PORTFOLIO

Professional background:

I am working as a trainee for a data analytics position at "Trainity" at the moment. This role is new to me. Before this, I spent six months working as a "BPO" for DXC Technology as a "data technician". I have strong communication skills and am proficient in **Excel, SQL, data visualization, and POWER BI** (intermediate). I worked with "Trainity" on eight real-time projects. I would love to learn about the actual issues faced by the business sector as a new employee and gain insight into how things operate there. Even though I'm new, I can learn new things extremely easily and adaptably. I possess a theoretical understanding. I'm holding off on putting my theoretical understanding to use, though. Additionally, I shall learn by making a major effort.

Table of contents

Project-1 Instagram User Analytics

PROJECT DESCRIPTION:

The **overview** of this project is to analyze user engagement and interactions with the Instagram app to produce insightful data that will aid in the expansion of the Instagram app. User analysis is used to track how users interact with digital products like software applications or mobile apps.

Problem Statement:

Analyzing user interactions and engagements with the Instagram app to provide valuable insights that can help to grow the business.

Approach:

I created a necessary database for the project by running necessary SQL commands and added values to it provided by the management team using MySQL workbench in the MySQL database. After the database was created, I used to run SQL queries in MySQL workbench and, I used to extract the necessary insights from the database table.

I connected to MySQL workbench as a software and server with the service name mysql@localhost:3306 with version- 8.0.23 with MySQL community server – GPL. I used MySQL workbench because it is an open-source relational database design tool. And it is a friendly user interface**.**

FINDINGS:

A) MARKETING ANALYSIS

1. Loyal user reward: Identify the five oldest users on Instagram from the provided database.

SQL query:

```
USE ig_clone;
SELECT
    username, created_at
FROM
    users
ORDER BY created_at
LIMIT 5;
```

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

2. Inactive user engagement: Identify users who have never posted a single photo on Instagram.

SQL query:

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

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

3. Contest winner declaration: determine the winner of the contest and provide their details to the team.

SQL query:

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

id	image_url	total_likes
145	https://jarret.name	48

4. Hashtag research: identify and suggest the top five most used hashtags on the platform.

SQL query:

```
SELECT
    tags.tag_name, COUNT(*) AS no_of_hastags
FROM
    photo_tags
        JOIN
    tags ON photo_tags.tag_id = tags.id
GROUP BY tags.id
ORDER BY no_of_hastags desc
LIMIT 5;
```

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

5. Ad campaign launch: determine the day of the week when most users register on Instagram. Provide insights on when to schedule an ad campaign**.**

SQL query:

```
SELECT
    DAYNAME(created_at) AS `day`, COUNT(*) AS total_reg
FROM
    users
GROUP BY `day`
ORDER BY total_reg DESC
LIMIT 2;
```

day	total_reg
Thursday	16
Sunday	16

B) INVESTOR METRICS:

1. user engagement: calculate the average number of posts per user on Instagram. also, provide the total number of photos on Instagram divided by the total number of users**.**

SQL query:

```
SELECT
    (SELECT
            COUNT(id)
        FROM
            photos) / (SELECT
            COUNT(DISTINCT user_id)
        FROM
            photos) AS avg_posts,
    (SELECT
            COUNT(id)
        FROM
            photos) / (SELECT
            COUNT(id)
        FROM
            users) AS ratio_pho_users;
```

avg_posts	ratio_pho_users
3.4730	2.5700

2. BOTS & FAKE ACCOUNTS: Identify users who have liked every single photo on the site.

SQL query:

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

username	num_likes
Aniya_Hackett	257
Bethany20	257
Duane60	257
Jaclyn81	257
Janelle.Nikolaus81	257
Julien_Schmidt	257
Leslie67	257
Maxwell.Halvorson	257
Mckenna17	257
Mike.Auer39	257
Nia_Haag	257
Ollie_Ledner37	257
Rocio33	257

Analysis & Insights:

By using SQL queries, I extracted meaningful insights from the database. Through the result table, I tracked how users engage and interact with our digital platform Instagram app. Also, I tried to produce meaningful business insights for various teams such as marketing, product, and development teams.

Marketing analysis

• From marketing analysis insights, the Instagram team can give a reward to the most loyal users. Those who have been using the platform for the longest time.

• The marketing team can start sending emails to encourage inactive users.

• By conducting a contest, users can declare a winner to those who got the most likes on a single photo on Instagram.

• By researching the most popular hashtags on Instagram that which a partner brand can use in their post to reach the most people.

• By conducting an ad campaign launch can suggest the marketing team schedule an ad campaign at most users register on the day of the weak.

Investor metrics

• On calculating the average number of posts for users on Instagram, the investor team can know the user engagement on the Instagram app.

• By identifying bots & fake accounts, through this investor team can control or delete fake and dummy accounts**.**

CONCLUSION:

These meaningful insights can help the product manager and the rest of the team to make informed decisions about the future direction of the Instagram app.

Project-2 Operation Analytics & Investigating Metric Spike

Project description:

Analyzing a company's end-to-end operations is a critical step in the process of operational analytics. this analysis helps identify areas of improvement within the company. metric spike investigation is one of the main applications of operational analytics. this entails figuring out why and how important metrics suddenly change. as a decline in sales or a reduction in daily user engagement.

Problem statement:

Operational analytics is a crucial process that involves analyzing a company's end-to-end operations. One of the key aspects is investigating metric spikes. As a data analyst, it involves understanding and explaining sudden changes in key metrics such as a dip in daily user engagement or a drop in sales.

Approach:

I used SQL commands to create the databases that were required for the project.

These are:1 oaim_spike; 2 ope_ana_spike.

Case study 1: uses the first database oaim_spike, to work on job data analysis using job-data dataset. I used the appropriate SQL queries to add values to the table.

Case study 2: uses the second database ope_ana_spike, to work on the metric spike investigation with the "users"," events", and" email_events" tables. After that imported csv file into my SQL workbench to add values to the tables. after it has been created the necessary insights are derived from the tables by running SQL queries.

I connected to MySQL workbench as a software and server to do data analysis

FINDINGS:

Case Study 1 – Job Data Analysis

A. Jobs Reviewed Over Time:

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

Query:

```
SELECT
    ds AS date,
    COUNT(job_id) AS total_jobs_reviewed_per_hour,
    SUM(time_spent) / 3600 AS total_hours_spent_per_day
FROM
    job_data
GROUP BY ds
ORDER BY ds;
```

date	total_jobs_reviewed_per_hour	total_hours_spent_per_day
2020-11-25	1	0.0125
2020-11-26	1	0.0156
2020-11-27	1	0.0289
2020-11-28	2	0.0092
2020-11-29	1	0.0056
2020-11-30	2	0.0111

B. Throughput Analysis:

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

Query:

```
SELECT
    ds AS 'date',
    ROUND(COUNT(event) / SUM(time_spent), 3) AS daily_throughtput,
    (SELECT
            ROUND(COUNT(event) / SUM(time_spent), 3)
        FROM
            job_data) AS 7day_rolling_avg
FROM
    job_data
GROUP BY ds
ORDER BY ds;
```

date	daily_throughtput	7day_rolling_avg
2020-11-25	0.022	0.027
2020-11-26	0.018	0.027
2020-11-27	0.010	0.027
2020-11-28	0.061	0.027
2020-11-29	0.050	0.027
2020-11-30	0.050	0.027

C. Language Share Analysis: Task: write an SQL query to calculate the percentage share of each language over the last 30 days.

