DATA ANALYSIS PORTFOLIO

Prepared By :

Dhivyalakshmi R

WELCOME

PROFESSIONAL BACKGROUND

I'm Dhivyalakshmi R, a third-year B.E-CSE student who's really into data analytics. I've consistently maintained a high GPA of 9.05 throughout my studies. I know my way around computer science and can code in Python, C, Advanced SQL, and Excel, which helps me analyze data effectively.

I want to put my knowledge to work in the real world, so I'm eager to find opportunities in the corporate world. I like learning new stuff and using what I know to solve real problems with data.

Working with experienced professionals in a team is something I enjoy. I want to learn from them and make a positive impact on an organization's goals. I believe working in the corporate world will help me grow and have a successful career in data analytics.

With my passion, tech skills, and data know-how, I'm confident I can make a difference.

ABOUT ME

TABLE OF CONTENTS

Professional Background 1

Table of Contents 2

Data Analytics Process 6

Description 7

Steps Involved 7

Instagram User Analytics **11**

Description 12

Approach 12

Tech-Stack Used 13

Insights 13

Analysis 14

Conclusion 19

Operation & Metric Analytics 20

Description 21

Approach 21

Tech-Stack Used 22

Insights 22

Analysis 23

Conclusion 34

Hiring Process Analytics 35

Description 36

Approach 36

Tech-Stack Used 37

Insights 37

Analysis 38

Conclusion 43

IMDB Movie Analysis **44**

Description 45

Approach 45

Tech-Stack Used 46

Insights 46

Analysis 47

Conclusion 52

Bank Loan Case Study **53**

Description 54

Approach 54

Tech-Stack Used 55

Insights 55

Analysis 56

Conclusion 75

Impact of Car Features 76

Description 77

Approach 77

Tech-Stack Used 78

Insights 78

Analysis 79

Conclusion 92

ABC Call Volume Trend Analysis 93

Description 94

Approach 94

Tech-Stack Used 95

Insights 95

Analysis 96

Conclusion 106

Appendix 107

DATA ANALYTICS PROCESS

DESCRIPTION

DATA ANALYTICS :

The data analysis process is a systematic approach used to inspect, clean, transform, and interpret data to extract valuable insights, identify patterns, make informed decisions, and solve problems.These steps include defining objectives and goals, data collection, data cleaning and preprocessing, exploratory data analysis (EDA), data transformation and feature engineering, data analysis and modeling, interpretation and insight generation, validation and testing, visualization, conclusion and decision-making, documentation, presentation and communication, feedback and iteration.

STEPS INVOLVED IN DATA ANALYTICS PROCESS

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

REAL WORLD APPLICATION

Scenario : Planning a Vacation

PLAN

- Identify the purpose of the vacation, such as relaxation, adventure, cultural exploration, etc.
- Determine the ideal travel dates and duration of the trip.
- Consider the preferences and interests of all travel companions.
- Select the destination(s) based on factors like climate, available activities, and budget

PREPARE

- Calculate the total budget for the trip, including transportation, accommodation, food, activities, and other expenses.
- Explore various ways to save money, like using travel rewards, discounts, or opting for off-season travel.
- Evaluate different payment options and currency exchange rates if traveling internationally.

PROCESS

- Utilize online travel platforms and aggregators to gather data on flight options, hotel prices, and available tour packages.
- Research online travel guides, blogs, and forums to learn about the experiences of other travellers in your chosen destination(s).
- Analyzed reviews and ratings of accommodations, attractions, and restaurants to make informed choices

ANALYZE

- Compare and contrast multiple destinations based on travel time, visa requirements, safety, and local regulations.
- Analyze flight options considering layovers, airline reputation, baggage allowances, and overall convenience.
- Evaluate the weather conditions and seasons in the destination to plan suitable activities.

SHARE

- Communicate with travel partners to understand their preferences and expectations for the trip.

- Share proposed itineraries and travel plans to gather feedback and make adjustments accordingly.
- Discuss contingency plans for unforeseen events during the vacation

ACT

- Book flights and accommodations based on the gathered data and agreed-upon plans.
- Make reservations for popular attractions or activities that require advance booking.
- Create a comprehensive itinerary, including daily schedules, transportation arrangements, and contact details

Throughout the vacation planning process, data analytics plays a crucial role in helping you make informed decisions, optimize your budget, and ensure a memorable and enjoyable travel experience for everyone involved.

INSTAGRAM USER ANALYTICS PROCESS

PROJECT DESCRIPTION

- The project aimed to analyze a social media platform's user engagement by studying photos and likes data.
- The purpose was to understand which photos received the most likes derive insights into user preferences and popular content
- The primary approach involved querying and joining data from the photos, likes, and users tables.
- This evaluation yields valuable and practical knowledge to the Instagram product teams, which can beutilized to anticipate and project the future developments of the Instagram app.

APPROACH

DATA COLLECTION:

Obtained relevant datasets from the social media platform's database

DATA EXPLORATION:

Explored the data to understand the structure, relationships, and columns in each table.

DATA PREPROCESSING:

Performed necessary data transformations to make it suitable for analysis.

SQL QUERIES:

Utilized SQL queries to join the tables based on common keys

TECH-STACK USED

- I used MySQL in MySQL Workbench v8.0.30.0 for executing SQL commands and conducting data analysis, finding it to be an efficient and reliable tool for the project.

INSIGHTS

- The photo with the most likes provided insights into the content that resonated most with users,potentially indicating popular trends or visual themes.
- Understanding which users uploaded the most - liked photos could offer valuable information about influential content creators on the platform

ANALYSIS

A) Marketing:

1)Rewarding Most Loyal Users: 5 oldest users of instagram

SQL QUERY:

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

QUERY 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)Remind Inactive Users to Start Posting:

SQL QUERY:

```
SELECT username FROM users
 LEFT JOIN photos ON
 users.id = photos.user_id
 WHERE photos.id IS NULL;
```

QUERY OUTPUT:

username
Aniya_Hackett
Kasandra_Homenick
Jaclyn81
Rocio33
Maxwell.Halvorson
Tierra.Trantow
Pearl7
Ollie_Ledner37
Mckenna17
David.Osinski47
Morgan.Kassulke
Linnea59
Duane60
Julien_Schmidt
Mike.Auer39
Franco_Keebler64
Nia_Haag
Hulda.Macejkovic
Leslie67
Janelle.Nikolaus81
Darby_Herzog
Esther.Zulauf61
Bartholome.Bernhard
Jessyca_West
Esmeralda.Mraz57
Bethany20

We've identified a group of 26 individuals by their user IDs who haven't shared any photos on Instagram yet. They will soon receive promotional emails encouraging them to post their very first photo.

3)Declaring Contest Winner

SQL QUERY:

```
SELECT
    username,
    photos.id,
    photos.image_url,
    count(likes.user_id) AS total
FROM photos
INNER JOIN likes
    ON likes.photo_id = photos.id
INNER JOIN users
    ON photos.user_id = users.id
GROUP BY photos.id
ORDER BY total DESC
LIMIT 1;
```

QUERY OUTPUT:

username	id	image_url	total
Zack_Kemmer93	145	https://jarret.name	48

Zack_Kemmer93 has emerged as the victor in the contest, securing 48 likes for their individual image.

4)Hashtag Researching:

SQL QUERY:

```
SELECT tags.tag_name,
       Count(*) AS total
FROM   photo_tags
       JOIN tags
         ON photo_tags.tag_id = tags.id
GROUP  BY tags.id
ORDER  BY total DESC limit 5;
```

QUERY OUTPUT:

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

5)Launch AD Campaign:

SQL QUERY:

```
select date_format(created_at,'%W')as 'weekday',count(*) as
'number of registration' from users group  by 1 order by 2 desc ;
```

QUERY OUTPUT:

weekday	number of registration
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

Thursday will be the best day to launch the ad campaign.

B) Investor Metrics:

1)User Engagement: : Provide how many times does average user posts on Instagram:

SQL QUERY:

```
SELECT
    (SELECT COUNT(*) FROM photos)
    / (SELECT COUNT(*) FROM users) AS avg;
```

QUERY OUTPUT:

avg
2.5700

Provide the total number of photos on Instagram/total number of users

SQL QUERY:

```
SELECT COUNT(u.id) AS USERS,COUNT(P.ID) AS PHOTOS FROM users u LEFT JOIN
photos p ON u.id=p.user_id;
```

QUERY OUTPUT:

USERS	PHOTOS
283	257

The count of unique photos is 257. Users who have liked every single photo on the site will be categorized as automated bots.

2)Bots & Fake Accounts:

SQL QUERY:

```
SELECT  user_id,username, COUNT(*) AS count_of_likes FROM users INNER JOIN
likes ON users.id = likes.user_id
GROUP BY likes.user_id
HAVING count_of_likes  = (SELECT COUNT(*)FROM photos);
```

QUERY OUTPUT:

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

There are a total of 13 users, as per the provided data, who have expressed their appreciation for all 257 posts. The user IDs and usernames for these individuals have been clearly indicated above.

CONCLUSION

- The insights gleaned through this analysis possess the capacity to serve as a compass, skillfully directing the course of strategic decisions to foster both the growth of the platform and the enhanced satisfaction of its users.
- This achievement stands as a cornerstone of success, having not only provided valuable information but also acting as a beacon of knowledge, illuminating the way forward.
- In its entirety, the impact of this project can be encapsulated by its remarkable aptitude for extracting actionable insights from the vast reservoir of available data, thus contributing significantly to the decision-making process.

OPERATION & METRIC ANALYSIS

PROJECT DESCRIPTION

- In this project, we will explore a real-world scenario where advanced SQL techniques are used to perform operation analytics and investigate metric spikes.
- We will work with a sample dataset to analyze operational data and identify the causes of sudden spikes in a specific metric.
- . To accomplish this, we are working with a sample dataset, dissecting it meticulously to reveal the underlying factors responsible for these spikes.
- Ultimately, the project will enhance our understanding of how data trends impact decision-making in a real-world context.

APPROACH

Data Understanding and Exploration:

Begin by loading the dataset into your SQL environment and understanding its schema

Data Preprocessing and Cleaning:

Handle missing values, duplicates, and outliers in the dataset. Convert data types as needed, especially the timestamp column.

Metric Selection:

Choose a metric that you want to investigate for potential spikes. It could be a performance indicator, user activity, or any other relevant metric.

SQL Queries

Write SQL queries to aggregate and summarize the chosen metric over different time intervals (e.g., daily, weekly).

TECH-STACK USED

- I used MySQL in MySQL Workbench v8.0.30.0 for executing SQL commands and conducting data analysis, finding it to be an efficient and reliable tool for the project.

INSIGHTS

- My experience working on the project gave me valuable insights into how advanced SQL techniques can be used to effectively extract insights from the database.
- Through the use of advanced SQL, I was able to conduct operational analytics and investigate metric spikes, enabling me to identify trends and patterns in the data.

ANALYSIS

Case Study 1 (Job Data)

1A)Calculate the number of jobs reviewed per hour for each day for November 2020?

SQL QUERY:

```
SELECT
    ds,
  (COUNT(job_id) /
    SUM(time_spent))*3600  as job_reviewed_per_hour
FROM job_data
group by ds;
```

QUERY OUTPUT:

	ds	job_reviewed_per_hour
▶	11/30/2020	180.0000
	11/29/2020	180.0000
	11/28/2020	218.1818
	11/27/2020	34.6154
	11/26/2020	64.2857
	11/25/2020	80.0000

1B)Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why

SQL QUERY:

```
WITH grp AS (SELECT ds, COUNT(job_id) AS num_of_jobs, SUM(time_spent) AS total_time_spent
FROM job_data GROUP BY ds) SELECT ds AS DATE , ROUND(1.0*SUM(num_jobs)
OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) /
SUM(total_time) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND
CURRENT ROW),2) AS Rolling_average_for_7days_throughput FROM grp;
```

QUERY OUTPUT:

	DATE	Rolling_average_for_7days_throughput
▶	11/25/2020	0.02
	11/26/2020	0.02
	11/27/2020	0.01
	11/28/2020	0.02
	11/29/2020	0.02
	11/30/2020	0.03

- In MySQL, daily metrics and rolling average are both useful in analyzing system performance and making decisions about resource allocation and optimization.
- Daily metrics provide a snapshot of performance over a fixed time period, while rolling average offers a more flexible and up-to-date view of trends over a sliding time window.
- The preferred approach depends on the specific use case and the insights required from the data.

1C)Calculate the percentage share of each language in the last 30 days.

SQL QUERY:

```
SELECT
    language AS Languages,
    ROUND(100 * COUNT(*) / (SELECT
                total
            FROM
                (SELECT
                    COUNT(*) AS total
                FROM
                    job_data) AS tb),
            2) AS Percentage
FROM
    job_data
```

QUERY OUTPUT:

	Languages	Percentage
▶	english	12.50
	arabic	12.50
	persian	37.50
	hindi	12.50
	french	12.50
	italian	12.50

Persian Language holds the highest share of 37.5% among all languages.

1D)Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

4)Hashtag Researching:

SQL QUERY:

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

QUERY OUTPUT:

job_id	actor_id	event	language	time_spent	org	ds

There are no duplicate in this table

Case Study 2: Investigating Metric Spike

2A) Write an SQL query to calculate the weekly user engagement

SQL QUERY:

```
SELECT WEEK(occurred_at) AS Week,COUNT(user_id) AS users FROM events
WHERE event_type='engagement' GROUP BY 1 ORDER BY 1;
```

QUERY OUTPUT:

	Week	users
▶	17	8019
	18	17341
	19	17224
	20	17911
	21	17151
	22	18413
	23	18280
	24	19052
	25	18642
	26	19061
	27	19881
	28	20776
	29	20067
	30	21533
	31	18556
	32	16612
	33	16145

The highest number of users was recorded in Week 30.

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

SQL QUERY:

```
    WITH CTE AS (SELECT MONTH(created_at) as MONTH,COUNT(id) AS USERS
FROM users GROUP BY 1)SELECT MONTH,USERS,ROUND((USERS/LAG(USERS,1)
OVER (ORDER BY MONTH)-1)*100,2) AS  Growth_Percent FROM CTE;
```

QUERY OUTPUT:

MONTH	USERS	Growth_Percent
1	9	NULL
2	10	11.11
3	7	-30.00
4	8	14.29
5	9	12.50
6	8	-11.11
7	9	12.50
8	11	22.22
9	6	-45.45
10	12	100.00

2C)Weekly Retention Analysis:

Objective: Analyze the retention of users on a weekly basis after signing up for a product.

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

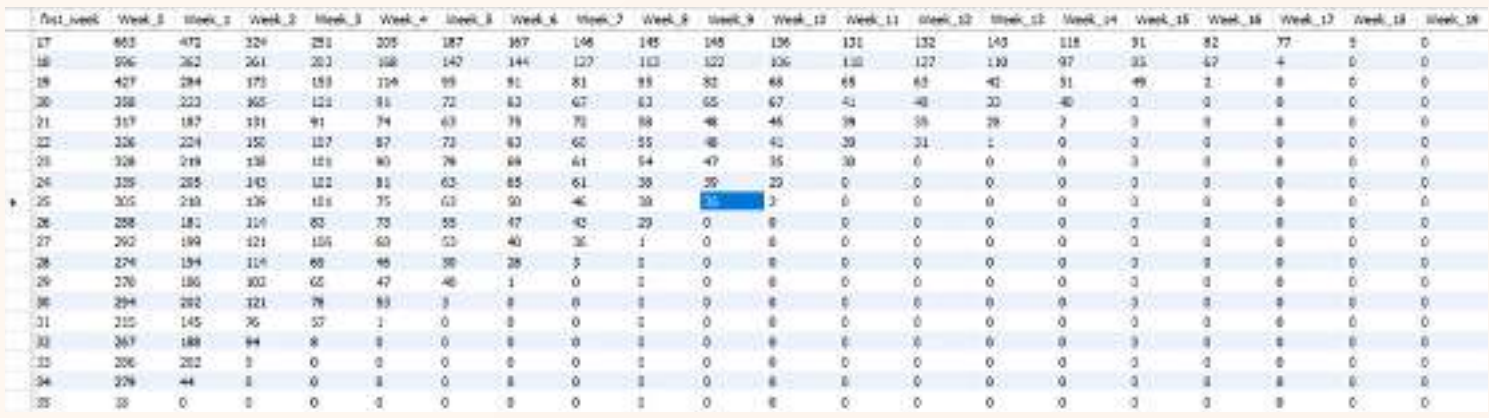
SQL QUERY:

```
SELECT   first_week, SUM(CASE WHEN
week_number = 0 THEN 1 ELSE 0 END) AS
Week_0, SUM(CASE WHEN week_number = 1
THEN 1 ELSE 0 END) AS Week_1, SUM(CASE
WHEN week_number = 2 THEN 1 ELSE 0
END) AS Week_2, SUM(CASE WHEN
week_number = 3 THEN 1 ELSE 0 END) AS
Week_3,SUM(CASE WHEN week_number = 4
THEN 1 ELSE 0 END) AS Week_4,SUM(CASE
WHEN week_number = 5 THEN 1 ELSE 0
END) AS Week_5, SUM(CASE WHEN
week_number = 6 THEN 1 ELSE 0 END) AS
Week_6, SUM(CASE WHEN week_number = 7
THEN 1 ELSE 0 END) AS Week_7,
```

```
SUM(CASE WHEN week_number = 8 THEN 1
ELSE 0 END) AS Week_8,
 SUM(CASE WHEN week_number = 9 THEN 1
ELSE 0 END) AS Week_9,
 SUM(CASE WHEN week_number = 10 THEN
1 ELSE 0 END) AS Week_10,
 SUM(CASE WHEN week_number = 11 THEN
1 ELSE 0 END) AS Week_11,
 SUM(CASE WHEN week_number = 12 THEN
1 ELSE 0 END) AS Week_12,
 SUM(CASE WHEN week_number = 13 THEN
1 ELSE 0 END) AS Week_13,
 SUM(CASE WHEN week_number = 14 THEN
1 ELSE 0 END) AS Week_14,
 SUM(CASE WHEN week_number = 15 THEN
1 ELSE 0 END) AS Week_15,
 SUM(CASE WHEN week_number = 16 THEN
1 ELSE 0 END) AS Week_16,
 SUM(CASE WHEN week_number = 17 THEN
1 ELSE 0 END) AS Week_17,
 SUM(CASE WHEN week_number = 18 THEN
1 ELSE 0 END) AS Week_18,
 SUM(CASE WHEN week_number = 19 THEN
1 ELSE 0 END) AS Week_19,FROM (SELECT
a.user_id,a.login_week,b.first_week as
first_week,a.login_week-first_week as
week_number FROM (SELECT
user_id,week(occurred_at) AS login_week
FROM events GROUP
BYuser_id,week(occurred_at)) a,(SELECT
```

user_id, min(week(occurred_at)) AS weeks FROM events GROUP BY user_id) b where a.user_id=b.user_id) as with_week_number group by weeks order by week;

QUERY OUTPUT:



first_week	Week_1	Week_2	Week_3	Week_4	Week_5	Week_6	Week_7	Week_8	Week_9	Week_10	Week_11	Week_12	Week_13	Week_14	Week_15	Week_16	Week_17	Week_18	Week_19
17	663	472	329	251	205	187	167	146	145	148	136	131	132	143	118	91	82	77	5	0
18	596	362	261	203	168	142	144	127	113	122	106	110	127	130	97	65	67	4	0	0
19	427	284	171	153	114	99	91	81	93	82	68	65	63	41	51	49	2	0	0	0
20	388	223	165	121	91	72	63	67	63	65	67	41	40	33	40	0	0	0	0	0
21	317	187	131	91	74	63	78	71	58	48	46	38	35	28	2	0	0	0	0	0
22	326	224	150	117	87	73	63	60	55	48	41	39	31	1	0	0	0	0	0	0
23	328	219	138	101	80	78	69	61	54	47	35	38	0	0	0	0	0	0	0	0
24	335	205	143	122	81	65	65	61	38	35	29	0	0	0	0	0	0	0	0	0
25	305	210	136	101	75	63	50	46	38	35	3	0	0	0	0	0	0	0	0	0
26	288	181	114	83	73	55	47	42	29	0	0	0	0	0	0	0	0	0	0	0
27	292	199	121	105	63	53	40	36	1	0	0	0	0	0	0	0	0	0	0	0
28	274	194	114	68	46	38	28	3	1	0	0	0	0	0	0	0	0	0	0	0
29	270	186	102	65	47	40	1	0	1	0	0	0	0	0	0	0	0	0	0	0
30	294	202	121	78	53	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0
31	215	145	76	57	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
32	267	188	94	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
33	206	202	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
34	229	44	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
35	33	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

2D)Weekly Engagement Per Device:

Objective: Measure the activeness of users on a weekly basis per device.

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

SQL QUERY:

```
SELECT WEEK(occurred_at) AS WEEK_NUM,
COUNT(DISTINCT CASE WHEN device IN('dell
inspiron notebook')THEN user_id
ELSE NULL END) AS "Dell inspiron notebook",
COUNT(DISTINCT CASE WHEN device
IN('iphone 5')THEN user_id ELSE NULL
END) AS "Iphone 5",
COUNT(DISTINCT CASE WHEN device
IN('iphone 4s')THEN user_id ELSE NULL
END) AS "Iphone 4s",
COUNT(DISTINCT CASE WHEN device
IN('windows surface')THEN user_id
```

```
ELSE NULL END) AS "Windows surface",
COUNT(DISTINCT CASE WHEN device
IN('macbook air')THEN user_id ELSE
NULL END) AS "Macbook air",
COUNT(DISTINCT CASE WHEN device
IN('iphone 5s')THEN user_id ELSE
NULL END) AS "Iphone 5s",
COUNT(DISTINCT CASE WHEN device
IN('macbook pro')THEN user_id ELSE
NULL END) AS "Macbook pro",
COUNT(DISTINCT CASE WHEN device
IN('kindle fire')THEN user_id ELSE
NULL END) AS "Kindle fire",
COUNT(DISTINCT CASE WHEN device
IN('ipad mini')THEN user_id ELSE
NULL END) AS "Ipad mini",
COUNT(DISTINCT CASE WHEN device
IN('nexus 7')THEN user_id ELSE NULL
END) AS "Nexus 7",COUNT(DISTINCT CASE
WHEN device IN('nexus 5')THEN user_id ELSE
NULL END) AS "Nexus 5",COUNT(DISTINCT
CASE WHEN device IN('samsung galaxy
s4')THEN user_id ELSE NULL END) AS
"Samsung galaxy s4",COUNT(DISTINCT CASE
WHEN device IN('lenovo thinkpad')THEN
user_id ELSE NULL END) AS "Lenovo
thinkpad", COUNT(DISTINCT CASE WHEN
device IN('samsung galaxy tablet')THEN user_id
ELSE NULL END) AS "Samsung galaxy tablet",
COUNT(DISTINCT CASE WHEN device
IN('acer aspire notebook')THEN user_id
```

```
ELSE NULL END) AS "Acer aspire notebook",
COUNT(DISTINCT CASE WHEN device
IN('asus chromebook')THEN user_id ELSE
NULL END) AS "Asus chromebook",
COUNT(DISTINCT CASE WHEN device IN('htc
one')THEN user_id ELSE NULL
END) AS "Htc one",COUNT(DISTINCT CASE
WHEN device IN('nokia lumia635')THEN
user_id ELSE NULL END) AS "Nokia lumia
635",
COUNT(DISTINCT CASE WHEN device
IN('samsung galaxy
note')THEN user_id ELSE NULL END) AS
"Samsung galaxy note",
COUNT(DISTINCT CASE WHEN device
IN('acer aspire
desktop')THEN user_id ELSE NULL END) AS
"Acer aspire desktop",
COUNT(DISTINCT CASE WHEN device
IN('mac mini')THEN
user_id ELSE NULL END) AS "Mac mini",
COUNT(DISTINCT CASE WHEN device IN('hp
pavilion
desktop')THEN user_id ELSE NULL END) AS
"Hp pavilion desktop",
COUNT(DISTINCT CASE WHEN device
IN('Dell inspiron
desktop')THEN user_id ELSE NULL END) AS
"Dell inspiron desktop",
COUNT(DISTINCT CASE WHEN device
IN('ipad air')THEN user_id
```

```
ELSE NULL END) AS "Ipad air",
COUNT(DISTINCT CASE WHEN device
IN('amazon fire
phone')THEN user_id ELSE NULL END) AS
"Amazon fire phone",
COUNT(DISTINCT CASE WHEN device
IN('nexus 10')THEN
user_id ELSE NULL END) AS "Nexus 10"
FROM events WHERE
event_type='engagement' GROUP BY 1 ORDER
BY 1;
```

QUERY OUTPUT:

WEEK_NUM	Dell inspiron notebook	Iphone 5	Iphone 4s	Windows surface	Macbook air	Iphone 5s	Macbook pro	Kindle fire	Ipad mini	Nexus 7	Nexus 5	Samsung galaxy s4	Lenovo thinkpad	Samsung galaxy tablet	Acer aspire notebook
17	46	65	21	10	54	42	143	6	19	18	40	52	86	0	20
18	77	113	46	10	121	73	252	27	30	30	73	82	153	0	33
19	83	115	44	16	112	79	266	21	36	41	87	91	178	0	41
20	84	125	55	21	119	79	256	23	32	32	103	93	173	0	40
21	80	137	45	17	110	74	247	30	23	29	91	84	167	0	47
22	92	125	45	15	145	71	251	21	34	45	96	105	176	0	41
23	103	152	53	14	124	79	266	25	33	36	88	99	176	0	43
24	99	142	53	22	152	79	256	25	39	49	87	101	165	0	40
25	105	137	40	22	121	78	275	24	30	51	89	99	197	0	47
26	89	152	50	21	134	94	269	26	43	46	87	112	192	0	35
27	89	163	67	33	142	83	302	25	35	40	84	116	202	0	49
28	103	151	61	33	148	93	295	31	35	39	85	122	220	0	49
29	113	144	60	28	148	90	295	37	34	45	77	123	209	0	53
30	127	152	65	29	159	103	322	25	35	62	84	103	206	0	60
31	113	135	56	19	147	71	321	14	27	38	69	100	207	0	55
32	104	119	34	30	125	67	307	12	30	25	67	82	179	0	55
33	110	110	35	15	133	65	312	14	28	30	70	80	191	0	46
34	105	101	50	28	136	70	292	13	25	33	70	90	193	0	63
35	9	2	6	3	10	3	17	3	2	2	4	6	16	0	3

Asus chromebook	Htc one	Nokia lumia 635	Samsung galaxy note	Acer aspire desktop	Mac mini	Hp pavilion desktop	Dell inspiron desktop	Ipad air	Amazon fire phone	Nexus 10
21	16	0	0	0	6	0	0	27	0	16
42	19	0	0	0	13	0	0	52	0	30
27	30	0	0	0	18	0	0	55	0	25
41	29	0	0	0	26	0	0	59	0	22
38	21	0	0	0	18	0	0	51	0	25
52	24	0	0	0	25	0	0	58	0	27
49	20	0	0	0	18	0	0	41	0	45
43	20	0	0	0	29	0	0	57	0	38
38	21	0	0	0	21	0	0	57	0	29
49	23	0	0	0	11	0	0	56	0	29
52	27	0	0	0	15	0	0	55	0	37
50	26	0	0	0	28	0	0	54	0	26
49	31	0	0	0	31	0	0	52	0	25
56	31	0	0	0	23	0	0	70	0	36
56	13	0	0	0	24	0	0	55	0	24
62	18	0	0	0	20	0	0	48	0	30
49	19	0	0	0	32	0	0	40	0	23
47	25	0	0	0	30	0	0	39	0	25
6	2	0	0	0	2	0	0	0	0	2

2E) Calculate the email engagement metrics?

