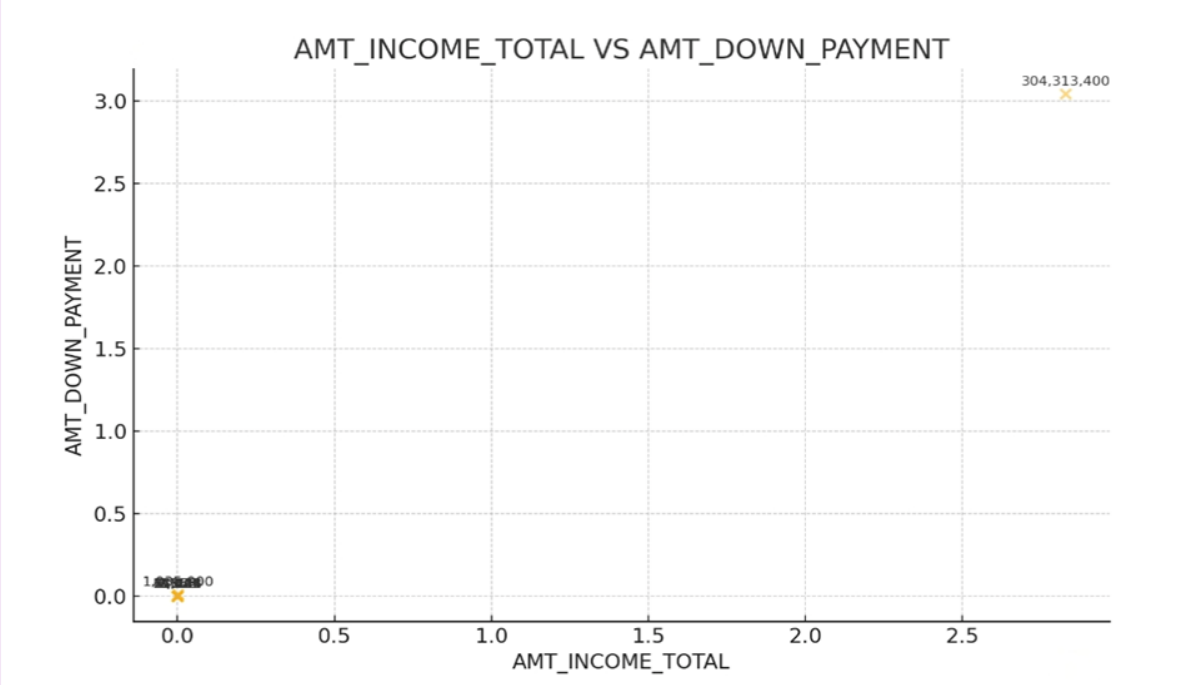
The descriptive analysis performed to the rest of the columns is given in the below link :

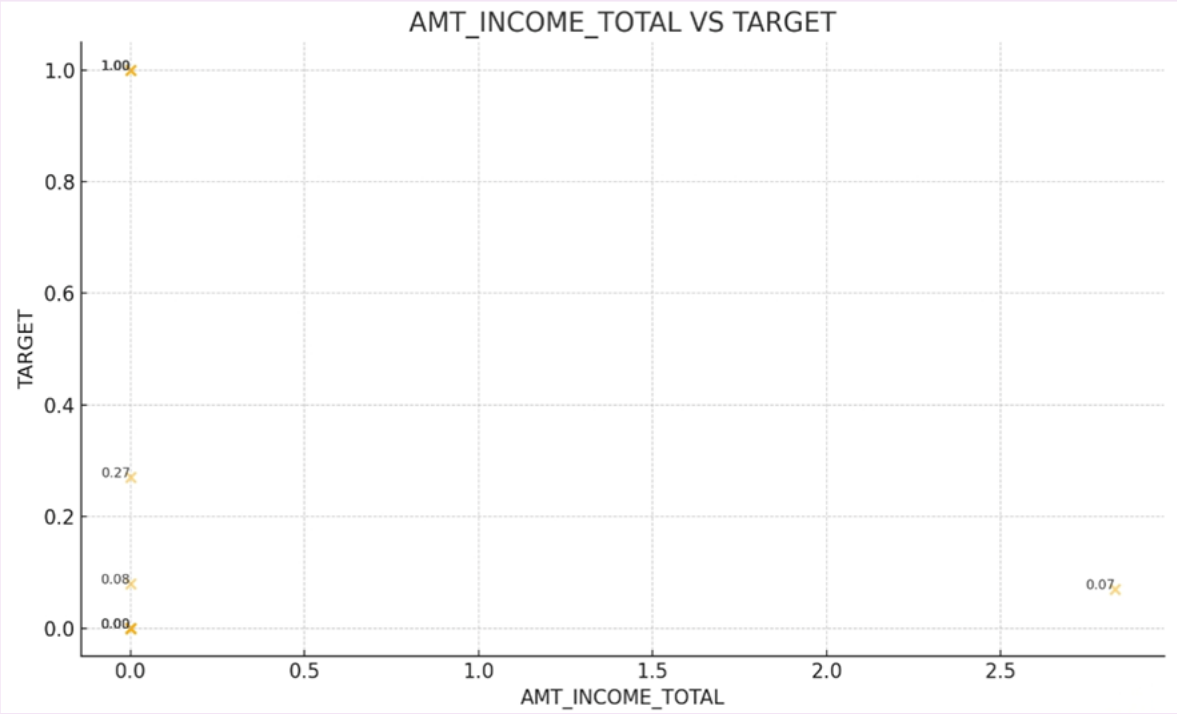
https://docs.google.com/spreadsheets/d/1QXr5Pp7rNYbdqPP3zGXIkr_lf9_Ddb0P/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

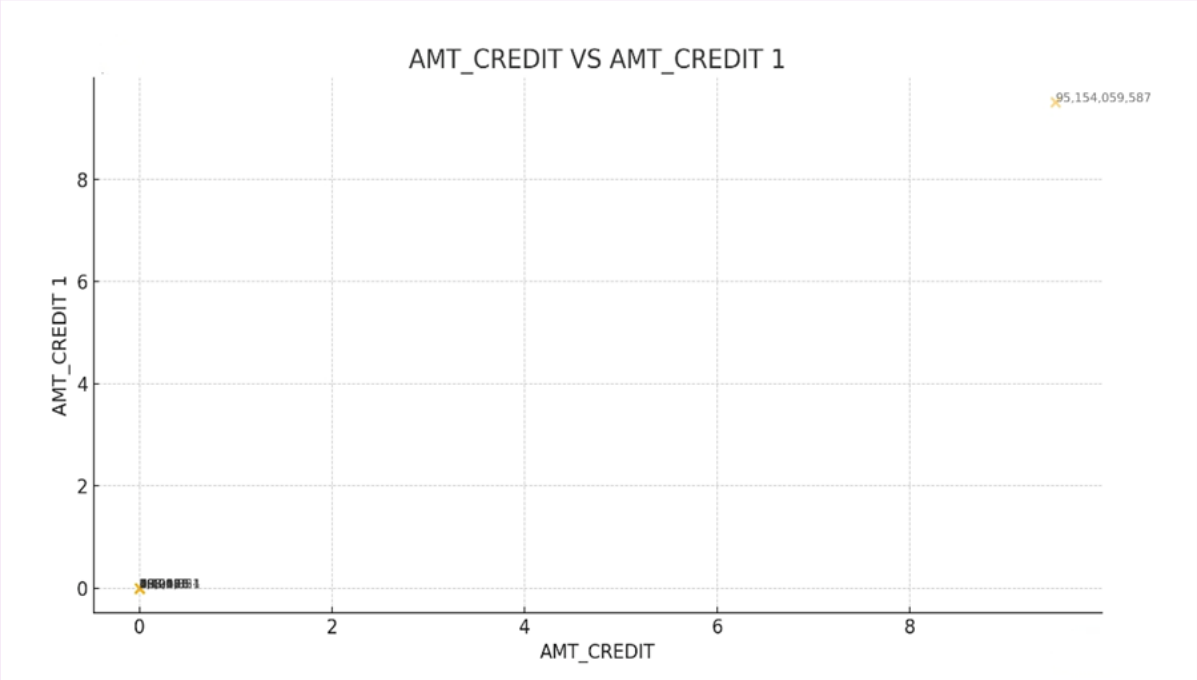
The scatter plots plotted to visualize the relationships between variables and target variable are given below

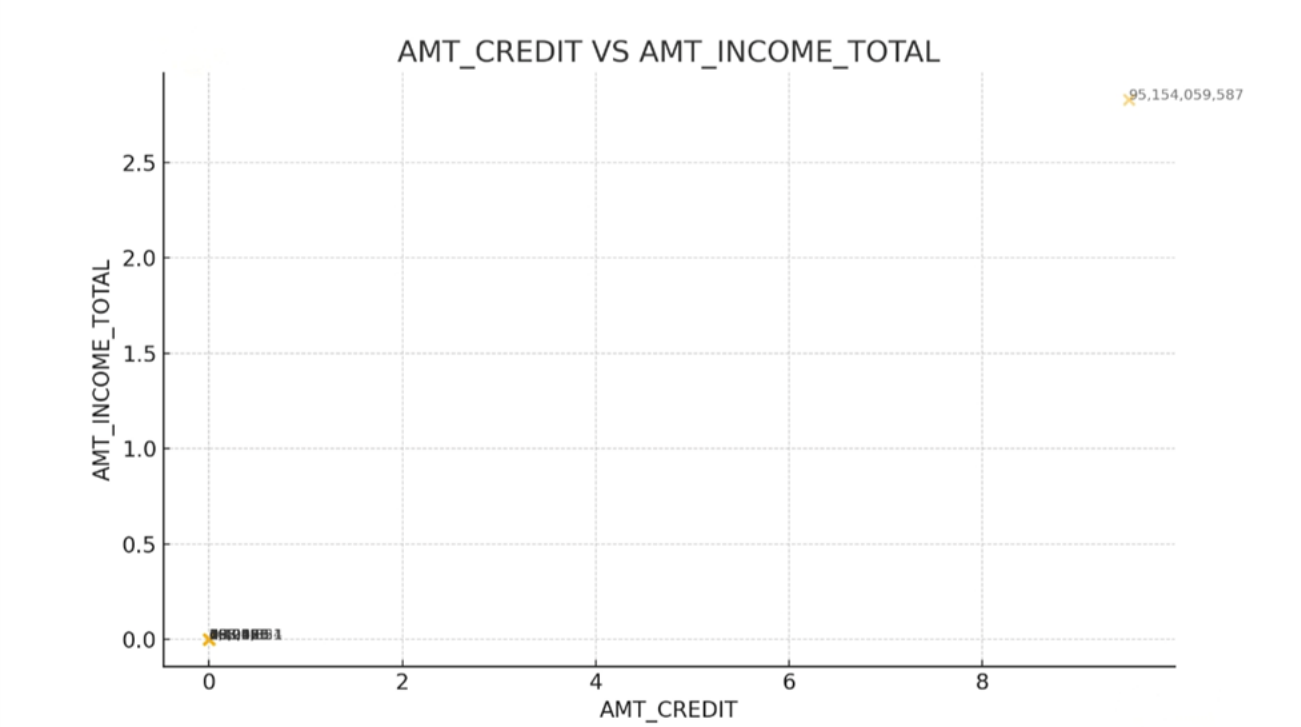
SCATTER PLOTS :