Query:

```
SELECT
    language,
    SUM(time_spent) / (SELECT
            SUM(time_spent)
        FROM
            job_data) * 100 AS percentage_share
FROM
    job_data
GROUP BY language
ORDER BY 2 DESC;
```

language	percentage_share
french	34.8993
persian	32.8859
italin	15.1007
arabic	8.3893
english	5.0336
hindi	3.6913

D. Duplicate Rows Detection:

Task: write an SQL query to display duplicate rows from the job_data table.

Query:

```
SELECT
    *
FROM
    job_data
WHERE
    job_id IN (SELECT
            job_id
        FROM
            job_data
        GROUP BY job_id
        HAVING COUNT(*) > 1)
ORDER BY ds;
```

ds	job_id	actor_id	event	language	time_spent	org
2020-11-26	23	1004	skip	persian	56	A
2020-11-28	23	1005	transfer	persian	22	D
2020-11-29	23	1003	decision	persian	20	C

Case Study 2 - Investigating Metric Spike:

A. Weekly User Engagement:

Task – write an SQL query to calculate the weekly user engagement.

Query:

```
SELECT
    EXTRACT(WEEK FROM occured_at) AS week_number,
    COUNT(DISTINCT user_id) as activeness_users
FROM
    `events`
GROUP BY week_number;
```

Result Table:

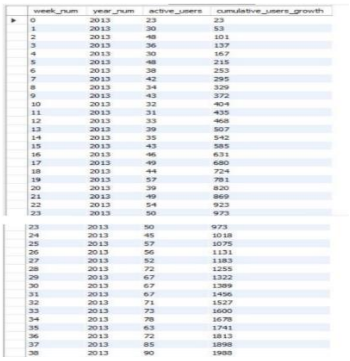
week_number	activeness_users
17	663
18	1068
19	1113
20	1154
21	1121
22	1186
23	1232
24	1275
25	1264
26	1302
27	1372
28	1365
29	1376
30	1467
31	1299
32	1225
33	1225
34	1204
35	104

B. User Growth Analysis:

Task – write an SQL query to calculate the user growth for the product.

Query:

```
select week_num,year_num,active_users,sum(active_users)
over(order by week_num,year_num) as cumulative_users_growth
from
(
select  extract(week from created_at)as week_num,
extract(year from created_at)as year_num,
count(distinct user_id)as active_users from users
WHERE state ='active'
group by week_num,year_num
ORDER BY week_num,year_num
);
```



it continues……

C. Weekly Retention Analysis:

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

Query:

```sql
SELECT first AS 'week_number',
    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'
FROM
(SELECT
        m.user_id,
            m.signup_week,
            n.first,
            m.signup_week - n.first AS week_number
    FROM
        (SELECT
        user_id, EXTRACT(WEEK FROM occured_at) AS signup_week
    FROM
        events
    GROUP BY 1 , 2) m, (SELECT
        user_id, MIN(EXTRACT(WEEK FROM occured_at)) AS first
    FROM
        events
    GROUP BY 1) n
    WHERE
        m.user_id = n.user_id) sub
GROUP BY first
ORDER BY first;
```

Result Table:

week_number	Week 0	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
17	663	472	324	251	205	187	167	146	145	145	136	131	132	143	116	91	82	77	5
18	596	362	261	203	168	147	144	127	113	122	106	118	127	110	97	85	67	4	0
19	427	284	173	153	114	95	91	81	95	82	68	65	63	42	51	49	2	0	0
20	358	223	165	121	91	72	63	67	63	65	67	41	40	33	40	0	0	0	0
21	317	187	131	91	74	63	75	72	58	48	45	39	35	28	2	0	0	0	0
22	326	224	150	107	87	73	63	60	55	48	41	39	31	1	0	0	0	0	0
23	328	219	138	101	90	79	69	61	54	47	35	30	0	0	0	0	0	0	0
24	339	205	143	102	81	63	65	61	38	39	29	0	0	0	0	0	0	0	0
25	305	218	139	101	75	63	50	46	38	35	2	0	0	0	0	0	0	0	0
26	288	181	114	83	73	55	47	43	29	0	0	0	0	0	0	0	0	0	0
27	292	199	121	106	68	53	40	36	1	0	0	0	0	0	0	0	0	0	0
28	274	194	114	69	46	30	28	3	0	0	0	0	0	0	0	0	0	0	0
29	270	186	102	65	47	40	1	0	0	0	0	0	0	0	0	0	0	0	0
30	294	202	121	78	53	3	0	0	0	0	0	0	0	0	0	0	0	0	0
31	215	145	76	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	267	188	94	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	286	202	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	279	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

D. Weekly Engagement Per Device:

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

Query:

```sql
SELECT
    EXTRACT(WEEK FROM occured_at) AS week,
    EXTRACT(YEAR FROM occured_at) AS year,
    device,
    COUNT(DISTINCT user_id) AS weekly_engagement_device
FROM
    events
WHERE
    event_type = 'engagement'
GROUP BY week , year , device
ORDER BY week , year , device;
```

Result Table:

week	year	device	weekly_engagement_device
17	2014	acer aspire desktop	9
17	2014	acer aspire notebook	20
17	2014	amazon fire phone	4
17	2014	asus chromebook	21
17	2014	dell inspiron desktop	18
17	2014	dell inspiron notebook	46
17	2014	hp pavilion desktop	14
17	2014	htc one	16
17	2014	ipad air	27
17	2014	ipad mini	19
17	2014	iphone 4s	21
17	2014	iphone 5	65
17	2014	iphone 5s	42
17	2014	kindle fire	6
17	2014	lenovo thinkpad	86
17	2014	mac mini	6
17	2014	macbook air	54
17	2014	macbook pro	143
17	2014	nexus 10	16
17	2014	nexus 5	40
17	2014	nexus 7	18
17	2014	nokia lumia 635	17
17	2014	samsumg galaxy tablet	8
17	2014	samsung galaxy note	7
17	2014	samsung galaxy s4	52
17	2014	windows surface	10
18	2014	acer aspire desktop	26
18	2014	acer aspire notebook	33
18	2014	amazon fire phone	9
18	2014	asus chromebook	42
18	2014	dell inspiron desktop	58
18	2014	dell inspiron notebook	77
18	2014	hp pavilion desktop	37
18	2014	htc one	19
18	2014	ipad air	52
18	2014	ipad mini	30
18	2014	iphone 4s	46
18	2014	iphone 5	113
18	2014	iphone 5s	73
18	2014	kindle fire	27
18	2014	lenovo thinkpad	153
18	2014	mac mini	13
18	2014	macbook air	121
18	2014	macbook pro	252
18	2014	nexus 10	30
18	2014	nexus 5	73
18	2014	nexus 7	30
18	2014	nokia lumia 635	33
18	2014	samsumg galaxy tablet	11
18	2014	samsung galaxy note	15
18	2014	samsung galaxy s4	82