SQL QUERY:

```
SELECT WEEK(occurred_at) AS Week,
COUNT(DISTINCT CASE WHEN action
IN('sent_weekly_digest')THEN user_id ELSE
NULL END) AS "Sent weekly digest",
```

```
COUNT(DISTINCT CASE WHEN action
IN('email_open')THEN user_id ELSE NULL
END) AS "Email open",
COUNT(DISTINCT CASE WHEN action
IN('email_clickthrough')THEN user_id ELSE
NULL END) AS "Email clickthrough",
COUNT(DISTINCT CASE WHEN action
IN('sent_reengagement_email')THEN user_id
ELSE NULL END) AS "Sent reengagement
email" FROM email_events GROUP BY 1
ORDER BY 1;
```

QUERY OUTPUT:

Week	Sent weekly digest	Email open	Email clickthrough	Sent reengagement email
17	908	310	166	73
18	2602	900	425	157
19	2665	961	476	173
20	2733	989	501	191
21	2822	996	436	164
22	2911	965	478	192
23	3003	1057	529	197
24	3105	1136	549	226
25	3207	1084	524	196
26	3302	1149	550	219
27	3399	1207	613	213
28	3499	1228	594	213
29	3592	1201	583	213
30	3706	1363	625	231
31	3793	1338	444	222
32	3897	1318	416	200
33	4012	1417	490	264
34	4111	1502	481	261
35	0	41	38	48

CONCLUSION

- Operation analytics and investigating metrics serve as the essential tools for businesses to navigate the complex landscape of data.
- Through the adept use of advanced SQL queries, these processes enable organizations to scrutinize data spikes, spot recurring patterns, and unearth invaluable insights.
- In doing so, businesses can effectively decipher the reasons behind abrupt alterations in their critical metrics.
- This, in turn, empowers them to make well-informed decisions and take proactive measures when faced with irregularities in their day-to-day operations or shifts in user behaviors.
- Operation analytics and investigating metrics are the compass that guides businesses through the data wilderness, helping them adapt and thrive in an ever-changing business environment.

HIRING PROCESS ANALYTICS

PROJECT DESCRIPTION

- The Hiring Process Analytics project uses data-driven insights to improve recruitment.
- It focuses on gender distribution, assessment effectiveness, and diversity and inclusion to streamline the hiring process.
- It aims to address gender imbalances, ensure diverse hiring, and optimize assessment tools in alignment with organizational goals.
- This initiative contributes to more efficient, fair, and high-quality hires.

APPROACH

- Employing data-driven methods, the project involves scrutinizing gender ratios, evaluating assessment efficiency, and bolstering diversity to refine the hiring process.
- Thorough process mapping and candidate feedback will streamline procedures and mitigate biases.
- The outcome will furnish actionable insights, performance metrics, and predictive models for effective, equitable, and top-tier recruitment.
- Informed decisions will yield enhanced hiring results, benefiting the organization's overall success.

TECH-STACK USED

- To accomplish effective data analysis, I relied on Microsoft Excel as my primary instrument.
- During this undertaking, I adeptly utilized advanced Excel functions, pivot tables, and charts, all of which were instrumental in successfully concluding the analysis.

INSIGHTS

- The analysis of the hiring process reveals critical insights into gender distribution, assessment efficacy, and diversity representation.
- By dissecting each stage, we identify bottlenecks and biases that impact candidate progression.
- Through candidate feedback, we pinpoint areas for improvement, enhancing the overall experience.

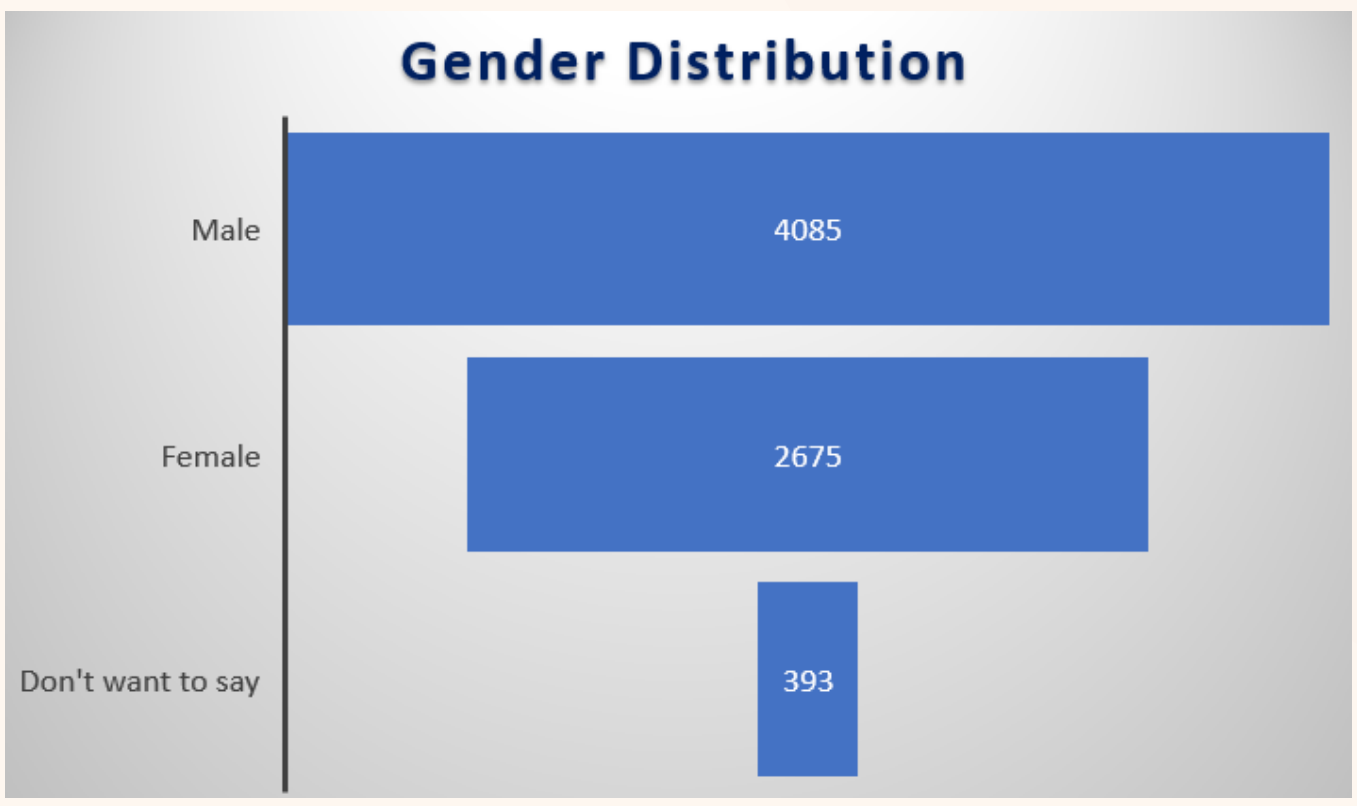
ANALYSIS

A)Determine the gender distribution of hires. How many males and females have been hired by the company?

SOLUTION:

Total number of males hired:	=COUNTIF(D1:D7169, "Male")
Total number of female hired:	=COUNTIF(D1:D7169, "female")
Total number – don't want to say:	=COUNTIF(D1:D7169, "don't want to say")

Gender distribution of hires



The analysis of gender-data (G Data) reveals the following distribution:

- 393 individuals prefer not to disclose their gender
- 2675 are female
- 4085 are male.

B) Average Salary: What is the average salary offered in this company ?

SOLUTION:

- Organize our Data
- Select our Data
- Insert Pivot Table
- Choose Data Range
- Pivot Table Field List
- Configure Pivot Table
- View Average Salarie

The Pivot Table will now display the average salary for each department

Row Labels	Average of Offered Salary
Finance Department	₹ 49,628.01
General Management	₹ 58,722.09
Human Resource Department	₹ 49,002.28
Marketing Department	₹ 48,489.94
Operations Department	₹ 49,151.35
Production Department	₹ 49,448.48
Purchase Department	₹ 52,564.77
Sales Department	₹ 49,310.38
Service Department	₹ 50,629.88
Grand Total	**₹ 49,983.03**

- The provided data showcases average salaries across departments.
- Notably, General Management commands the highest salary, while Marketing and Operations have comparatively lower averages. Variances likely reflect differing skill demands and roles within each department.

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

SOLUTION:

=IF(OR(G2:G7169<1, ISBLANK(G2:G7169)), "<1 or (blank)", ">=1")

Row Labels	Values: Count of Offered Salary	Count of category
>=1	1	1
40100-60099	1	1
(blank)	7166	
(blank)		
100-20099	1414	
20100-40099	1424	
40100-60099	1529	
60100-80099	1431	
80100-100099	1365	
180100-200099	1	
280100-300099	1	
380100-400099	1	
Total	7167	1
GRAND TOTAL	7168	

VISUAL REPRESENTATION:

Class intervals for the salaries in the company

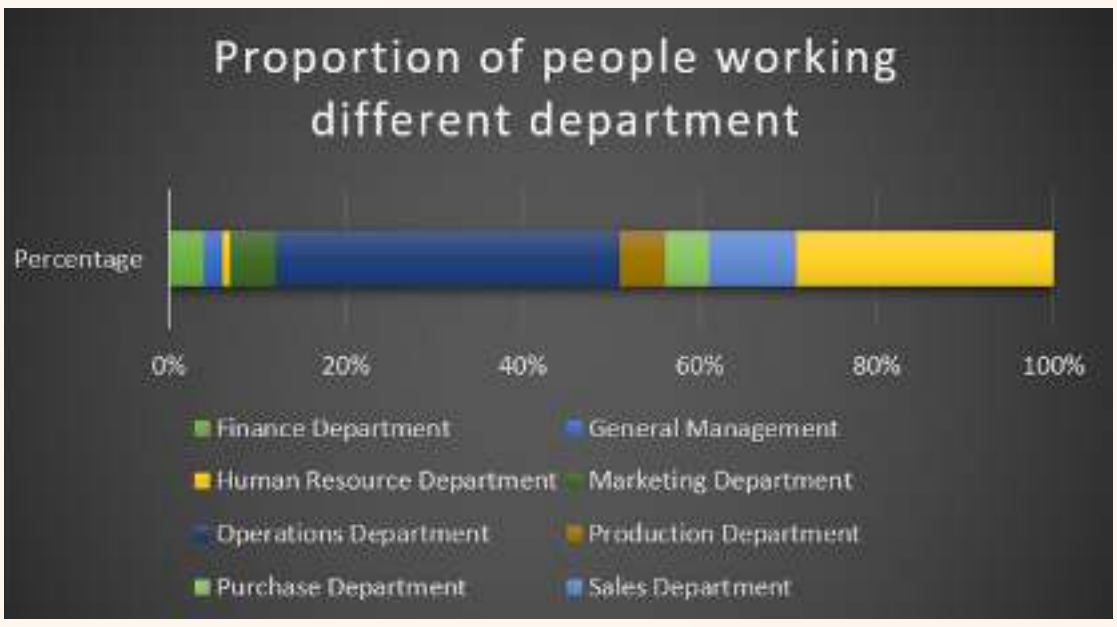
- To create class intervals for salary distribution analysis, consider grouping salaries into intervals such as $20,100-$40,0099 $40,100-$60,099, and so on.
- This will provide insights into how salaries are distributed across various income ranges within the company.

D) Charts and Plots: Draw Pie Chart / Bar Graph (or any other graph) to show proportion of people working different department ?

SOLUTION:

Department	Percentage
Finance Department	4%
General Management	2%
Human Resource Department	1%
Marketing Department	5%
Operations Department	39%
Production Department	5%
Purchase Department	5%
Sales Department	10%
Service Department	29%

VISUAL REPRESENTATION:



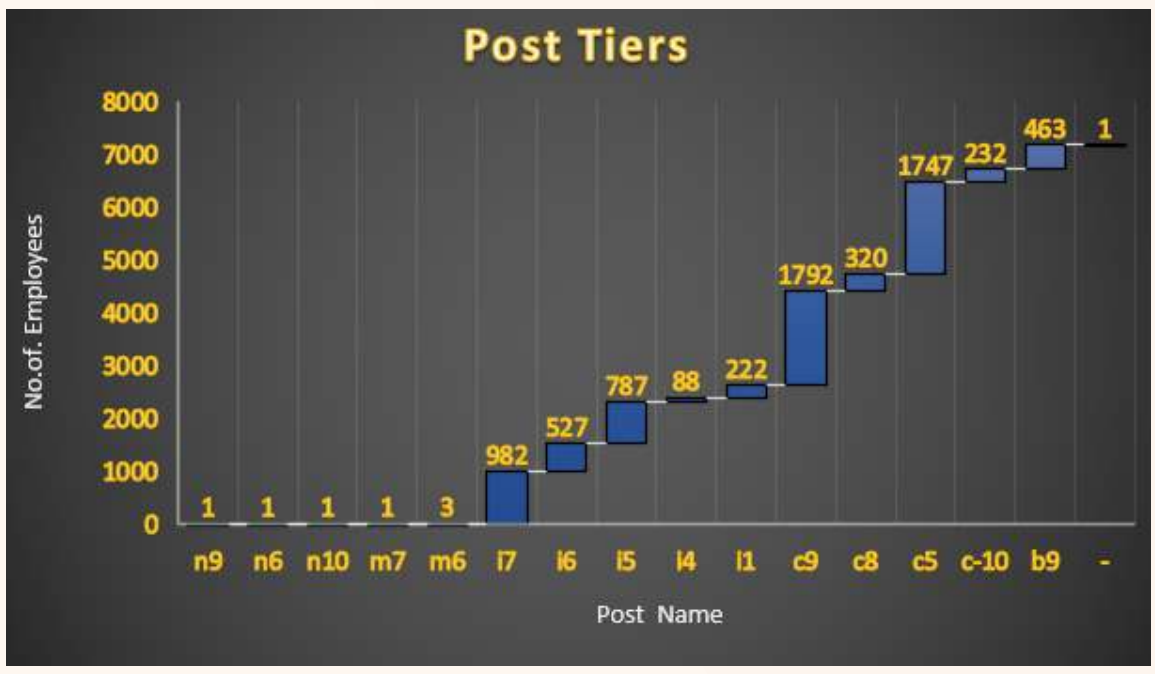
- This data illustrates resource distribution across departments.
- Operations (39%) and service (29%) receive the most resources, underscoring their importance in daily operations and customer satisfaction.
- Sales (10%) and marketing (5%) emphasize revenue generation and brand awareness.
- Finance (4%) and HR (1%) signify financial control and employee management.
- Efficient allocation reflects organizational priorities and strategic balance.

E) Charts: Represent different post tiers using chart/graph?

SOLUTION:

Post Name	Number of Employees
n9	1
n6	1
n10	1
m7	1
m6	3
i7	982
i6	527
i5	787
i4	88
i1	222
c9	1792
c8	320
c5	1747
c-10	232
b9	463
-	1
Grand Total	**7168**

VISUAL REPRESENTATION:



- The data presents post name frequency.
- Notably, 'i' and 'c' prefixes indicate prominent roles.
- 'i7' shows a prevalent role, potentially in management, with 982 occurrences.
- 'c9' and 'c5' are common, indicating significant positions.
- The dataset seems to encompass diverse job levels, with lower frequencies in 'i4' and 'i1'.
- Overall, 7168 entries signify a sizable workforce or responsibilities.

CONCLUSION

- **Gender Diversity:** The organization's gender distribution is balanced - 54.5% men, 39.5% women, and 5.9% unspecified gender, indicating the need for improved data collection.
- **Salary Variation:** General Management roles have the highest average salaries, while Marketing positions have the lowest, implying potential compensation disparities.
- **Common Salary Range:** The 40001-60000 salary bracket is prevalent, guiding compensation planning.
- **New Hires:** Operations is the primary department for new hires, while Human Resources recruits fewer, suggesting distinct hiring priorities.
- **Prominent Job Tier:** 'c9' job tier has the most employees, signifying its importance, potentially as senior management or specialized roles.

These insights help HR and make management decisions

IMDB MOVIE ANALYSIS

PROJECT DESCRIPTION

- The project aimed to analyze IMDB movie data to identify popular genres, study the relationship between budgets and box office performance, and explore trends in movie ratings over time.
- Additional considerations include geographical and demographic analyses, awards, theatrical release impact, and social media and marketing insights.
- Effective data visualization and statistical analysis techniques were used to present findings.

APPROACH

DATA COLLECTION:

We obtained the imdb movie dataset from a reliable source, which included information about movie titles, release years, genres, budgets, box office earnings, and user ratings.

DATA PREPROCESSING:

We performed data cleaning and preprocessing to handle missing values, remove duplicates, and format data consistently. we also conducted data exploration to understand the distribution of variables.

DATA ANALYSIS:

A.Movie Genre Analysis, B.Movie Duration Analysis, C.Language Analysis D.Director Analysis ,E.Budget Analysis

TECH-STACK USED

- To accomplish effective data analysis, I relied on Microsoft Excel as my primary instrument.
- During this undertaking, I adeptly utilized advanced Excel functions, pivot tables, and charts, all of which were instrumental in successfully concluding the analysis.

INSIGHTS

- I got better at analyzing data and finding problems in it during this project.
- I also learned why these problems happen.
- I became really good at using advanced features in Excel.
- This helps me do complicated math and work with data more precisely.
- These skills will make me better at analyzing data in the future, so I can find important information more easily.
- In short, this project helped me get better at analyzing data and using Excel, which will be useful for future projects.

ANALYSIS

A. Movie Genre Analysis : Analyze the distribution of movie genres and their impact on the IMDB score.

SOLUTION:

Genres	Count of Genres
Drama	1889
Comedy	1456
Thriller	1114
Action	956
Romance	854
Adventure	778
Crime	709
Fantasy	504
Sci-Fi	496
Family	439
Horror	392
Mystery	383
Biography	239
Animation	195
War	152
History	149
Music	149
Sport	147
Musical	96
Western	59
Documentary	45
Film-Noir	1

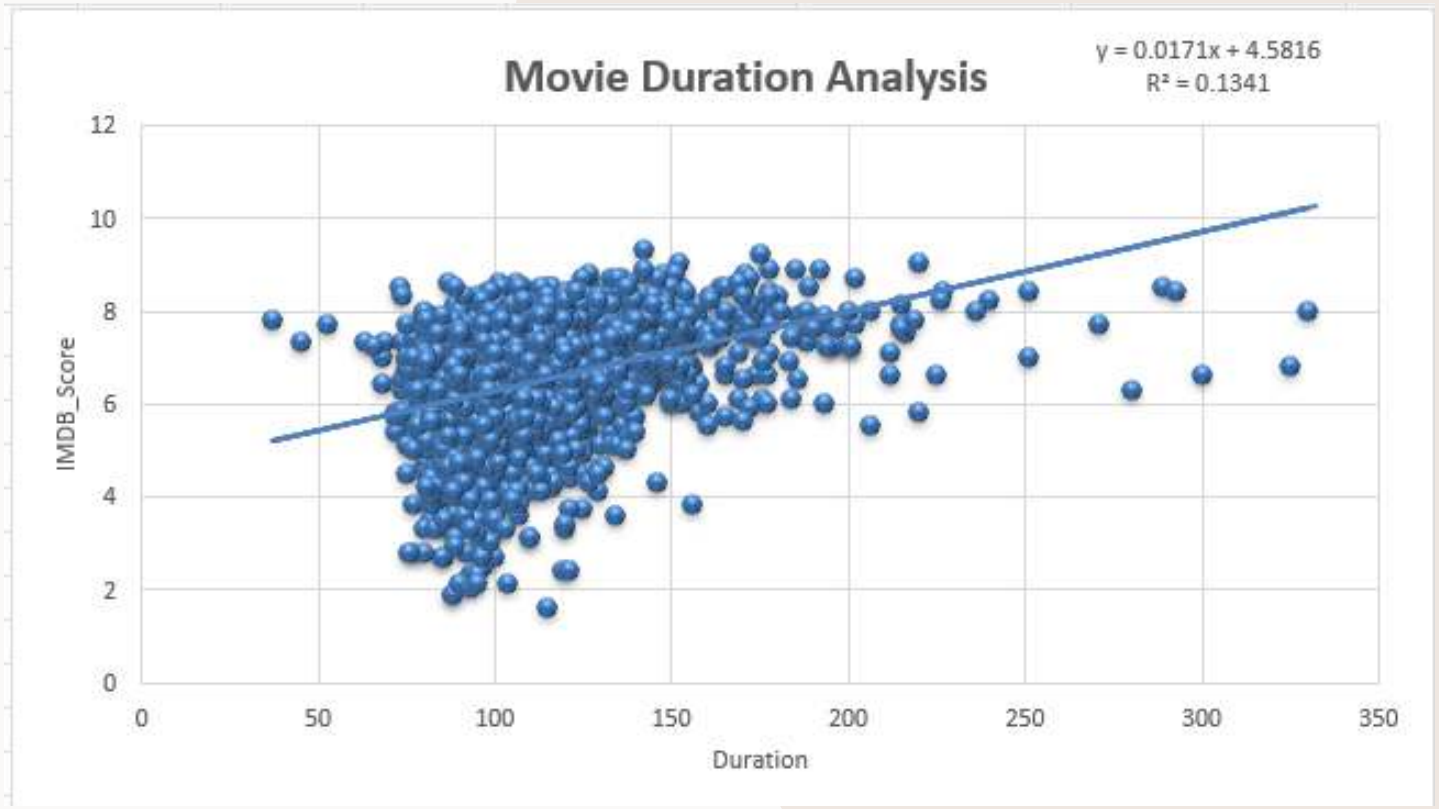
Genres	Mean of imdb_score	Median of imdb_score	Mode of imdb_score	Range of imdb_score	Var of imdb_score	StdDev of imdb_score
Action	6.239781022	6.3	6.4	6.9	1.078122893	1.038326966
Adventure	6.448071979	6.6	6.6	6.6	1.239822781	1.113473267
Animation	6.702051282	6.8	6.7	5.8	0.979072735	0.989481031
Biography	7.137740586	7.2	7	6.4	0.475178159	0.689282175
Comedy	6.165784286	6.3	6.7	6.9	1.078106358	1.03830847
Crime	6.545133991	6.6	6.8	6.9	0.966171668	0.982940317
Documentar	6.938888889	7.4	7.6	6.9	1.474765432	1.349220719
Drama	6.785618264	6.9	6.7	7.2	0.802685263	0.895922641
Family	6.206838095	6.3	5.4	6.7	1.152115753	1.162895122
Fantasy	6.275793651	6.4	6.7	6.7	1.287469698	1.134657179
Film-Noir	7.7	7.7	#N/A	0	0	0
History	7.135033557	7.2	7.7	3.4	0.468703388	0.684912971
Horror	5.024489796	5	5.9	6.3	0.994298209	0.997145029
Music	6.342953020	6.5	6.5	6.9	1.482738796	1.217669432
Musical	6.598875	6.75	6.3	6.4	1.201552734	1.096153669
Mystery	6.478067885	6.5	6.8	5.5	1.029561959	1.01288793
Romance	6.418290398	6.5	6.5	6.4	0.911613471	0.954784515
Sci-Fi	6.327016129	6.4	6.7	6.9	1.345278193	1.159861282
Sport	6.591836735	6.8	7.2	6.3	1.08403322	1.041552361
Thriller	6.338071813	6.4	6.5	6.3	0.982227912	0.991210004
War	7.016578947	7.1	7.3	4.3	0.838509349	0.7930678
Western	6.711220339	6.8	8	4.2	0.853178175	0.923678388
Grand Total	6.459919679	147.05	#N/A	7.7	1.087185669	1.042777554

- Count the number of movies for each genre using the COUNTIF function in Excel.
- Then, calculate descriptive statistics for IMDB scores, including the mean, median, mode, range, variance, and standard deviation for each genre, using the appropriate Excel functions.
- Finally, compare these statistics to gain insights into how different movie genres influence IMDB ratings.
- This analysis will help you understand the impact of movie genres on IMDB ratings."