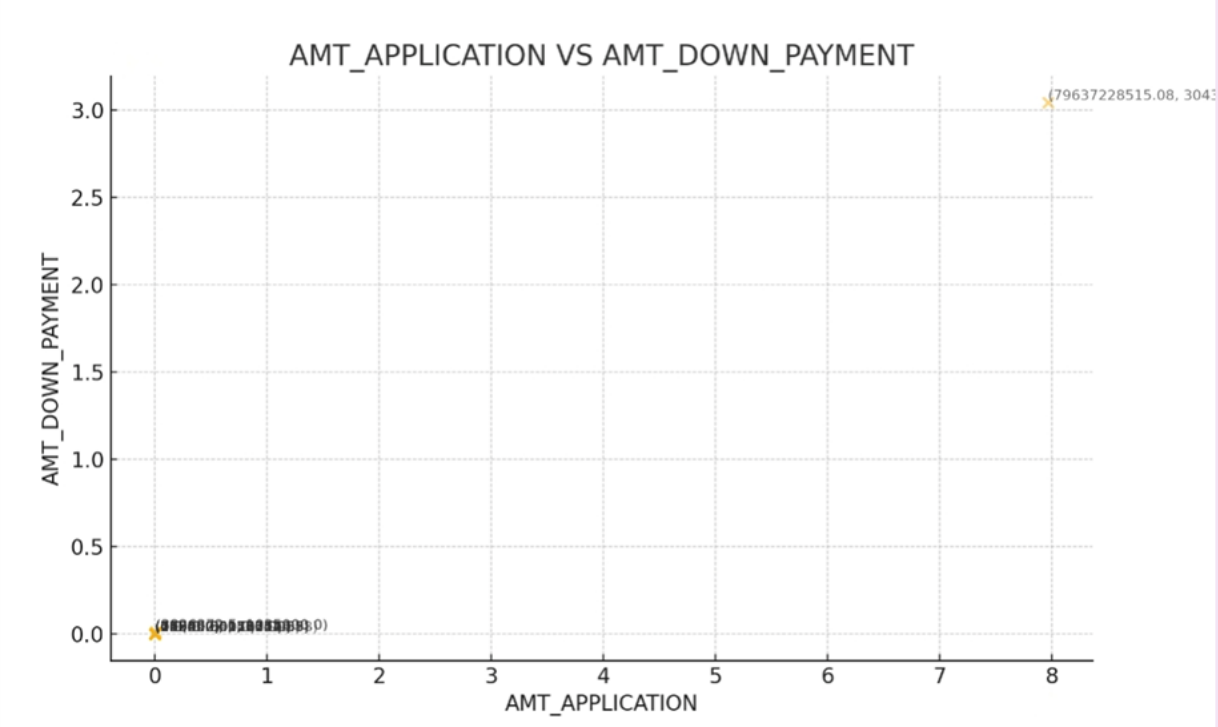


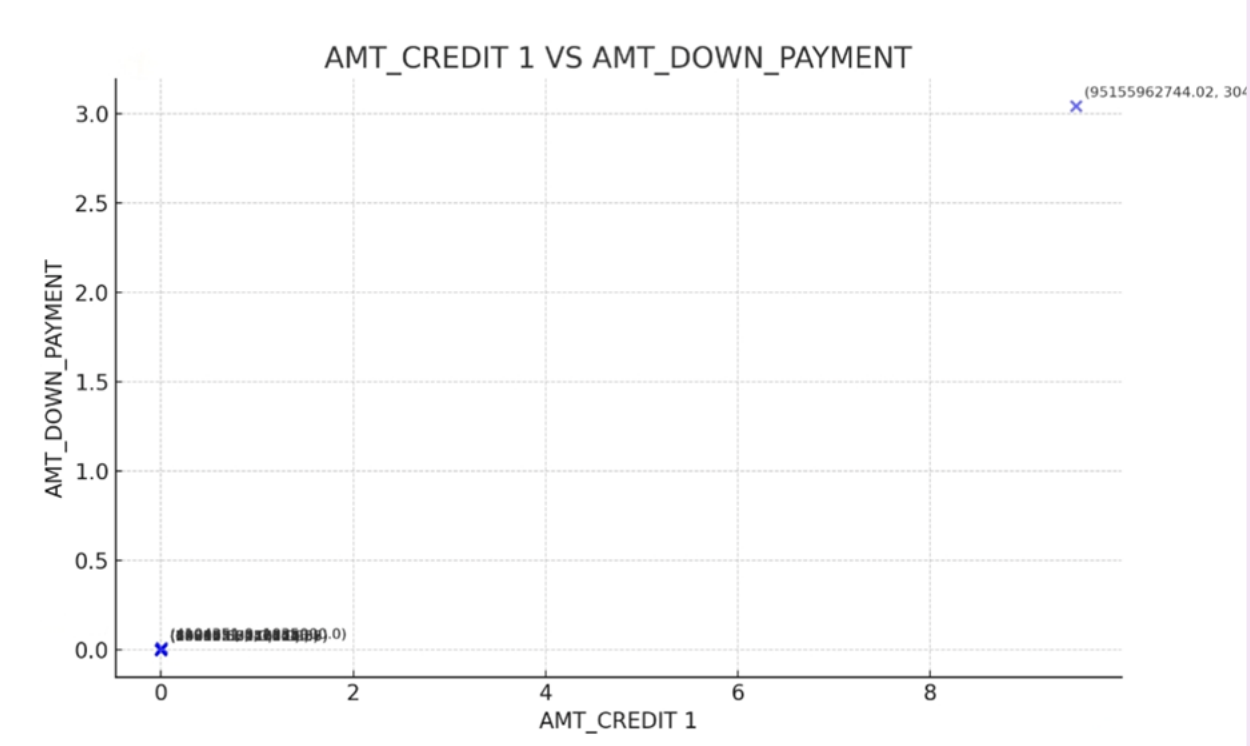


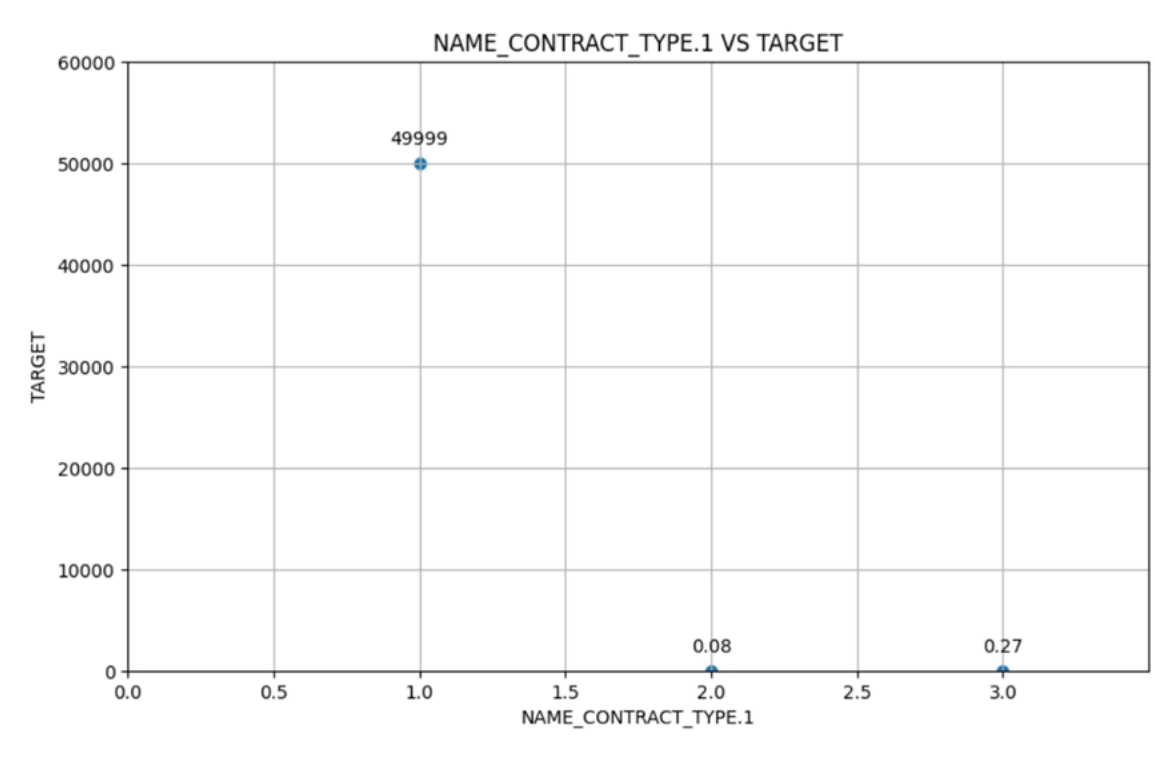

AMT_INCOME_TOTAL VS TARGET
TARGET
AMT_INCOME_TOTAL
1.00
0.27
0.08
0.00
0.07
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5


AMT_CREDIT VS AMT_CREDIT 1
AMT_CREDIT 1
AMT_CREDIT
95,154,059,587
0
2
4
6
8
0
2
4
6
8


AMT_CREDIT VS AMT_INCOME_TOTAL
95,154,059,587
AMT_INCOME_TOTAL
AMT_CREDIT
0.0
0.5
1.0
1.5
2.0
2.5
0
2
4
6
8


AMT_APPLICATION VS AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
AMT_APPLICATION
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
7
8


AMT_CREDIT 1 VS AMT_DOWN_PAYMENT
AMT_DOWN_PAYMENT
AMT_CREDIT 1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8


NAME_CONTRACT_TYPE.1 VS TARGET
TARGET
NAME_CONTRACT_TYPE.1
60000
50000
40000
30000
20000
10000
0
49999
0.08
0.27
0.0
0.5
1.0
1.5
2.0
2.5
3.0

E. Identify Top Correlations for Different Scenarios:

5. Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions. Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

OUTPUT :

The segments found are :

1) CNT_CHILDREN

2) AMT_INCOME_TOTAL

3) AMT_CREDIT

4) REGION_POPULATION_RELATIVE

5) DAYS_BIRTH

6) DAYS_EMPLOYED

7) DAYS_ID_PUBLISH

8) REGION_RATING_CLIENT

9)AMT_GOODS_PRICE

10)AMT_REQ_CREDIT_BUREAU_YEAR

Including Target variable taken

	A	B	C	D	E	F	G	H
1	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH
2	1	0	202500	406597.5	0.018801	-9461	-637	-2120
3	1	0	270000	1293502.5	0.003541	-16765	-1188	-291
4	1	0	67500	135000	0.010032	-19046	-225	-2531
5	1	0	135000	312682.5	0.008019	-19005	-3039	-2437
6	1	0	121500	513000	0.028663	-19932	-3038	-3458
7	1	0	99000	490495.5	0.035792	-16941	-1588	-477
8	1	1	171000	1560726	0.035792	-13778	-3130	-619
9	1	0	360000	1530000	0.003122	-18850	-449	-2379
10	1	0	112500	1019610	0.018634	-20099	365243	-3514
11	1	0	135000	405000	0.019689	-14469	-2019	-3992
12	1	1	112500	652500	0.0228	-10197	-679	-738
13	1	0	38419.155	148365	0.015221	-20417	365243	-2512
14	1	0	67500	80865	0.031329	-13439	-2717	-3227
15	1	1	225000	918468	0.016612	-14086	-3028	-4911
16	1	0	189000	773680.5	0.010006	-14583	-203	-2056
17	1	0	157500	299772	0.020713	-8728	-1157	-1368
18	1	0	108000	509602.5	0.018634	-12931	-1317	-3866

The columns taken to find correlation coefficient including target variable (1) are given in the link below

https://docs.google.com/spreadsheets/d/15-QGp86IugTP0eNVSf_9fqLi1r9YWuQC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=correl(A2:A50000,B2:B50000)

The Correal Function is used between target variable (1) and the other variables to find the coefficients

TARGET (1)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009588558	0.00497156	0.002085777	-0.025555665	0.329263754	-0.239693041	-0.032115773	0.025913889	0.005882852
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	-0.000190769	0.029841469	0.016002774	-0.031615555	0.003506846	-0.038188511	-0.005089418
AMT_CREDIT	0.00497156	0.069315897	1	0.013258331	0.095111221	-0.059342658	-0.070471393	-0.012228765	-0.100507425	0.003898337
AMT_GOODS_PRICE	0.002085777	-0.000190769	0.013258331	1	0.00576511	-0.003153225	-0.00424617	-0.00093887	-0.00141928	-0.004311374
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	0.00576511	1	-0.032513748	-0.004101686	-0.004345136	-0.532667302	0.002375614
DAYS_BIRTH	0.329263754	0.016002774	-0.059342658	-0.003153225	-0.032513748	1	-0.613553972	0.270825141	0.016779196	0.00771461
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393	-0.00424617	-0.004101686	-0.613553972	1	-0.270382022	0.034321673	-0.006483373
DAYS_ID_PUBLISH	-0.032115773	0.003506846	-0.012228765	-0.00093887	-0.004345136	0.270825141	-0.270382022	1	-0.002307011	0.012634162
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.00141928	-0.532667302	0.016779196	0.034321673	-0.002307011	1	0.004039346
AMT_REQ_CREDIT_BUREAU_YEAR	0.005882852	-0.005089418	0.003898337	-0.004311374	0.002375614	0.00771461	-0.006483373	0.012634162	0.004039346	1

The Full data of coefficients found are in the link below :

https://docs.google.com/spreadsheets/d/1mkxx0A5Vz-E0xnF9Yu6W-utq0qtGXUGj/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The conditional formatting rules are used to highlight Correlations

Conditional Formatting Rules

Maximum = 1

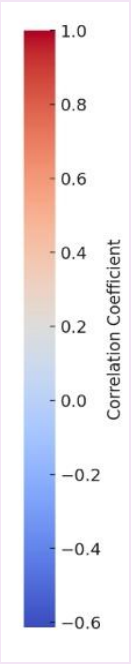
Minimum = - 1

Midpoint = 0

THE CORRELATION MATRICES FOR TARGET VARIABLE (1) done using conditional formatting

TARGET (1)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009588558	0.00497156	0.002085777	-0.025555665	0.329263754	-0.239693041	-0.032115773	0.025913889	0.005882852
AMT_INCOME_TOTAL	0.009588558	1	0.069315897	-0.000190769	0.029841469	0.016002774	-0.031615555	0.003506646	-0.038188511	-0.005089418
AMT_CREDIT	0.00497156	0.069315897	1	0.013258331	0.095111221	-0.059342658	-0.070471393	-0.012228765	-0.100507425	0.003898937
AMT_GOODS_PRICE	0.002085777	-0.000190769	0.013258331	1	0.00576511	-0.003153225	-0.00424617	-0.00093887	-0.00141928	-0.004311374
REGION_POPULATION_RELATIVE	-0.025555665	0.029841469	0.095111221	0.00576511	1	-0.032513748	-0.004101686	-0.004345136	-0.532667302	0.002975614
DAYS_BIRTH	0.329263754	0.016002774	-0.059342658	-0.003153225	-0.032513748	1	-0.613553972	0.270825141	0.016779196	0.00771461
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393	-0.00424617	-0.004101686	-0.613553972	1	-0.270382022	0.034321673	-0.006483373
DAYS_ID_PUBLISH	-0.032115773	0.003506646	-0.012228765	-0.00093887	-0.004345136	0.270825141	-0.270382022	1	-0.002307011	0.012634162
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.00141928	-0.532667302	0.016779196	0.034321673	-0.002307011	1	0.004039346
AMT_REQ_CREDIT_BUREAU_YEAR	0.005882852	-0.005089418	0.003898937	-0.004311374	0.002975614	0.00771461	-0.006483373	0.012634162	0.004039346	1

LEGEND :



Color	Explanation
RED	Negative correlation
BLUE	positive correlation
WHITE	No correlation

Explanation of Color Coding:

1.Blue:

- Explanation: positive correlation.
- Example: As one variable increases, the other also increases strongly.

2. White:

- Explanation: No correlation.
- Example: No linear relationship between the variables.

3. Red:

- Explanation: negative correlation.
- Example: As one variable increases, the other shows a minimal decrease.

The Clear Correlation Matrices for Target variable (1) is in the given link below :

https://docs.google.com/spreadsheets/d/1vayEb52rByP9-xnlNrS1Qzc8fsZ-YD6C/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The segments found are :

1) CNT_CHILDREN

2) AMT_INCOME_TOTAL

3) AMT_CREDIT

4) REGION_POPULATION_RELATIVE

5) DAYS_BIRTH

6) DAYS_EMPLOYED

7) DAYS_ID_PUBLISH

8) REGION_RATING_CLIENT

9)AMT_GOODS_PRICE

10)AMT_REQ_CREDIT_BUREAU_YEAR

Including Target variable taken

	A	B	C	D	E	F	G	H
1	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH
2	0	0	202500	406597.5	0.018801	-9461	-637	-2120
3	0	0	270000	1293502.5	0.003541	-16765	-1188	-291
4	0	0	67500	135000	0.010032	-19046	-225	-2531
5	0	0	135000	312682.5	0.008019	-19005	-3039	-2437
6	0	0	121500	513000	0.028663	-19932	-3038	-3458
7	0	0	99000	490495.5	0.035792	-16941	-1588	-477
8	0	1	171000	1560726	0.035792	-13778	-3130	-619
9	0	0	360000	1530000	0.003122	-18850	-449	-2379
10	0	0	112500	1019610	0.018634	-20099	365243	-3514
11	0	0	135000	405000	0.019689	-14469	-2019	-3992
12	0	1	112500	652500	0.0228	-10197	-679	-738
13	0	0	38419.155	148365	0.015221	-20417	365243	-2512
14	0	0	67500	80865	0.031329	-13439	-2717	-3227
15	0	1	225000	918468	0.016612	-14086	-3028	-4911
16	0	0	189000	773680.5	0.010006	-14583	-203	-2056
17	0	0	157500	299772	0.020713	-8728	-1157	-1368

The columns taken to find correlation coefficient including target variable (0) are given in the link below

https://docs.google.com/spreadsheets/d/1pilMPVgwnAurs6V9GU8HMX67tpJCjaQy/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

=correl(A2:A50000,B2:B50000)

The Correal Function is used between target variable (0) and the other variables to find the coefficients

	TARGET VARIABLE (0)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	ION_POPULATION_RELA	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	GION_RATING_CLIE	AMT_GOODS_PRICE	AMT_REQ_CREDIT_BUREAU_YEAR
62	CNT_CHILDREN	1	0.009389816	0.006515123	-0.026689704	0.328293814	-0.239627071	-0.032268389	0.025908972	0.004130779	-0.001437756
63	AMT_INCOME_TOTAL	0.009389816	1	0.066433368	0.028425341	0.01586795	-0.03052779	0.003309333	-0.03657852	-0.000172738	0.000185516
64	AMT_CREDIT	0.006515123	0.066433368	1	0.095127525	-0.058641286	-0.070718891	-0.011346152	-0.099112917	0.011234511	-0.001744445
65	REGION_POPULATION_RELATIVE	-0.026689704	0.028425341	0.095127525	1	-0.030369054	-0.005920011	-0.002298217	-0.534389649	0.003609725	0.003684573
66	DAYS_BIRTH	0.328293814	0.01586795	-0.058641286	-0.030369054	1	-0.61451668	0.271792027	0.017541173	-0.002299615	-0.000614776
67	DAYS_EMPLOYED	-0.239627071	-0.03052779	-0.070718891	-0.005920011	-0.61451668	1	-0.270558903	0.034195347	-0.006196234	-0.00744605
68	DAYS_ID_PUBLISH	-0.032268389	0.003309333	-0.011346152	-0.002298217	0.271792027	-0.270558903	1	-0.003836517	0.001301719	0.004114932
69	REGION_RATING_CLIENT	0.025908972	-0.03657852	-0.099112917	-0.534389649	0.017541173	0.034195347	-0.003836517	1	0.003321731	-0.002785354
70	AMT_GOODS_PRICE	0.004130779	-0.000172738	0.011234511	0.003609725	-0.002299615	-0.006196234	0.001301719	0.003321731	1	0.00371245
71	AMT_REQ_CREDIT_BUREAU_YEAR	-0.001437756	0.000185516	-0.001744445	0.003684573	-0.000614776	-0.00744605	0.004114932	-0.002785354	0.00371245	1