still result continues……

E. Email Engagement Analysis:

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

Query:

```
SELECT
    100 * SUM(CASE
        WHEN email_category = 'email_open' THEN 1
        ELSE 0
    END) / SUM(CASE
        WHEN email_category = 'email_sent' THEN 1
        ELSE 0
    END) AS email_open_rate,
    100 * SUM(CASE
        WHEN email_category = 'email_click' THEN 1
        ELSE 0
    END) / SUM(CASE
        WHEN email_category = 'email_sent' THEN 1
        ELSE 0
    END) AS email_click_rate
FROM (SELECT *,
          CASE
              WHEN action IN ('sent_weekly_digest' , 'sent_reengagement_email') THEN 'email_sent'
              WHEN action IN ('email_open') THEN 'email_open'
              WHEN action IN ('email_clickthrough') THEN 'email_click'
          END AS email_category
    FROM   email_events) AS t;
```

email_open_rate	email_click_rate
33.5834	14.7899

Analysis & insights:

Here are some of the insights I have drawn by analyzing the data.

- I have calculated no. of jobs reviewed over time for November month. More jobs were reviewed on November 28 &30 compared to any other dates.
- I preferred to calculate the 7-day rolling average of throughput over the daily metric. The reason is the daily measurements are subject to fluctuations. Therefore, since the 7-day rolling average is less affected by the aforementioned factors and can provide a realistic sense of the data.
- The percentage share of each language over the last 30 days says that the **French** language has the longest share of 34.8993 and the **Hindi** language has the lowest percentage share of 3.6913. also, I found three actor- IDs with duplicate rows present in the data.
- When calculating weekly user engagement, I found that, at week 35, there are 104 fewer active user engagement. And at week 30, 1467 active users were engaged which is the highest engagement.
- The product's user growth in the 50th week of 2013 with 124 active users has the best. And 18 users were the least active in the 35th week of 2014.
- Based on the sign-up cohort, there were more retained users on the 17th week number compared to other week numbers.
- On the 17th week of 2014, the MacBook Pro had the greatest engagement rate with 302 users.
- The email engagement metrics i.e., the email open rate is 33.583 % and the email click rate is 14.789%.

Conclusion:

Using my extensive SQL skills, I was able to extract valuable information from two databases. It is possible to enhance business operations and comprehend abrupt shifts in important metrics by using result tables.

This project has improved my understanding of the value of operational analytics. This effort has helped me to comprehend how the business employs metric spikes as a weapon of mass destruction. Management can use insights to create data-driven decisions by taking a proactive and educated approach.

PROJECT – 3 Hiring Process Analytics

Project description:

A company hiring procedure is an essential component. It may have an impact on the organization's total productivity. This results in a company's expansion. Gaining insights into hiring patterns, such as the quality of interviews, job kinds, number of hires/rejections, and openings, can be beneficial for the hiring department.

Problem statement:

The hiring process is a crucial function of a company and understanding trends, analyzing the company's hiring process, and drawing meaningful insights to answer the questions that can help the company improve the hiring process.

Data cleaning & Approach:

I started by downloading the dataset that the management had provided. Then by eliminating duplicate entries, and outliers this way I clean the data. To analyze data bar graphs, pivot tables, excel formulas, functions, and statistical measurements were all utilized in Excel. My understanding of insights is made easier by this analysis.

The conclusions drawn from this analysis are documented in Microsoft Word and saved as PDF. I used Microsoft excel for drawing insights and do analysis.

FINDINGS:

A. Hiring Analysis:

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

B. Salary Analysis

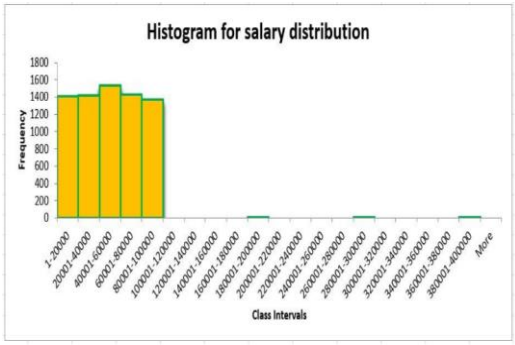
Task- What is the average salary offered by this company? Use Excel functions to calculate this.

Departmental-wise average salary:

C. Salary distribution:

Task- Create class intervals for the salaries in the company. This will help us to understand the salary distribution

Class Intervals	Salary Distribution
1-20000	1410
20001-40000	1421
40001-60000	1531
60001-80000	1432
80001-100000	1370
100001-120000	0
120001-140000	0
140001-160000	0
160001-180000	0
180001-200000	1
200001-220000	0
220001-240000	0
240001-260000	0
260001-280000	0
280001-300000	1
300001-320000	0
320001-340000	0
340001-360000	0
360001-380000	0
380001-400000	1
More	0



D. Departmental analysis:

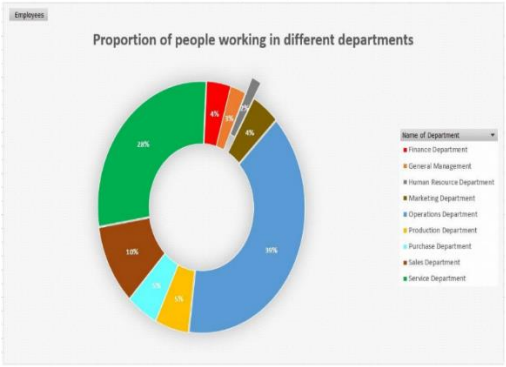
Task- Use a pie chart, bar graph, or any other suitable visualization to show the proportion of people working in different departments.

To find no. of employees working in each department I used the below formula for different department names.

```
=COUNTIFS(E:E,"Service Department",C:C,"Hired")
=COUNTIFS(E:E,"Operations Department",C:C,"Hired")
=COUNTIFS(E:E,"Purchase Department",C:C,"Hired")
=COUNTIFS(E:E,"Finance Department",C:C,"Hired")
=COUNTIFS(E:E,"Marketing Department",C:C,"Hired")
=COUNTIFS(E:E,"Human Resource Department",C:C,"Hired")
=COUNTIFS(E:E,"Production Department",C:C,"Hired")
=COUNTIFS(E:E,"Sales Department",C:C,"Hired")
=COUNTIFS(E:E,"General Management",C:C,"Hired")
```

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

Row Labels	No.of Employees	
Finance Department	176	4%
General Management	113	3%
Human Resource Department	70	2%
Marketing Department	202	4%
Operations Department	1843	39%
Production Department	246	5%
Purchase Department	230	5%
Sales Department	485	10%
Service Department	1332	28%
Grand Total	**4697**	



E. Position Tier Analysis:

Task- use a chart or graph to represent different position tiers within the company. This will help us to understand the distribution of positions across the tiers.