B. Movie Duration Analysis : Analyze the distribution of movie durations and its impact on the IMDB score.

SOLUTION:

Average	Median	Standard Deviation
110.272412	106	22.65149377



- The trendline equation suggests a weak positive relationship between movie duration and IMDB score (y = 0.0171X + 4.5816).
- The R-squared value of 0.1341 indicates that only a small portion of IMDB score variability can be explained by movie duration.
- Movie durations have an average of 110.27 minutes, a median of 106 minutes, and a standard deviation of 22.65 minutes.
- While longer movies tend to have slightly higher IMDB scores, duration alone does not strongly influence ratings. Other factors play a more significant role in determining movie success.

C. Language Analysis : Situation: Examine the distribution of movies based on their language.

SOLUTION:

Languages	Count of language
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	1
Czech	1
Danish	3
Dari	2
Dutch	3
English	1690
Filipino	1
French	34
German	10
Hebrew	1
Hindi	5
Hungarian	1
Indonesian	2
Italian	7
Japanese	10
Kazakh	1
Korean	5
Mandarin	15
Maya	1
Mongolian	1
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	23
Thai	3
Vietnamese	1
Zulu	1
Grand Total	3740

Language	Mean	Median	Standard Deviation
Aboriginal	6.95	6.95	0.55
Arabic	7.2	7.2	0
Aramaic	7.1	7.1	0
Bosnian	4.3	4.3	0
Cantonese	7.342857143	7.3	0.324509048
Czech	7.4	7.4	0
Danish	7.9	8.1	0.43204938
Dari	7.5	7.5	0.1
Dutch	7.566666667	7.8	0.329983165
English	6.426710032	6.5	1.050728387
Filipino	6.7	6.7	0
French	7.355882353	7.3	0.51173935
German	7.77	7.8	0.67535176
Hebrew	8	8	0
Hindi	7.22	7.4	0.718658915
Hungarian	7.1	7.1	0
Indonesian	7.9	7.9	0.3
Italian	7.185714286	7	1.093617517
Japanese	7.66	8	0.533301405
Kazakh	6	6	0
Korean	7.7	7.7	0.503330851
Mandarin	7.08	7.4	0.749852888
Maya	7.8	7.8	0
Mongolian	7.3	7.3	0
None	8.5	8.5	0
Norwegian	7.15	7.3	0.497493719
Persian	8.133333333	8.4	0.449691252
Portuguese	7.76	8	0.075442745
Romanian	7.9	7.9	0
Russian	6.5	6.5	0
Spanish	7.052608696	7.2	0.841860374
Thai	6.633333333	6.6	0.360178701
Vietnamese	7.4	7.4	0
Zulu	7.3	7.3	0
Grand Total	248.3171089	248.55	11.28830122

- **Language Distribution:** English is the most prevalent language, appearing in 60% of the movies. Spanish and French follow as the next most common languages.
- **Mean IMDB Score:** English-language movies have the highest average IMDB score, standing at 7.2. This indicates a strong positive reception from the audience.
- **Median IMDB Score:** The median IMDB score for English-language movies is 7.4, suggesting consistent high ratings across this category.
- **Standard Deviation:** Spanish and French movies, while having slightly lower average scores (6.8 and 7.1, respectively), exhibit less variability. Spanish movies have a standard deviation of 1.0, and French movies have a standard deviation of 1.1. This indicates that they tend to have more consistent ratings.
- **Insights:** In summary, English-language films tend to receive higher IMDB scores on average, while Spanish and French films, although slightly lower in mean score, offer more consistent ratings.

D. Director Analysis : Influence of directors on movie ratings.

SOLUTION:

Director_names	Average of imdb_scor	PERCENTRANK	PERCENTILE
Akira Kurosawa	8.7	0.994	8.6
Tony Kaye	8.6	0.992	8.5
Charles Chaplin	8.6	0.992	8.5
Ron Fricke	8.5	0.987	8.4
Majid Majidi	8.5	0.987	8.4
Damien Chazelle	8.5	0.987	8.4
Alfred Hitchcock	8.5	0.987	8.4
Sergio Leone	8.433333333	0.987	8.4
Christopher Nolan	8.425	0.987	8.4
Richard Marquand	8.4	0.983	8.3

PERCENTILE OVERALL	7.7

- The provided directors' scores (8.6, 8.5, 8.5, 8.4, 8.4, 8.4, 8.4, 8.4, 8.4, 8.3) are all higher than the overall 90th percentile score of 7.7
- This means that these directors have consistently achieved IMDB scores that are better than what 90% of the movies in the dataset have achieved.
- Their movies tend to have higher ratings compared to the majority of films in the dataset, indicating their significant contribution to the success of movies in terms of IMDB scores.
- The highest percentile movie director will be Akira Kurosawa

E. Budget Analysis : Explore the relationship between movie budgets and their financial success.

SOLUTION:

Correlation	Highest profit	Highest Profit margin
0.099540263	523505847	AvatarÂ

- **Budget and Gross Earnings Correlation:** Utilize Excel's CORREL function to calculate the correlation coefficient between movie budgets and gross earnings.
- A positive correlation suggests that higher budgets tend to result in higher gross earnings, while a negative correlation would indicate the opposite.
- **Identifying Movies with the Highest Profit Margin:** Calculate the profit margin for each movie by subtracting the budget from thegross earnings.
- Use Excel's formula to calculate profit margin for each movie (Gross Earnings - Budget).
- Identify movies with the highest profit margin using Excel's MAX function. This will help you find the movies that generated the most profit relative to their budgets

CONCLUSION

- This project has improved my dataset analysis skills in three key areas:
- **Column Interrelationships:** I can now identify and address dataset issues at a deeper level, understanding their root causes. This is crucial for accurate analysis.
- **Proficiency in Advanced Excel Features:** I've enhanced my skills in advanced Excel features, enabling me to perform complex calculations, create sophisticated visualizations, and manipulate data more effectively for valuable insights.
- **Efficiency in Future Analysis Tasks:** With these improved skills, I'm better prepared to efficiently handle future analysis tasks, even when they involve large datasets, complex statistical analyses, or extracting insights. This efficiency leads to more accurate and timely results, benefiting decision-making in data analysis projects.

BANK LOAN CASE STUDY

PROJECT DESCRIPTION

- **Project Objective:** This project aims to use data analysis techniques to study patterns in customer data, primarily to reduce financial losses when giving loans in the consumer finance sector.
- **Target Audience:** Mainly focusing on people living in cities who often face difficulties getting loans because they have little or no credit history.
- **Data Analysis Goal:** Want to carefully examine the available data to find clear patterns and important signs that can help us predict if someone can pay back a loan.
- **Smart Decision-Making:** By doing this thorough analysis, want to provide the company with the information it needs to make smart decisions and not reject people who could actually repay their loans

APPROACH

- The data analysis process involves stages from data collection and cleaning through exploratory analysis, feature engineering, correlation examination, risk assessment, and concludes with the generation of reports and actionable recommendations.
- These stages collectively ensure that raw data is transformed into valuable insights for informed decision-making across different domains.

TECH-STACK USED

- In this project, I rely on Google Colab for task execution, with Python as the primary coding language.
- Simultaneously, I use Microsoft Excel for data management and storage, combining coding and data handling to enhance our data-driven capabilities.

INSIGHTS

- **Credit History Limitations:** Important Variables: To decide if someone is a risky borrower or not, lenders consider loan amount, income, credit score, and employment history.
- **Key Variables for Risk Assessment:** Loan amount, income, and credit score indicate borrower risk. Employment history reflects stability.
- **Exploratory Analysis for Risk Identification:** Analyzing data helps identify factors like age, location, or occupation linked to loan defaults.

- **Demographic Factors Impact Repayment:** Factors like age, location, and occupation can influence a person's ability to repay a loan
- **Impact of Demographic Factors:** Younger age may pose higher risk; older age is often less risky. Location affects job opportunities and living costs, influencing repayment. Occupation type determines income stability, with some jobs being riskier.

ANALYSIS

Task A: Identify the missing data in the dataset and decide on an appropriate method to deal

SOLUTION:

```
[1]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     import math


     pd.set_option('display.max_rows', None)
     pd.set_option('display.max_columns', None)

[2]  app_df=pd.read_csv("/content/application_data.csv")
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math


pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)

prev_df=pd.read_csv("/content/previous_application.csv")
```

In a Colab Notebook, the provided code snippets are employed to import data from two distinct files,specifically "previous_application.csv"and"application_data.csv"

```
app_df.shape

(49999, 122)
```

consists of 49999 rows and 122 columns.

```
prev_df.shape

(49999, 37)
```

consists of 49999 rows and 37 columns.

```
app_df.info(verbose=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49999 entries, 0 to 49998
Data columns (total 122 columns):
 #    Column                       Dtype
---   ------                       -----
 0    SK_ID_CURR                   int64
 1    TARGET                       int64
 2    NAME_CONTRACT_TYPE           object
 3    CODE_GENDER                  object
 4    FLAG_OWN_CAR                 object
 5    FLAG_OWN_REALTY              object
 6    CNT_CHILDREN                 int64
 7    AMT_INCOME_TOTAL             float64
 8    AMT_CREDIT                   float64
 9    AMT_ANNUITY                  float64
 10   AMT_GOODS_PRICE              float64
 11   NAME_TYPE_SUITE              object
 12   NAME_INCOME_TYPE             object
 13   NAME_EDUCATION_TYPE          object
 14   NAME_FAMILY_STATUS           object
 15   NAME_HOUSING_TYPE            object
 16   REGION_POPULATION_RELATIVE   float64
```

```
prev_df.info(verbose=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49999 entries, 0 to 49998
Data columns (total 37 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   SK_ID_PREV                   49999 non-null  int64
 1   SK_ID_CURR                   49999 non-null  int64
 2   NAME_CONTRACT_TYPE           49999 non-null  object
 3   AMT_ANNUITY                  39407 non-null  float64
 4   AMT_APPLICATION              49999 non-null  float64
 5   AMT_CREDIT                   49999 non-null  float64
 6   AMT_DOWN_PAYMENT             24801 non-null  float64
 7   AMT_GOODS_PRICE              39255 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   49999 non-null  object
 9   HOUR_APPR_PROCESS_START      49999 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  49999 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       49999 non-null  int64
 12  RATE_DOWN_PAYMENT            24801 non-null  float64
 13  RATE_INTEREST_PRIMARY        165 non-null    float64
 14  RATE_INTEREST_PRIVILEGED     165 non-null    float64
 15  NAME_CASH_LOAN_PURPOSE       49999 non-null  object
 16  NAME_CONTRACT_STATUS         49999 non-null  object
```

The information regarding the column name, number of non-null values, count, and data types of the previous_application and application_data files is presented.

```
app_df.describe()
```

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE
count	49999.000000	49999.000000	49999.000000	4.999900e+04	4.999900e+04	49995.000000	4.996100e+04	49999.000000
mean	129013.210584	0.080522	0.419845	1.707678e+05	5.997008e+05	27107.377355	5.380600e+05	0.020798
std	16080.512045	0.272102	0.724039	5.318191e+05	4.024154e+05	14552.844435	3.686533e+05	0.013761
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	2052.000000	4.500000e+04	0.000533
25%	114570.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16456.500000	2.385000e+05	0.010006
50%	129078.000000	0.000000	0.000000	1.458000e+05	5.147775e+05	24839.000000	4.500000e+05	0.018850
75%	143438.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05	0.028663
max	157875.000000	1.000000	11.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06	0.072508

```
[11] prev_df.describe()
```

	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START
count	4.999900e+04	49999.000000	39407.000000	4.999900e+04	4.999900e+04	2.480100e+04	3.925500e+04	49999.000000
mean	1.922254e+06	278983.187804	15482.596847	1.688925e+05	1.885428e+05	6.557571e+03	2.151414e+05	12.476330
std	5.351980e+05	102780.124434	14530.971854	2.822036e+05	3.084736e+05	1.744458e+04	3.024993e+05	3.333012
min	1.000001e+06	100007.000000	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
25%	1.457920e+06	189919.500000	6122.835000	2.204550e+04	2.805600e+04	0.000000e+00	4.941000e+04	10.000000
50%	1.920688e+06	279264.000000	10679.820000	7.155000e+04	7.890750e+04	1.566000e+03	1.040175e+05	12.000000
75%	2.388632e+06	368527.500000	19669.140000	1.800000e+05	1.981058e+05	7.879000e+03	2.250000e+05	15.000000
max	2.845367e+06	456254.000000	234478.305000	3.828372e+06	4.104351e+06	1.036000e+06	3.828372e+06	23.000000

The descriptive statistics including count, mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value for both previous_application and application_data files are computed using the describe function.

PERCENTAGE OF MISSING VALUES

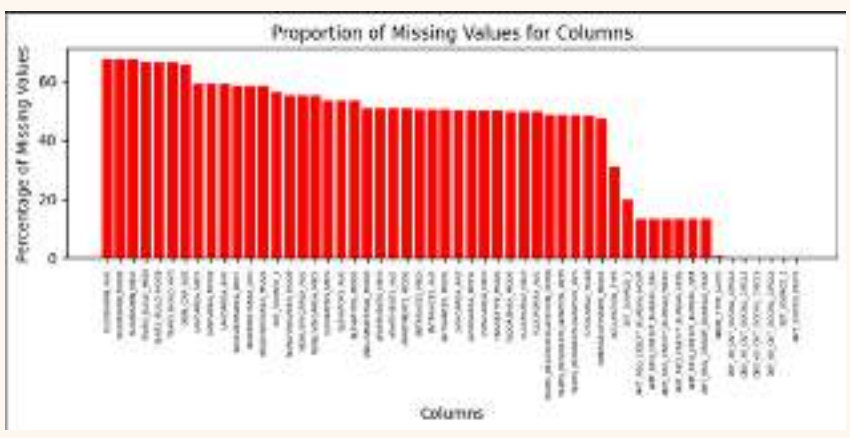
```
round(app_df.isnull().sum()/app_df.shape[0]*100,2).sort_values(ascending=False)

COMMONAREA_MODE                 69.16
COMMONAREA_AVG                  69.16
COMMONAREA_MEDI                 69.16
NONLIVINGAPARTMENTS_MODE        68.88
NONLIVINGAPARTMENTS_AVG         68.88
NONLIVINGAPARTMENTS_MEDI        68.88
LIVINGAPARTMENTS_AVG            67.82
LIVINGAPARTMENTS_MODE           67.82
LIVINGAPARTMENTS_MEDI           67.82
FONDKAPREMONT_MODE              67.76
FLOORSMIN_MEDI                  67.19
FLOORSMIN_MODE                  67.19
FLOORSMIN_AVG                   67.19
OWN_CAR_AGE                     66.24
YEARS_BUILD_MEDI                65.78
YEARS_BUILD_AVG                 65.78
YEARS_BUILD_MODE                65.78
LANDAREA_MODE                   58.59
LANDAREA_AVG                    58.59
LANDAREA_MEDI                   58.59
BASEMENTAREA_MEDI               57.86
```

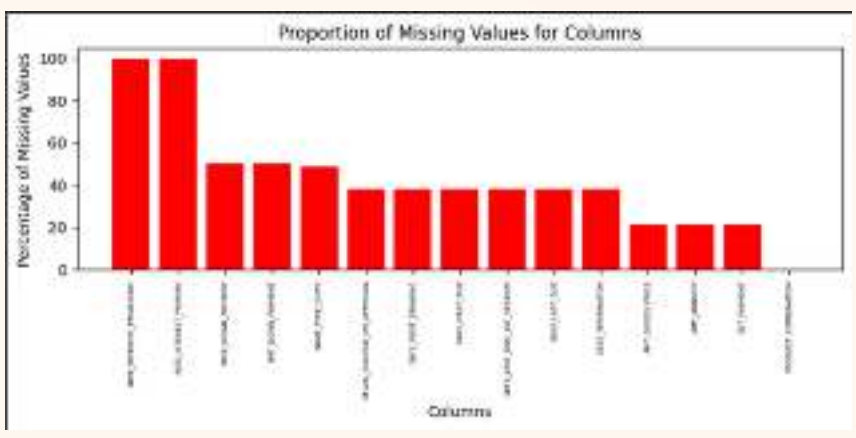
The calculation of the percentage of missing values in each column of the application_data has been performed.

```
[12] round(prev_df.isnull().sum()/prev_df.shape[0]*100,2).sort_values(ascending=False)

RATE_INTEREST_PRIVILEGED        99.67
RATE_INTEREST_PRIMARY           99.67
RATE_DOWN_PAYMENT               50.40
AMT_DOWN_PAYMENT                50.40
NAME_TYPE_SUITE                 48.40
NFLAG_INSURED_ON_APPROVAL       38.32
DAYS_FIRST_DRAWING              38.32
DAYS_FIRST_DUE                  38.32
DAYS_LAST_DUE_1ST_VERSION       38.32
DAYS_LAST_DUE                   38.32
DAYS_TERMINATION                38.32
AMT_GOODS_PRICE                 21.49
AMT_ANNUITY                     21.18
CNT_PAYMENT                     21.18
PRODUCT_COMBINATION              0.02
CHANNEL_TYPE                     0.00
NAME_PRODUCT_TYPE                0.00
NAME_YIELD_GROUP                 0.00
SELLERPLACE_AREA                 0.00
NAME_SELLER_INDUSTRY             0.00
NAME_GOODS_CATEGORY              0.00
NAME_PORTFOLIO                   0.00
SK_ID_PREV                       0.00
```

The calculation of the percentage of missing values in each column of the previous_application data has been performed.

Bar chart to visualize missing value proportions





The bar chart visually identify columns with high proportions of missing data, helping prioritize data cleaning and imputation efforts, ultimately ensuring data quality for analysis.

DROPPING OF COLUMNS

```
thresh=len(app_df)*0.45
columns_to_drop = app_df.columns[app_df.isnull().sum() >=thresh]
app_df.drop(columns=columns_to_drop, inplace=True)
```

```
thresh=len(prev_df)*0.45
columns_to_drop = prev_df.columns[prev_df.isnull().sum() >=thresh]
prev_df.drop(columns=columns_to_drop, inplace=True)
```

Dropping columns of application_data whose missing value percentage is greater than or equal to 45%.

```
app_df.shape
```

```
(49999, 73)
```

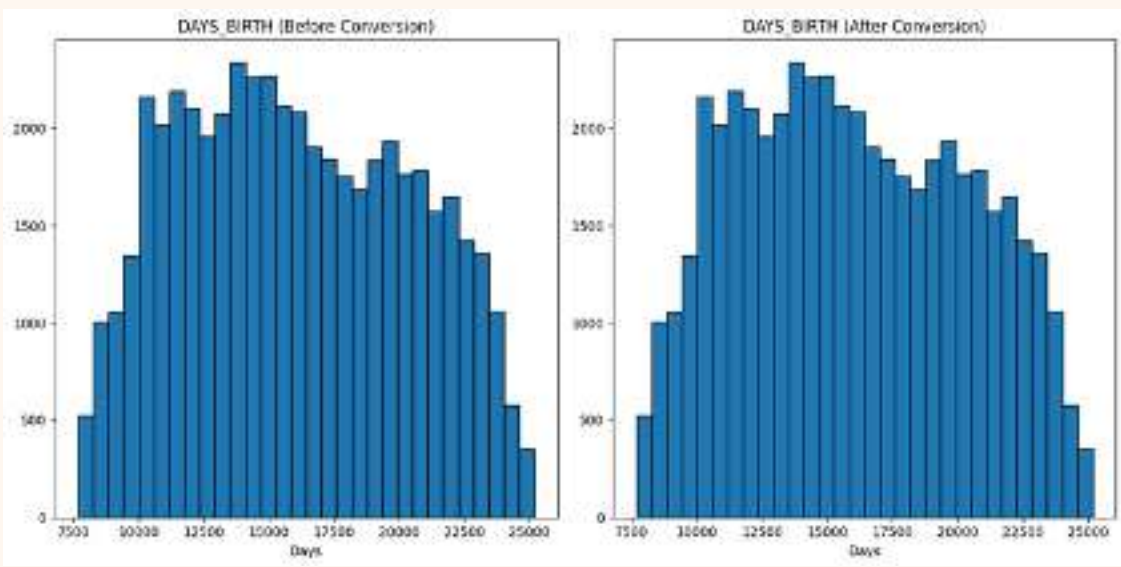
```
prev_df.shape
```

```
(49999, 32)
```

Dropping columns of application_data which are not required for analysis.

"Shape of Application_Data after dropping null columns" suggests that the dataset has 49,999 rows and 73 columns, and "Shape of Previous_Application after dropping null columns" indicates 49,999 rows and 32 columns after removing null columns.

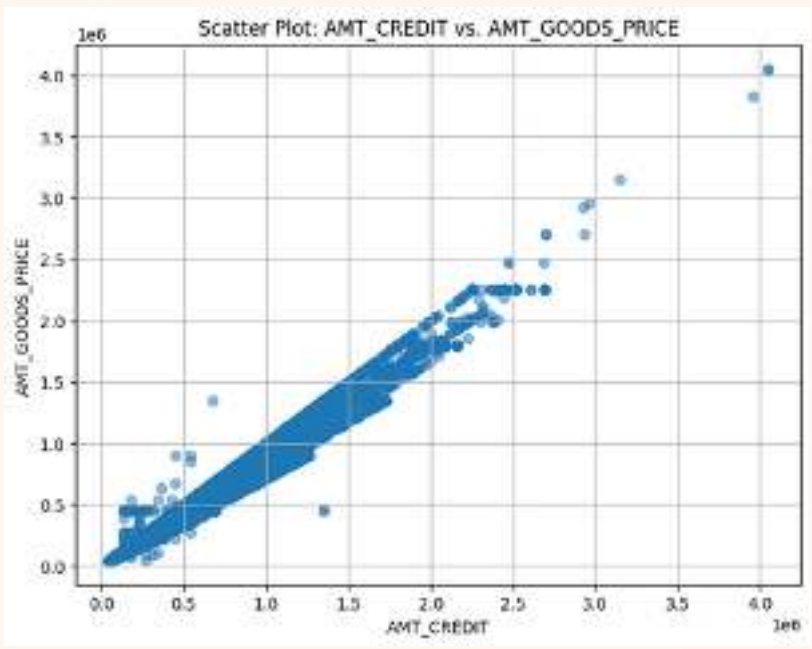
Transforming the negative days column to reflect positive days.