The Full data of coefficients found are in the link below :

https://docs.google.com/spreadsheets/d/1YmOyXEuEbEVg40XHfrVaRl2wqXoAc6V9/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The conditional formatting rules are used to highlight Correlations

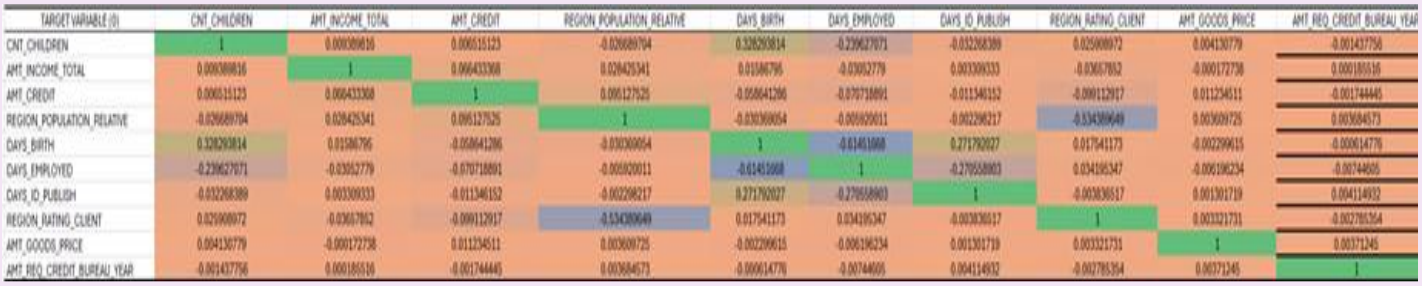
Conditional Formatting Rules

Maximum = 1

Minimum = - 1

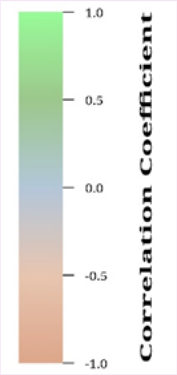
Midpoint = 0

THE CORRELATION MATRICES FOR TARGET VARIABLE (0) done using conditional formatting



TARGET VARIABLE (0)	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_GOODS_PRICE	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.009389816	0.006515123	-0.026689704	0.328293814	-0.239627071	-0.032268389	0.025908972	0.004130779	-0.001437756
AMT_INCOME_TOTAL	0.009389816	1	0.066433368	0.028425341	0.01586795	-0.03052779	0.003309333	-0.03657852	-0.000172738	0.000185516
AMT_CREDIT	0.006515123	0.066433368	1	0.095127525	-0.058641286	-0.070718891	-0.011346152	-0.099112917	0.011234511	-0.001744445
REGION_POPULATION_RELATIVE	-0.026689704	0.028425341	0.095127525	1	-0.030369054	-0.005920011	-0.002298217	-0.534389649	0.003609725	0.003684573
DAYS_BIRTH	0.328293814	0.01586795	-0.058641286	-0.030369054	1	-0.61451668	0.271792027	0.017541173	-0.002299615	-0.000614776
DAYS_EMPLOYED	-0.239627071	-0.03052779	-0.070718891	-0.005920011	-0.61451668	1	-0.270558903	0.034195347	-0.006196234	-0.00744605
DAYS_ID_PUBLISH	-0.032268389	0.003309333	-0.011346152	-0.002298217	0.271792027	-0.270558903	1	-0.003836517	0.001301719	0.004114932
REGION_RATING_CLIENT	0.025908972	-0.03657852	-0.099112917	-0.534389649	0.017541173	0.034195347	-0.003836517	1	0.003321731	-0.002785354
AMT_GOODS_PRICE	0.004130779	-0.000172738	0.011234511	0.003609725	-0.002299615	-0.006196234	0.001301719	0.003321731	1	0.00371245
AMT_REQ_CREDIT_BUREAU_YEAR	-0.001437756	0.000185516	-0.001744445	0.003684573	-0.000614776	-0.00744605	0.004114932	-0.002785354	0.00371245	1

LEGEND :



Color	Explanation
Brown	No correlation
Green	Positive correlation
Blue	negative correlation

Explanation of Color Coding:

1.Green :

- Explanation: positive correlation.
- Example: As one variable increases, the other also increases strongly.

2.Brown :

- Explanation: No correlation.
- Example: No linear relationship between the variables.

3. Blue :

- Explanation: negative correlation.
- Example: As one variable increases, the other shows a minimal Decrease

The Clear Correlation Matrices for Target variable (0) is in the given link below :

https://docs.google.com/spreadsheets/d/1fT7Ao_s8Oj6tvAdj3JQhGi4pxx_Ebz7q/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

ANALYSIS

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

Why is it that the target_variable is of so much importance?

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

Why is it that proportion of Female clients more than that of the Male clients?

---> In countries like India especially there have been laws made by the Government for Women who want to establish their own Start-up, Business or their own classes, catering services, etc.;

These laws offer loans to women clients at a relatively low interest rates; Also in some cases people purposely use their retired/household mother or household wife so that they can get some sort of concession i.e. low interest rates while applying for Home loans

Why should bank prefer other Housing type clients though House/Apartments Housing type clients have the highest proportion of non-defaulters?

----> Cause people in other groups like Municipal Apartment,

Co-op Apartment, Rented Apartment, with Parents are in the search of their own house of their own name plate; Also now a days in India the joint family system is declining and the future generations opt to live in their own 1/2 BHK's rather than living together will all family members in big Family Apartments

Why should bank opt for working class clients more than the state-government class clients though state-government employees enjoy a lot of benefits and regular salary?

----> It is true that state government employee enjoy a lot of benefits but they also get housing allowances greater than that of working class and in some cases they even get an apartment to live with their families as long as they work for the state government;

On the other hand the working class don't enjoy such housing allowances or get very less of it, also the working class don't get an apartment to live in for their entire professional life(i.e. until retirement) and so working class opt for purchasing their own house by taking house loan

Why should Bank not go for approving loans to 'Laborers' occupation_type clients though they have the highest non- defaulters count?

-----> Laborers take only personal loans for marriage or house repair purpose and their loan amount is also less and the interest on such loans is also less as compared to home loan, car loan, etc. which in turn will cause less profits to the bank

Why is it that females with low income group have the lowest count of defaulters?

----> Females belonging to such groups take loan of small amounts just for starting their own start-ups, business or catering/ parlor services and they usually enjoy benefit from government schemes for such purpose

CONCLUSION

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

- The proportion/percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92%

- The Bank generally lends more loan to Female clients as compared to Males clients as the count of Female clients in the defaulter's list is less than that of Males. Still Bank can look for more Male clients if their credit amount is satisfied

- Also the clients who belong to Working class tend to pay their loans on time followed by the clients who fall under Commercial Associate

- Clients having Education status like Secondary/ Higher Secondary or more tend to pay loan on time so bank can prefer lending loans to clients having such Education Status

- Clients who fall in the Age Group 31-40 have the highest count for paying off their loans on time followed by the clients who fall in the Age Groups 41-60

- Clients having LOW credit amount range tend to pay off their loans on time than compared to HIGH and MEDIUM credit range

- Clients living with their Parents tend to pay off their loans quickly as compared to other housing type. So Bank can lend loan to clients having housing type → Living with Parents

- Clients taking loan for purchasing New Home i.e. clients taking Home Loans or purchasing New Car i.e. Car Loans and clients who have a income type as State Servant tend to pay their loans on time and hence Bank should prefer clients having Such background

- The Bank should be more cautious when lending money to clients with Repairs purpose because they have high count of Defaulters along with High count of Defaulters

ANALYZING THE IMPACT OF CAR FEATURES ON PRICE & PROFITABILITY

DESCRIPTION

The dataset includes variables such as car's make, model, year, fuel type, engine power, transmission, wheels, number of doors, market category, size, style, estimated miles per gallon, popularity, and manufacturer's suggested retail price (MSRP).

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. It is important to know the impact of car features on price and profitability in the automotive industry. The purpose is to analyze 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.

By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

THE PROBLEM

The automotive industry is highly competitive, with manufacturers constantly striving to offer innovative features that enhance the appeal and performance of their vehicles. However, determining the optimal combination of car features that maximizes both price and profitability poses a significant challenge. Manufacturers must balance the cost of incorporating advanced features with the potential increase in vehicle price and profitability.

This project aims to analyze the impact of various car features on vehicle price and profitability. By leveraging data analytics, the objective is to identify key features that significantly influence the market value and profitability of cars. The analysis will provide valuable insights for automotive manufacturers to make data-driven decisions in feature selection, pricing strategies, and market positioning.

DESIGN

Before starting the actual analysis, I have: -

• First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.

• Then Il removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.

• I removed rows having blank spaces and NULL values.

• Then removed duplicate rows from the datasets.

Software used for doing the overall Analysis: -

----> Microsoft Excel

FINDIINGS

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

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

OUTPUT :

To find how the popularity of a car model vary across different market categories we first have to use pivot table to find the popularity of a car model across different market categories

The columns used to create a pivot table are Model.Market category & Popularity columns

	A	B	C
1	Model	Market Category	Popularity
2	1 Series M	Factory Tuner,Luxury,High-Performance	3916
3	1 Series	Luxury,Performance	3916
4	1 Series	Luxury,High-Performance	3916
5	1 Series	Luxury,Performance	3916
6	1 Series	Luxury	3916
7	1 Series	Luxury,Performance	3916
8	1 Series	Luxury,Performance	3916
9	1 Series	Luxury,High-Performance	3916
10	1 Series	Luxury	3916
11	1 Series	Luxury	3916

The columns used to make pivot tables are given in the Sheet 1 in link below :

Kindly open sheet 1 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

3	Row Labels	Count of Model	Average of Popularity
4	Crossover	1103	1529.0
5	Crossover,Diesel	7	873.0
6	Crossover,Exotic,Luxury,High-Performance	1	238.0
7	Crossover,Exotic,Luxury,Performance	1	238.0
8	Crossover,Factory Tuner,Luxury,High-Performance	26	1823.5
9	Crossover,Factory Tuner,Luxury,Performance	5	2607.4
10	Crossover,Factory Tuner,Performance	4	210.0
11	Crossover,Flex Fuel	64	2073.8
12	Crossover,Flex Fuel,Luxury	10	1173.2
13	Crossover,Flex Fuel,Luxury,Performance	6	1624.0
14	Crossover,Flex Fuel,Performance	6	5657.0
15	Crossover,Hatchback	72	1675.7
16	Crossover,Hatchback,Factory Tuner,Performance	6	2009.0

The pivot table created are given in the Sheet 3 in link below :

Kindly open sheet 3 given link below in Microsoft Excel to view charts & Data

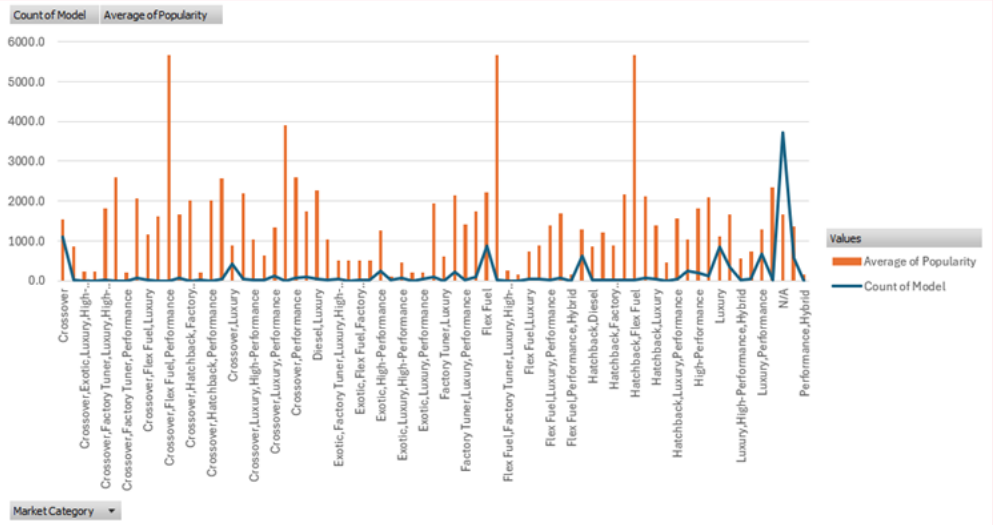
https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

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

OUTPUT :

The Combo chart is created to visualize the relationship between market category & popularity using the pivot table columns

The combo chart created :



The combo chart which shows the relationship between market category & popularity created is given in the Sheet 3 in link below :

Kindly open sheet 3 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The gretest popularity was for “crossover.Flex,fuel,performance”,”Flex Fuel,Diesel”,”Hatchback,Flex Fuel” markets

This concludes that the varied degrees of popularity of different automobile modes in different market segments offers insights into customer preferences in this sector.

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

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

OUTPUT :

To find the relationship between a car's engine power and its price we first have to use the scatter plot to find the relationship between a car's engine power & its price

The columns used to plot scatter plot are engine HP & MSRP columns

	A	B
1	Engine HP	MSRP
2	335	46135
3	300	40650
4	300	36350
5	230	29450
6	230	34500
7	230	31200
8	300	44100
9	300	39300
10	230	36900
11	230	37200
12	300	39600

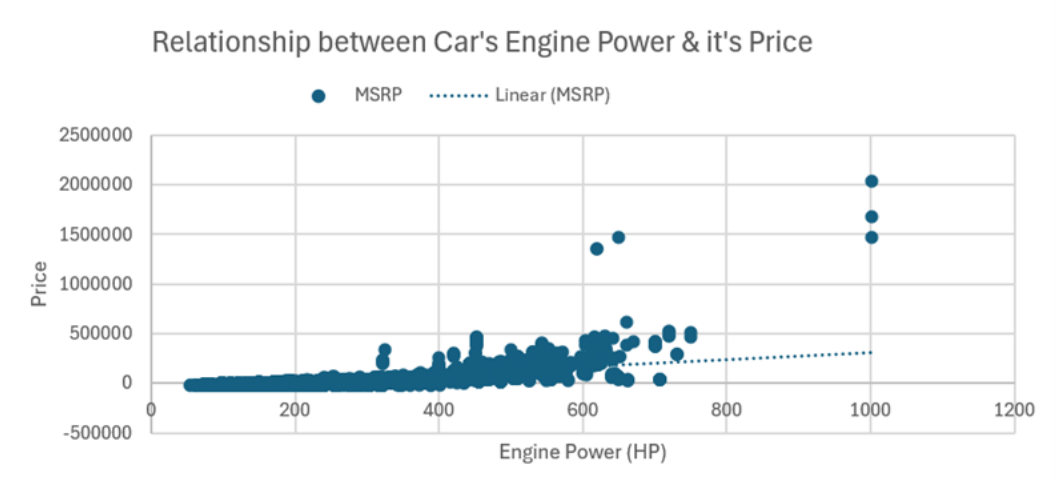
The columns taken to plot a scatter plot are in the sheet 4 in the given link below :

Kindly open sheet 4 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The scatter plot is created to visualize the relationship between engine power & price using the columns Engine HP & MSRP

The scatter plot created :



The Scatter plot chart which shows the relationship between Engine power & price created is given in the Sheet 4 in the link below :

Kindly open sheet 4 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The scatter plot shows a general upward trend. This indicates that as the

engine power of a car increases, its price also tends to increase.

Correlation: The trendline further supports this positive correlation. The upward slope of the trendline suggests that there's a positive association between engine power and price.

The scatter plot reveals a positive correlation between engine power and price, meaning that generally, cars with higher engine power tend to be more expensive. However, the spread of data points indicates that other factors also contribute to the price variation.