For finding different position tiers I used the formula below for different tier names.

```
=COUNTIFS(F:F,"b9",C:C,"Hired")
```

Post Name	Employees
-	1
b9	308
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
m7	0
n10	0
n6	1
n9	0
Grand Total	**4697**

Analysis & Insights:

- While calculating the gender distribution of hires the company employed 1856 women and 2563 men out of a total of 7167 applicants.
- The average salary offered by the company is 49983.03/- and I also calculated department-wise average salary.
- The salary distribution among employees in the organization is highly correlated with the class interval range between 40001 – 60000.
- The largest proportion of employees working in the operation department was 39% In terms of numbers, 1843 employees are working out of 4697. Out of nine departments, the human resource department had the fewest employees (2%), with 70 people.
- The distribution of position across different tiers. It clearly states that there were 1239 individuals working under post designation 'c9'. And it goes as follows in descending order c5, i7, i5, i6, b9, c8, i1, c-10, i4, m6, n6, -.

Conclusion:

I was able to extract significant information from the given dataset hiring process by utilizing my extensive Excel abilities. And I gain valuable insights from it. This project taught me how to deal with missing data, and outliers, and detect outliers from a given dataset. With the use of Excel functions, pivot tables, charts, bar graphs, and pie charts, I gained valuable insights from this project. I also learned how a data analyst may assist a company in analyzing data from the hiring process and understanding trends such as the number of rejections, hires, interviews, and job kinds for the hiring department. Using this analysis, the hiring department may make important decisions for the company's growth.

PROJECT – 4 IMDB Movie Analysis

Project Description:

The data supplied is related to IMDB movies. One such problem to study is "What factors influence the success of a movie on IMDB?". In this context, success is determined by high IMDB ratings. This issue has a huge impact on movie producers, directors, and financiers who want to understand what makes a film successful so that they can make informed judgments about future projects.

Problem statement:

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?".

Data cleaning & Approach:

❖ First, I downloaded the dataset and made a copy of it. Then, I read over the problem statement and tasks that I needed to complete to gain insights. The given dataset was imported into Excel. After importing the data, I cleaned it up to make it acceptable for analysis.

❖ It includes removing unneeded data for specific tasks- The initial dataset had 28 columns and 5044 rows. All columns are not required to obtain insights. So, I removed unnecessary columns and decreased to 8 columns. I also added 2 new columns, one for genre(separated), which is divided using flash fill, and another for profit. I now have 10 columns.

❖ Then after I deleted duplicate rows using the Remove Duplicates from data tab.

❖ After that, I used 'find & select' followed by 'go to special 'to handle all empty rows before selecting the 'blanks' option. This operation highlighted all empty rows. Afterward, I used the shortcut 'ctrl + - 'to choose the full rows option. This approach resulted in the removal of all empty entries.

❖ from the dataset. Finally, now have a cleaned dataset with 10 columns and 3790 rows.

❖ With the cleaned dataset, I began looking for insights into certain jobs. To do the analysis, I utilized Excel 365 and pivot tables, graphs, charts, and other tools to visualize data and create patterns. After that, I concluded and offered recommendations based on the patterns I discovered, and I also used Microsoft Word to create a complete project report.

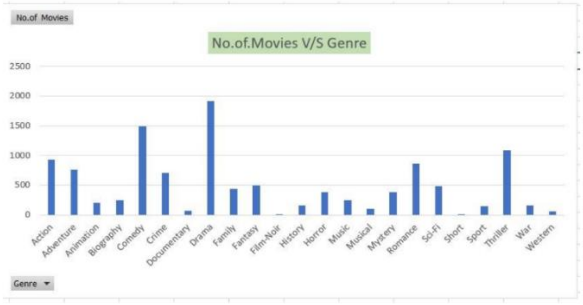
FINDINGS:

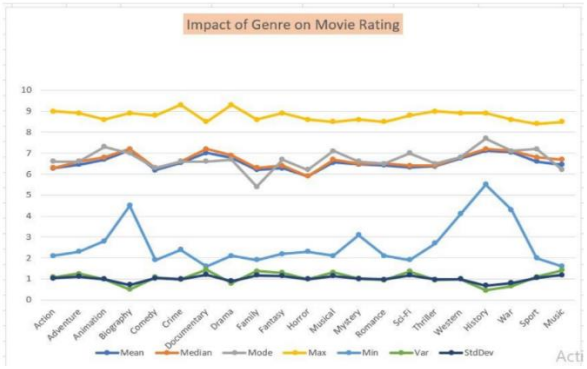
A. Movie Genre Analysis: Analyse the distribution of movie genres and their impact on IMDB score.

Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics.

The most prevalent movie genres are mentioned below along with descriptive statistics.

Genre	No.of Movies	Mean	Median	Mode	Max	Min	Var	StdDev
Action	935	6.285989305	6.3	6.6	9	2.1	1.078186788	1.038357736
Adventure	766	6.454960836	6.6	6.6	8.9	2.3	1.247524378	1.116926308
Animation	197	6.700507614	6.8	7.3	8.6	2.8	0.987295659	0.993627525
Biography	242	7.140082645	7.2	7	8.9	4.5	0.504237338	0.71009671
Comedy	1492	6.183310992	6.3	6.3	8.8	1.9	1.081431552	1.039919012
Crime	703	6.545661451	6.6	6.6	9.3	2.4	0.97143058	0.98561178
Documentary	67	7.011940299	7.2	6.6	8.5	1.6	1.439855269	1.199939694
Drama	1914	6.787774295	6.9	6.7	9.3	2.1	0.795417626	0.891861887
Family	442	6.202262443	6.3	5.4	8.6	1.9	1.3670697	1.169217559
Fantasy	496	6.285080645	6.4	6.7	8.9	2.2	1.30054464	1.140414241
Horror	379	5.903957784	5.9	6.2	8.6	2.3	0.982127152	0.991023285
Musical	102	6.550980392	6.7	7.1	8.5	2.1	1.307672297	1.143535
Mystery	377	6.469496021	6.5	6.6	8.6	3.1	1.014838309	1.007391835
Romance	867	6.425490196	6.5	6.5	8.5	2.1	0.938321786	0.968670112
Sci-Fi	484	6.327272727	6.4	7	8.8	1.9	1.362318841	1.16718415
Thriller	1087	6.372309108	6.4	6.5	9	2.7	0.939112803	0.969078327
Western	60	6.756666667	6.8	6.8	8.9	4.1	0.982158192	0.991038946
History	153	7.122875817	7.2	7.7	8.9	5.5	0.460196938	0.678378168
War	159	7.048427673	7.1	7.1	8.6	4.3	0.652386753	0.80770462
Sport	147	6.601360544	6.8	7.2	8.4	2	1.09876526	1.048220043
Music	247	6.456680162	6.7	6.2	8.5	1	1.413359666	1.188848041
Short	2	6.8	6.8	#N/A	7.1	6.5	0.18	0.424264069
Film-Noir	1	7.7	7.7	#N/A	7.7	7.7	#DIV/0!	#DIV/0!