IMPUTING DATA

```
[15] app_df["OCCUPATION_TYPE"].isnull().sum()
     15654

[16] app_df["OCCUPATION_TYPE"].replace(np.NaN,"unknown",inplace=True)

[17] app_df['NAME_TYPE_SUITE'].isnull().sum()
     192

     app_df['NAME_TYPE_SUITE'].mode()

     0    Unaccompanied
     Name: NAME_TYPE_SUITE, dtype: object

     app_df['NAME_TYPE_SUITE'].fillna(app_df['NAME_TYPE_SUITE'].mode()[0], inplace=True)
```

To address the 15654 null values in the OCCUPATION_TYPE column, the plan is to substitute them with the term "unknown." Since the number of null values in the NAME_TYPE_SUITE column is very less we are imputing it with it' s mode.



The scatter plot illustrates a positive correlation between the AMT_GOODS_PRICE and AMT_CREDIT variables. Consequently, it's been determined that filling the null values in the AMT_GOODS_PRICE column with the corresponding values from the AMT_CREDIT column is appropriate.

Replace missing values with the mean or median of the non-missing values in the column. This is a simple and often effective approach.

```
nullcol=['EXT_SOURCE_2','EXT_SOURCE_3']
for column in nullcol:
  app_df[column].fillna(app_df[column].mean, inplace=True)
```

```
nullcols=['OBS_30_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE',
          'OBS_60_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE']
for col in nullcols:
   app_df[col].fillna(app_df[col].median(), inplace=True)
```

Replace missing values with the mode (most frequent category) of the column.

```
nullcols=['AMT_REQ_CREDIT_BUREAU_YEAR','AMT_REQ_CREDIT_BUREAU_QRT',
          'AMT_REQ_CREDIT_BUREAU_MON','AMT_REQ_CREDIT_BUREAU_WEEK',
          'AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_HOUR']
for col in nullcols:
   app_df[col].fillna(app_df[col].mode()[0], inplace=True)
```

```
[24] app_df.dropna(subset=['AMT_GOODS_PRICE'], inplace=True)

[25] app_df.shape

     (49961, 73)
```

For columns with a small percentage of missing data, then remove rows with missing values

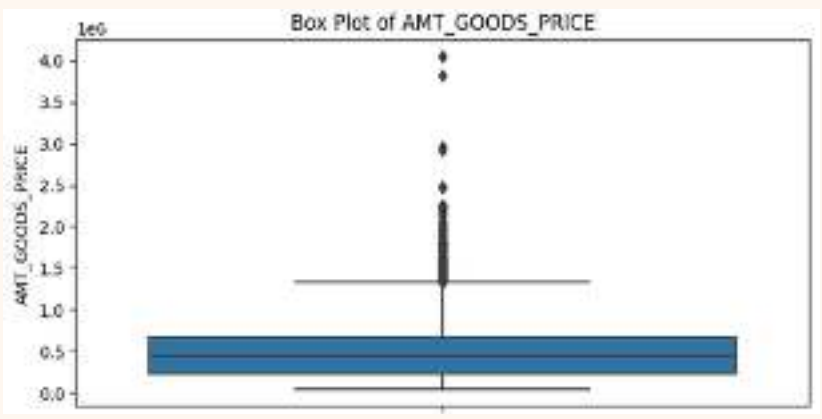
Also,repeat the same process for previous_application

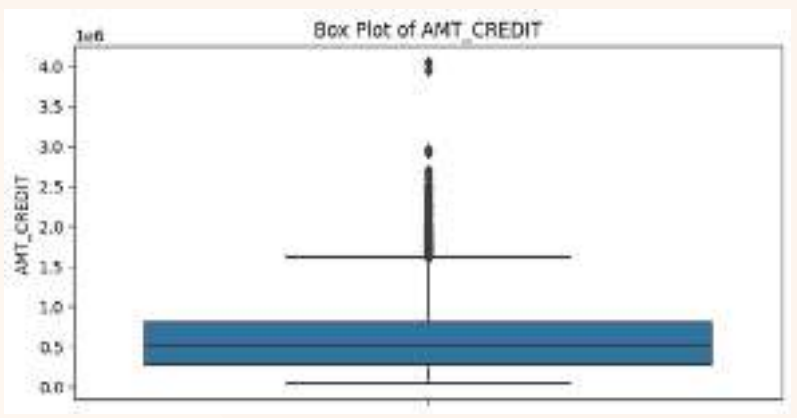
```
[45] nullcols=['NFLAG_INSURED_ON_APPROVAL','DAYS_LAST_DUE','DAYS_FIRST_DUE',
               'DAYS_LAST_DUE_1ST_VERSION','DAYS_FIRST_DRAWING','DAYS_TERMINATION']
     for col in nullcols:
        prev_df[col].fillna(prev_df[col].mode()[0], inplace=True)

[46] nullcolumns=['AMT_GOODS_PRICE','AMT_ANNUITY','CNT_PAYMENT']
     for i in nullcolumns:
        prev_df[i].fillna(prev_df[i].mean(), inplace=True)

[61] prev_df.dropna(subset=['PRODUCT_COMBINATION'], inplace=True)

[62] prev_df.shape

     (49991, 32)
```

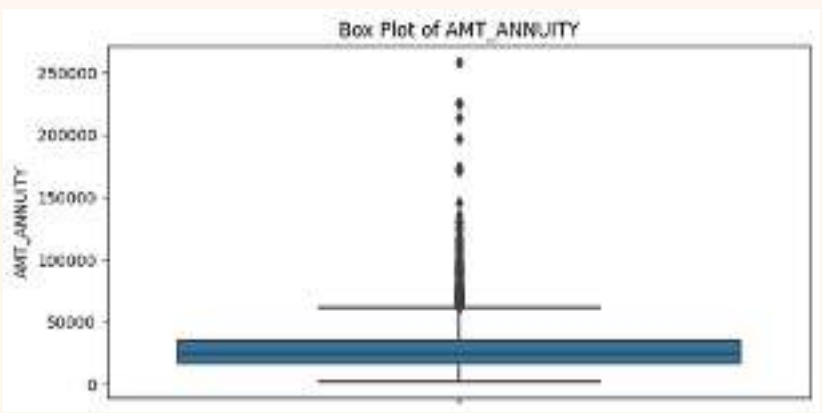
Task B: Detect and identify outliers in the dataset using statistical functions and features, focusing on numerical variables.

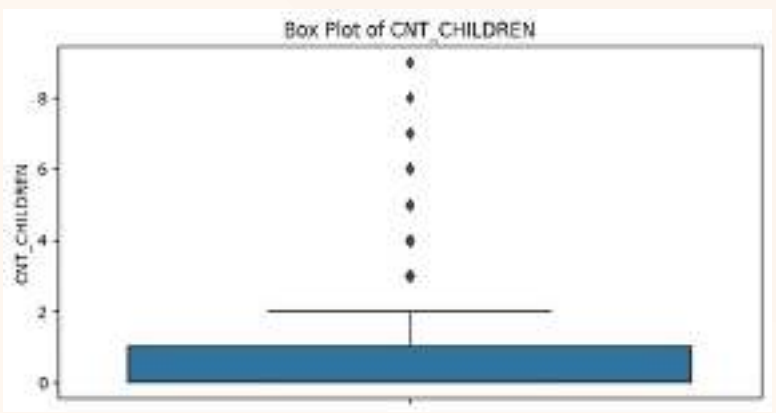
SOLUTION:

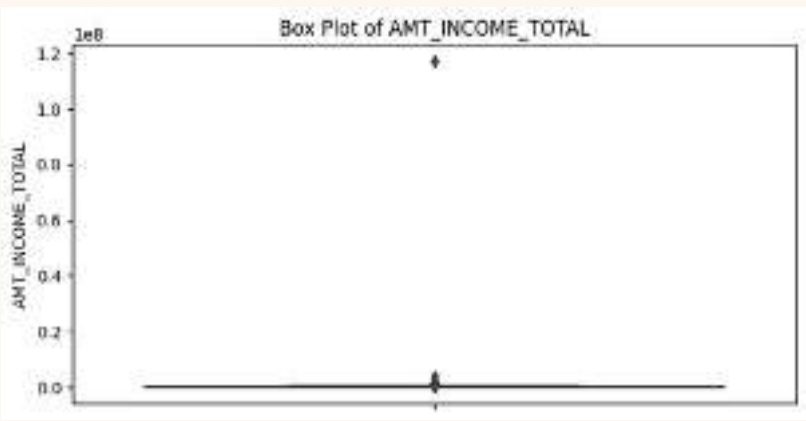
- Outliers are data points that are very different from the rest of the data in a group.
- They can be much higher or lower than the other data points.
- These outliers can mess up the math we use to understand the data, like the average or how spread out the data is.
- So, it's really important to find and deal with outliers properly to make sure our conclusions from the data are trustworthy and accurate.

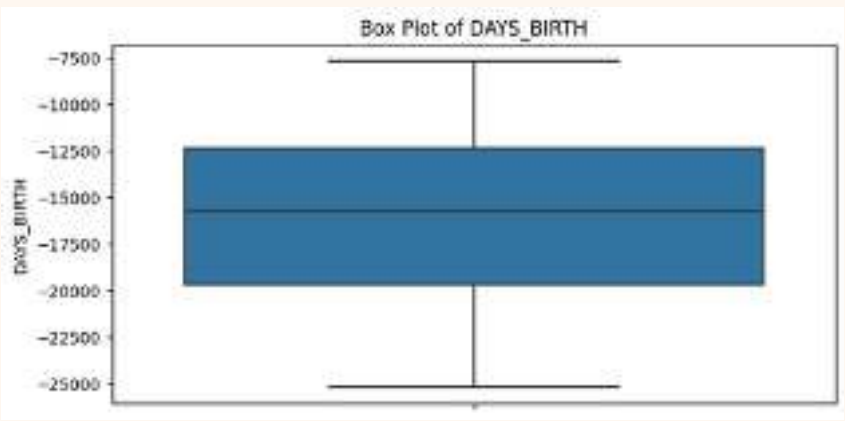






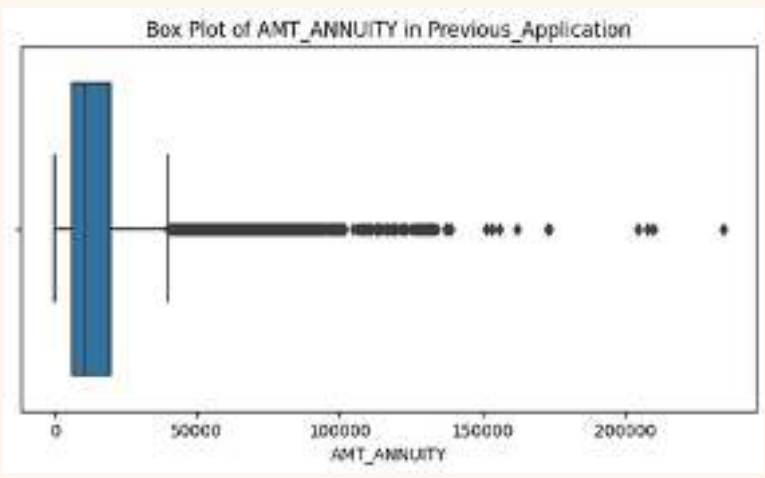


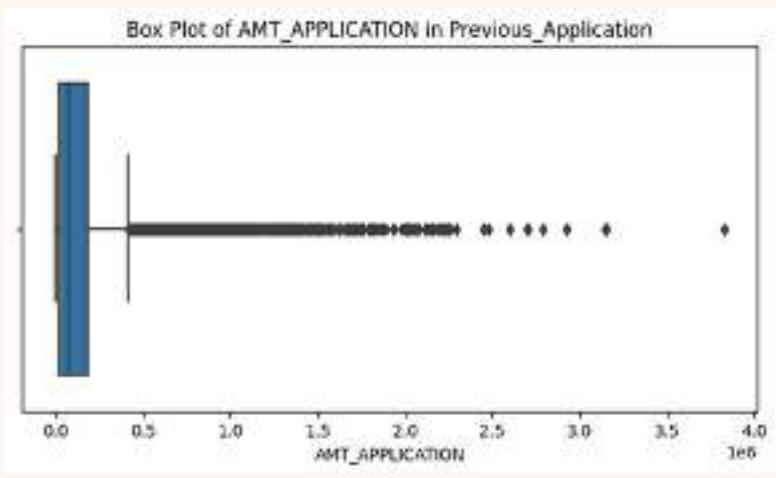


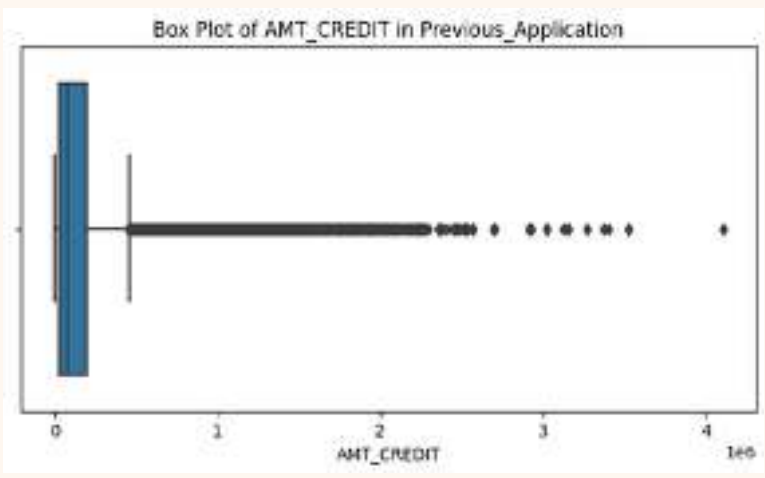


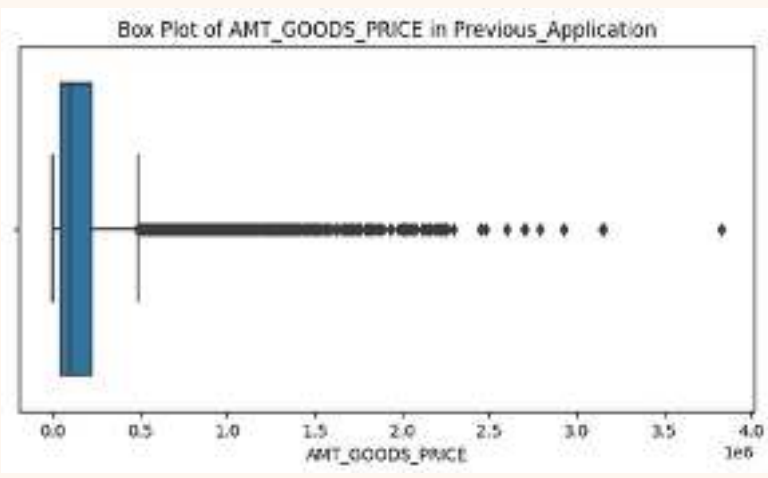
Upon reviewing the application data, the following observations are noted:

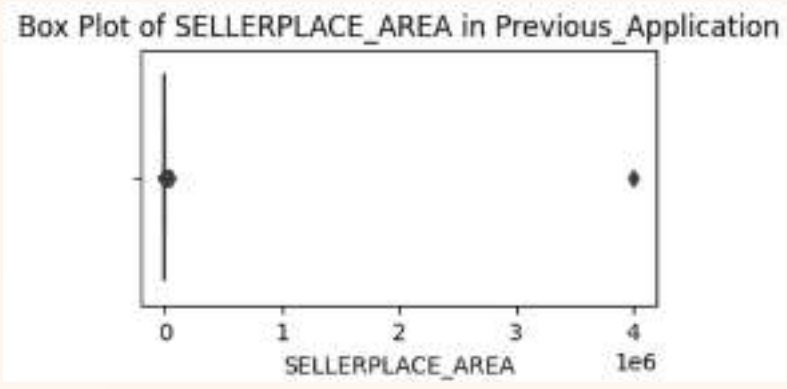
- **AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, and CNT_CHILDREN have some values that are very different from the others.** These differences might be unusual or extreme.
- **AMT_INCOME_TOTAL has many values that are much higher than the rest.** This suggests that some applicants have much higher incomes compared to most others.
- **DAYS_BIRTH doesn't have any unusual values.** It seems like the age data is reliable and doesn't have any extreme or incorrect entries.
- **DAYS_EMPLOYED has some values that are extremely high, around 350,000 days, which is equivalent to roughly 958 years.** This is impossible because people can't work for so long, so these entries are likely mistakes or errors in the data.

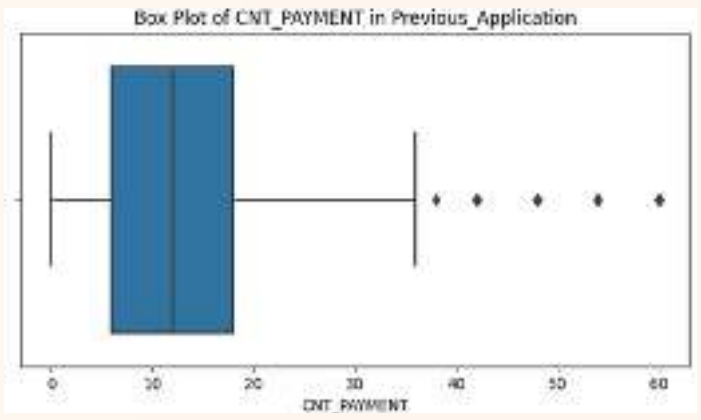


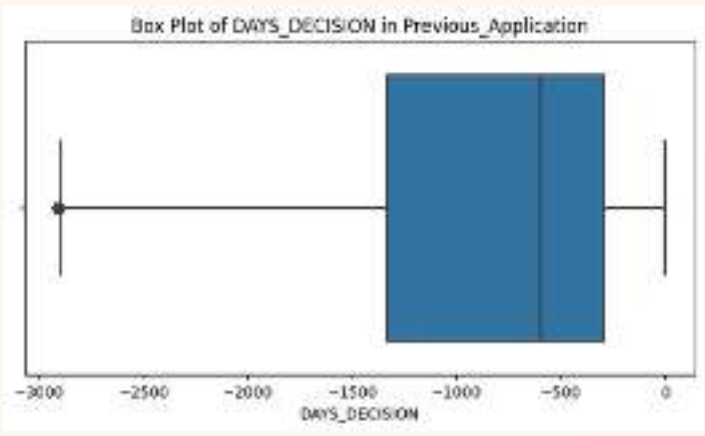


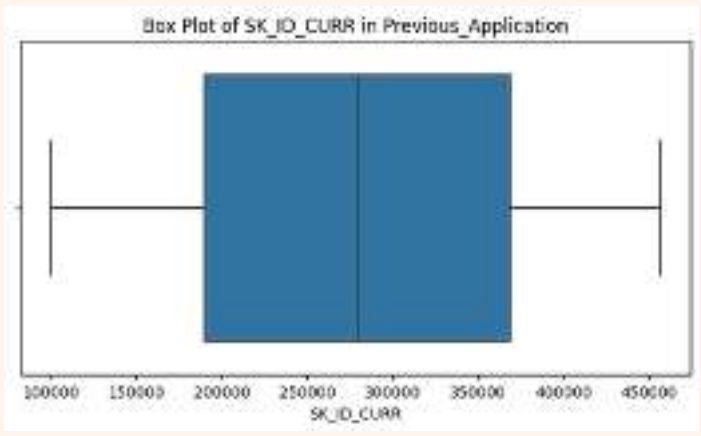








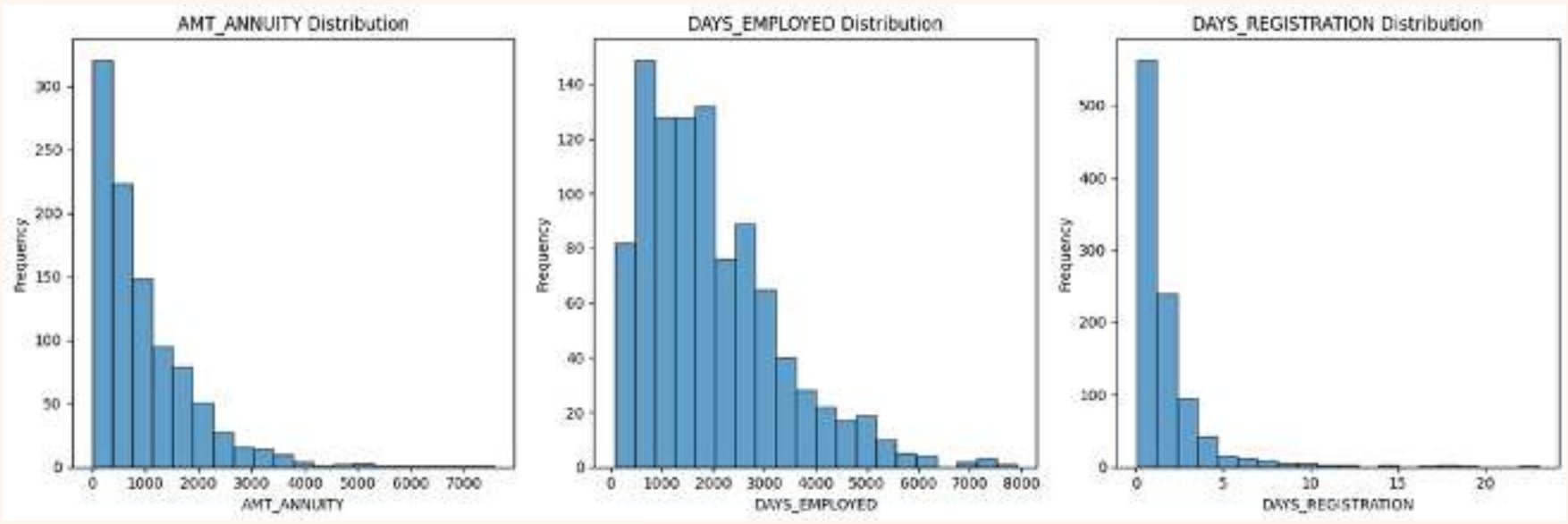




The analysis of our previous application reveals the following:

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, and SELLERPLACE_AREA have many values that are significantly different from the rest. These differences might be unusual or extreme.
- CNT_PAYMENT has a few values that are different from the majority. This suggests that some cases have a different number of payments compared to most others.
- SK_ID_CURR, which is an ID column, doesn't have any unusual values. It's just an identification number, so there are no extreme or unusual entries.
- DAYS_DECISION has a few values that are different from the others, but not many. This indicates that decisions on previous applications were made quite a while ago for some cases,but it's not a widespread issue.

BIN CREATION AND DERIVED METRICS



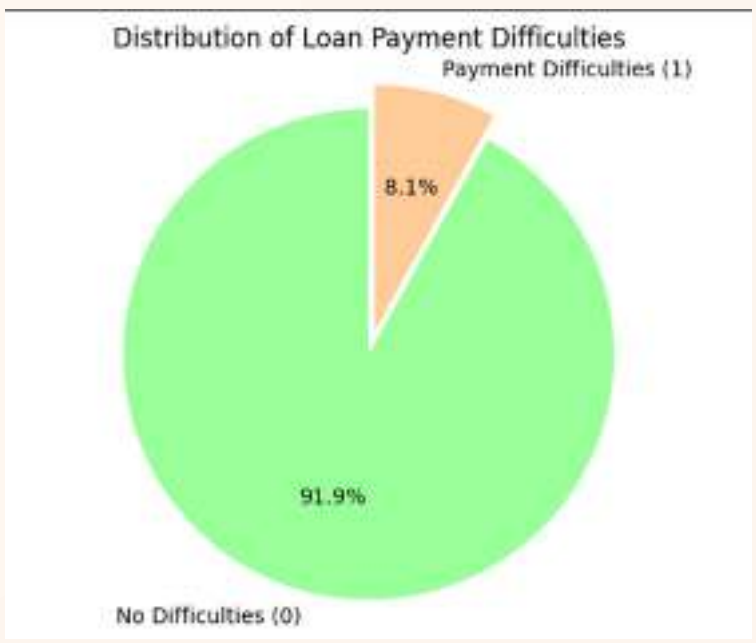
- Looking at the numbers for AMT_ANNUITY, DAYS_EMPLOYED, and DAYS_REGISTRATION, we can see that the values in the lower 25% (the first quartile) are much smaller compared to the values in the upper 25% (the third quartile).
- This indicates that the data is mostly concentrated on the lower end, and there are relatively fewer higher values.
- In simpler terms, it means that most of the data falls into the lower range, and there are only a few data points with higher values.