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

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

OUTPUT :

To find which car feature is most important in determining a car's price we first use Regression analysis to identify the variables that have the strongest relationship with a car's price

The columns used to do regression analysis are Engine HP,Engine Cylinders,Number of Doors,Highway MPG,City Mpg,MSRP columns

	A	B	C	D	E	F
1	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg	MSRP
2	335	6	2	26	19	46135
3	300	6	2	28	19	40650
4	300	6	2	28	20	36350
5	230	6	2	28	18	29450
6	230	6	2	28	18	34500
7	230	6	2	28	18	31200
8	300	6	2	26	17	44100

The columns taken to do Regression analysis are given in Sheet 5 in the link below :

Kindly open sheet 5 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Regression Analysis which is performed on Engine HP,Engine Cylinders,Number of Doors,Highway MPG,City Mpg,MSRP columns :

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.680708139				
5	R Square	0.46336357				
6	Adjusted R Square	0.463136297				
7	Standard Error	44170.77827				
8	Observations	11812				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	5	1.98891E+13	3.97782E+12	2038.799457	0
13	Residual	11806	2.30342E+13	1951057653		
14	Total	11811	4.29233E+13			
15						

16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	-101601.736	3684.351697	-27.57655738	2.765E-162	-108823.673	-94379.799	-108823.673	-94379.79896
18	Engine HP	322.7465574	6.01767382	53.63310924	0	310.9509241	334.5421906	310.9509241	334.5421906
19	Engine Cylinders	6989.177662	439.6449924	15.89732121	2.53591E-56	6127.400961	7850.954363	6127.400961	7850.954363
20	Number of Doors	-4472.158125	465.7180593	-9.602715711	9.35015E-22	-5385.042338	-3559.27391	-5385.042338	-3559.273912
21	highway MPG	570.1808088	105.7839778	5.390048859	7.17937E-08	362.826764	777.5348535	362.826764	777.5348535
22	city mpg	1163.755457	121.9978136	9.539150113	1.72109E-21	924.61962	1402.891294	924.61962	1402.891294

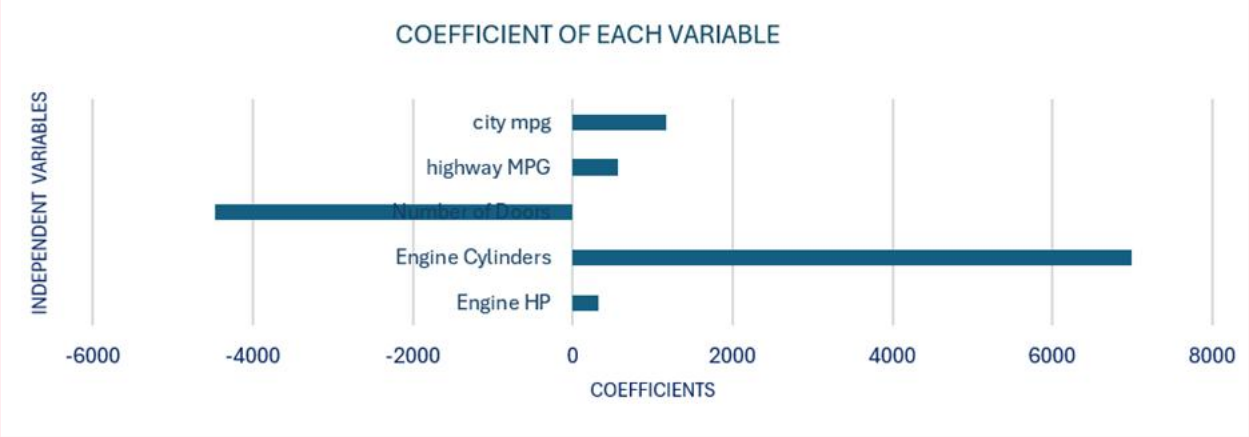
The Regression Analysis is done in the sheet 6 in the given link below :

Kindly open sheet 6 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

A bar chart is created to visualize the variables relative importance & the bar chart is created using columns Engine HP, Engine cylinders ,no of doors, highway MPG,city MPG& MSRP

The bar chart created :



The Bar chart which shows the variables relative importance created is given in the Sheet6 in the link below :

Kindly open sheet 6 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

- Engine Cylinders: With a high coefficient of 6989.177662 , the number of engine cylinders has the largest correlation with a car's price. This means that as the number of engine cylinders increases, the car's price tends to be higher.

- Number of Doors: This factor has a negative coefficient of -4472.158125, indicating that it has the least correlation with car pricing. More doors are associated with a decrease in the car's price.

the price of a car is mostly influenced by the number of engine cylinders, whereas the number of doors has a minimal impact.

Insight Required: 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.

OUTPUT :

To find how the average price of a car vary across different manufacturers we first have to use pivot table to visualize the average price of the cars for each manufacturer

The columns taken to create the pivot table are Make ,MSRP columns

	A	B
1	Make	MSRP
2	BMW	46135
3	BMW	40650
4	BMW	36350
5	BMW	29450
6	BMW	34500
7	BMW	31200
8	BMW	44100
9	BMW	39300
10	BMW	36900
11	BMW	37200
12	BMW	39600
13	BMW	31500
14	BMW	44400
15	BMW	37200

The columns taken to create a pivot table is given in Sheet 7 in the link below :

Kindly open sheet 7 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

3	Make	Average of MSRP
4	Acura	34887.5873
5	Alfa Romeo	61600
6	Aston Martin	197910.3763
7	Audi	53452.1128
8	Bentley	247169.3243
9	BMW	61546.76347
10	Bugatti	1757223.667
11	Buick	28206.61224
12	Cadillac	56231.31738
13	Chevrolet	28273.35695
14	Chrysler	26722.96257
15	Dodge	22390.05911
16	Ferrari	237383.8235
17	FIAT	22206.01695
18	Ford	27393.42051
19	Genesis	46616.66667
20	GMC	30493.29903
21	Honda	26629.81879
22	HUMMER	36464.41176
23	Hyundai	24597.0363
24	Infiniti	42394.21212

The pivot table created is in **Sheet 8** in the given link below :

Kindly open sheet 8 given link below in Microsoft Excel to view charts

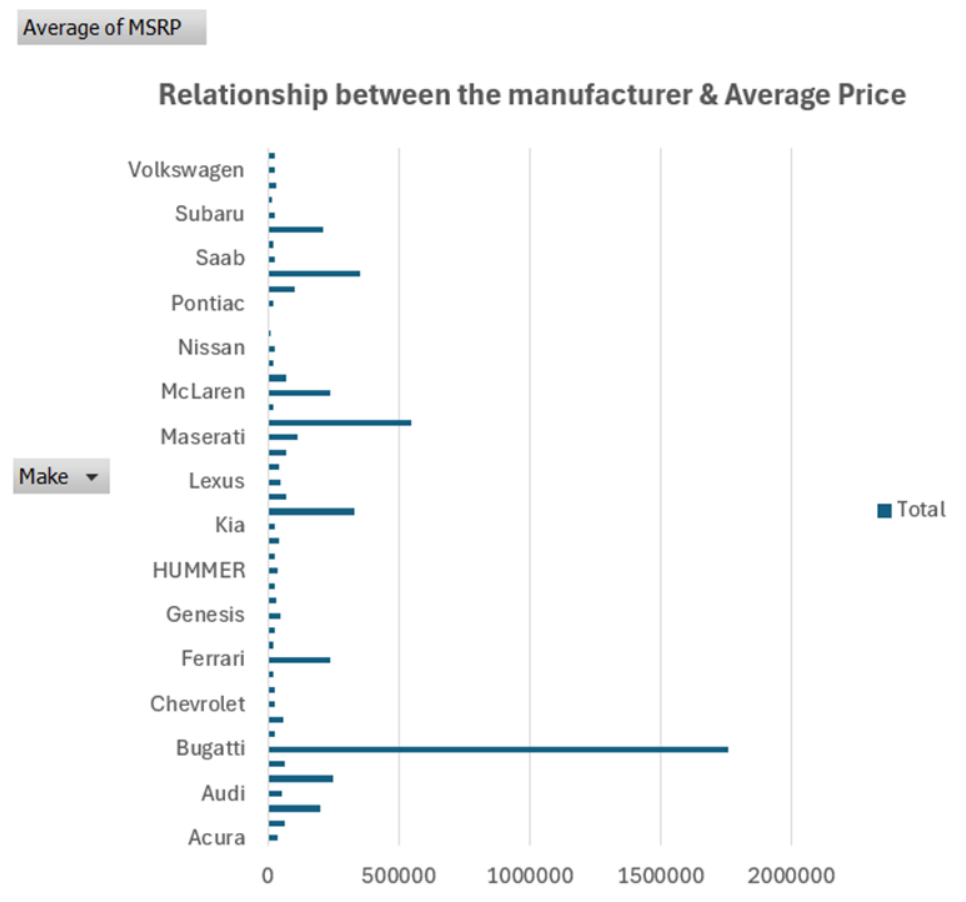
https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

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

OUTPUT :

A bar chart is created to visualize the relationship between manufacturer and average price using the pivot table columns .

The bar chart created :



A bar chart which is created to visualize the relationship between manufacturer and average price Is in sheet 8 in the given link below :

Kindly open sheet 8 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

- Bugatti: Has the highest average price of any brand, which is

  $1,757,223.667

- Maybach: Follows with an average price of $546,221.875

This information can help car manufacturers with pricing and positioning initiatives.

By analyzing average car prices, manufacturers can gain insights into the automotive industry's pricing landscape. This helps them position their products strategically, adjust pricing to align with consumer expectations and competitor prices, and ultimately enhance their competitiveness and profitability.

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

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

OUTPUT :

To find the relationship between fuel efficiency and the no of cylinders in a car’s engine we have to first use scatter plot to visualize the relationship between the no of cylinders & highway MPG and use a trendline to estimate the slope of the relationship and assess its significance

The Columns used to plot a scatter plot are Engine Cylinders & Highway MPG columns

	A	B
1	Engine Cylinders	highway MPG
2	6	26
3	6	28
4	6	28
5	6	28
6	6	28
7	6	28
8	6	26
9	6	28
10	6	28
11	6	27
12	6	28
13	6	28
14	6	28
15	6	28
16	6	28
17	6	25

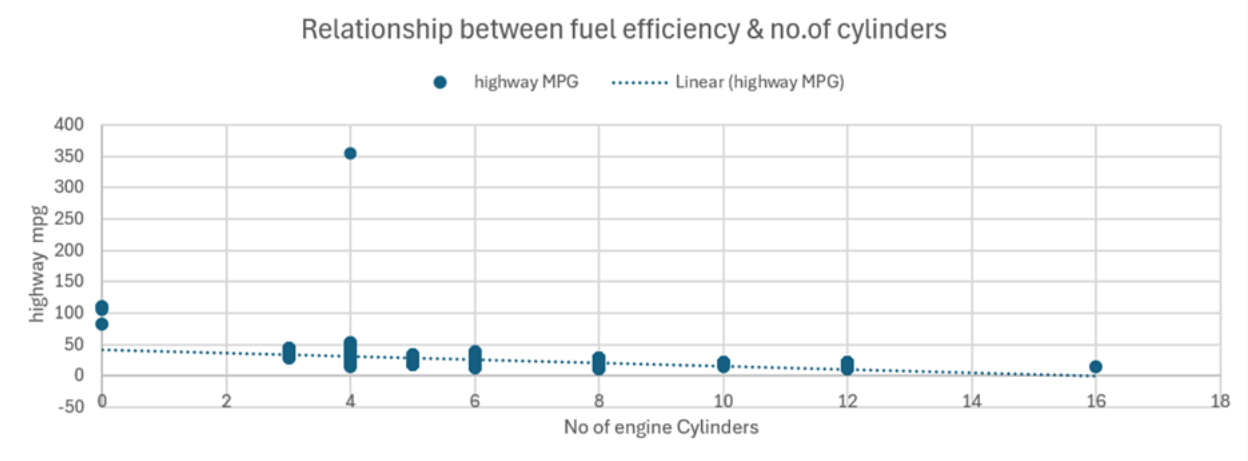
The Columns used to plot a scatter plot are in the Sheet 9 in the given link below :

Kindly open sheet 9 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

A scatter plot is created to visualize the relationship between the no of cylinders & highway MPG and a trendline on the scatter plot is created to estimate the slope of the relationship and assess its significance & the scatter plot is created using Engine cylinder & Highway MPG columns

The scatter plot created :



A Scatter plot which is created to visualize the relationship between no of cylinders and highway MPG Is in sheet 9 in the given link below :

Kindly open sheet 9 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the correlation coefficient is found between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

The Correlation Coefficient of Engine Cylinders & Highway MPG column is found using CORREL function :

= CORREL(A2:A11813,B2:B11813)

= -0.620312551

The Correlation Coefficient of Engine Cylinders & Highway MPG column is -0.620312551

The investigation shows a negative relationship between highway MPG and the number of engine cylinders, meaning that as the cylinder count increases, highway MPG decreases. For example, cars with 16 cylinders have the lowest highway MPG of 14 illustrating the impact of engine capacity on fuel economy.

A correlation coefficient of -0.62 confirms this significant negative relationship, indicating that higher cylinder counts generally lead to lower highway fuel efficiency.

This analysis quantifies how higher cylinder counts typically result in lower highway fuel efficiency.

Building the Dashboard :

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

OUTPUT :

To find How the distribution of car prices vary by brand and body style we first have to use pivot tables to find the total MSRP for each brand and body style

The columns used to create pivot tables are Make,Vehicle Style,MSRP columns

	A	B	C
1	Make	Vehicle Style	MSRP
2	BMW	Coupe	46135
3	BMW	Convertible	40650
4	BMW	Coupe	36350
5	BMW	Coupe	29450
6	BMW	Convertible	34500
7	BMW	Coupe	31200
8	BMW	Convertible	44100
9	BMW	Coupe	39300
10	BMW	Convertible	36900
11	BMW	Convertible	37200
12	BMW	Coupe	39600

The columns used to create pivot tables are in the Sheet 10 in the given link below :

Kindly open sheet 10 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

Sum of MSRP	Column Labels															
Make	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regular Cab Pickup	Sedan	Wagon	Grand Total
Acura	480917		357440	2663505					793748						4294702	201360	8791672
Alfa Romeo							129800		178200								308000
Aston Martin							7321655		9635275						1448735		18405665
Audi	4000			2674900			3291405		3556290						7158348	847350	17532293
Bentley							6012870		6356760						5920900		18290530
BMW	80097		1144950	3160950			4502671		3419051						7989300	259600	20556619
Bugatti									5271671								5271671
Buick				2141770			179325		18534			330065			2850590	8212	5528496
Cadillac				7182555			985607		2953574	599150					9418847	1184100	22323833
Chevrolet	8000	213310	1209735	6569568	420150	78688	2953245	106300	3504525	5927617	3117951	1178515	607670	2260032	3068812	300675	31524793
Chrysler	98805			250545			630105		114510			922295			2479859	501075	4997194
Dodge	48000	44000	18000	2572405	60520	338497	12000		3264627	2235775	864172	557425	70708	719408	2417585	793055	14016177
Ferrari							4723811		11418289								16142100
FIAT	325315			369305			327965									287570	1310155
Ford	36000	479873	480155	4370871	680770	566351	730007		1398144	3812353	2285584	1271330	2431898	1299240	2299348	1635565	23777489
Genesis															139850		139850