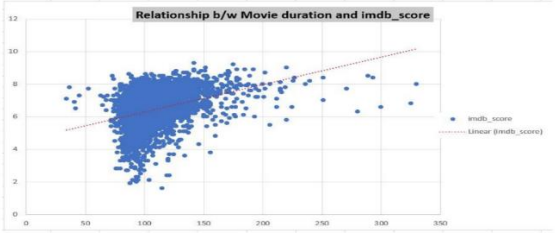




B. Movie Duration Analysis: analyze the distribution of movie durations and its impact on IMDB score.

Task: analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

operations	values
Mean	109.8029
Median	105
Mode	101
Variance	517.8906
Std Deviation	22.75721



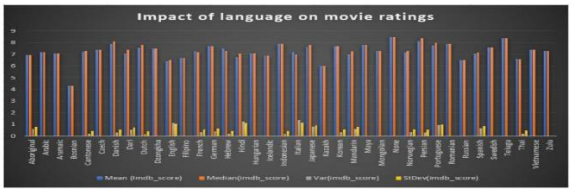
C. Language Analysis: situation: examine the distribution of movies based on their languages.

Task: determine the most common languages used in movies and analyze their

impact on IMDB score using descriptive statistics.

descriptive statistics for language analysis

Languages	No.of.Movies	Mean (Imdb_score)	Median(imdb_score)	Var(imdb_score)	StDev(imdb_score)
Aboriginal	2	6.95	6.95	0.605	0.777817459
Arabic	1	7.2	7.2	#DIV/0!	#DIV/0!
Aramaic	1	7.1	7.1	#DIV/0!	#DIV/0!
Bosnian	1	4.3	4.3	#DIV/0!	#DIV/0!
Cantonese	8	7.2375	7.3	0.194107143	0.440575922
Czech	1	7.4	7.4	#DIV/0!	#DIV/0!
Danish	3	7.9	8.1	0.28	0.529150262
Dari	2	7.08125	7.4	0.536291667	0.732319375
Dutch	3	7.566666667	7.8	0.163333333	0.404145188
Dzongkha	1	7.5	7.5	#DIV/0!	#DIV/0!
English	3606	6.421436495	6.5	1.107753941	1.052498903
Filipino	1	6.7	6.7	#DIV/0!	#DIV/0!
French	37	7.286486486	7.2	0.31509009	0.561328861
German	13	7.692307692	7.7	0.410769231	0.640912811
Hebrew	3	7.5	7.3	0.19	0.435889894
Hindi	10	6.76	7.05	1.236	1.111755369
Hungarian	1	7.1	7.1	#DIV/0!	#DIV/0!
Icelandic	1	6.9	6.9	#DIV/0!	#DIV/0!
Indonesian	2	7.9	7.9	0.18	0.424264069
Italian	7	7.185714286	7	1.334761905	1.155318962
Japanese	12	7.625	7.8	0.809318182	0.899621132
Kazakh	1	6	6	#DIV/0!	#DIV/0!
Korean	5	7.7	7.7	0.325	0.570087713
Mandarin	14	7.021428571	7.25	0.586428571	0.765786244
Maya	1	7.8	7.8	#DIV/0!	#DIV/0!
Mongolian	1	7.3	7.3	#DIV/0!	#DIV/0!
None	1	8.5	8.5	#DIV/0!	#DIV/0!
Norwegian	4	7.15	7.3	0.33	0.574456265
Persian	3	8.133333333	8.4	0.303333333	0.550757055
Portuguese	5	7.76	8	0.958	0.978774744
Romanian	1	7.9	7.9	#DIV/0!	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!	#DIV/0!
Spanish	26	7.05	7.15	0.6826	0.826196103
Swedish	1	7.6	7.6	#DIV/0!	#DIV/0!
Telugu	1	8.3	8.3	#DIV/0!	#DIV/0!
Thai	3	6.633333333	6.6	0.203333333	0.450924975
Vietnamese	1	7.4	7.4	#DIV/0!	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!	#DIV/0!



D. Director analysis: influence of directors on movie ratings**.**

Task: identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

Approach for this task- I calculated the average IMDB score for each director. I also used the percentile function to determine which directors had the highest rating. Following this, I utilized a number filter to determine the top 12 directors based on percentile data.

Top 12 directors according to their average imdb scores with percentile:

director_name	No.of Movi	Average(imdb_scor	Percent
Tony Kaye	1	8.6	0.998
Charles Chaplin	1	8.6	0.998
Alfred Hitchcock	1	8.5	0.996
Ron Fricke	1	8.5	0.996
Damien Chazelle	1	8.5	0.996
Majid Majidi	1	8.5	0.996
Sergio Leone	3	8.433333333	0.996
Christopher Nolan	8	8.425	0.995
S.S. Rajamouli	1	8.4	0.993
Richard Marquand	1	8.4	0.993
Asghar Farhadi	1	8.4	0.993
Marius A. Markevicius	1	8.4	0.993

E. Budget analysis: explore the relationship between movie budget and their financial success.

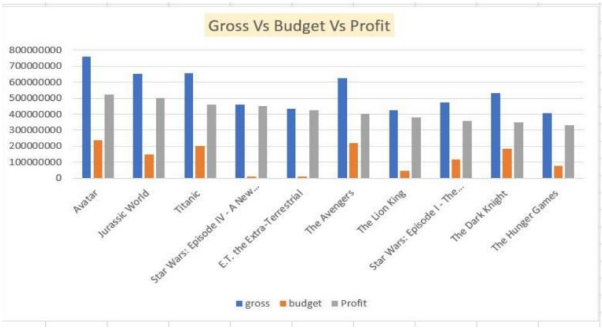
Task: analyse the correlation between movie budgets and gross earnings, identify the movies with the highest profit margin.

The top 10 movies with the highest profit margin are shown in the graph.

correlation coefficient: 0.0966483

Movie_Title	gross	budget	Profit
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	11000000	449935665
E.T. the Extra-Terrestrial	434949459	10500000	424449459
The Avengers	623279547	220000000	403279547
The Lion King	422783777	45000000	377783777
Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677
The Dark Knight	533316061	185000000	348316061
The Hunger Games	407999255	78000000	329999255

The relation between gross, budget, and profit:



Analysis & insights:

- Drama is the most common movie genre in films, the genre with the fewest number of films were film-noir (1), and short (2). Biography, history, documentary, and war have the highest mean, and median, impact on movie ratings.
- Crime (9.3) has received maximum rating among the other genres. Whereas documentary and music received the minimum ratings.
- The movie duration data has no outliers. It's because the mean and median differ so little. The standard deviation is 22.75721, which is large and shows that the data values deviate significantly from the mean. By displaying the relationship between movie duration and IMDB score, it is obvious that the trend line in the scatterplot is slightly upward, indicating a moderate positive link between duration and IMDB score.
- English is the most used common language in movies (3606), Telugu and Persian languages have the highest mean (8.4,8.1) and median (8.4,8.4) IMDB score, while Bosnian has the lowest mean.
- The standard deviation of the IMDB score is less than the mean of the IMDB score for each language and is positive.
- The highest percentile is 0.998 for directors Tony Kaye and Charles Chaplin.
- The correlation coefficient of 0.096514 is near zero, implying the correlation between budget and profit is weak.
- Avatar has the greatest profit margin of 523505847 million dollars,

recommendations:

• Although drama is the most prevalent genre, I would recommend making a film containing the biography, history, documentary, and war genres. It's because these genres have the highest average IMDB rating. These genres can have a significant impact on movie ratings.