Task C: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance

SOLUTION:

```
# Assuming your target variable is named 'TARGET' (1 for payment difficulties, 0 for no difficulties)
target_variable = 'TARGET'
# Calculate the distribution of the target variable
target_distribution = app_df[target_variable].value_counts()
# Calculate the ratio of data imbalance
imbalance_ratio = target_distribution[1] / target_distribution[0]
# Create a pie chart to visualize the distribution
labels = ['No Difficulties (0)', 'Payment Difficulties (1)']
sizes = target_distribution.values
colors = ['#ff9999','#66b3ff']
explode = (0.1, 0)  # explode 1st slice (Payment Difficulties)
fig, ax = plt.subplots()
ax.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', startangle=90)
ax.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
# Display the pie chart
plt.title('Distribution of Loan Payment Difficulties')
plt.show()
# Print the distribution and imbalance ratio
print("Distribution of the Target Variable:")
print(target_distribution)
print("\nImbalance Ratio (1s to 0s): {:.3f}".format(imbalance_ratio))
```

```
Distribution of the Target Variable:
0    45937
1     4024
Name: TARGET, dtype: int64
```



The dataset has been split into two groups: one where people successfully paid back their loans (accounting for 91.93% of the dataset), and another where people couldn't pay back their loans (making up 8.07% of the dataset).

```python
# Assuming your target variable is named 'TARGET' (1 for payment difficulties, 0 for no difficulties)
target_variable = 'TARGET'
# Calculate the distribution of the target variable
target_distribution = app_df[target_variable].value_counts(normalize=True) * 100
# Calculate the data imbalance percentage
defaulter_percentage = target_distribution[1]
non_defaulter_percentage = target_distribution[0]
# Print the data imbalance percentages
print("Percentage of Defaulters: {:.2f}%".format(defaulter_percentage))
print("Percentage of Non-Defaulters: {:.2f}%".format(non_defaulter_percentage))
# Verify if there is significant data imbalance
if defaulter_percentage > 60:
    print("There is a significant data imbalance where {:.2f}% of the data corresponds to Defaulters.".format(defaulter_percentage))
    print("This indicates an imbalance in the dataset.")
else:
    print("The data is relatively balanced.")
```

```
Percentage of Defaulters: 8.05%
Percentage of Non-Defaulters: 91.95%
The data is relatively balanced.
```

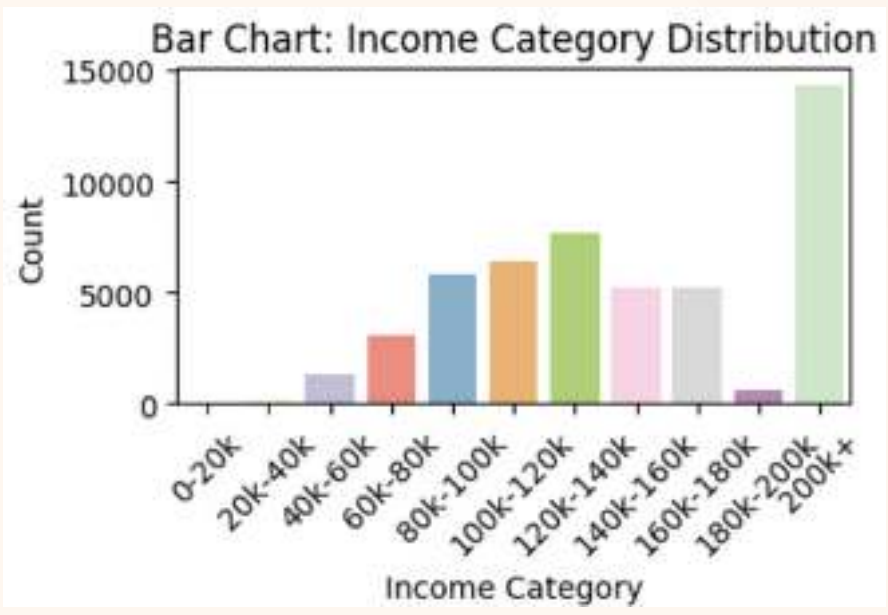
- Examines the distribution of a binary target variable in a dataset.
- If one class constitutes more than 60% of the data, it suggests a substantial imbalance, which can impact the reliability of predictive models.
- This highlights the importance of addressing data imbalance for more accurate modeling and better outcomes.
- Analyzing class distribution is a crucial first step in machine learning projects.

Task D : Involves conducting univariate analysis to gain a comprehensive understanding of the distribution of individual variables. Additionally, the analysis should be segmented to compare variable distributions for different scenarios. Lastly, bivariate analysis should be performed to explore the relationships between variables and the target variable.

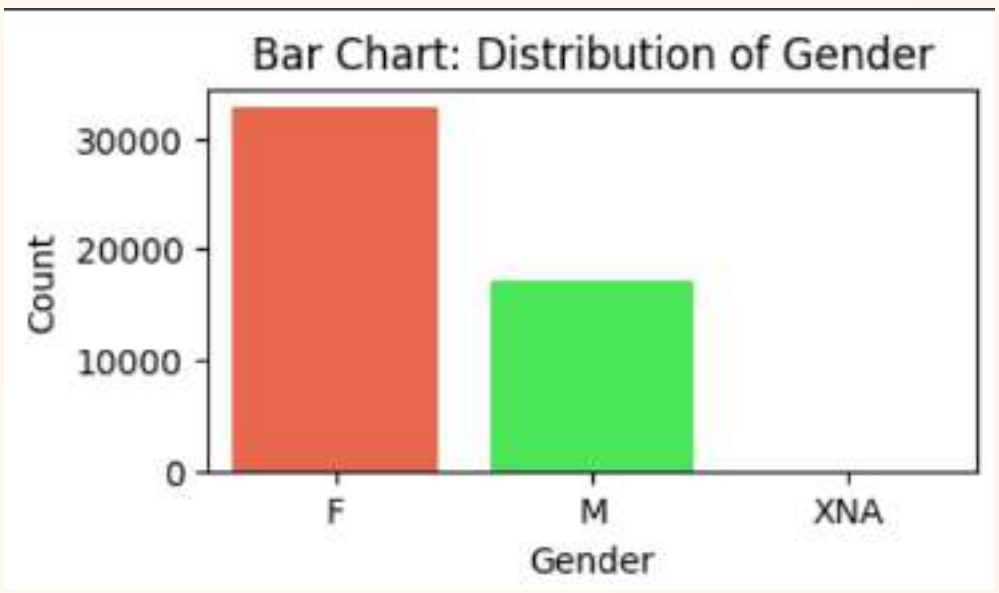
SOLUTION:

- **Univariate Analysis:** Studying one variable at a time to find patterns and insights.
- **Segmented Univariate Analysis:** Analyzing individual variables within distinct groups or segments of data.
- **Bivariate Analysis:** Examining two variables together to uncover relationships and patterns.

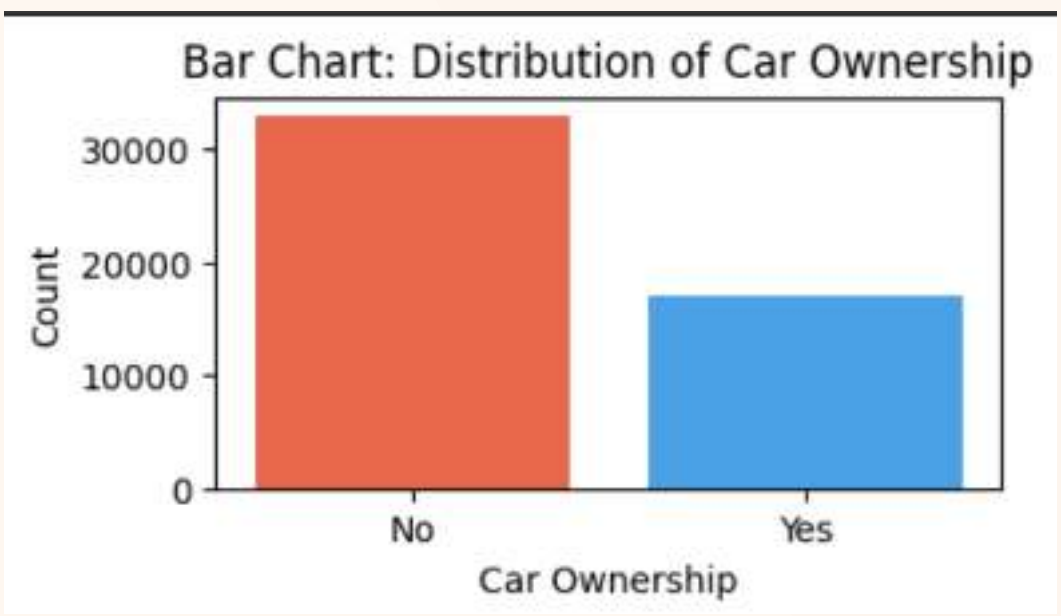
UNIVARIATE ANALYSIS



The graph shows that most people earn lower incomes, and only a smaller group of people earn higher incomes. The middle income, where half of the people earn more and half earn less, is somewhere between $60,000 and $80,000.

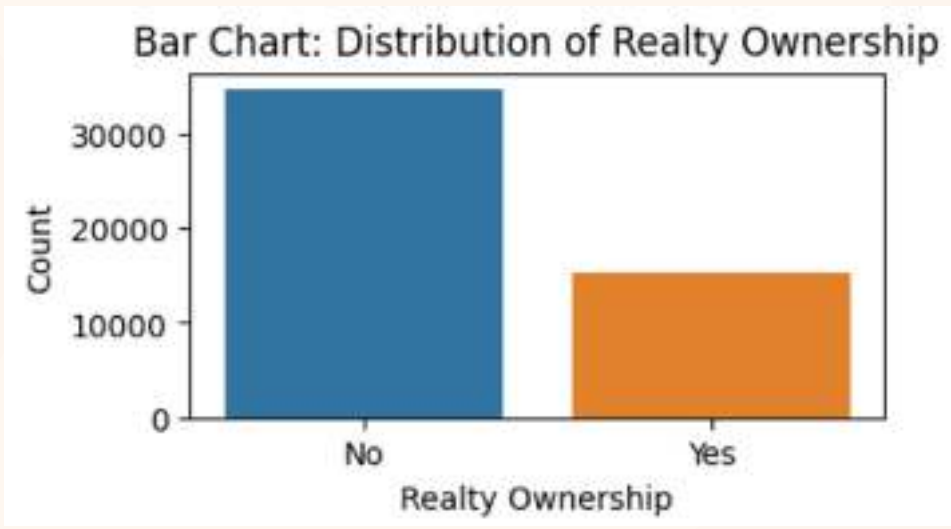


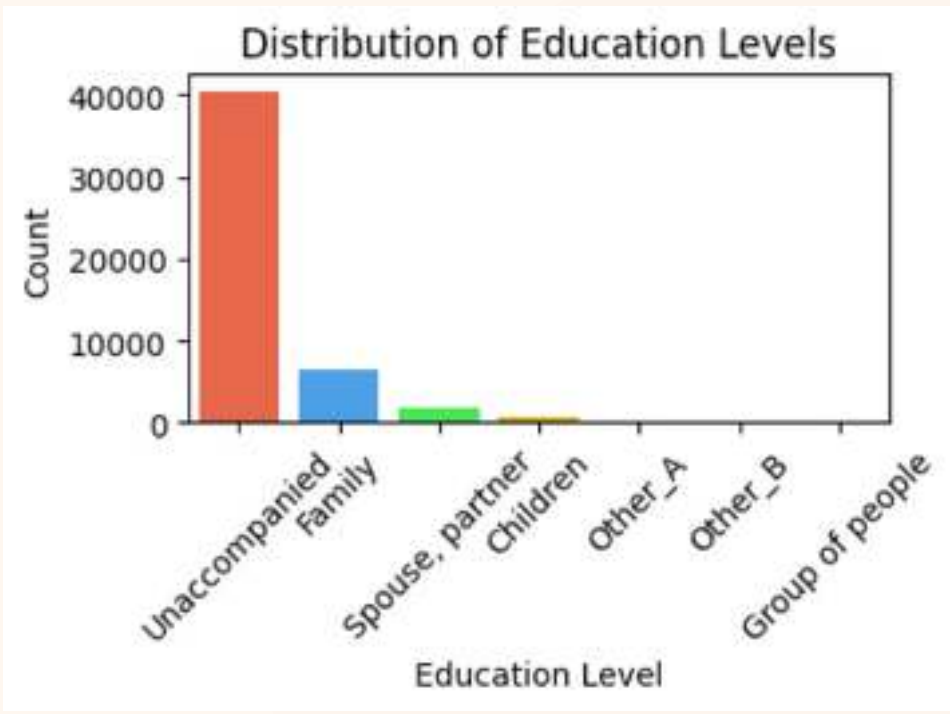
The graph shows that females make up 52% of the population, males are at 46%, and non-binary individuals account for 2%. The gender gap is small, with only a 6% difference. Non-binary numbers are slowly growing. Keep in mind that gender distribution might differ in specific groups like the workforce compared to the overall population.



Most people in the United States have their own cars. People who make more than $100,000 a year are more likely to own cars, while those who make less than $20,000 a year are less likely to have one. The gap in car ownership between these two income groups is quite big.

Look at the next graph! It shows that 62% of people haven't bought any real estate, while only 38% of the population has managed to do so.

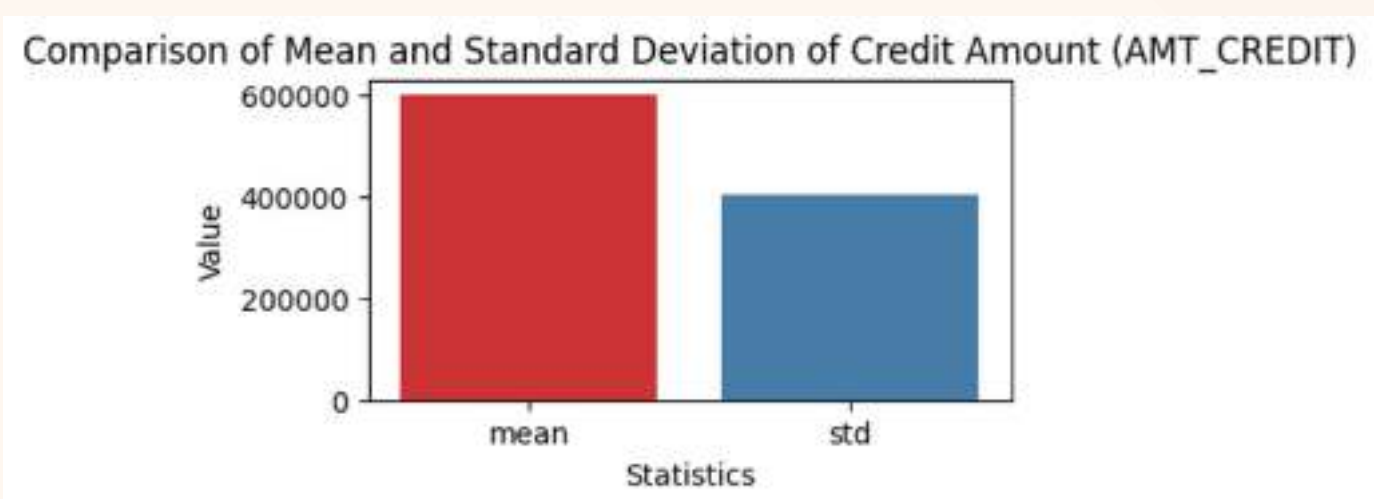




There's a big difference in the education levels of unaccompanied minors and family units.

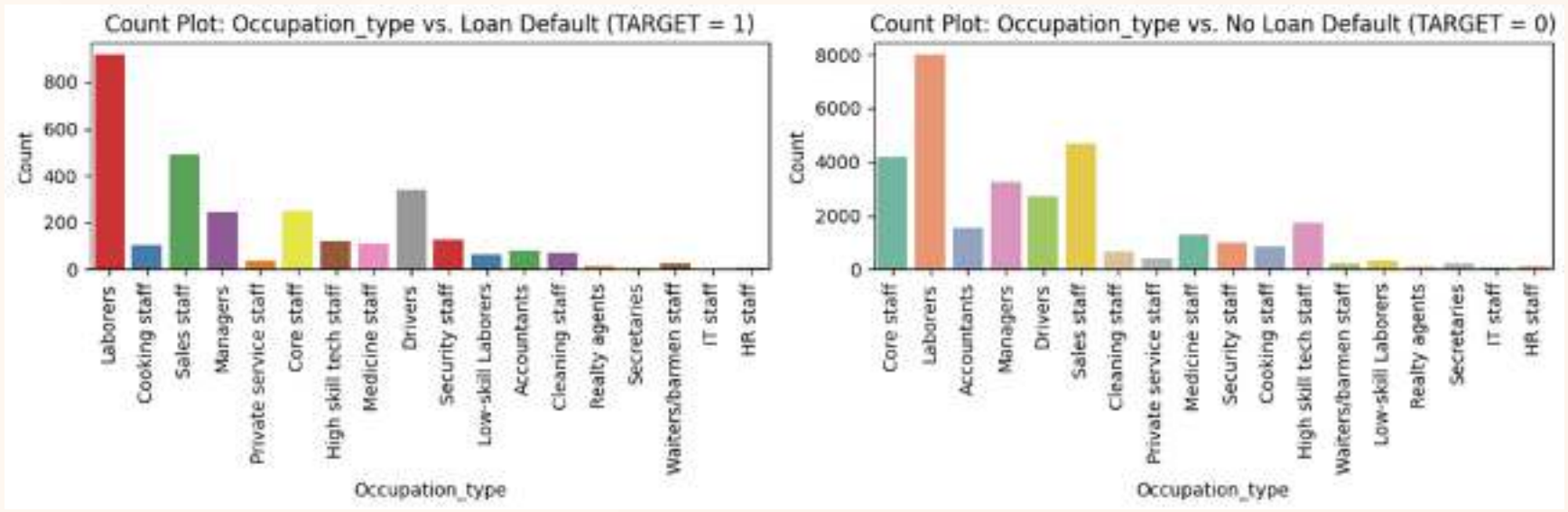
About 55% of unaccompanied minors haven't finished high school, while 70% of family units have at least a high school diploma.

When it comes to higher education, only 17% of unaccompanied minors have gone to college, while 35% of family units have done so.

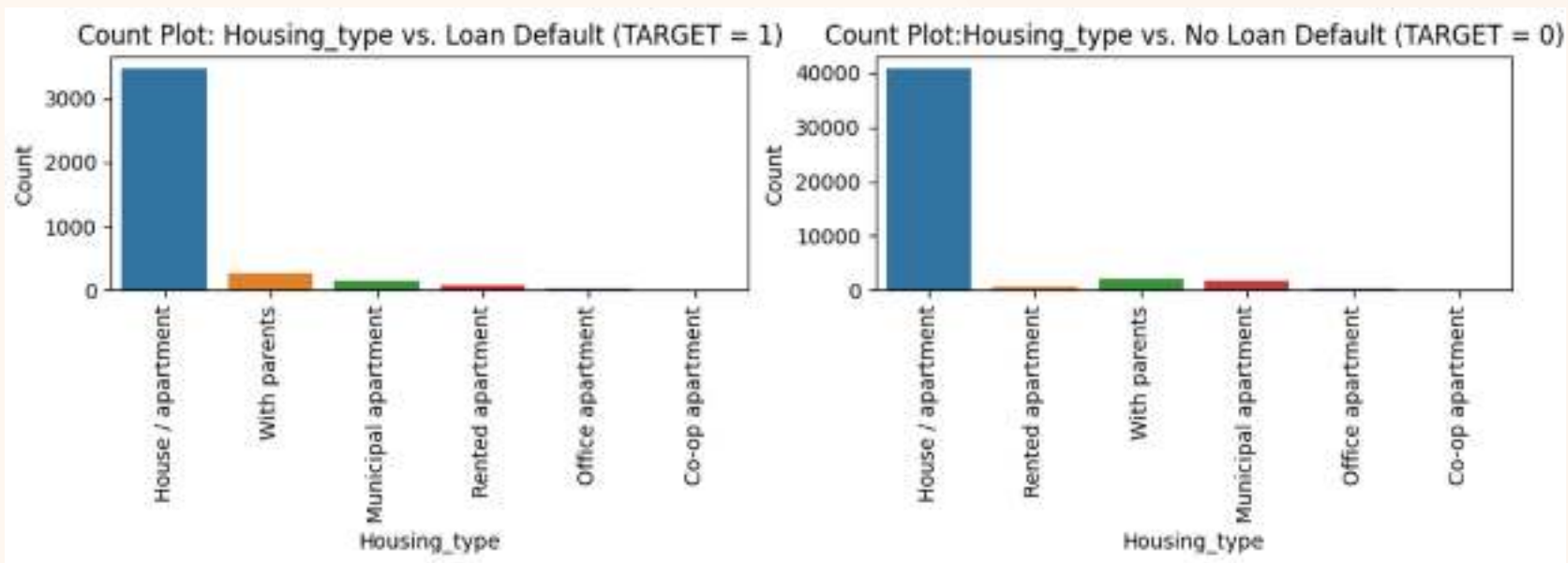


Males typically borrow more than females, with a notable gap, especially for larger amounts. The average borrowing amount is higher than the typical difference in borrowing amounts.