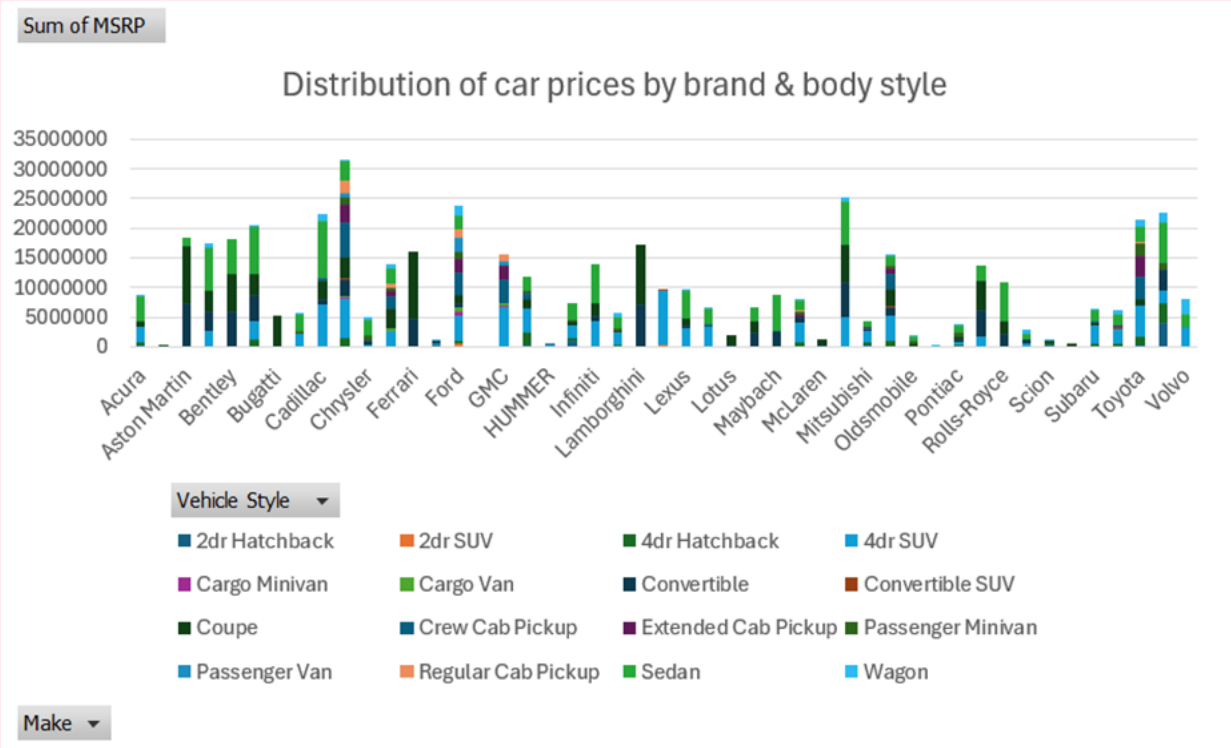
The pivot table is in the sheet 11 in the given link below

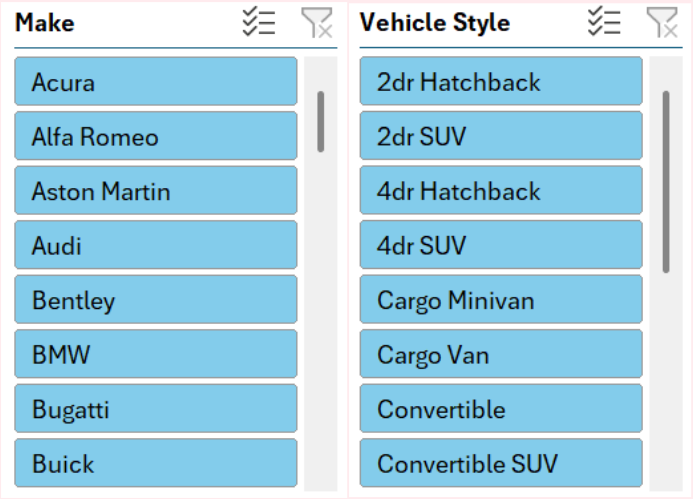
Kindly open sheet 11 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Stacked column chart is created to show the distribution of car prices by brand and body style and slicers are used to make the chart interactive and Stacked column chart is created using the pivot table columns

Stacked column created with sliders :





Stacked column chart which is created to show the distribution of car prices by brand and body style and slicers is in the sheet 11 in the given link below :

Kindly open sheet 11 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Chevrolet leads with the highest total prices amounting to $31524793

- Mercedes-Benz follows with a total price of $25181309

- Sedan have the highest total cost among car types, totaling $117474790

- Four-door SUV come next with a total of $100258517

These insights are valuable for the car industry's marketing strategies, product development, and pricing decisions.

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

OUTPUT :

To find Which car brands have the highest and lowest average MSRPs, and how does this vary by body style we have to first use Pivot tables to find the average MSRP for each brand and body style

The columns used to create pivot table are Make,Vehicle Style & MSRP columns

	A	B	C
1	Make	Vehicle Style	MSRP
2	BMW	Coupe	46135
3	BMW	Convertible	40650
4	BMW	Coupe	36350
5	BMW	Coupe	29450
6	BMW	Convertible	34500
7	BMW	Coupe	31200
8	BMW	Convertible	44100
9	BMW	Coupe	39300
10	BMW	Convertible	36900
11	BMW	Convertible	37200
12	BMW	Coupe	39600

The columns used to create a pivot table is in sheet 12 in the given link below :

Kindly open sheet 12 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

3	Average of MSRP	Column Labels													
4	Make	2dr Hatchback	2dr SUV	4dr Hatchback	4dr SUV	Cargo Minivan	Cargo Van	Convertible	Convertible SUV	Coupe	Crew Cab Pickup	Extended Cab Pickup	Passenger Minivan	Passenger Van	Regu
5	Acura	17175.60714		51062.85714	42959.75806					39687.4					
6	Alfa Romeo							64900		59400					
7	Aston Martin							203379.3056		192705.5					
8	Audi	2000			48634.54545			70029.89362		93586.57895					
9	Bentley							250536.25		254270.4					
10	BMW	26699		54521.42857	58536.11111			63417.90141		51803.80303					
11	Bugatti									1757223.667					
12	Buick				33996.34921			25617.85714		2059.333333			30005.90909		
13	Cadillac				72551.06061			70400.5		45439.6	66572.22222				
14	Chevrolet	2000	8887.916667	18329.31818	32046.67317	20007.14286	7153.454545	62835	17716.66667	38939.16667	39255.74172	24170.16279	24552.39583	24306.8	
15	Chrysler	32935			35792.14286			24234.80769		19085			29751.45161		
16	Dodge	2000	2000	2000	30992.83133	20173.33333	12536.92593	2000		45980.66197	31052.43056	13938.25806	25337.5	14141.6	
17	Ferrari							214718.6818		248223.6739					
18	FIAT	19136.17647			24620.33333			23426.07143							
19	Ford	2000	13710.65714	18467.5	42027.60577	21274.0625	17698.46875	34762.2381		34101.07317	41438.61957	23808.16667	23115.09091	32425.30667	

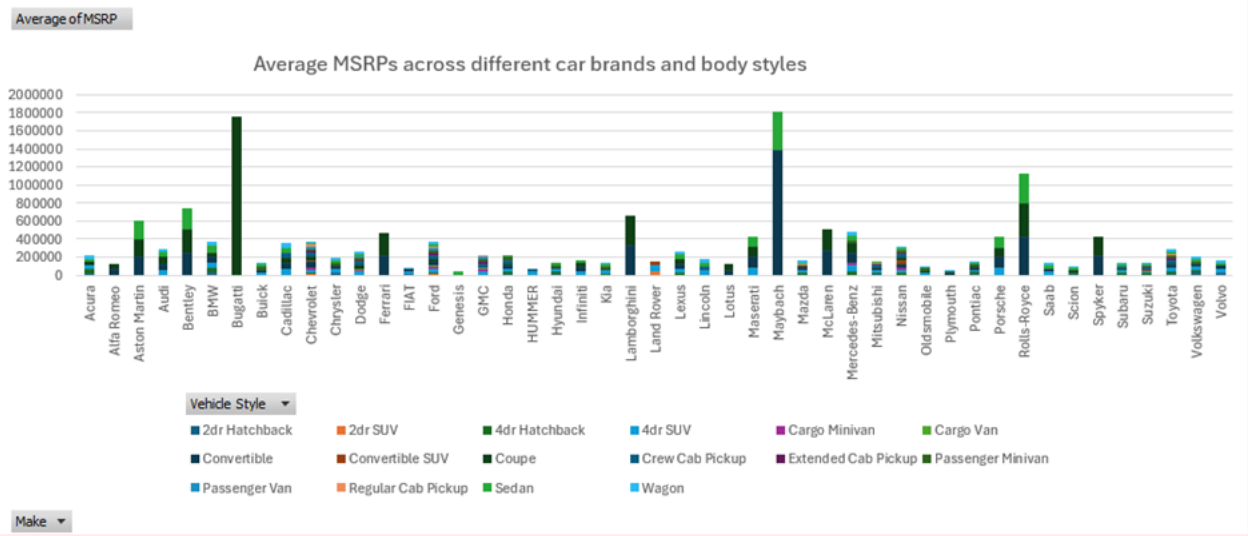
The pivot table is in the sheet 13 in the given link below :

Kindly open sheet 13 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Clustered column chart is created to compare the average MSRPs across different car brands and body styles and Clustered column chart is created using the pivot table columns

The clustered column chart created:



The Clustered column chart which is created to compare the average MSRPs across different car brands and body styles are in the sheet 13 in the given link below :

Kindly open sheet 13 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Brand Analysis:

- Bugatti : Highest average price of $1757223.667
- Maybach: Second highest average price at $546221.875
- Plymouth: Lowest average price per brand at $3122.902439
- Oldsmobile: Second lowest average price at $11542.54

- Body Style Analysis:
  - Convertibles: Highest average price at $84224.28499
  - Coupes: Second highest average price at $76900.70504
  - 2-door SUVs: Most affordable at $. 10115.18841
  - 2-door Hatchbacks: Second most affordable at $16778.65408

This research provides valuable insights into price patterns across various car markets.

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

OUTPUT :

To find How different feature such as transmission type affect the MSRP, and how does this vary by body style we have to first use Pivot tables to find the average MSRP for each combination of transmission type and body style

The columns used to create pivot tables are Transmission type ,MSRP & vehicle style columns

	A	B	C
1	Transmission Type	MSRP	Vehicle Style
2	MANUAL	46135	Coupe
3	MANUAL	40650	Convertible
4	MANUAL	36350	Coupe
5	MANUAL	29450	Coupe
6	MANUAL	34500	Convertible
7	MANUAL	31200	Coupe
8	MANUAL	44100	Convertible
9	MANUAL	39300	Coupe
10	MANUAL	36900	Convertible
11	MANUAL	37200	Convertible
12	MANUAL	39600	Coupe
13	MANUAL	31500	Coupe
14	MANUAL	44400	Convertible
15	MANUAL	37200	Convertible
16	MANUAL	31500	Coupe
17	MANUAL	48250	Convertible

The columns used to create a pivot table are in sheet 14 in the given link below :

Kindly open sheet 14 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

Average of MSRP	Column Labels					
VEHICLE STYLE	AUTOMATED_MANUAL	AUTOMAT	DIRECT_D	MANUAL	UNKNOWI	Grand Total
2dr Hatchback	27180.96491	20926.5		13353.7	7361.5	16778.65408
2dr SUV		18615.2		6303.81	2371	10115.18841
4dr Hatchback	29249.07407	23833.7	34511.9	17594.4		22086.30236
4dr SUV	40451.15385	41555.2		15426.5		40426.82137
Cargo Minivan		20910.9				20910.85714
Cargo Van		15280.2				15280.22105
Convertible	121256.6444	90637.4		62357.8	5783.5	84224.28499
Convertible SUV		38925.5		9233.14		17424.13793
Coupe	245588.3571	63852		51070.5	2000	76900.70504
Crew Cab Pickup		37744.1		28360.5		37220.46696
Extended Cab Pickup		30637.3		10884.2		22488.77689
Passenger Minivan		26392		4405.33		25591.51214
Passenger Van		29015.2				29015.20313
Regular Cab Pickup		28536.8		7557.77	2000	15953.70918
Sedan	47498.70813	43794.4	27822.5	17119.2	2000	38989.30966
Wagon	31985.27778	27613.2		17844.1		25483.90119
Grand Total	**99195.584**	**41137.3**	**33620**	**26671.4**	**3040.74**	**40559.93532**

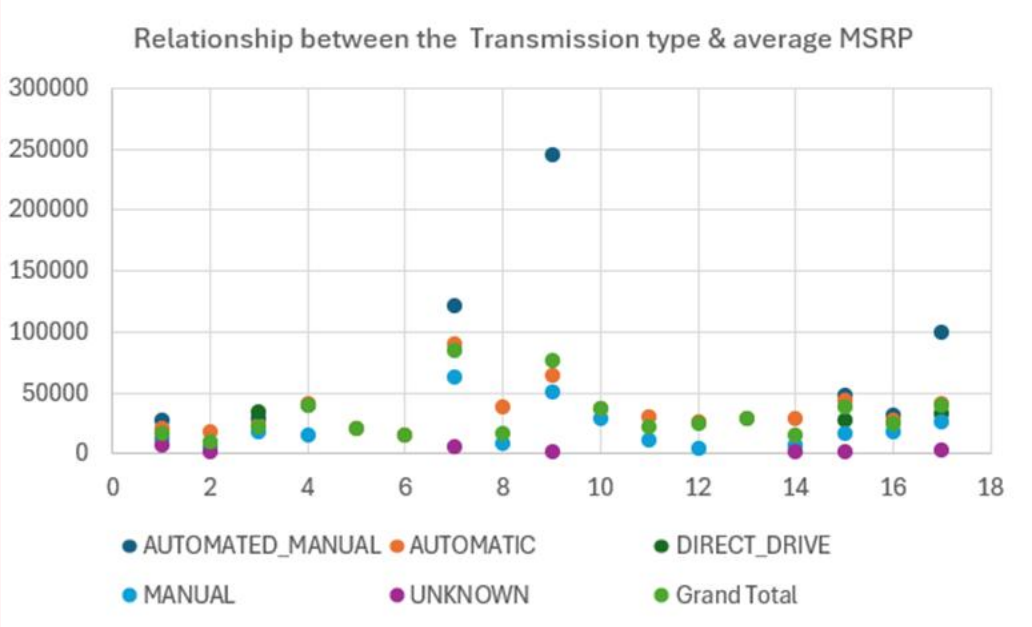
The pivot table is in the sheet 15 in the given link below :

Kindly open sheet 15 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Scatter plot chart is created to visualize the relationship between MSRP and transmission type, with different symbols for each body style and the scatter plot chart is created using the pivot table columns

The scatter plot chart created :



The Scatter plot chart which is created to visualize the relationship between MSRP and transmission type, with different symbols for each body style Is in the sheet 15 in the given link below :

Kindly open sheet 15 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Transmission Types

- Direct Drive: Average price of $33620
- Automated Manual: Highest average price at $99195.584
- Manual : Second highest average price at $26671.4
- Unknown: Lowest average price at $3040.737

Body Shapes

- Coupes: Average price of $76900.70504
- Convertibles: Highest average price at $84224.28499
- 2-Door SUVs: Least expensive at $10115.18841
- 2-Door Hatchbacks: Next most affordable at $16778.65408

These insights provide useful information for understanding price trends in the car industry across various body shapes and gearbox types.

When considering transmission types, automated manual transmission has the highest average price, closely followed by direct drive transmission. In terms of body shape, convertibles have the highest average price, with coupes .Manual transmission ranks second, while unknown transmission has the lowest average price. Among body shapes, 2-door SUVs are the least expensive, followed by 2-door hatchbacks.