• I discovered that movies with a duration of 80 to 200 minutes have the most influence. If the movie storyline, cast, and crew are good, and the duration is between 80-200 minutes, or although the duration is longer, the IMDB score can be greater than 7.5, and the movie's success rate is high.

• The most used language in movies is English, followed by French, Spanish, Mandarin, German, Japanese, and Hindi. Movies made in other languages can have a significant impact on IMDB scores if they are dubbed into the most widely used languages or have subtitles added to these movies in popular languages. Profits are also possible.

• Tony Koye and Charlie Chaplin are the two filmmakers with the highest average IMDB movie rating. I advise that movie makers get inspiration from the top ten directors. In addition, film production companies can collaborate with these top directors.

• The most profitable films are between $100 and $300 million. To reduce the risk of loss, film production business should maintain their budgets within the range.

Conclusion:

- By working on this project, I learned how to handle big datasets, clean data, and use editing functions like Find & Select, Sort & Filter, as well as the flash fill option, percentile, correlation, and descriptive statistics in Excel.
- By working as, a data analyst, I learned to study audience preferences and identify target audiences, as well as trends in the film industry based on insights. I also learned how to record a looming video to explain this project.

PROJECT – 5 Bank Loan Case Study

Project Description:

Analyzing data from a bank loan case study is the goal of this project. It's difficult for any finance company to lend different kinds of loans to urban customers.

Companies face two risks when a customer applies for a loan:

1. The firm suffers if the loan applicant can repay it but is denied approval.

2. The company suffers a loss of revenue if the loan is accepted but the applicant is unable to repay it.

By taking advantage of this some borrowers with insufficient credit histories fail to make loan payments. There are four possible results when a customer applies for a loan:

Approved: The company has approved the loan application.

Cancelled: The customer cancelled the application during the approval process.

Refused: The company rejected the loan.

Unused offer: The loan was approved but the customer did not use it.

To draw insights, I employ **Exploratory data analysis** to examine the trends and determine how loan and customer attributes affect the probability of default.

Problem statement:

As a data analyst at a finance company that specializes in lending various types of loans to urban customers face some challenges. To use EDA to analyse the patterns in the data and ensure that capable applicants are not rejected.

Data cleaning & Approach:

I started by downloading the stakeholder-provided dataset. It has two datasets in it.

1. Application_data, 2. previous _application. I decided to use Excel as my data analysis tool for this project. Second, after creating a copy of the data, began cleaning it. Eliminating columns that contain missing data by using the **"count blank"** function to determine the percentage of blank shells in a column. then I dealt with the missing data using a suitable manner. after that, I began my analysis, gleaned significant insights, and then made recommendations. I used Microsoft Excel 2021 MSO (version 2403).

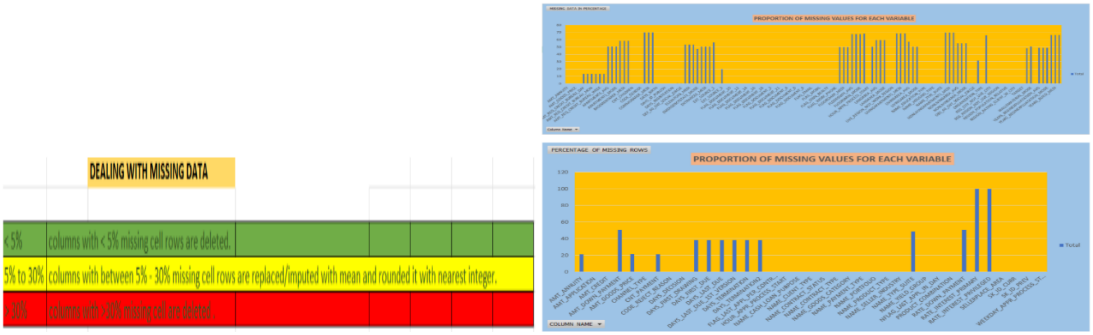
FINDINGS:

A. Identify Missing Data and Deal with it Appropriately

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

Dealing with missing data

The original dataset included 50,000 rows and 122 columns, many of which had missing values. So, the dataset was trimmed down to 77 columns. This made the dataset easier to handle and more trustworthy for analysis. I coloured – and highlighted the relevant column in the original dataset to help to understand data handling and cleaning. I then looked for duplicate values. In the dataset, there are no duplicate values. I organized the Days_Birth column into Days_Birth_Years and like that some more columns. Additionally, I used the interquartile range approach while taking relevant variables into account to identify outliers. That's how I approached the data. Following this, I began understanding analysis to identify trends.



Here missing data is identified and visually represented as the proportion of missing values for each column with column charts for both data sets.

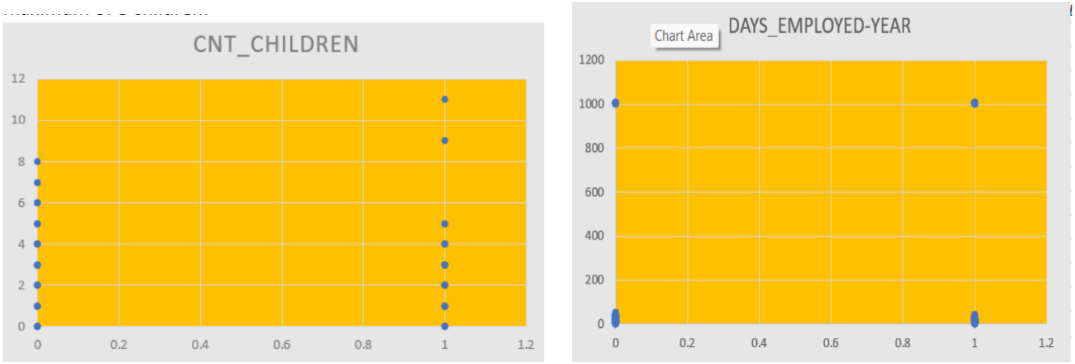
B. Identify outliers in the Dataset:

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

• I use a scatter plot to show the numerical variable distribution in the dataset and draw attention to the presence of outliers.

• Within the provided dataset, I discovered two outliers in the column names **"CNT_CHILDREN"** and "**DAYS_EMPLOYED".**

• Where the xy plotter shows that the applicant with target variable "1" has a maximum of 11 children, which is extremely unusual for a modern application. Whereas candidates with target variable "0" have a maximum of 8 children.