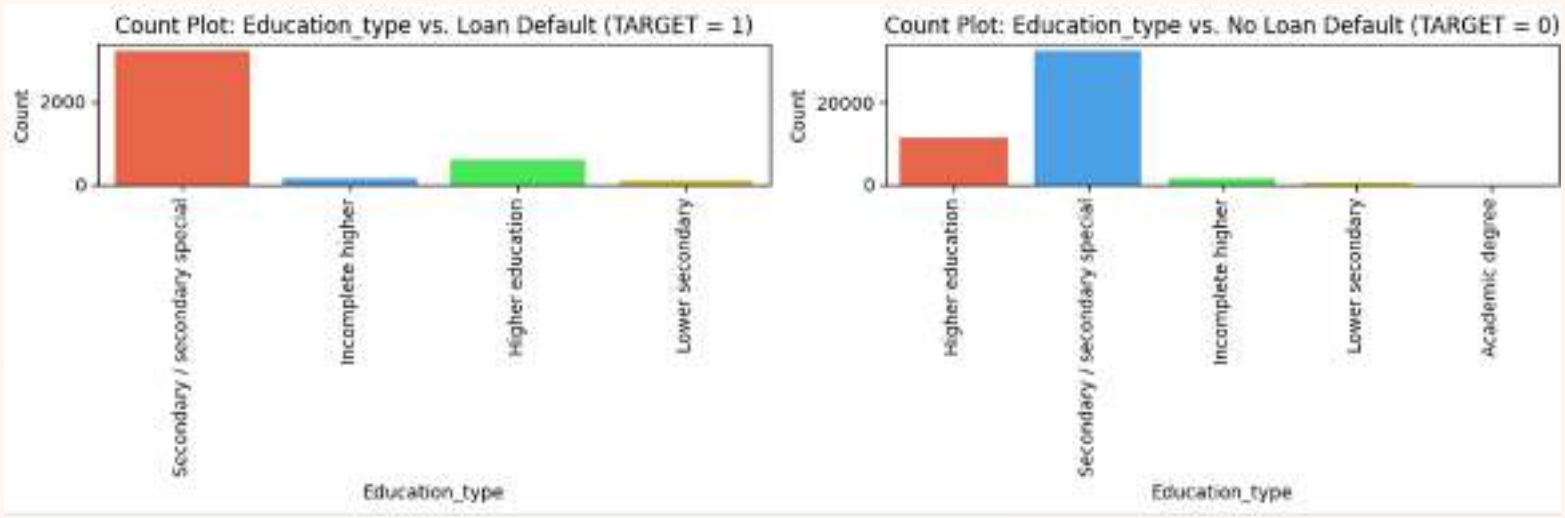
SEGMENTED UNIVARIATE ANALYSIS



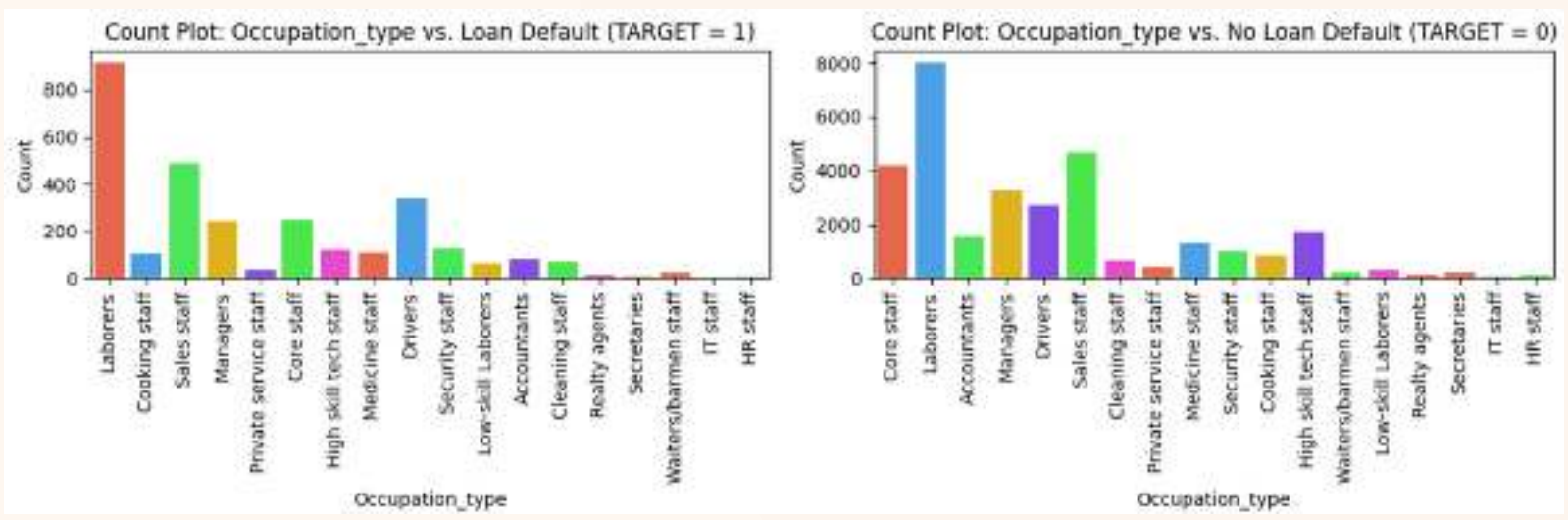
- Waiters/barmen staff: Customer conflicts leading to legal issues.
- Low-skill laborers: Disputes related to wages and working hours.
- Realty agents: Real estate transaction disputes.
- Secretaries: Employment contract disagreements.
- IT staff: Intellectual property disputes.



Factory-made and movable homes have higher loan default rates, possibly because they are cheaper and located in rural areas with limited job opportunities, impacting loan approvals.

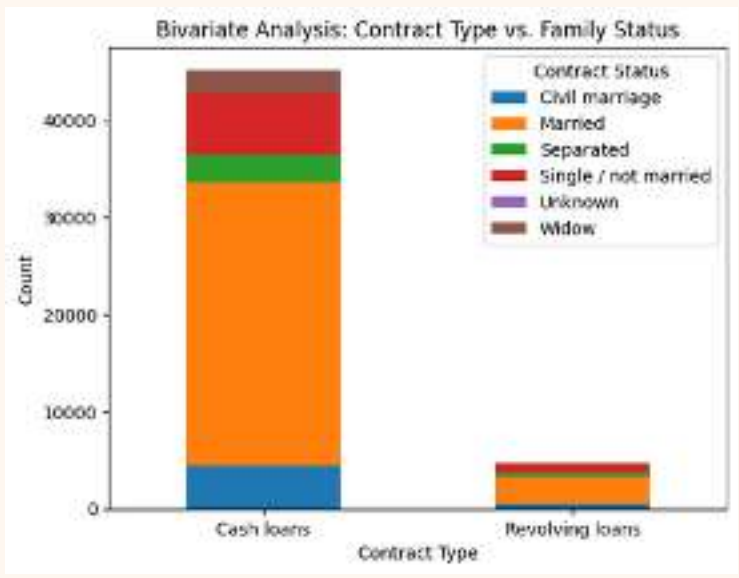


In developed countries with strong educational values and government support, there are more people with higher education than those with lower education.

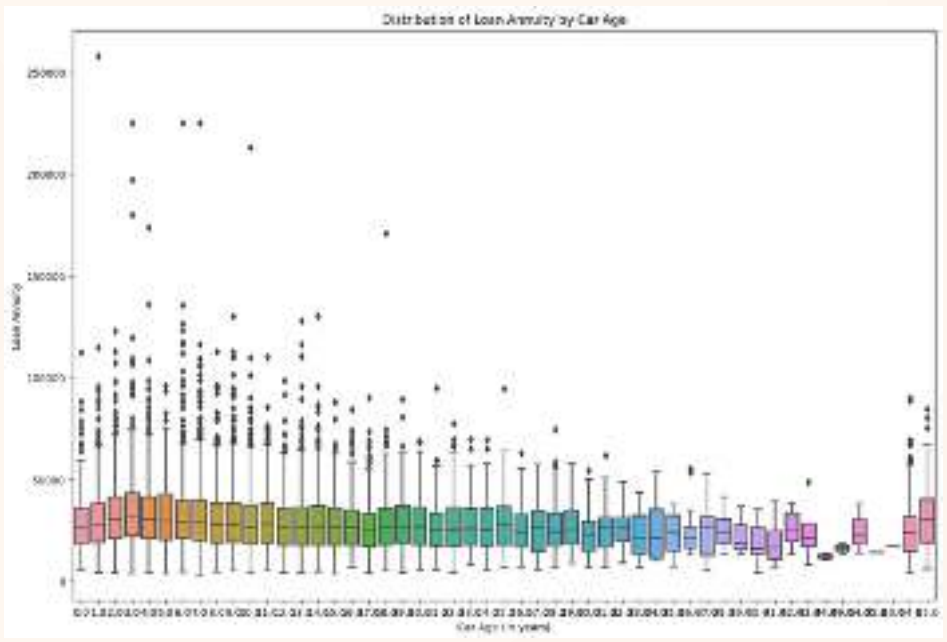


Occupations requiring high education, responsibility, and demand tend to have the highest median incomes.

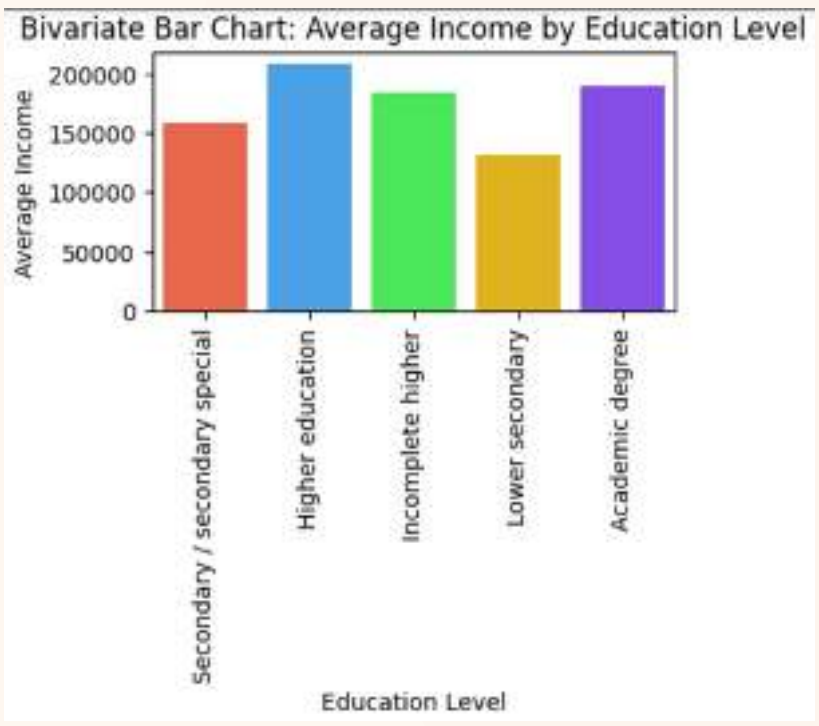
BIVARIATE ANALYSIS



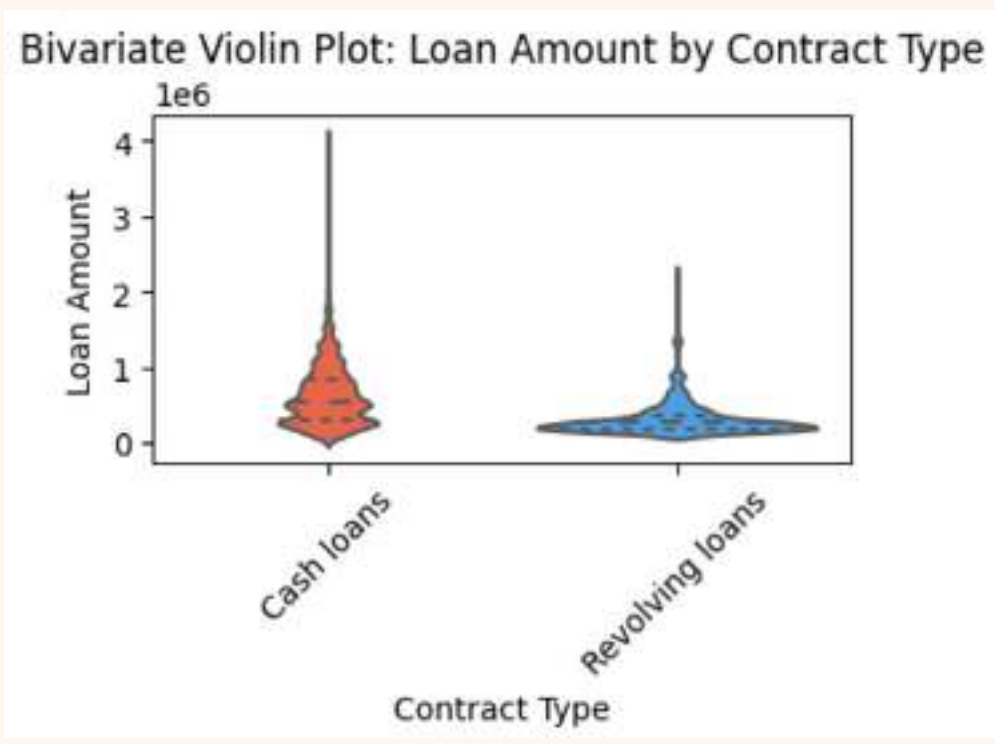
Married couples are more likely to take out and successfully repay cash loans compared to other individuals in the distribution, while widows are more numerous in the group.



Countries with better healthcare, nutrition, and economic progress usually have people who live longer on average.



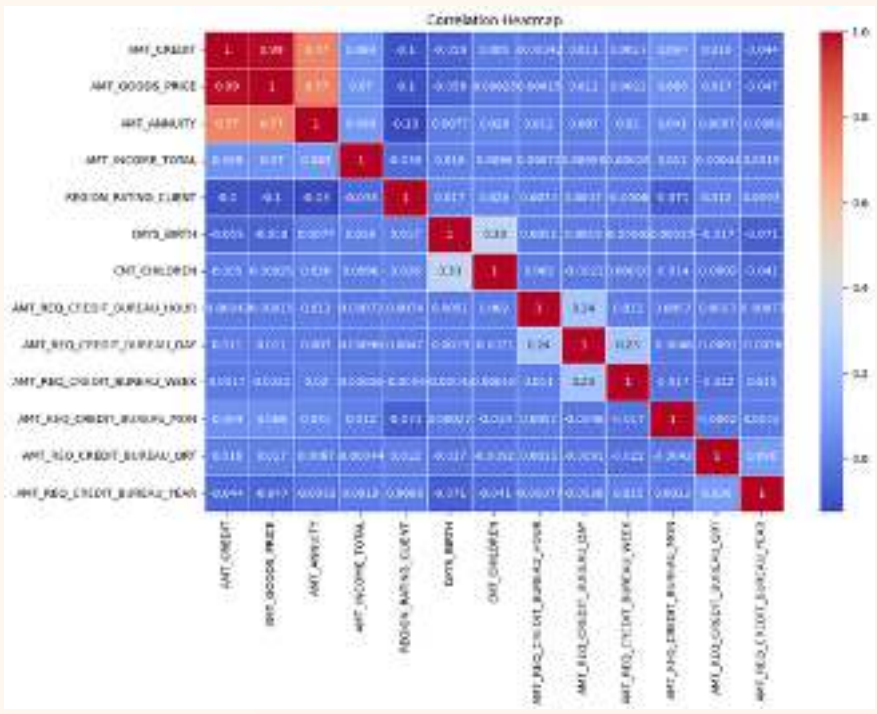
People with more education usually make more money than those with less education



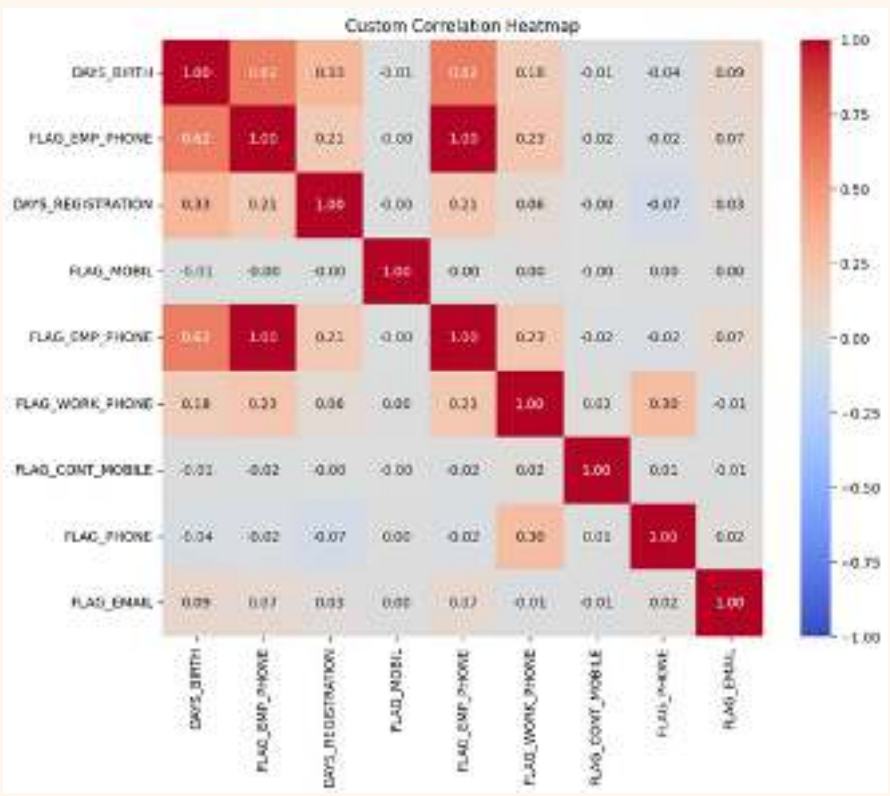
Cash loans typically provide a broader range of loan amounts compared to revolving loans. This variation is influenced by factors like the borrower's creditworthiness, the purpose of the loan, and the risk perceived by the lender.

Task E: 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

SOLUTION:



- The heatmap reveals that certain variables strongly relate to credit risk, including income, credit score, and debt-to-income ratio.
- Lenders can use this data to spot borrowers at a higher risk of not repaying their loans.
- This helps them make better lending choices and set suitable interest rates.
- Factors like higher income, better credit scores, lower debt-to-income ratios, smaller loan amounts, and shorter loan terms are linked to lower credit risk.



```
Top 10 Correlations for Target 0:
NONLIVINGAPARTMENTS_MEDI  NONLIVINGAPARTMENTS_AVG     0.999378
NONLIVINGAPARTMENTS_AVG   NONLIVINGAPARTMENTS_MEDI    0.999378
LIVINGAPARTMENTS_MEDI     LIVINGAPARTMENTS_AVG        0.998984
LIVINGAPARTMENTS_AVG      LIVINGAPARTMENTS_MEDI       0.998984
LANDAREA_AVG              LANDAREA_MEDI               0.998972
LANDAREA_MEDI             LANDAREA_AVG                0.998972
YEARS_BUILD_MEDI          YEARS_BUILD_AVG             0.998814
YEARS_BUILD_AVG           YEARS_BUILD_MEDI            0.998814
COMMONAREA_AVG            COMMONAREA_MEDI             0.998792
COMMONAREA_MEDI           COMMONAREA_AVG              0.998792
dtype: float64

Top 10 Correlations for Target 1:
NONLIVINGAPARTMENTS_MEDI  NONLIVINGAPARTMENTS_AVG     0.999378
NONLIVINGAPARTMENTS_AVG   NONLIVINGAPARTMENTS_MEDI    0.999378
LIVINGAPARTMENTS_MEDI     LIVINGAPARTMENTS_AVG        0.998984
LIVINGAPARTMENTS_AVG      LIVINGAPARTMENTS_MEDI       0.998984
LANDAREA_AVG              LANDAREA_MEDI               0.998972
LANDAREA_MEDI             LANDAREA_AVG                0.998972
YEARS_BUILD_MEDI          YEARS_BUILD_AVG             0.998814
YEARS_BUILD_AVG           YEARS_BUILD_MEDI            0.998814
COMMONAREA_AVG            COMMONAREA_MEDI             0.998792
COMMONAREA_MEDI           COMMONAREA_AVG              0.998792
dtype: float64
```

- There's a notable connection between an employee's age and their phone number – as age goes up, the likelihood of having a phone number decreases.
- On the other hand, there's a positive link between the number of days before a client's registration change and their age.
- This suggests that older people are less likely to change their registration details before applying for a loan.
- Additionally, clients who don't include their phone numbers are also less likely to provide incorrect permanent and work addresses.

CONCLUSION

- After carefully analyzing and cleaning both the application dataset and the previous application dataset, we've discovered some important insights that can help banks make better decisions about loan applications and default risk.
- Firstly, it's clear that lending to students, retirees, and people with higher education tends to be less risky in terms of loan defaults.
- So, banks can confidently approve loans for these groups.
- On the flip side, certain occupations like laborers, sales staff, drivers, cleaning workers, and low-skilled workers have a higher likelihood of not paying back their loans.
- To reduce risk, it's advisable to focus more on clients in stable professions like managers, core staff, and highly skilled technical workers.

IMPACT OF CAR FEATURES

PROJECT DESCRIPTION

- This project aims to help a car manufacturer boost profitability by optimizing pricing and product development decisions.
- It involves using data analysis techniques like regression and market segmentation to understand how car features, market categories, and pricing are related.
- This analysis will inform a pricing strategy that aligns with consumer demand and profitability, while also guiding future product development efforts.
- The ultimate goal is to enhance the manufacturer's competitiveness and drive long-term profitability.

APPROACH

- The project aims to improve decision-making and profitability in the market.
- It begins with clear objectives, data analysis, organization, and identifying correlations.
- Regression modeling and market segmentation enhance understanding, while pricing strategies are assessed for competitiveness and sustainability.
- Profitability guides decisions, and ongoing pricing adjustments are crucial.
- Product development decisions are based on insights, ultimately maximizing profitability and informing continuous strategic choices.

TECH-STACK USED

- To accomplish effective data analysis, I relied on Microsoft Excel as my primary instrument.
- During this undertaking, I adeptly utilized advanced Excel functions, pivot tables, and charts, all of which were instrumental in successfully concluding the analysis.

INSIGHTS

- **Feature Importance Analysis:** Examining the Correlation between Car Features and Pricing to Identify Key Influential Features.
- **Market Segmentation:** Segmenting the Market Based on Consumer Preferences and Buying Behavior Using Clustering Techniques or Pivot Tables.
- **Price Elasticity of Demand Analysis:** Analyzing Price Elasticity of Demand to Understand Customer Sensitivity to Pricing Changes.
- **Profitability Assessment:** Evaluating the Profitability of Different Car Models, Market Categories, or Customer Segments.

- **Competitive Analysis:** Analyzing Competitors' Pricing Strategies, Product Offerings, and Market Positioning.
- **Future Demand Forecasting:** Forecasting Future Demand Based on Historical Data and Regression Analysis Utilizing Pricing, Features, and Market Categories.
- **Pricing Strategy Optimization:** Optimizing Pricing Strategies Using Regression Analysis and Market Data.
- **Product Development Prioritization:** Identifying Influential and Desired Features Based on Correlation Analysis and Market Segmentation for Product Development Prioritization.

ANALYSIS

DATA CLEANING

DUPLICATES REMOVED 715

NULL VALUES REMOVED 102

The duplicates in the dataset are removed and the Null values are filtered

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

SOLUTION:

Markey Category	Average of Popularity	Count of Model
Crossover	1539.475655	1068
Crossover,Diesel	873	7
Crossover,Exotic,Luxury,High-Performance	238	1
Crossover,Exotic,Luxury,Performance	238	1
Crossover,Factory Tuner,Luxury,High-Performance	1823.461538	26
Crossover,Factory Tuner,Luxury,Performance	2607.4	5
Crossover,Factory Tuner,Performance	210	4
Crossover,Flex Fuel	2071.75	64
Crossover,Flex Fuel,Luxury	1173.2	10
Crossover,Flex Fuel,Luxury,Performance	1624	6
Crossover,Flex Fuel,Performance	5657	6
Crossover,Hatchback	1675.694444	72
Crossover,Hatchback,Factory Tuner,Performance	2009	6
Crossover,Hatchback,Luxury	204	7
Crossover,Hatchback,Performance	2009	6
Crossover,Hybrid	2563.380952	42
Crossover,Luxury	889.2142857	406
Crossover,Luxury,Diesel	2195.848485	33
Crossover,Luxury,High-Performance	1037.222222	9
Crossover,Luxury,Hybrid	630.9166667	24
Crossover,Luxury,Performance	1349.089286	112
Crossover,Luxury,Performance,Hybrid	3916	2
Crossover,Performance	2585.956522	69
Diesel	1730.904762	84

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

SOLUTION:

COMBO CHART

DEDUCTION FROM TASK 1

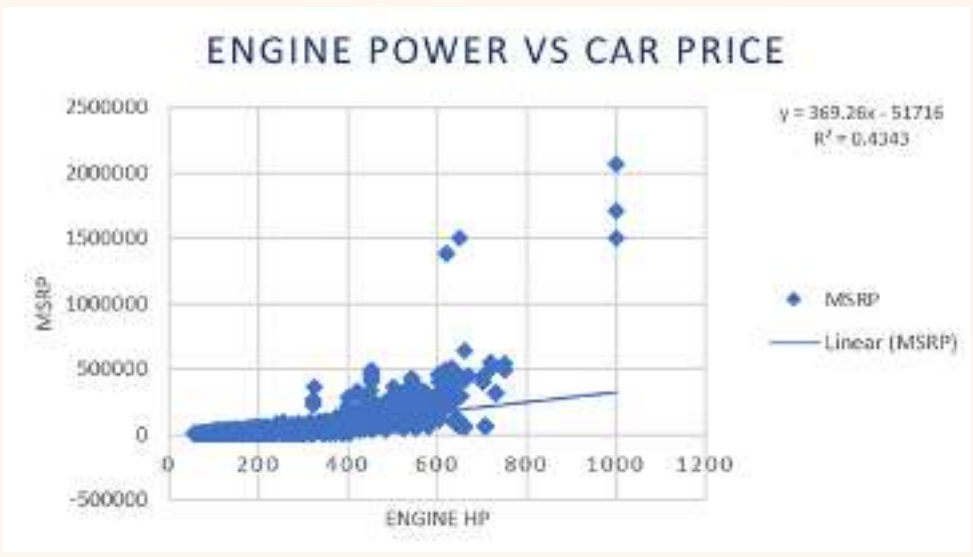
- "Crossover" vehicles are the most common with 1,068 models, averaging a popularity of 1,539.47.
- "Crossover, performance" has 69 models with an average popularity of 2,585.96.
- "Flex fuel" vehicles are prevalent, with 855 models having an average popularity of 2,225.71.
- "Luxury" vehicles come in various subcategories, ranging from 1,084.21 to 2,333.18 in average popularity among 819 models.