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

OUTPUT :

To find How the fuel efficiency of cars vary across different body styles and model years first we have to use Pivot tables to find the average MPG for each combination of body style and model year

The columns used to create pivot table are Year, vehicle style, Highway MPG & city MPG columns

	A	B	C	D
1	Year	Vehicle Style	highway MPG	city mpg
2	2011	Coupe	26	19
3	2011	Convertible	28	19
4	2011	Coupe	28	20
5	2011	Coupe	28	18
6	2011	Convertible	28	18
7	2012	Coupe	28	18
8	2012	Convertible	26	17
9	2012	Coupe	28	20
10	2012	Convertible	28	18
11	2013	Convertible	27	18
12	2013	Coupe	28	20
13	2013	Coupe	28	19
14	2013	Convertible	28	19
15	2013	Convertible	28	19
16	2013	Coupe	28	19

The columns used to create pivot table are in the sheet 16 in the given link below :

Kindly open sheet 16 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

3	Year	Average of highway MPG
4	⊞1990	23.1
5	⊞1991	22.2
6	⊞1992	24.1
7	⊞1993	24.2
8	⊞1994	23.9
9	⊞1995	23.2
10	⊞1996	23.7
11	⊞1997	22.3
12	⊞1998	21.9
13	⊞1999	23.0
14	⊞2000	24.0
15	⊞2001	23.7
16	⊞2002	22.8
17	⊞2003	22.7
18	⊞2004	23.1
19	⊞2005	23.6
20	⊞2006	23.4
21	⊞2007	21.9
22	⊞2008	23.0
23	⊞2009	23.9
24	⊞2010	24.2
25	⊞2011	25.2
26	⊞2012	26.3
27	⊞2013	27.4
28	⊞2014	26.9
29	⊞2015	28.5
30	⊞2016	28.5
31	⊞2017	28.3
32	Grand Total	26.3

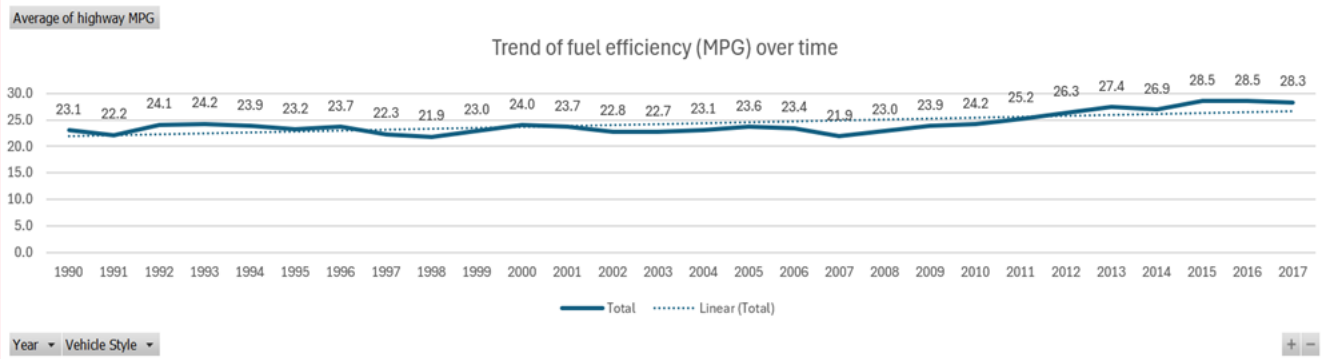
The pivot table is in the sheet 17 in the given link below :

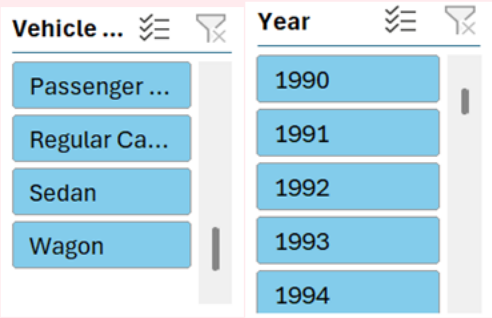
Kindly open sheet 17 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Line chart is created to visualize the trend of fuel efficiency (MPG) over time for each body style and the line chart is created using the pivot table columns

The line chart created with Sliders :





The Line chart which is created to visualize the trend of fuel efficiency (MPG) over time for each body style is in the sheet 17 in the given link below :

Kindly open sheet 17 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The data indicates a trend of increased fuel efficiency as car model years progress, likely due to advancements in automotive technology and regulations focused on reducing emissions and enhancing fuel economy. In 2016, the highest average highway fuel efficiency, with an average MPG was 28.5 , achieved by four-door hatchbacks, reflecting efforts to improve fuel efficiency. Conversely, cargo vans had the lowest average highway MPG at 16, highlighting variations in fuel efficiency across different vehicle categories.

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

OUTPUT :

To find How the car's horsepower, MPG, and price vary across different Brands we first have to use pivot tables to find average horsepower, MPG, and MSRP for each car brand

The columns used to create pivot table are Make,Engine HP,Highway MPG,MSRP columns

	A	B	C	D
1	Make	Engine HP	highway MPG	MSRP
2	BMW	335	26	46135
3	BMW	300	28	40650
4	BMW	300	28	36350
5	BMW	230	28	29450
6	BMW	230	28	34500
7	BMW	230	28	31200
8	BMW	300	26	44100
9	BMW	300	28	39300
10	BMW	230	28	36900
11	BMW	230	27	37200
12	BMW	300	28	39600
13	BMW	230	28	31500
14	BMW	300	28	44400

The columns used to create a pivot table is in the sheet 18 in the given link below :

Kindly open sheet 18 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The pivot table created :

3	Make	Average of Engine HP	Average of highway MPG	Average of MSRP
4	Acura	244.797619	28.11111111	34887.5873
5	Alfa Romeo	237	34	61600
6	Aston Martin	484.3225806	18.89247312	197910.3763
7	Audi	277.695122	28.82317073	53452.1128
8	Bentley	533.8513514	18.90540541	247169.3243
9	BMW	326.9071856	29.24550898	61546.76347
10	Bugatti	1001	14	1757223.667
11	Buick	219.244898	26.94897959	28206.61224
12	Cadillac	332.3098237	25.23677582	56231.31738
13	Chevrolet	247.0565022	25.6690583	28273.35695
14	Chrysler	229.1390374	26.36898396	26722.96257
15	Dodge	244.4153355	22.34504792	22390.05911
16	Ferrari	509.9117647	15.72058824	237383.8235
17	FIAT	143.559322	33.91525424	22206.01695
18	Ford	243.0979263	23.74078341	27393.42051
19	Genesis	347.3333333	25.33333333	46616.66667
20	GMC	259.8446602	21.4038835	30493.29903
21	Honda	195.7494407	32.25055928	26629.81879
22	HUMMER	261.2352941	17.29411765	36464.41176
23	Hyundai	201.9174917	30.39273927	24597.0363
24	Infiniti	310.0666667	24.77878788	42394.21212
25	Kia	206.8274336	29.29646018	25112.38938

The Pivot table created is in the sheet 19 in the given link below :

Kindly open sheet 19 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Bubble chart is created to visualize the relationship between horsepower, MPG, and price across different car brands. different colors has been assigned to each brand and the bubbles have been labelled with the car model name. and the bubble chart is created using pivot table columns

The bubble chart created :

The Bubble chart which is created to visualize the relationship between horsepower, MPG, and price across different car brands is in the sheet 19 in the given link below :

Kindly open sheet 19 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/15EpHN5oN-UiHcCwMF3B81wkcF797FYg-/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Brand Comparison

Bugatti:

- Greatest Average Engine Horsepower: 1001
- Highest Average Price: $ 1757223.667
- Highest Average Highway MPG : 14

Alfa Romeo:

- Highest Average Highway MPG: 34

- Greatest Average Engine Horsepower: 237

- Highest Average Price: $ 61600

- Bugatti is Known for high-performance cars with top-tier engine power and pricing, appealing to those seeking unmatched performance.

- Alfa Romeo emphasizes performance and style, with a range of gasoline and plug-in hybrid vehicles. They offer powerful engines and sporty designs, appealing to those who enjoy driving dynamics.

Alfa Romeo appeals to driving enthusiasts who value performance and design.

The comparison illustrates different goals and preferences within the automobile industry, catering to a broad spectrum of customers with varying requirements and interests.

BUILDING THE DASHBOARD:

you need to create the Interactive Dashboard.

OUTPUT :

To create a interactive dashboard filters and slicers are used in the dashboard .

DASHBOARD :

The interactive dashboard is created by combining all the last 5 charts which are :

1.Distribution of car prices by brand & body style Chart

2.Trend of Fuel efficiency over Time Chart

3.Average MSRPs across different car brands & body styles Chart

4.Relationship between Transmission type & average MSRP Chart

5.Relationship between Horsepower, MPG and price across different car brands chart

Along with their respective sliders attached and the necessary filters can be applied

The Interactive Dashboard created :

The Interactive Dashboard which is created by combining all the 5 last charts with sliders attached Is in the sheet 22 in the given link below :

Kindly open sheet 22 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1dGZTFve0RqLlgmviySxQ3sDRz2nMmot4/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

ANALYSIS

Car Features Impacting Price:

1. Engine Type and Performance: Advanced engines, such as hybrid or electric, command higher prices due to superior performance and fuel efficiency.

2. Safety Features: Comprehensive safety features like airbags and ABS increase the perceived value, leading to higher prices.

3. Infotainment and Connectivity: State-of-the-art infotainment systems and connectivity options cater to tech-savvy consumers, allowing for higher pricing.

4. Interior and Comfort: Premium materials and comfort features like heated seats contribute to a higher price tag.

5. Design and Aesthetics: Stylish designs and unique elements attract a wider audience, justifying higher prices.

Car Features Impacting Profitability:

1. Fuel Efficiency: High fuel efficiency attracts consumers, increasing sales volumes and profitability.

2. Maintenance Costs: Low maintenance costs enhance customer satisfaction and brand loyalty, leading to repeat purchases.

3. Resale Value: Features that improve resale value, such as reliability, contribute to increased demand and profitability.

4. Production Costs: Features with a high value-to-cost ratio enhance profit margins for manufacturers.

5. Market Demand: Prioritizing high-demand features maximizes profitability and market success.

CONCLUSION

Analyzing the impact of car features on price and profitability reveals critical insights for automotive manufacturers. Key features such as advanced engines, comprehensive safety systems, state-of-the-art infotainment, and premium interior comfort significantly influence vehicle prices. These features enhance the perceived value and market appeal, allowing manufacturers to command higher prices.

Moreover, features like fuel efficiency, low maintenance costs, and high resale value drive long-term profitability. Fuel-efficient vehicles attract more consumers, leading to increased sales volumes and brand loyalty. Low maintenance costs enhance customer satisfaction, resulting in repeat purchases and positive referrals. High resale value maintains a car's market appeal, reducing depreciation and boosting profitability.

By understanding consumer preferences and leveraging data analytics, manufacturers can optimize feature offerings, pricing strategies, and market positioning. This approach not only enhances their competitive edge but also ensures sustained financial performance in the dynamic automotive market.

Additionally, adopting a data-driven approach enables manufacturers to anticipate market trends and consumer demands more accurately, fostering innovation and better product development.

Ultimately, a strategic focus on integrating desirable car features with efficient cost management will create value for both manufacturers and consumers, driving overall growth and success in the competitive automotive industry.

ABC CALL VOLUME TREND ANALYSIS

DESCRIPTION

A customer experience (CX) team consists of professionals who analyze customer feedback and data, and share insights with the rest of the organization. Typically, these teams fulfil various roles and responsibilities such as: Customer experience programs (CX programs), Digital customer experience, Design and processes, Internal communications, Voice of the customer (VoC), User experiences, Customer experience management, Journey mapping, Nurturing customer interactions, Customer success, Customer support, Handling customer data, Learning about the customer journey. Let's look at some of the most impactful AI-empowered customer experience tools you can use today: Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, Intelligent Routing In a Customer Experience team there is a huge employment opportunities for Customer service representatives A.k.a. call centre agents, customer service agents.

Some of the roles for them include: Email support, Inbound support, Outbound support, social media support.

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

THE PROBLEM

ABC Insurance Company operates a call center that handles customer inquiries and service requests. The center currently operates during daytime hours (9 AM to 9 PM) but receives calls throughout the 24-hour period. The company is experiencing significant operational challenges that are negatively impacting customer service quality and resource utilization.

High Call Abandonment Rate

- The current call abandonment rate stands at approximately 30%, meaning nearly one-third of customers disconnect before their calls are answered

- Night-time calls (9 PM to 9 AM) go completely unanswered due to lack of agent availability

- For every 100 calls during daytime hours, an additional 30 calls are received during night hours

Resource Allocation Inefficiencies

- Lack of optimal staffing levels during different time periods
- Inconsistent call handling times across different time buckets
- No structured approach to handling night-time call volumes
- Ineffective distribution of available agent resources during peak and low-volume periods

Service Quality Issues

- Poor customer experience due to high abandonment rates
- Complete lack of service during night hours
- Varying call duration patterns indicating potential inconsistencies in service delivery
- Peak hour congestion leading to longer wait times

DESIGN

Before starting the actual analysis, I have: -

• First, I made a copy of the raw data where I can perform the Analysis so that the changes, I make it will not affect the original data.

• Then Il removed the irrelevant columns(data) from the dataset which was not necessary for doing the analysis.

• I removed rows having blank spaces and NULL values.

• Then removed duplicate rows from the datasets.

Software used for doing the overall Analysis: -

----> Microsoft Excel

FINDINGS

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

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

OUTPUT :

to find the average duration of calls for each time bucket

I have taken Time_ bucket ,call_seconds(s), call_status columns for analysis

	A	B	C
1	Time_Bucket	Call_Seconds (s)	Call_Status
2	9_10	96.00	answered
3	9_10	140.00	answered
4	9_10	85.00	answered
5	9_10	91.00	answered
6	9_10	165.00	answered
7	9_10	0.00	abandon
8	9_10	85.00	answered
9	9_10	0.00	abandon
10	9_10	65.00	answered
11	9_10	180.00	answered
12	9_10	108.00	answered
13	9_10	186.00	answered
14	9_10	0.00	abandon
15	9_10	100.00	answered
16	9_10	75.00	answered
17	9_10	0.00	abandon

The Rest of the data of Time_bucket,call_seconds(s),call_status columns used for analysis is in the given link below :

Kindly open sheet 1 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqileOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have created a pivot table using theTime_bucket,call_seconds(s),call_status columns to get the average call duration for each time bucket

	A	B
1	Call_Status	answered
2		
3	**Time_Bucket**	**Average of Call_Seconds (s)**
4	10_11	203.3310302
5	11_12	199.2550234
6	12_13	192.8887829
7	13_14	194.7401744
8	14_15	193.6770755
9	15_16	198.8889175
10	16_17	200.8681864
11	17_18	200.2487831
12	18_19	202.5509677
13	19_20	203.4060725
14	20_21	202.845993
15	9_10	199.0691057
16	**Grand Total**	**198.6227745**

19_20 is the highest average call duration recorded

12_13 is the lowest average call duration recorded

The pivot table created using the Time_bucket,call_seconds(s),call_status columns to get the average call duration for each time bucket is in the given link below :

Kindly open sheet 2 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqiIeOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Thus by finding the average duration of call for each time bucket .The analysis also reveals significant differences in call durations throughout various times of the day. Calls tend to last the longest between 7 and 8 PM, and the shortest between 12 and

1 PM.

This suggests that call length patterns vary with different times of the day. Factors like fewer available agents, more complex calls, or increased client interaction might contribute to longer call durations at night. Conversely, lower call volume or quicker responses to consumer concerns might explain the shorter average call duration in the afternoon.

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

OUTPUT :

to visualize the number of calls received in each time bucket I have taken Time_ bucket & call_status columns for analysis

	A	B
1	Time_Bucket	Call_Status
2	9_10	answered
3	9_10	answered
4	9_10	answered
5	9_10	answered
6	9_10	answered
7	9_10	abandon
8	9_10	answered
9	9_10	abandon
10	9_10	answered
11	9_10	answered
12	9_10	answered
13	9_10	answered
14	9_10	abandon
15	9_10	answered
16	9_10	answered

The Rest of the data of Time_bucket,call_status columns used for analysis is in the given link below :

Kindly open sheet 3 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqiIeOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have created a pivot table using the Time_bucket,call_status columns to get the no of calls received for each time bucket

3	Row Labels	Count of Call_Status
4	10_11	13313
5	11_12	14626
6	12_13	12652
7	13_14	11561
8	14_15	10561
9	15_16	9159
10	16_17	8788
11	17_18	8534
12	18_19	7238
13	19_20	6463
14	20_21	5505
15	9_10	9588
16	**Grand Total**	**117988**

The pivot table created using the Time_bucket,call_status columns to get the no of calls received for each time bucket is in the given link below :

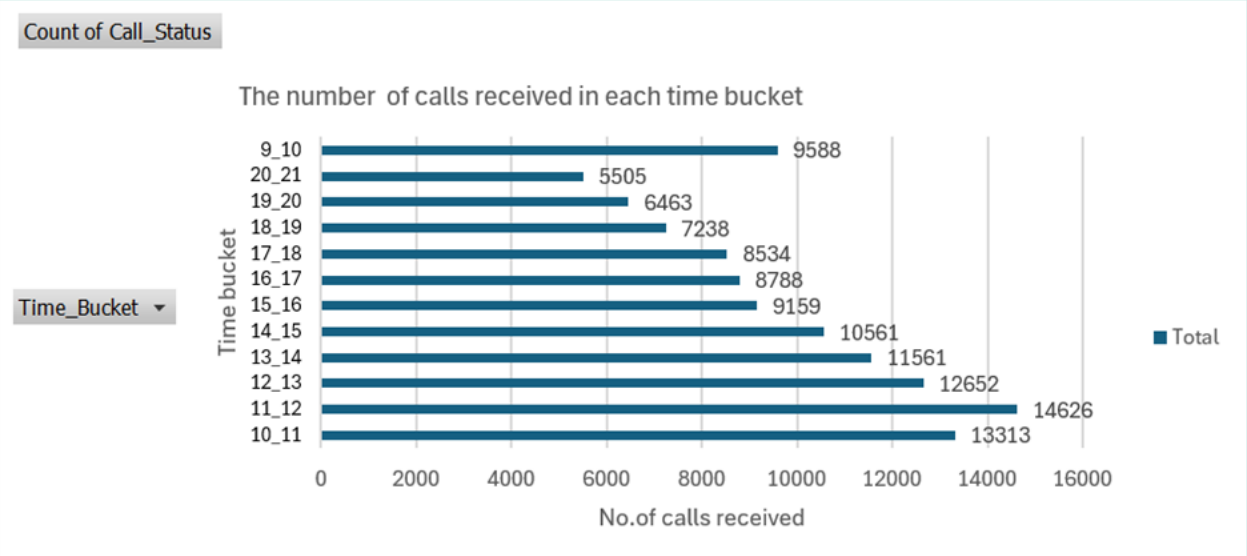
Kindly open sheet 4 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqiIeOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Thus we have found the no of calls received for each time bucket

To visualize the no of calls received for each time bucket I have created a clustered bar chart by using the pivot table columns Row labels and count of call_status

clustered bar chart by using the pivot table columns Row labels and count of call_status



In the chart ,The highest no of calls received is from 11am to 12 pm that is 14626 calls and the lowest no of calls received is from 8 am to 9 pm that is 5505 calls received by the agents

The clustered chart used to visualize the no of calls received for each time bucket by using the pivot table columns Row labels and count of call_status is in the given link below :

Kindly open sheet 4 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqiIeOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Thus by visualizing the no of calls received for each time bucket through the clustered bar chart

The study also analyzes variations in call volume across different times of the day. Calls peak between 11:00AM and 12:00 PM and are least frequent between 8:00 PM and 9:00 PM.

These findings emphasize the importance of understanding temporal patterns in call traffic to allocate resources effectively and ensure prompt customer service. By aligning staffing levels with peak call hours, organizations can enhance customer satisfaction and operational efficiency.

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

Your Task: What is the minimum number of agents required in each time bucket to

reduce the abandon rate to 10%?

OUTPUT :

to find the minimum number of agents required in each time bucket to reduce the abandon rate to 10%

Atleast 90 calls should be consumed out of 100 to reduce the abundant percent from 30% to 10 %

I have taken the Date_&_Time, Duration(hh:mm:ss), Call_Status columns for analysis

	A	B	C
1	Date_&_Time	Duration(hh:mm:ss)	Call_Status
2	1/1/2022	0:01:36	answered
3	1/1/2022	0:02:20	answered
4	1/1/2022	0:01:25	answered
5	1/1/2022	0:01:31	answered
6	1/1/2022	0:02:45	answered
7	1/1/2022	0:00:00	abandon
8	1/1/2022	0:01:25	answered
9	1/1/2022	0:00:00	abandon
10	1/1/2022	0:01:05	answered
11	1/1/2022	0:03:00	answered
12	1/1/2022	0:01:48	answered

The rest of the data of the Date_&_Time, Duration(hh:mm:ss), Call_Status columns

For analysis are in the given link below :

Kindly open sheet 5 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqileOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have created a pivot table using the Date_&_Time, Duration(hh:mm:ss), Call_Status columns

Count of Duration(hh:mm:ss)	Column Labels			
Date_&_Time	abandon	answered	transfer	Grand Total
1-Jan	684	3883	77	4644
2-Jan	356	2935	60	3351
3-Jan	599	4079	111	4789
4-Jan	595	4404	114	5113
5-Jan	536	4140	114	4790
6-Jan	991	3875	85	4951
7-Jan	1319	3587	42	4948
8-Jan	1103	3519	50	4672
9-Jan	962	2628	62	3652
10-Jan	1212	3699	72	4983
11-Jan	856	3695	86	4637
12-Jan	1299	3297	47	4643
13-Jan	738	3326	59	4123
14-Jan	291	2832	32	3155
15-Jan	304	2730	24	3058
16-Jan	1191	3910	41	5142
17-Jan	16636	5706	5	22347
18-Jan	1738	4024	12	5774
19-Jan	974	3717	12	4703
20-Jan	833	3485	4	4322
21-Jan	566	3104	5	3675
22-Jan	239	3045	7	3291
23-Jan	381	2832	12	3225
Grand Total	34403	82452	1133	117988
Average no. of call status	1496	3585	49	5130
Call status in %	29%	70%	1%	
Agent's working hour	4.5			
Average of call duration in sec	198.62			
Hours needed for 90%	255			
Total no.of agents required is	57			

The pivot table created using the Date_&_Time, Duration(hh:mm:ss), Call_Status columns is in the given link below :

Kindly open sheet 5 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqiIeOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

To know the no of calls answered ,abundant and transfer I have used the average function from 6 to 28 cell to find the the average for abandon column

=AVERAGE(I6:I28)

in the same way I have found averages for transfer and answered columns in the pivot table in order to know the call status in percentage I have just divided

The average of abandon by grand total that gives call percentage of abandon as 29%, and the average of answered by grand total that gives the call percentage of answered as 70%, and the average of transfer by grand total that gives the call percentage of transfer as 1%

1.In order to lower the abandon rate to 10% how many agents must

be present in each time bucket at minimum ?

To reduce the abandonment rate to 10%, the minimum number of agents required per time bucket must be determined

The working conditions and schedule of an agent, including:

- Work Schedule: Agents work 6 days a week.
- Leaves: Agents take an average of 4 unscheduled leaves per month.
- Daily Working Hours: Agents have a total working time of 9 hours per day.
- Breaks: 1.5 hours are allocated for lunch and snacks.

- Customer Interaction: Agents typically engage with customers for 60% of their working hours.
- Monthly Consideration: The month is considered to have 30 days.

Considering these factors, the goal is to calculate the necessary number of agents per time bucket to ensure that at least 90 out of 100 calls are answered, taking into account the actual time agents spend communicating with customers during their work hours.

The agent works 60 % of 7.5 hours that is

=(60/100)*7.5 = 4.5

Within 7.5 hours they have worked only for 60 % so the actual time period they have worked is 4.5 hours daiy

The average of call seconds which is calculated in task 1 is 198.62

In order to increase the call duration of answered column from 70% to 90%

To find the hours needed for 90%

I have Divided the grand total no of calls received into average call duration in seconds into 0.9 divided by 3600

=L30*I33*0.9/3600

I have divided by 3600 because I want to convert the seconds into hours

Thus the hours needed for 90% is 255 which explains it takes 255 hours taken for workers to get answered rate from 70% to 90%

To find the total no of agents required I have divided the hours needed for 90% into agents working hour which is 57

Thus the answer is to get 90% of the work done we require atleast 57 no of agents to complete the task

The workforce planning and resource allocation for a call center. It states that each agent works six days a week for 7.5 hours a day, with 60% of that time spent answering calls, equating to about 4.5 hours per day on calls. It was determined that approximately 57 agents are needed to ensure coverage and maintain a service level where at least 90 out of 100 calls are answered. This conclusion was reached after calculating the total number of hours required to manage incoming calls across all time periods, which totaled 254.7 hours. The analysis aims to meet call volume expectations and guarantee customer satisfaction.

4.Night Shift Manpower Planning: Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9pm, they also make 30 calls at night between 9 pm and 9 am. Your Task: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

OUTPUT :

to propose a manpower plan for each time bucket throughout the day keeping the maximum abandon rate at 10%

a strategy to maintain a maximum abandonment rate of 10% for call centers. It emphasizes adjusting staffing levels for different time buckets to handle call volumes effectively. the proposed plan includes:

- Daytime (9:00 AM to 9:00 PM): Allocate agents based on current call volume, ensuring staffing levels meet demand to keep the abandonment rate below 10%.

- Nighttime (9:00 PM to 9:00 AM the next day): Currently, there are no agents available during these hours, resulting in unanswered calls and poor customer experience. To address this, consider hiring overnight agents or setting up call forwarding to handle the additional 30 calls made overnight. Ensure these additional personnel are equipped to handle inquiries promptly to maintain customer satisfaction.

By implementing this strategy, the goal is to maintain an abandonment rate of 10% across all time periods, ensuring a positive customer experience around the clock.

I have taken the date_&_time, duration(hh:mm:ss), call_status columns for analysis as before in the previous task

	A	B	C
1	Date_&_Time	Duration(hh:mm:ss)	Call_Status
2	1/1/2022	0:01:36	answered
3	1/1/2022	0:02:20	answered
4	1/1/2022	0:01:25	answered
5	1/1/2022	0:01:31	answered
6	1/1/2022	0:02:45	answered
7	1/1/2022	0:00:00	abandon
8	1/1/2022	0:01:25	answered
9	1/1/2022	0:00:00	abandon
10	1/1/2022	0:01:05	answered
11	1/1/2022	0:03:00	answered
12	1/1/2022	0:01:48	answered

The rest of the data of the Date_&_Time, Duration(hh:mm:ss), Call_Status columns

For analysis are in the given link below :

Kindly open sheet 8 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqileOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

I have taken the same pivot table used in the previous task and I have added night calculations to it

Count of Duration(hh:mm:ss)	Column Labels			
Date_&_Time	abandon	answered	transfer	Grand Total
1-Jan	684	3883	77	4644
2-Jan	356	2935	60	3351
3-Jan	599	4079	111	4789
4-Jan	595	4404	114	5113
5-Jan	536	4140	114	4790
6-Jan	991	3875	85	4951
7-Jan	1319	3587	42	4948
8-Jan	1103	3519	50	4672
9-Jan	962	2628	62	3652
10-Jan	1212	3699	72	4983
11-Jan	856	3695	86	4637
12-Jan	1299	3297	47	4643
13-Jan	738	3326	59	4123
14-Jan	291	2832	32	3155
15-Jan	304	2730	24	3058
16-Jan	1191	3910	41	5142
17-Jan	16636	5706	5	22347
18-Jan	1738	4024	12	5774
19-Jan	974	3717	12	4703
20-Jan	833	3485	4	4322
21-Jan	566	3104	5	3675
22-Jan	239	3045	7	3291
23-Jan	381	2832	12	3225
Grand Total	34403	82452	1133	117988
Average no. of call status	1496	3585	49	5130
Call status in %	29%	70%	1%	
Agent's working hour	4.5			
Average of call duration in sec	198.62			
Average of no .of calls at night	1539			
For 90% call rate at night	76			
Total no of agents needed in the night shift	17			

The pivot table used in the previous task along with the night calculations are in the given link below :

Kindly open sheet 7 given link below in Microsoft Excel to view charts & Data

https://docs.google.com/spreadsheets/d/1Z8RKNlaZAfqileOzd0wcdwX_rJ1Sajbr/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

To know the no of calls answered ,abundant and transfer I have used the average function from 6 to 28 cell to find the the average for abandon column

=AVERAGE(I6:I28)

in the same way I have found averages for transfer and answered columns in the pivot table in order to know the call status in percentage I have just divided

The average of abandon by grand total that gives call percentage of abandon as 29%, and the average of answered by grand total that gives the call percentage of answered as 70%, and the average of transfer by grand total that gives the call percentage of transfer as 1%

The agent works 60 % of 7.5 hours that is

=(60/100)*7.5 = 4.5

Within 7.5 hours they have worked only for 60 % so the actual time period they have worked is 4.5 hours daiy

As the same as The average of call seconds which is calculated in task 1 is 198.62

As mentioned there is 30 calls received during the night

So to find the Average of no .of calls at night

The calculation is

=0.3*F30

Which is 30 into grand total of no of calls

Which gives 1539

To find 90% call rate at night

I have divided Average of call duration in sec into Average of no .of calls at night

Which is =C34*C33*0.9/3600

In the above calculation

I have multiplied it by 0.9 to get 90% and divided it by 3600 to get the calculation in hours

Which gives average of no of calls at night as 76

To find the total no of agents needed in the night shift

I have divided 90% call rate at night into Agent's working hour

Which is =C35/C32

Which gives 17

Thus the Total no of agents needed in the night shift is 17 agents

17 agents needed to work during the night shift so that 90 out of 100 calls can be answered during that time and the company profit will be increased and the customer satisfaction will arise

Thus analyzing the need for night shift staffing, estimating that an average of 1,539 calls are made during these hours. To achieve a 90% call rate, approximately 17 agents are required. This approach ensures effective scheduling, sufficient coverage to address call volume needs, and maintains service quality during the night shift.

ANALYSIS

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

Why is that the average call answered were more in count in the time bucket of 10_11, 18_19, 19_20 and 20_21 as compared to other time buckets?

---> Most of the customers are office people and they need to reach office by 10 AM or 11 AM, so these customers call during 10_11 time bucket i.e. while they in transit to office or have reached office and have some free time before they start their work; During the time bucket 18_19, 19_20 and 20_21 the customers have either left their office and reached home or they are in the transit to reach home and during

these time period i.e. 6 Pm to 9 pm people have free time where they can share their concern to the customer service.

During these time buckets most of the calls are from individual people with small problems which can be resolved quickly

Why is it that the time bucket 11_12 has the highest number of incoming calls but it does not have the highest number of average answered calls?

---> Maybe there were more number of incoming calls in the time bucket 11_12 and there were not enough personnel to handle most of the queries of the customers during the 11_12 time bucket

Why is it that the total number of incoming calls reached it's peak value during the time bucket 11_12 and got decreased from time bucket 12_13 onwards?

---> It is a general tendency of the customers(people) that they want their query/complaint get resolved on that particular day itself when they called the customer center; so most of the customers try to place their complaint/query

before 12 Pm so that by the end of the day their complaint gets resolved depending upon the complexity of the problem faced by the customer

Why is proportion if the monthly transfer rate is less than compared to monthly answered and abandon rate?

---> In most of the customer service centers they have the

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

Why is that one cannot provide the exact distribution of agents during the night time i.e. from 9 PM to 9 AM if the number of agents available during the night shift are already defined, so as to keep the abandon rate 10%?