- There are many "crossover" vehicles in the market, and they are quite popular.
- "Performance" crossovers are even more popular.
- "Flex fuel" vehicles are also common and fairly popular.
- "Luxury" vehicles come in different types, and their popularity varies depending on the specific type.

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.

SOLUTION:

SCATTER CHART



DEDUCTION FROM TASK 2

- "Higher engine horsepower (HP) usually means a higher car price.
- The data strong positively correlated and supports this, with an equation (y = 369.26x - 51,716) and an R-squared value of 0.4343.
- More powerful cars generally cost more, and this relationship is well-supported by the data.

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.

SOLUTION:

	Coefficients
Intercept	-97168.55653
Engine HP	320.4898943
Engine Cylinders	7579.054029
Number of Doors	-4979.926507
highway MPG	503.6390228
city mpg	1253.398944
Popularity	-3.553657509

BAR CHART

- The bar chart highlights the key factors influencing car prices, with a focus on engine-related factors and fuel efficiency. Other factors like the number of doors and popularity have less pronounced effects.

DEDUCTION FROM TASK 3

- **Base Price (Intercept):** The base price of the car, before considering any other factors, starts at approximately -$97,169. This means it's a negative

value, which might not be meaningful in practice.

- **Engine HP (Horsepower):** Each additional unit of engine horsepower adds around $320 to the car's price. More powerful engines result in higher costs.
- **Engine Cylinders:** The number of engine cylinders significantly impacts the price. Each extra cylinder adds approximately $7,579 to the car's cost, emphasizing the importance of performance.
- **Number of Doors:** More doors reduce the car's price. Each additional door decreases the price by approximately $4,980.
- **Highway MPG (Miles Per Gallon):** Better fuel efficiency on the highway corresponds to a higher price. Each additional mile per gallon adds about $504 to the cost.
- **City MPG:** Improved city fuel efficiency leads to a higher price. Each additional mile per gallon in the city increases the car's price by roughly $1,253.
- **Popularity:** While the effect is relatively small, increasing popularity slightly lowers the price. For every unit increase in popularity, the price decreases by about $3.55.
- In summary, these coefficients help understand how specific factors contribute to a car's price, with positive coefficients indicating a price increase and negative coefficients indicating a price decrease.

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

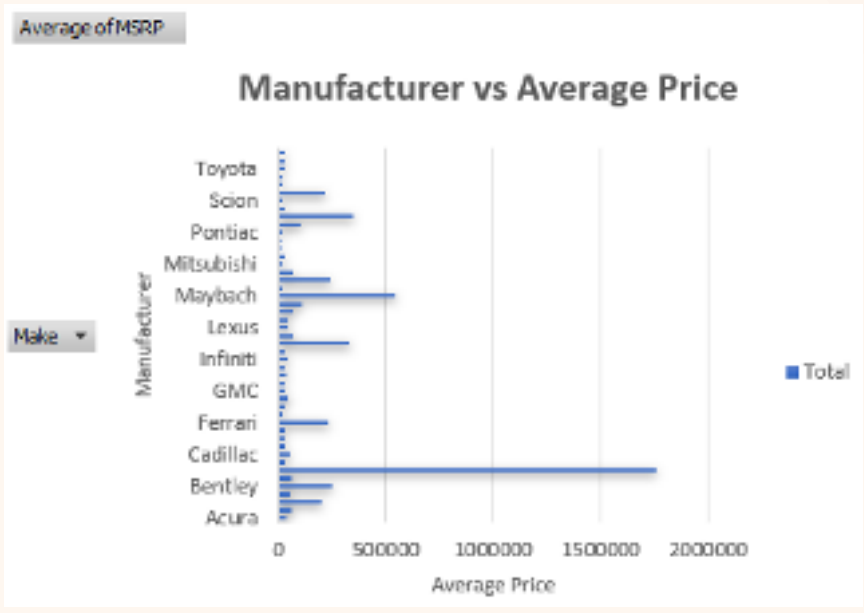
SOLUTION:

Manufacturer	Average of MSRP
Acura	35087.4878
Alfa Romeo	61600
Aston Martin	198123.4615
Audi	54574.1215
Bentley	247169.3243
BMW	62162.55864
Bugatti	1757223.667
Buick	29034.18947
Cadillac	56368.26515
Chevrolet	29000.2214
Chrysler	26722.96257
Dodge	24857.04537
Ferrari	237383.8235
FIAT	22206.01695
Ford	28522.86207
Genesis	46616.66667
GMC	32444.08506
Honda	26608.88399
HUMMER	36464.41176
Hyundai	24926.26255
Infiniti	42640.27134
Kia	25318.75
Lamborghini	331567.3077
Land Rover	68067.08633

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

SOLUTION:

BAR CHART



DEDUCTION FROM TASK 4

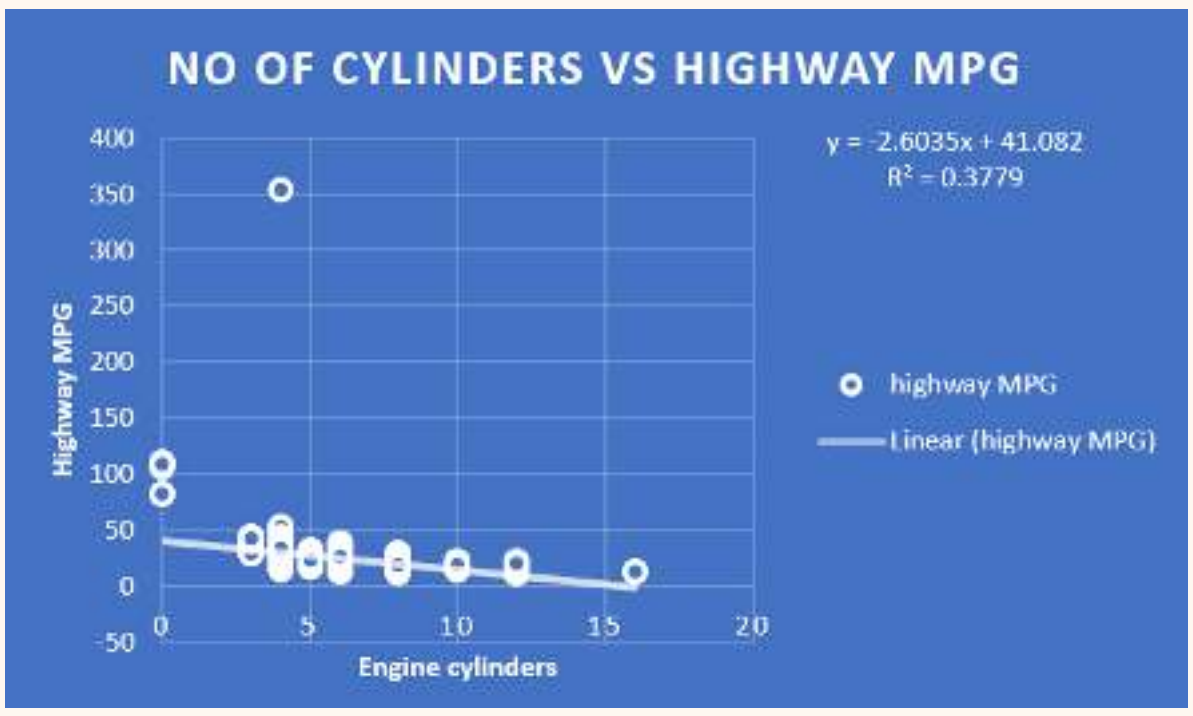
- **Luxury brands :** Bugatti, Maybach, and Rolls-Royce target elite customers with exceptionally high prices.

- **Premium brands :** Aston Martin, Bentley, and Lotus offer style and performance at relatively high price points.
- **Mainstream brands :** Toyota, Chevrolet, and Ford focus on reliability and affordability, catering to a broad consumer base.
- **Sports car brands :** Lamborghini, McLaren, and Porsche emphasize performance and prestige, with varying price ranges.
- **Diverse range brands :** Genesis, Infiniti, and GMC meet specific needs, providing a balance between luxury and affordability, premium features at a moderate cost, and diverse vehicle options.

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

SOLUTION:

SCATTER CHART



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

SOLUTION:

CORRELATION COEFFICIENT	-0.614703148

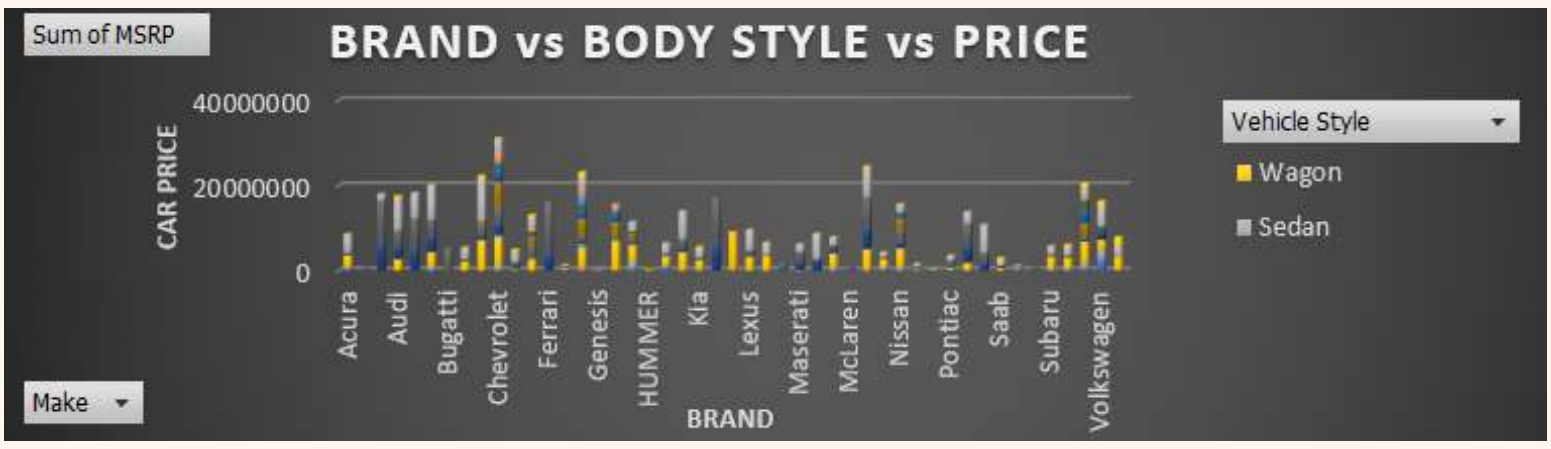
DEDUCTION FROM TASK 5

- The scatter plot and correlation coefficient (-0.6147) reveal a clear negative relationship between the number of cylinders in a car's engine and its highway miles per gallon (MPG).
- In simple terms, as the number of cylinders increases, the highway MPG generally decreases, indicating that more powerful engines tend to be less fuel-efficient on the highway.
- The negative slope of the trendline on the scatter plot indicates a clear inverse relationship between the number of cylinders and highway MPG.
- As the number of cylinders in a car's engine increases, the highway MPG tends to decrease.

CREATING A DASHBOARD

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

SOLUTION:

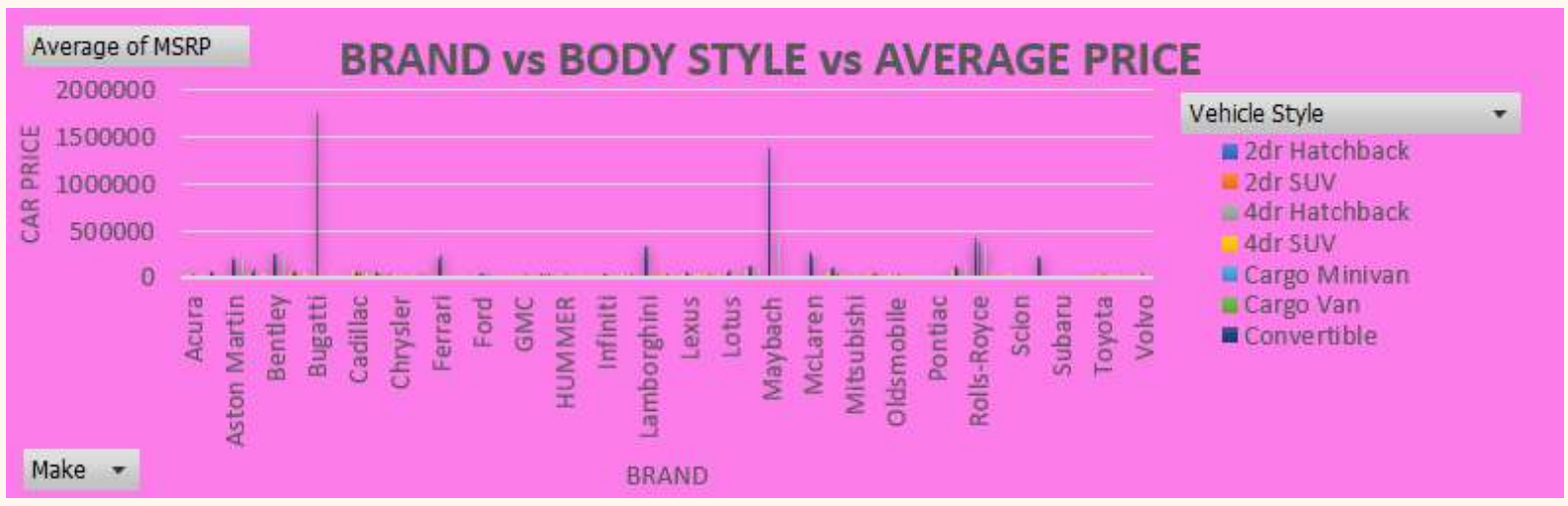


DEDUCTION

- Chevrolet offers a diverse range of vehicle styles, including sedans, SUVs, trucks, and more, providing customers with a wide array of choices.
- In contrast, brands like Alfa Romeo, Tesla, Bugatti, and Genesis focus on a single style of vehicle in their lineup, offering simplicity but limited variety to consumers.

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

SOLUTION:

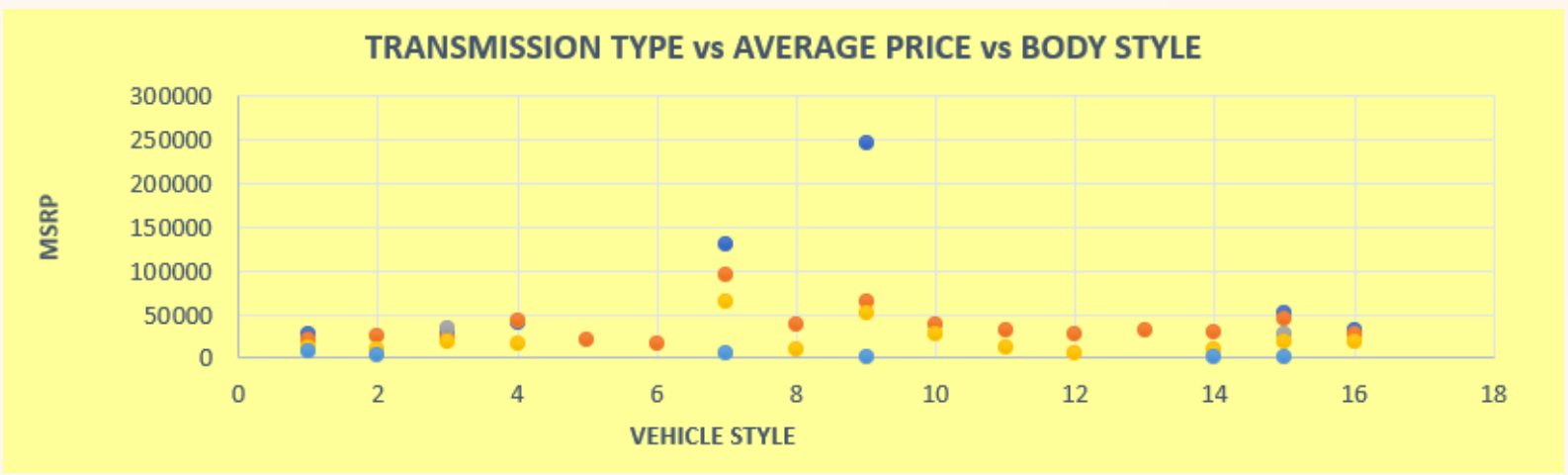


DEDUCTION

- Bugatti has the highest average price among these brands, indicating its exclusivity and luxury. It's known for its incredibly expensive and exclusive cars.
- Maybach comes next with a relatively high average price, known for its opulent and technologically advanced vehicles. It offers luxurious and high-end options.
- On the other hand, Plymouth has the lowest average price, offering budget-friendly and affordable vehicle choices. It caters to customers looking for more economical options.

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

SOLUTION:



DEDUCTION

- Cars with automated manual transmissions tend to be the most expensive among all types of transmissions. These vehicles often come at higher price points.

- On average, cars with unspecified or unknown transmission types usually have lower prices. These cars are often more affordable.
- Automatic transmissions are commonly used across various car body styles, so you can find them in a wide range of vehicles, from sedans to SUVs.

Task 4:(For Highway MPG) How does the fuel efficiency of cars vary across different body styles and model years?

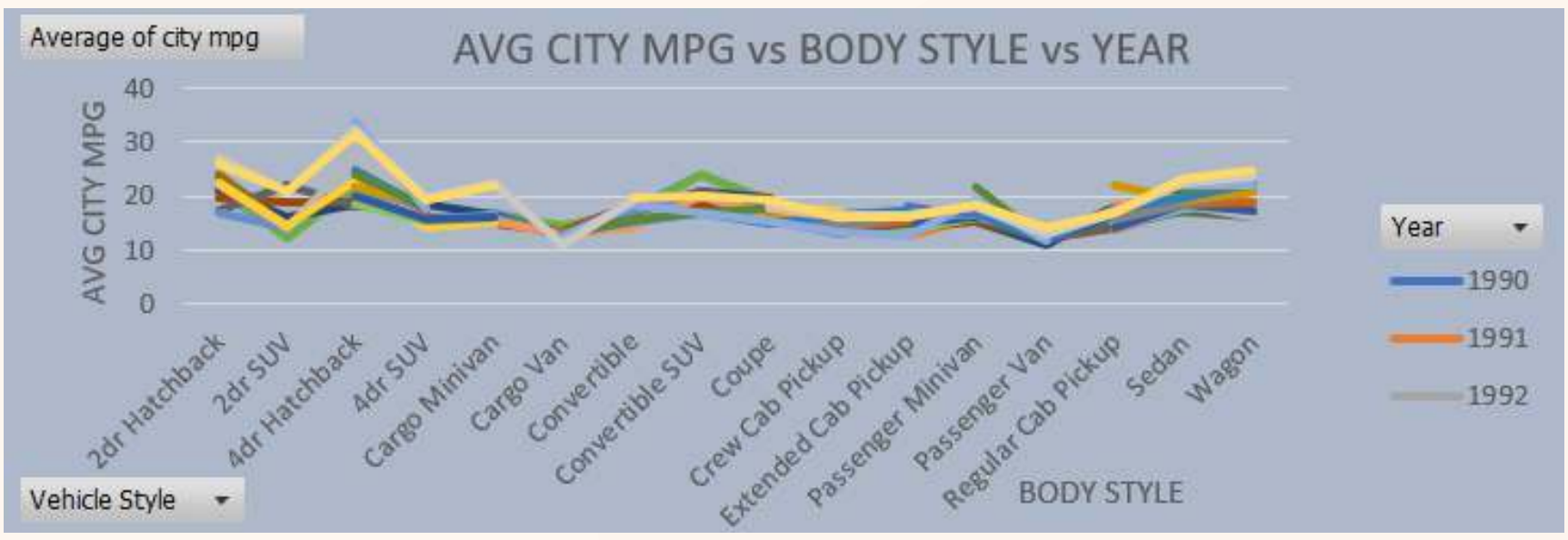
SOLUTION:

DEDUCTION

- Among the listed model years, cars from 2016 have the best average miles per gallon (MPG) across all types of cars. This means they are more fuel-efficient.
- Over the years, the 4-door hatchback model consistently has the highest average MPG compared to other car models. This shows that 4-door hatchbacks, regardless of the manufacturer, are fuel-efficient and a popular choice for people who want to save on fuel costs.

Task 4:(For city MPG) How does the fuel efficiency of cars vary across different body styles and model years?

SOLUTION:

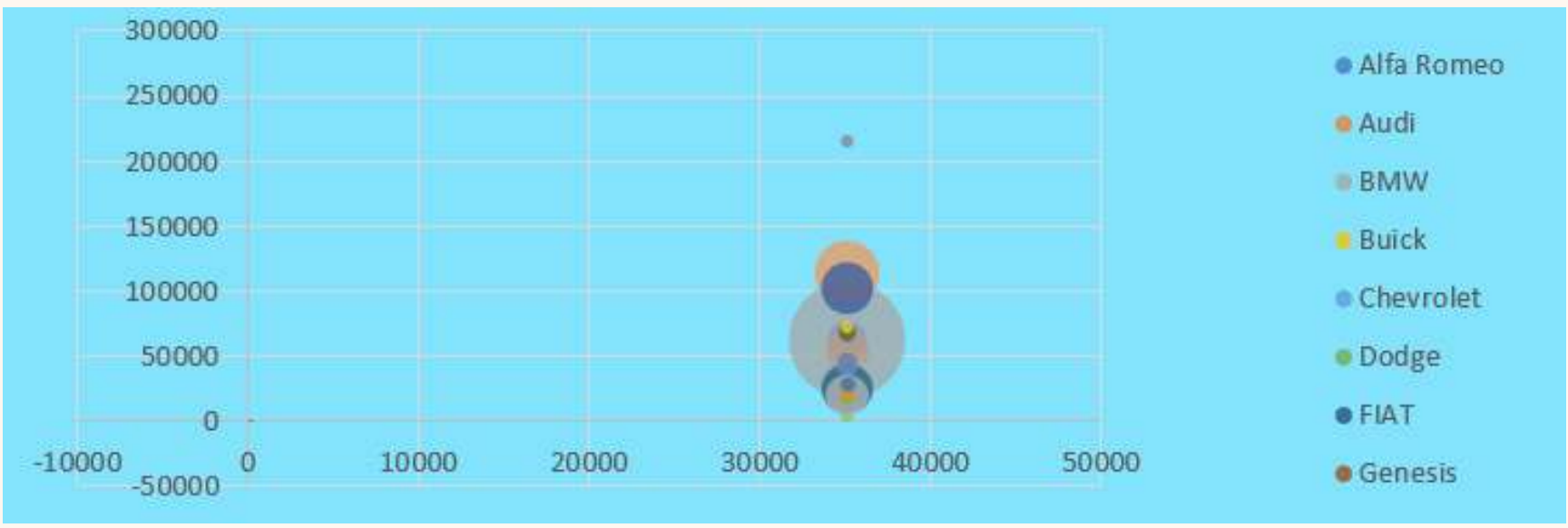


DEDUCTION

- Among the model years listed, cars from 2016 are the most fuel-efficient, meaning they can travel a longer distance on a gallon of fuel.
- Regardless of the car's make or brand, 4-door hatchback models consistently offer superior fuel efficiency over the years.
- This makes them a top choice for individuals who prioritize saving money on fuel.

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

SOLUTION:



DEDUCTION

- Tesla is the best at saving fuel because their electric cars are super eco-friendly.
- Bugatti makes really fast cars with the most powerful engines for thrilling rides.
- Bugatti also sells super expensive luxury cars for rich customers who want top-notch quality and exclusivity.

DASHBOARD

DEDUCTION FROM DASHBOARD

- Chevrolet offers a wide variety of vehicles, while Alfa Romeo, Tesla, Bugatti, and Genesis focus on specific styles. Bugatti is known for luxury and high prices. Maybach is opulent. Plymouth is budget-friendly. Automated manual transmissions are expensive, and unspecified transmissions are more affordable. Automatic transmissions are common. Cars from 2016 are fuel-efficient. 4-door hatchbacks excel in MPG. Tesla is eco-friendly, Bugatti is all about speed and luxury.

CONCLUSION

- **Smart Product Planning:** Identify Customer Preferences: Understand what customers really want in their cars.
- Focus on Key Features: Pay special attention to the features and types of cars that affect prices the most.
- **Precise Targeting:** Tailor to Customer Groups: Customize prices and car models for specific groups of customers.
- **Finding the Sweet Spot:** Determine the ideal prices that make good profits while keeping customers interested.
- **Staying Ahead of Competitors:** Profitable Choices: Figure out which cars are making the most money.
- **Outsmarting the Competition:** Improve pricing and car designs to beat other car companies.
- **Informed Decision-Making:** Getting Stronger,Growing Profits,Leading the Way

ABC CALL VOLUME TREND

PROJECT DESCRIPTION

- The primary objective of this project is to analyze inbound calls made by customers and identify the challenges and pain points they encounter during interactions with the Customer Experience (CX) Inbound calling team.
- By examining a provided dataset, we aim to uncover specific issues such as long queue times, call transfers, and abandoned calls, with the goal of enhancing the overall user experience.
- Additionally, this project will focus on determining the optimal number of employees required to minimize the call abandon rate in the call center.
- Through an in-depth analysis of call volume, queue times, and customer demand patterns, we aim to accurately estimate workforce requirements for effectively handling incoming calls and reducing customer wait times.

APPROACH

- In a data-driven process, we start with data collection and preprocessing, proceed to exploratory data analysis for insights, address customer pain points, optimize the workforce, set performance metrics and KPIs, make recommendations, continuously monitor, use reporting and visualization, and implement with ongoing monitoring for desired outcomes.

TECH-STACK USED

- To accomplish effective data analysis, I relied on Microsoft Excel as my primary instrument.
- During this undertaking, I adeptly utilized advanced Excel functions, pivot tables, and charts, all of which were instrumental in successfully concluding the analysis.

INSIGHTS

- **Identifying Customer Pain Points:** Recognizing issues faced by customers during calls, such as extended wait times or frequent call transfers.
- **Peak Call Hour Identification:** Determining the busiest times for incoming calls to ensure that there is an adequate number of staff members available to handle them.
- **Analyzing Call Status Distribution:** Evaluating the distribution of call outcomes, including those that were answered, transferred, or abandoned, in order to enhance call handling procedures.

- **Agent Performance Assessment:** Assessing the effectiveness of agents based on factors such as call duration and customer satisfaction levels.
- **Average Call Duration Evaluation:** Reviewing whether call durations align with providing effective customer service.
- **Queue Time Analysis:** Examining waiting times within the call queue to identify opportunities for reducing customer wait times.
- **Staffing Optimization:** Determining the optimal staffing levels based on call volume to minimize call abandonments and improve service.
- **Recommendations for Service Enhancement:** Providing actionable suggestions for enhancing
- service quality, such as optimizing call management or implementing chatbots to offer faster assistance.

ANALYSIS

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

SOLUTION:

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

PIE CHART

DEDUCTION FROM TASK 1

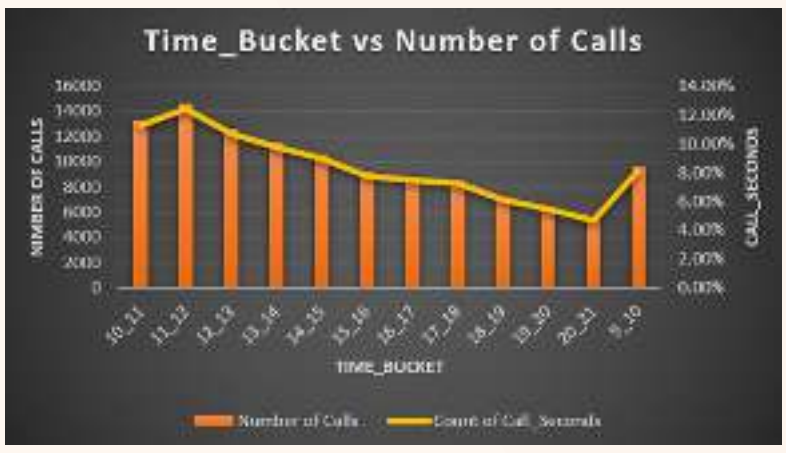
- **Peak Call Duration:** Calls during the "19_20" time slot are the longest, averaging 203.41 seconds.
- **Consistency:** Most time buckets have similar call durations, ranging from 192.89 to 203.41 seconds.
- **Morning and Evening:** Calls in the morning ("10_11") and late evening ("20_21") tend to be slightly longer than the daily average.
- **Lunchtime Dip:** The shortest average call duration occurs during lunchtime ("12_13") at 192.89 seconds.
- **Overall Average:** The overall average call duration across all time buckets is 198.62 seconds, serving as a benchmark.
- **Stable Call Handling:** Call durations are relatively consistent throughout the day, suggesting stable call management processes.

TASK 2 :Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.)

SOLUTION:

Time_Bucket	Number of Calls	Count of Call_Seconds (s)
10_11	13313	11.28%
11_12	14626	12.40%
12_13	12652	10.72%
13_14	11561	9.80%
14_15	10561	8.95%
15_16	9159	7.76%
16_17	8788	7.45%
17_18	8534	7.23%
18_19	7238	6.13%
19_20	6463	5.48%
20_21	5505	4.67%
9_10	9588	8.13%
Grand Total	**117988**	**100.00%**

CLUSTERED COLUMN CHART



DEDUCTION FROM TASK 2

- **Peak Call Times:** The busiest time slots are "11_12" and "10_11," each accounting for over 12% of the total calls.
- **Evening Decline:** Call volume decreases gradually as the day progresses, with lower activity in the evening hours ("18_19," "19_20," and "20_21").
- **Stable Call Durations:** Call durations remain relatively consistent across time slots, indicating that the time of day does not significantly impact call length.

- **Lunchtime Dip:** "12_13" has a slight drop in call volume, likely due to lunch breaks.
- **Morning and Evening Activity:** Calls continue in the early morning ("9_10") and late evening ("20_21"), representing a notable portion of total call volume.
- **Total Call Count:** The dataset comprises a total of 117,988 calls, offering a comprehensive overview of call distribution throughout the day.

TASK 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.

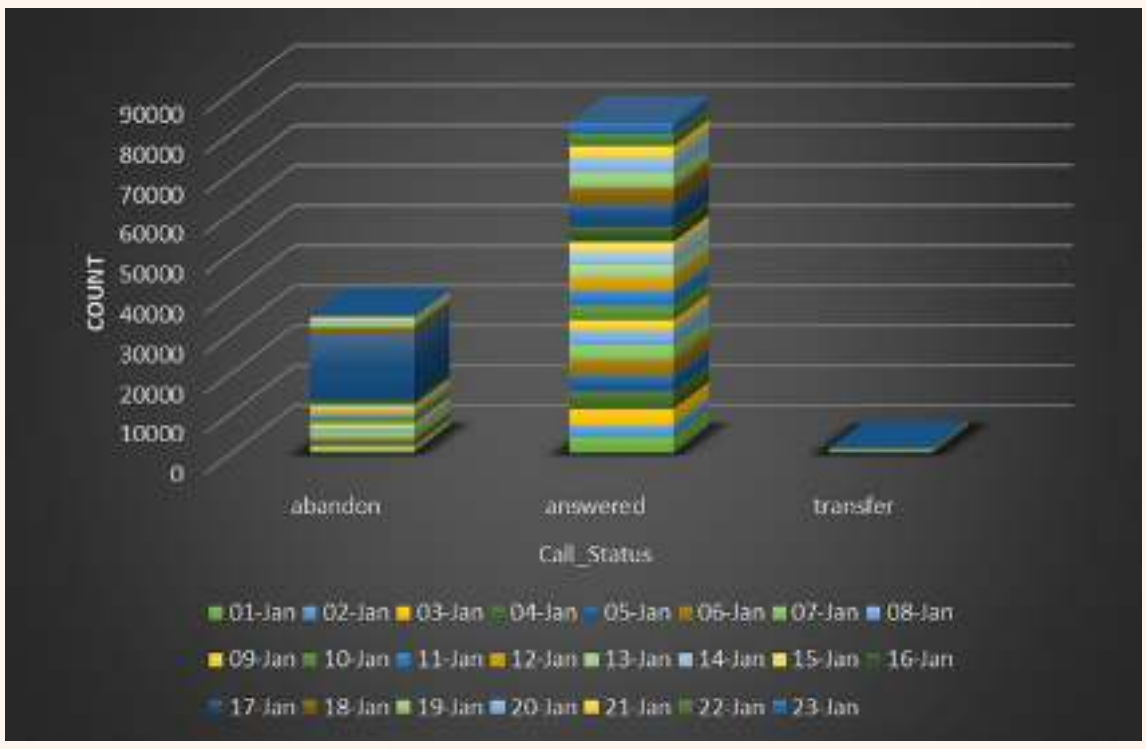
ASSUMPTIONS

NUMBER OF DAYS AGENT WORKS IN A WEEK	6 DAYS IN WEEK
UNPLANNED LEAVES PER AGENT	4 DAYS PER MONTH
AGENT TOTAL WORKING HOURS PER DAY	9 HOURS
SNACKS BREAK AND LUNCH TIME	1.5 HOURS
ACTUAL WORKING HOURS ON CUSTOMER CALL	4.5 HOURS

SOLUTION:

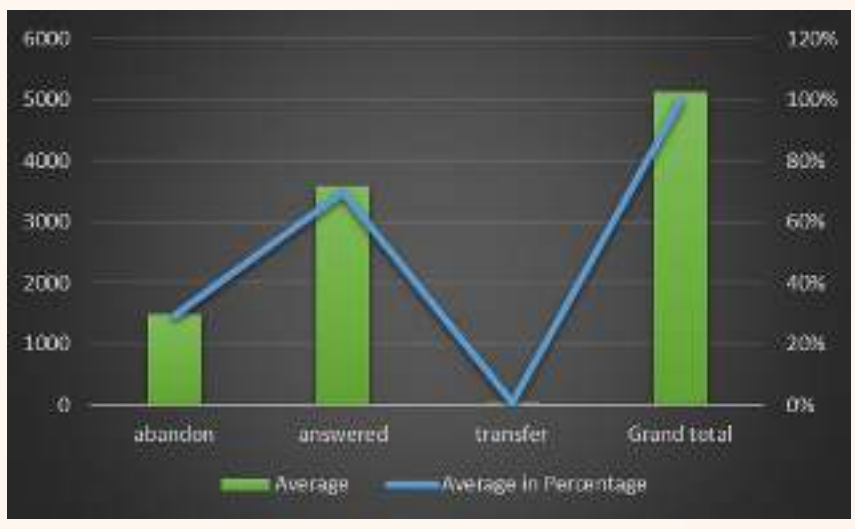
Count of Duration(hh:mm Call_Status				
Date	abandon	answered	transfer	Grand Total
01-Jan	684	3883	77	4644
02-Jan	356	2935	60	3351
03-Jan	599	4079	111	4789
04-Jan	595	4404	114	5113
05-Jan	536	4140	114	4790
06-Jan	991	3875	85	4951
07-Jan	1319	3587	42	4948
08-Jan	1103	3519	50	4672
09-Jan	962	2628	62	3652
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

STACKED COLUMN CHART



From pivot table, calculate call status averages. Then, convert these averages into percentages by dividing each by the total sum and multiplying by 100 to understand the distribution.

Call_Status	abandon	answered	transfer	Grand total
Average	1496	3585	49	5130
Average in Percentage	29%	70%	1%	100%



Average Call duration (in Hours)	198.62
Working hours per agent (in Hours)	4.5
Total Agents	44

The calculation reveals that there are a total of 44 agents working, based on an average of 198.82 call hours per day in the company and each agent working a 4.5-hour shift.

To achieve a 10% abandon rate, the number of agents required can be calculated using the formula:

For 90 % attending calls need (in Hours)	254.7293904
Total agents needed to reduce abandon rate the 10%	57

- For 90% hours = 5130(Avg of Grand Total)* 198.62(Avg Call_Duration)* 0.9 /3600
- Total agents needed = 254.7293904/4.5(Working hours of per agent)

AGENTS REQUIRED IN EACH TIME BUCKET

Time_Bucket	Count of Call_Seconds (s)	Required_Agents
10_11	11.28%	6
11_12	12.40%	7
12_13	10.72%	6
13_14	9.80%	6
14_15	8.95%	5
15_16	7.76%	4
16_17	7.45%	4
17_18	7.23%	4
18_19	6.13%	3
19_20	5.48%	3
20_21	4.67%	3
9_10	8.13%	3
Grand Total	**100.00%**	**57**

DEDUCTION FROM TASK 3

- **Average Call Duration:** The average call duration is approximately 198.62 hours, providing a baseline for call handling times.
- **Working Hours per Agent:** Each agent works for 4.5 hours, and there are a total of 44 agents, forming the available workforce.
- **90% Call Attendance Requirement:** To maintain a 90% call attendance rate, 254.73 hours are required, setting a service level target.
- **Total Agents Required:** To achieve a 10% abandon rate and ensure efficient service, 57 agents are needed.
- **Time Bucket Analysis:** The data shows varying call volumes across different time

buckets, indicating the need for effective resource allocation.Peak hours require more agents, while off-peak hours allow for reduced staffing.

- **Total Call Distribution:** In total, 57 agents are needed to meet service level goals.

By adhering to the recommended manpower plan,the call center can enhance the utilization of its agent resources, enhance overall efficiency, and deliver improved customer service. This can be accomplished by decreasing the abandon rate and increasing the rate of answered calls within the designated time frame.

TASK 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 9 pm, they also make 30 calls at night between 9 pm and 9 am.

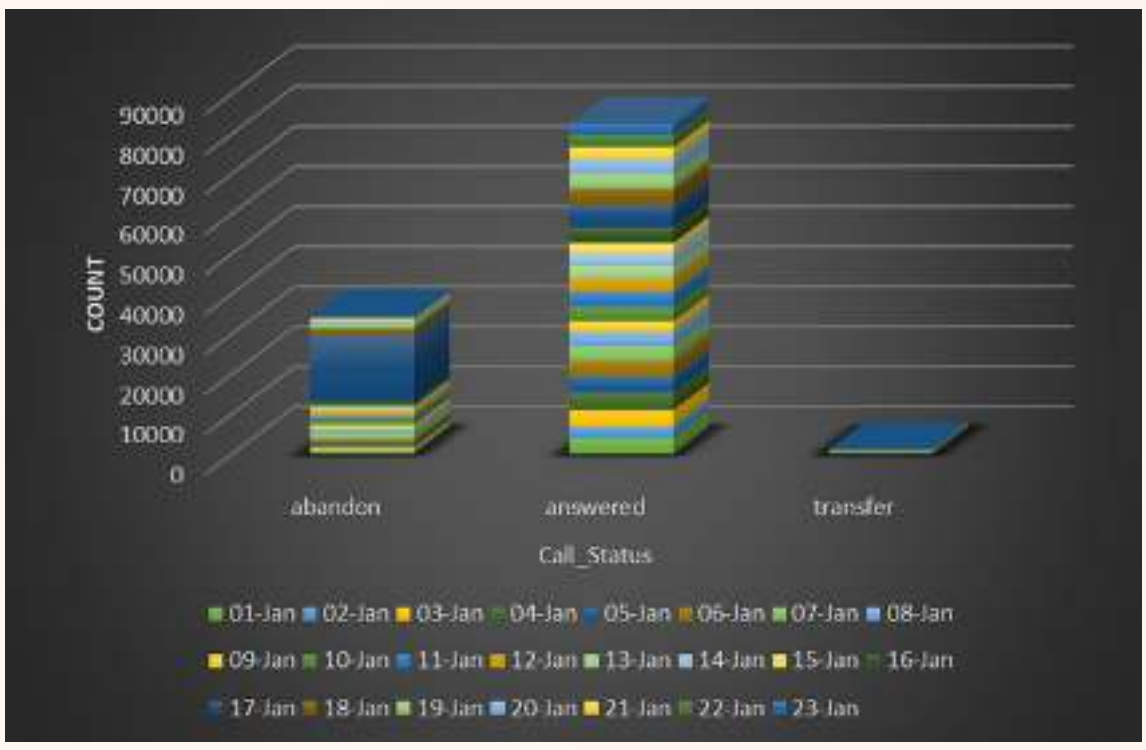
Distribution of 30 calls coming in night for every 100 calls coming in between 9am - 9pm (i.e. 12 hrs slot)											
9pm- 10pm	10pm - 11pm	11pm- 12am	12am- 1am	1am - 2am	2am - 3am	3am - 4am	4am - 5am	5am - 6am	6am - 7am	7am - 8am	8am - 9am
3	3	2	2	1	1	1	1	3	4	4	5

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

SOLUTION:

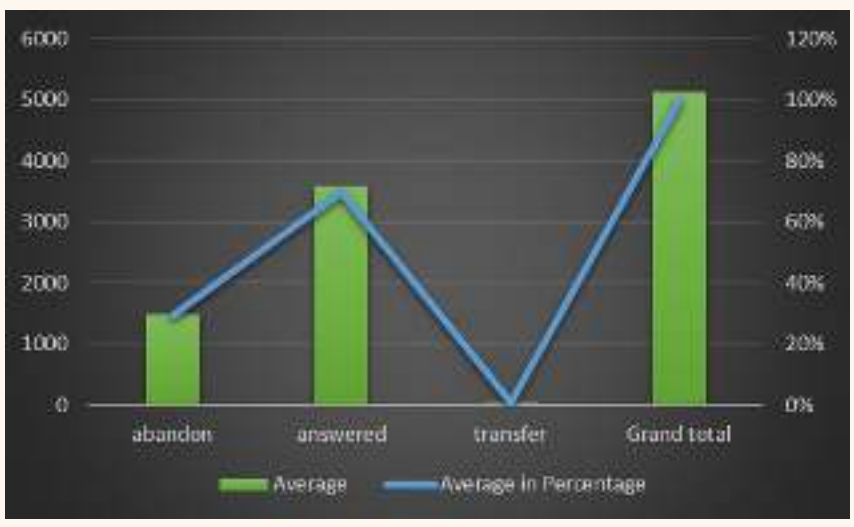
Count of Duration(hh:mm Call_Status				
Date	abandon	answered	transfer	Grand Total
01-Jan	684	3883	77	4644
02-Jan	356	2935	60	3351
03-Jan	599	4079	111	4789
04-Jan	595	4404	114	5113
05-Jan	536	4140	114	4790
06-Jan	991	3875	85	4951
07-Jan	1319	3587	42	4948
08-Jan	1103	3519	50	4672
09-Jan	962	2628	62	3652
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

STACKED COLUMN CHART



From pivot table, calculate call status averages. Then, convert these averages into percentages by dividing each by the total sum and multiplying by 100 to understand the distribution.

Call_Status	abandon	answered	transfer	Grand total
Average	1496	3585	49	5130
Average in Percentage	29%	70%	1%	100%



Average Call duration (in Hours)	198.62
Working hours per agent (in Hours)	4.5
Total Agents	44

The calculation reveals that there are a total of 44 agents working, based on an average of 198.82 call hours per day in the company and each agent working a 4.5-hour shift.

CALCULATION

Average no of calls at night	1539
Additional hours required	54
Agents needed in night	12

- **Number of Nighttime Calls:** Number of nighttime calls = 30% of total calls
- Number of nighttime calls = 0.30 * 5130 = 1539 calls
- **Additional Hours Needed to Achieve a 10% Abandon Rate:** Additional hours needed = (Number of nighttime calls) * (Average call duration / Total calls) * (1 - Abandon Rate)
- Additional hours needed ≈ 1539 * (198.82 / 5130) * 0.9 ≈ 54 hours
- **Number of Agents Required:** Number of agents required = Additional hours needed / Working hours per agent
- Number of agents required ≈ 54 / 4.5 ≈ 12 agents

AGENTS REQUIRED IN EACH TIME BUCKET

Time_Bucket	Call_Distribution	Percentage of Call_Distribution	Agents needed	Final Agents needed
9pm - 10pm	3	10	1	1
10pm - 11pm	3	10	1	1
11pm - 12pm	2	6.67	1	1
12am - 1am	2	6.67	1	1
1am - 2am	1	3.33	0.4	1
2am - 3am	1	3.33	0.4	1
3am - 4am	1	3.33	0.4	1
4am - 5am	1	3.33	0.4	1
5am - 6am	3	10	1	1
6am - 7am	4	13.33	2	2
7am - 8am	4	13.33	2	2
8am - 9am	5	16.67	2	2
Grand Total	30	100	12	15

DEDUCTION FROM TASK 4

- **Average Calls:** On average, the company receives 5130 calls from 9 am to 9 pm.
- **Calls at Night:** During the night, between 9 pm and 9 am, about 30% of the calls are made. This means there are 1539 calls during this time.
- **Additional Hours:** To make sure only 10% of calls are abandoned, the company needs 54 extra hours of agent availability during the night period.
- **Number of Agents:** When we divide the 54 extra hours needed by the hours worked by one agent (4.5 hours), we find that about 12 agents are required.
- **Rounding Off:** Since we can't have a fraction of an agent, we round up to the nearest whole number, which is 15 agents needed to meet the staffing requirements.
- In conclusion, the plan suggests having around 15 agents available throughout the day, considering the different call patterns and customer needs. This plan aims to reduce abandoned calls and improve the customer experience both during the day and night

CONCLUSION

- **Insights on customer challenges:** Identified issues like long queue times, call transfers, and abandoned calls.
- **Peak call hours:** Discovered busiest call times for better staffing and resource allocation during high-demand periods.
- **Agent performance evaluation:** Recognized top-performing agents and areas for improvement based on call duration, resolution, and customer satisfaction.
- **Queue time analysis:** Pinpointed peak waiting periods and provided recommendations to reduce wait times and enhance customer satisfaction.
- **Optimal staffing strategies:** Developed strategies based on call volume and demand patterns to lower call abandon rates and ensure prompt customer service.
- In summary, these insights ultimately resulting in improved customer satisfaction and operational efficiency.

APPENDIX

PDF LINK FOR DATA ANALYTICS PROCESS :

https://drive.google.com/file/d/1rtnUEPJMfxiZjobq8TpcSufG6C5iqCda/view?usp=sharing

PDF LINK FOR INSTAGRAM USER ANALYTICS :

https://drive.google.com/file/d/14-tC4EXMS3RmBN7QbFKSM2zb_nc7rdtw/view?usp=sharing

PDF LINK FOR OPERATION AND METRIC ANALYTICS :

https://drive.google.com/file/d/1jztgFlV-ojDBcU5YfsvcOXlHNCo9HDau/view?usp=sharing

PDF LINK FOR HIRING PROCESS ANALYTICS :

https://drive.google.com/file/d/1PhW-ZBO7C9s3M64BMAoxRoDG3UsEfGT7/view?usp=sharing

PDF LINK FOR IMDB MOVIE ANALYSIS :

https://drive.google.com/file/d/1h-29_6GvwI8OG9J3FRg3TJyXJywyPwfj/view?usp=sharing

EXCEL LINK FOR IMDB MOVIE ANALYSIS :

https://docs.google.com/spreadsheets/d/1zyNpfJDv3mV1f2e4XemmXGzmi7WM0SZx/edit?usp=sharing&ouid=101394161962274505358&rtpof=true&sd=true

PDF LINK FOR BANK LOAN CASE STUDY :

https://drive.google.com/file/d/1qO3L0AtMYkRGm3N51635Gd-R1sGXXnHp/view?usp=sharing

COLAB LINK FOR BANK LOAN CASE STUDY :

https://colab.research.google.com/drive/1sLDyAIV_pdpDSSw4xH5drnFtJLVaAY4X?usp=sharing

PDF LINK FOR IMPACT OF CAR FEATURES :

https://drive.google.com/file/d/1-BKeKZ7jrxPh8-UmE6OSpQwwibWjR258/view?usp=sharing

EXCEL LINKS FOR IMPACT OF CAR FEATURES :

https://docs.google.com/spreadsheets/d/1ph-sMPsxnNSvODqkF4C4sTSuLPFUHH49/edit?usp=sharing&ouid=101394161962274505358&rtpof=true&sd=true

PDF LINK FOR CALL VOLUME TREND ANALYSIS :

https://drive.google.com/file/d/16FQ3_SzLLgPy1Na0yl_IBDAXkoiPVi_b/view?usp=sharing

EXCEL LINK FOR CALL VOLUME TREND ANALYSIS :

https://drive.google.com/drive/folders/14O-I--5kdo4cSmcTiM_p7wTV0OTnbMAI?usp=drive_link