---> For this particular case, Since we have only 17 agents during night we need to distribute in an non analytical way

i.e. the agents who work in 19_20, 20_21 time bucket to wait and work in 21_22 and 22_23 time buckets as well. Also agents who work during 9_10, 10_11 time bucket can be asked to work for 7_8 and 8_9 time bucket as well. he agents who work

in the time bucket 1_2, 2_3, 3_4 and 4_5 can be asked to work in time buckets 6_7, 7_8 and 8_9 so as to keep the abandon rate at 10%. Also, the company needs to consider various factors like how far is the home of the agent if he/she is made to do night shift, Is the transport facility available during the night hours from the agent's home to company and many other factors and hence the exact distribution cannot be given using an analytical approach

CONCLUSION

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

- From the previous analysis we can derive that Avg calls answered per agent is 198.6 in each time bucket

- We need to reduce the abandon rate by 30%(current) - 10% (desired) = 20% İ.e. we need to increase call answered rate by 70% (current) + 20%(change) 90%.So, we need to have 90% of the total calls to be answered so as to reduce the abandon rate to 10%

- Total avg calls incoming per day= 5130

- Avg calls answered per second= 198.6

- Answered rate=90% i.e. 0.9

- Seconds per hour=3600

- So, time required to answer 90% of the incoming calls =5130 *198.6 * 0.9/3600

- So, new total number of agents working per day is 255 divided by the number of hours an agent actually works(on a consumer call) i.e. 4.5 = 255/4.5 = 56.67 == 57

Agents working per day 254.7001826

- So, to have a 10% abandon rate we need 57 Agents working per Day

- From the assumptions given the following points were noted:-

- In a day an agent work for 9 hours → Total Agent working hours = 9 HOURS

- Out of the total 9 hours, 1.5 hours goes for lunch and coffee/tea breaks; so remaining working hours =9-1.5 = 7.5 HOURS

- Out of the remaining 7.5 hours per day an agent is occupied with consumers call for only 60% of the time i.e. 60% of 7.5

i.e. 0.6 *7.5 = 4.5.

So, an agent spends only 4.5 hours per day out of total 7.5 hours on consumer calls. An agent works 6 days a week. In a month of 30 days 6 days per week; In a month of 30 there are 4 weeks; 7 days per week means total 28 days out of which 4 days

are unplanned leave

- Days of agent on floor = (20*7)/28 =5 days. Now, total days left 28 -4=24 days. Per week there is one Sunday which is an official holiday for all workplaces around the world; So in a month of 30 there are 4 Sundays.

Now total days left for work = 24 -4 = 20 days

So, an agent is available to work for 20 days in a month of 30 days

• In a certain scenario there are calls from consumers not only during the day time but also during the night time and if there are no agents available during the night time to answer the call then it creates a bad impression on the consumer regarding the company

• Now we need to give the distribution of the total manpower available for each time bucket right from 9AM to 9 PM and then from 9 PM to 9 AM, keeping the abandon rate at 10% i.e. keeping the answered rate at 90%

• For each 100 day calls there are 30 night calls; then for 5130 day calls there will be:5130*30/100 = 1539 night calls.

• So there are 1539 night calls for a total of 5130 day calls

• So, the additional working hours keeping the answered rate at 90% will be 1539 * 198.6(avg calls answered per sec) * 0.9 /3600(total seconds in each hour) = 76.41135

• So, additional agents needed by the company to answer night calls as well be 76.41135/4.5 = 16.98 == 17

• So, we need additional 17 agents to answer the night calls as well,making the total number of agents working per day keeping the answer rate to 90% will be 57(day call answer 90%) + 17(night call answer 90%) = 74 agents.So, we need 74 Agents per day to answer the consumer calls from day as well as the night time keeping the answered rate to 90%/ Abandon rate to 10%

SUMMARY OF LEARNINGS FROM ALL ABOVE PROJECTS

Gained technical skills such as Python, Tableau, Matplotlib, Seaborn, SQL and MySQL workbench, NumPy, pandas,Machine Learning, Data Analysis. I have also gained soft skills such as problem-solving, critical thinking,communication.

I have learned how to analyze data and extract useful insights to provide recommendations for price optimization & product development decisions to maximize profitability, manpower planning to reduce call abandon rate from 30% to 10% and improve customer experience, understand how customer attributes & loan attributes influence the likelihood of default and make decisions such as denying the loan, reducing loan amount etc.,

Find business insights that can be used by teams across the business to launch a new marketing campaign, decide on features to build for the app and measuring user engagement and improving the experience altogether while helping the business grow. You have also learned how to use Advanced SQL skills to analyze the data to answer questions posed by different departments within the company and provided valuable insights that can help in improving the company's operations and understand sudden changes in key metrics.

I have generated meaningful insights from the IMDB dataset to plan every move analytically based on data and helped the company to make movie that appeals to both Indian and global audience while minimizing the risk of losses. Finally, you have analyzed previous hiring dataset of an MNC to observe trends and drawing insights out of it useful for hiring department and helped the company to improve its hiring process.

APPENDIX

DATA ANALYTICS PROCESS :

- LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :
- **https://drive.google.com/file/d/1ekHxoQ_nMV1UKEv1HLnQV-TPHx9DV3H7/view?usp=sharing**

INSTAGRAM USER ANALYTICS :

- LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :
- **https://drive.google.com/file/d/1L_ugVPgvJ4AN3osd_pVtOqvNnDtaLwUw/view?usp=sharing**

OPERATION ANALYTICS & INVESTIGATING METRIC SPIKE ANALYSIS :

- LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :
- **https://drive.google.com/file/d/1yjxGuiFVX9lGzJvfmsuvxJpNmNBP_XCk/view?usp=sharing**

HIRING PROCESS ANALYTICS :

- LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :
- **https://drive.google.com/file/d/1Hv2PkQHhlkHxWw-7f7g_lDzUfT1Wr9MX/view?usp=sharing**
- LINK FOR **SHARED STATISTICS EXCEL SHEET FILE** (ANALYSIS) ON GOOGLE DRIVE :
- **https://docs.google.com/spreadsheets/d/1yl8TrZ-ohHXQQhkXLZ1qmy7ibq7EgfwD/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true**

IMDB MOVIE ANALYSIS :

- LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :
- https://drive.google.com/file/d/1pd2gDXUw4SmrNULxRGpqI5j0GkBDVgn0/view?usp=sharing
- EXCEL IMDB ANAYSIS **GIVEN WORKSHEET** LINK :
- https://docs.google.com/spreadsheets/d/1gtjCeY6VMr7u3A3GBv78SQIKw9E5GMh_/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- **EXCEL WORK SHEET** LINK :
- https://docs.google.com/spreadsheets/d/1rYhhik6kD7e6-FOKsE-GZjYG41RDQC8I/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- **POWER POINT** LINK :
- https://docs.google.com/presentation/d/1t8AI-BOaHYHUQDv5eWIRPhGaU_7cteJC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- **LOOM VIDEO** LINK :
- https://www.loom.com/share/0445c11c9811479c8e2f3d118bb2833d?sid=91f43aa2-29af-4d6d-9828-ee6d78b31f59

- IMDB MOVIE ANALYSIS **MS- WORD DOCUMENT** LINK :
- **https://docs.google.com/document/d/1xOwHno_P1QElCHKNjqsczPsNii-cmxkU/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true**

BANK LOAN CASE STUDY :

LINK FOR **SHARED PDF REPORT** ON GOOGLE DRIVE :

- https://drive.google.com/file/d/12oyPuOVzpnzI4MTTRjDNmiylSpIPSi5B/view?usp=sharing
- **https://drive.google.com/file/d/1BjsLh13IkVxgAd8XqgC_iw_YLbKf0WWv/view?usp=sharing**

BANK LOAN CASE STUDY **POWER-POINT** LINK :

- https://docs.google.com/presentation/d/1MP6SEf2__rH2GIKnrDY3N4vuqrOixOVZ/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

DEMOCREATOR BANK LOAN CASE STUDY **VIDEO** LINK :

- https://drive.google.com/file/d/1VoFdf4W7lgmrOGaCca5DgUy-m3ry5a8i/view?usp=sharing

THE **GIVEN DATASETS** FOR ANALYSIS LINK :

File 1:

https://docs.google.com/spreadsheets/d/1SqA29nV0MNZK-gNbi0JCR7zNITNOlg_w/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

File 2 :
https://docs.google.com/spreadsheets/d/18dCjncxQ5Tst27YOO6nKTDiWj6UhmjFb/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

File 3 :

https://docs.google.com/spreadsheets/d/1_oEJTXjHwLOAU7iNdTCA4vpYrZ-aUQbV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

THE EXCEL SHEET COLUMNS TAKEN FOR ALL **QUESTIONAIRE** LINK :

The columns used to find missing values link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns used to find outliers link :

https://docs.google.com/spreadsheets/d/11NXivWjahu5285-i-83vVTD8g9CE9Oxu/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to find data imbalance link :

https://docs.google.com/spreadsheets/d/1oVteOqRRD-49Eq0_ChBve4iZUBNufCLV/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform univariate analysis link :

https://docs.google.com/spreadsheets/d/1N8ho6bKiiMOMwOGUt60p_Hj9SGPq88_L/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform segmented univariate analysis link :

https://docs.google.com/spreadsheets/d/16sDbOypfawu5Guv0hyRxp8crEOJFisYP/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the columns chosen to perform bivariate analysis link :

https://docs.google.com/spreadsheets/d/1cTCFZvgNtDc8gfoOWMKdS_ehlnsNCWKh/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to perform correlation coefficient including target variable (1) are given in the link below

https://docs.google.com/spreadsheets/d/15-QGp86IugTP0eNVSf_9fqLi1r9YWuQC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The columns taken to perform correlation coefficient including target variable (0) are given in the link below

https://docs.google.com/spreadsheets/d/1pilMPVgwnAurs6V9GU8HMX67tpJCjaQy/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

THE **OUTPUT** FOR ALL QUESTIONAIRE LINK :

Outlier output link :

https://docs.google.com/spreadsheets/d/1b9cbpOIEyhBRMTom_q0YRpNZjvd5Zu1p/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

valid/invalid data point output link :

https://docs.google.com/spreadsheets/d/1TrCWbTN1DJhCJflxaRfpLZ9-ypGvzIi3/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

counts comparing target output link :

https://docs.google.com/spreadsheets/d/10QDDrkD32XXK70ZV8w6Vbq-emor5eCMB/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The descriptive analysis for univariate analysis output link :

https://docs.google.com/spreadsheets/d/1-cJcCWAY_A7eAVPrI7gKQY7S3k-6ho1m/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

segmented analysis segments found using pivot table output link :

https://docs.google.com/spreadsheets/d/19_q79GtofixKYSdwJPeJdAQ4VDfVO1jn/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

segmented analysis segment's descriptive analysis output link :

https://docs.google.com/spreadsheets/d/1ERFXED5g68gr3Vd_fwofd2Rog3rlfiHE/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

bivariate analysis segment's done using pivot tables output link :

https://docs.google.com/spreadsheets/d/1F0Fem3IY7msm2a3nPVK7lsPB7VBMjRG4/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

the descriptive analysis for bivariate analysis output link :

https://docs.google.com/spreadsheets/d/1QXr5Pp7rNYbdqPP3zGXIkr_lf9_Ddb0P/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The coefficients found for target variable 1 output link :

https://docs.google.com/spreadsheets/d/1mkxx0A5Vz-E0xnF9Yu6W-utq0qtGXUGj/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Matrices for Target variable (1) output link :

https://docs.google.com/spreadsheets/d/1vayEb52rByP9-xnlNrS1Qzc8fsZ-YD6C/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The coefficients found for target variable 0 output link :

https://docs.google.com/spreadsheets/d/1YmOyXEuEbEVg40XHfrVaRl2wqXoAc6V9/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

The Correlation Matrices for Target variable 0 output link :

https://docs.google.com/spreadsheets/d/1fT7Ao_s8Oj6tvAdj3JQhGi4pxx_Ebz7q/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

ANALYZING THE IMPACT OF CAR FEATURES ON PRICE & PROFITABILITY :

- LINK FOR SHARED **REPORT PDF FILE** ON GOOGLE DRIVE
- **https://drive.google.com/file/d/1ammW3cHxsH2qe5cCCX1CxbpzEHvs4sIf/view?usp=sharing**
- ANALYZING THE IMPACT OF CAR FEATURES ON PROFITABILITY & PRICE **GIVEN DATASET EXCEL FILE** LINK **:**
- https://docs.google.com/spreadsheets/d/1ybx8jsOaGGchvAyHhs-1BDWn9PlxV13N/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- ANALYZING THE IMPACT OF CAR FEATURES ON PROFITABILITY & PRICE **PDF REPORT** LINK :
- https://drive.google.com/file/d/196mbAqhI67fUpCsi6ubeEbjAVQZFWv28/view?usp=sharing
- ANALYZING THE IMPACT OF CAR FEATURES ON PROFITABILITY & PRICE EXCEL FILE **OUTPUT(REPORT)** LINK :

  Kindly open All the sheets in given link below in Microsoft Excel to view outputs

- https://docs.google.com/spreadsheets/d/1HLrxtV5M_DTvM3FmR9XLtVx3Ep9xnZfd/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- ANALYZING THE IMPACT OF CAR FEATURES ON PROFITABILITY & PRICE **POWERPOINT PRESENTATION** LINK **:**
- https://docs.google.com/presentation/d/1ih9XHR613brtLSOO2UviCFoq9iDvb5UK/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- ANALYZING THE IMPACT OF CAR FEATURES ON PROFITABILITY & PRICE **SCREENREC VIDEO PRESENTATION** LINK **:**
- https://drive.google.com/file/d/1_--sX-V941m4ggKpZ_aB9T2G8mUINjQ1/view?usp=sharing

Kindly open All the sheets in given hyperlink below in Microsoft Excel to view

ANALYSIS EXCEL FILE

PRESENTATION FILE

REPORT PDF LINK

VIDEO PRESENTATION FILE

ABC CALL VOLUME TREND ANALYSIS :

- LINK FOR THE SHARED **PDF REPORT FILE** ON GOOGLE DRIVE :
- **https://drive.google.com/file/d/1mPYQtyV7VDPzbsXluDooZsA3i5LZ5h1x/view?usp=sharing**
- ABC CALL VOLUME TREND ANALYSIS **GIVEN DATASET EXCEL FILE LINK:**
- https://docs.google.com/spreadsheets/d/1UN_oGGAcFvUSSBoQQTXGpsO9QaAc_s9S/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- ABC CALL VOLUME TREND ANALYSIS **REPORT PDF** LINK **:**
- https://drive.google.com/file/d/1KMGYYbE4pR84rNfGiJv2--i2RlrIH5CH/view?usp=sharing
- ABC CALL VOLUME TREND ANALYSIS **EXCEL SHEET FILE OUTPUT (REPORT)** LINK :
- **Kindly open All the sheets in given link below in Microsoft Excel to view outputs**
- **https://docs.google.com/spreadsheets/d/1W7zdGFkINLjea6rBhUOIPVWduFiNsamn/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true**
- ABC CALL VOLUME TREND ANALYSIS **POWER-POINT PRESENTATION** LINK**:**
- https://docs.google.com/presentation/d/1GEv1ywbOIRq_DJmSDj6qeaBtMXkiQ1Ep/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true
- ABC CALL VOLUME TREND ANALYSIS **POWER - POINT PRESENTATION VIDEO** LINK :
- **https://drive.google.com/file/d/1e6tSRFAu_R2n841gynuK_kyrN8kHbdlw/view?usp=sharing**

Kindly open All the sheets in given hyper link below in Microsoft Excel to view

ABC CALL VOLUME TREND ANALYSIS **EXCEL SHEET FILE OUTPUT (REPORT) LINK :**

PRESENTATION FILES

ABC CALL VOLUME TREND ANALYSIS **REPORT PDF LINK :**

ABC CALL VOLUME TREND ANALYSIS **POWER-POINT PRESENTATION LINK :**

ABC CALL VOLUME TREND ANALYSIS **POWER - POINT PRESENTATION VIDEO LINK :**