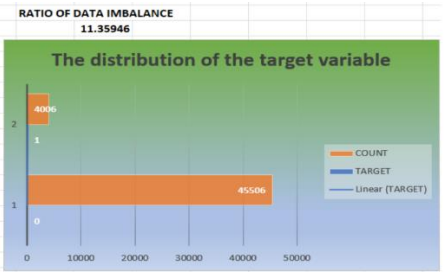


• Where days-employed shows that a few applicants have been employed for a millennium, and there are few candidates which, given that the majority of candidates have jobs that last between 80-90 years, is implausible.

c. Analyze Data Imbalance:

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

TARGE	COUN	PERCENTAGE
0	45506	91.90903215
1	4006	8.090967846
Total	49512	



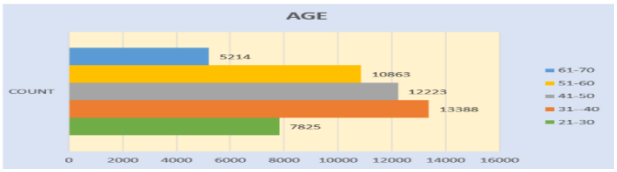
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

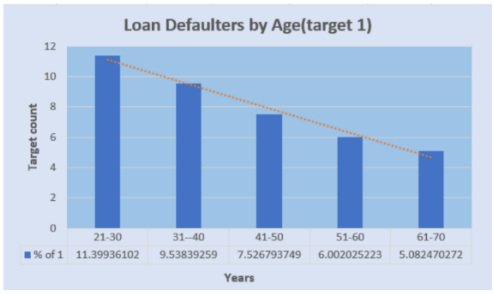
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

The below graphs show that univariate and segmented univariate for the same column names.

AGE:

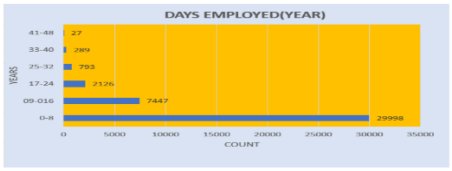
CLASS BINS	COUNT	TARGET 0	TARGET 1	% of 0	% of 1
21-30	7825	6933	892	88.6006	11.3994
31--40	13388	12111	1277	90.4616	9.53839
41-50	12223	11303	920	92.4732	7.52679
51-60	10863	10211	652	93.998	6.00203
61-70	5214	4949	265	94.9175	5.08247

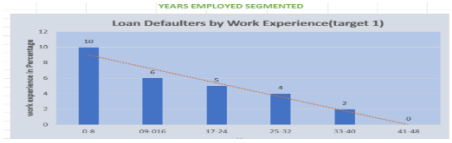




Days_Employed:

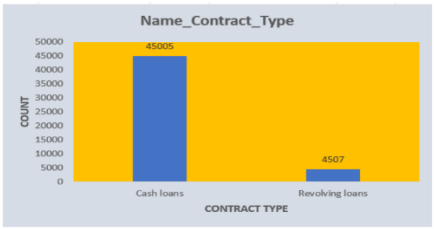
CLASS BINS	COUNT	TARGET 0	TARGET 1	% OF 0	% OF 1
0-8	29998	27056	2942	90	10
09-016	7447	7026	421	94	6
17-24	2126	2019	107	95	5
25-32	793	763	30	96	4
33-40	289	283	6	98	2
41-48	27	27	0	100	0

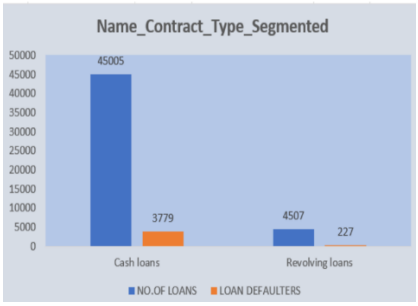




Name_Contract_Type:

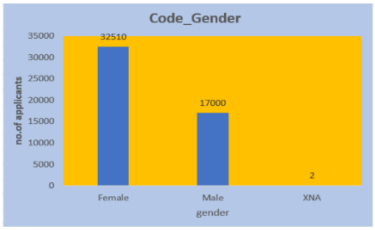
CONTRACT TYP	O.OF LO	LOAN DEFAULTERS
Cash loans	45005	3779
Revolving loans	4507	227

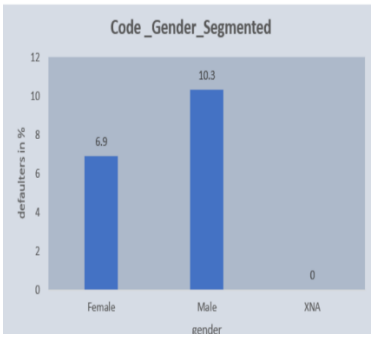




Code_Gender:

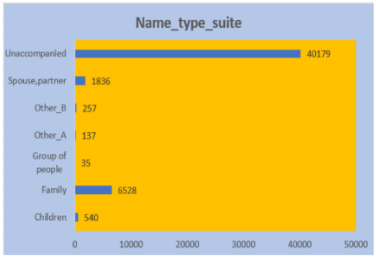
Gender	no.of Applicants	Loan Default
Female	32510	6.9
Male	17000	10.3
XNA	2	0

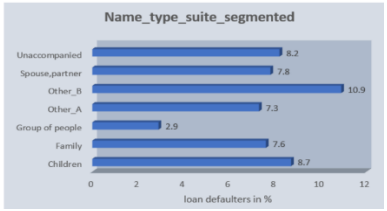




Name_Type_Suite:

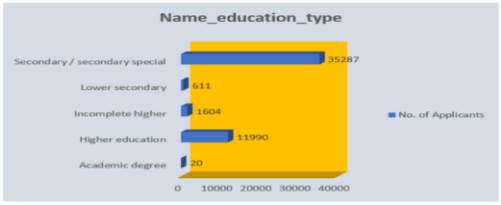
Name type suite	Applicants	Loan defaulters in %
Children	540	8.7
Family	6528	7.6
Group of people	35	2.9
Other_A	137	7.3
Other_B	257	10.9
Spouse,partner	1836	7.8
Unaccompanied	40179	8.2





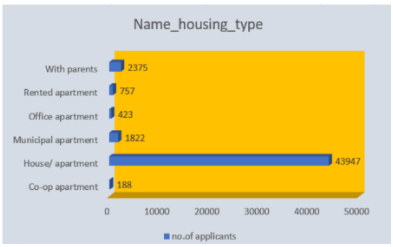
Name_Education_Type:

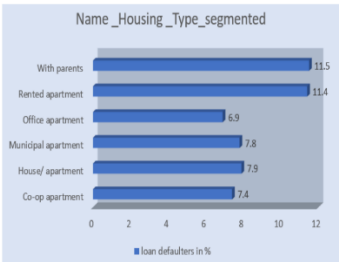
Education type	No. of Applicants	Loan defaulters in %
Academic degree	20	0
Higher education	11990	5
Incomplete higher	1604	8.5
Lower secondary	611	11.9
Secondary / secondary special	35287	9.1



Name_Housing_Type:

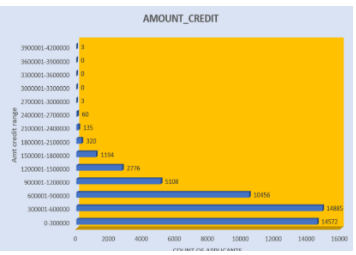
Name housing type	no.of applicants	loan defaulters in %
Co-op apartment	188	7.4
House/ apartment	43947	7.9
Municipal apartment	1822	7.8
Office apartment	423	6.9
Rented apartment	757	11.4
With parents	2375	11.5

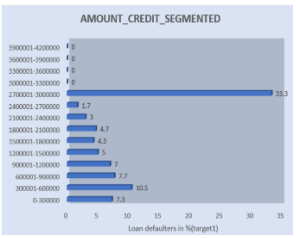




Amount_Credit:

Amt_credit_bins	count of applicants	LOAN DEFAULTERS IN %
0-300000	14572	7.3
300001-600000	14885	10.5
600001-900000	10456	7.7
900001-1200000	5108	7
1200001-1500000	2776	5
1500001-1800000	1194	4.3
1800001-2100000	320	4.7
2100001-2400000	135	3
2400001-2700000	60	1.7
2700001-3000000	3	33.3
3000001-3300000	0	0
3300001-3600000	0	0
3600001-3900000	0	0
3900001-4200000	3	0



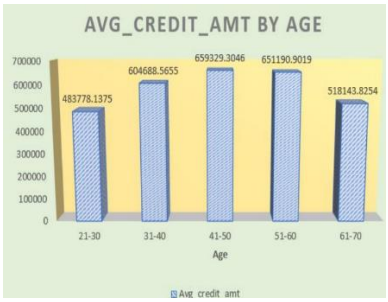


There are many attributes in the row to display.

Bivariate Analysis:

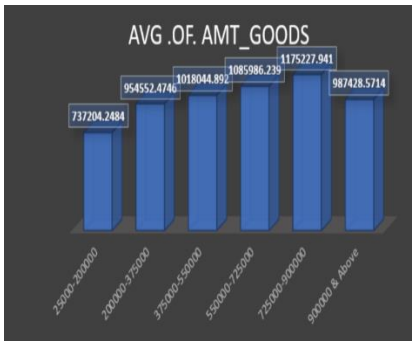
Avg_Credit_Amt_By Age:

Age	Avg_credit_amt
21-30	483778.1375
31-40	604688.5655
41-50	659329.3046
51-60	651190.9019
61-70	518143.8254

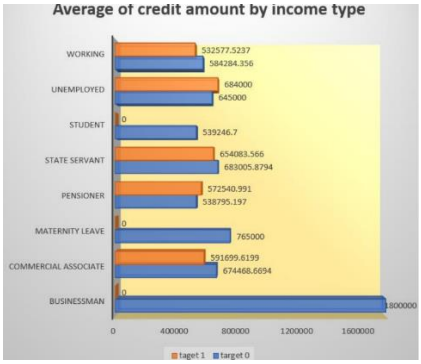


Avg_amt_goods:

Income bins	Avg .of. Amt_goods
25000-200000	737204.2484
200000-375000	954552.4746
375000-550000	1018044.892
550000-725000	1085986.239
725000-900000	1175227.941
900000 & Above	987428.5714



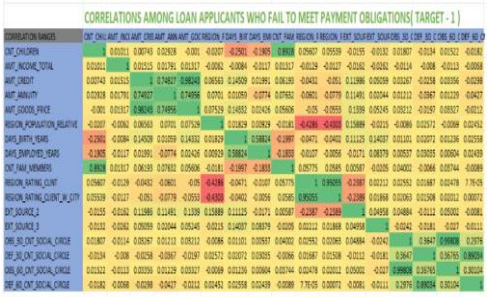
Income_type	Avg_credit_amt(target 0)	Ave_credit_amt(target 1)
Businessman	1800000	#DIV/0!
Commercial associate	674468.6694	591699.6199
Maternity leave	765000	#DIV/0!
Pensioner	538795.197	572540.991
State servant	683005.8794	654083.566
Student	539246.7	#DIV/0!
Unemployed	645000	684000
Working	584284.356	532577.5237



• The above graph shows the relationship between income_type and credit_amount (target 0 and target 1).

E. Identify Top Correlations for Different Scenarios:

• **Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.



The top 5 correlations for target 0 are:

• OBS_60_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE

• AMT_GOODS_PRICE and AMT_CREDIT

• REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT

• CNT_CHILDREN and CNT_FAMILY_MEMBERS

• DEF_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE

The top 5 correlations for target 1 are:

• OBS_60_CNT SOCIAL_CIRCLE and OBS_30_CNT SOCIAL_CIRCLE

• AMT_GOODS_PRICE and AMT_CREDIT

• REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT

• CNT_CHILDREN and CNT_FAMILY_MEMBERS

• DEF_60_CNT SOCIAL_CIRCLE and DEF_30_CNT SOCIAL_CIRCLE

Analysis &insights:

- Recognizing the outliers as invalid is crucial to ensuring the accuracy and reliability of the analytic results. In certain situations, further study and appropriate actions such as data cleaning, should be called for to maintain the integrity of the dataset. Two outliers which found to be **CNT_CHILDREN, DAYS_EMPLOYED** Other outliers were also present, but in the context of business terminology, they are regarded as legitimate data points rather than outliers.
- The ratio of all applicants experiencing payment issues (i.e. target 1) to all the applicants making timely payments (i.e. target 0) is 11.36 of the 45506 applicants, and 92% of them submit their applicants on the time, creating a majority class. Conversely, 8 % of applicants (4006) experience difficulties, creating a minor class.
- The majority of the loan applicants were in the age group between 31- 40 years. The majority of loan defaulters (target 1) are in the age group between 21-30 years.
- when the working years increase the loan defaulters decrease. And also the number of applicants applying for loans decreases. The number of applicants for contract type is high for cash loans and low for revolving loans. Majority of loan applicants are from females. But when comes to loan defaulter's females made payments on time compared to males.
- Those who have an income total between 2500-200000 are applying for loans is the highest number. and also highest loan defaulters are present in this income range only.

Conclusion:

The project offered helpful details on data analysis containing information about loan applications. With Excel's features and functions, I learned how to handle missing data, spot outliers, find data imbalances, perform statistical analyses like univariate, segmented, and bivariate analysis, and also identify correlations. I was able to gain a thorough understanding of the variables causing loan default by closely examining the relationships between different variables and loan default through this analysis. To forecast business decisions and manage risks, it is imperative to acquire this knowledge. The efficiency of data-driven techniques for reducing default risks and accelerating the loan approval process was also learned by this project**.**

PROJECT-6 Analyzing the Impact of Car Features on Price and Profitability

Project Description:

The goal of this research is to investigate how various car attributes impact the cost and profitability of automobiles within the automotive sector. through the examination of a dataset that includes details on

vehicle models, their specs, and market statistics. The intention is to provide information that will help automakers make more informed decisions about pricing and new product development.

Problem statement:

Manufacturers must comprehend what drives consumer demand in the rapidly evolving automobile market of today to properly set pricing and optimize profits. The difficulty of striking a balance between satisfying customer wants, adding appealing features, and determining profitable prices is tackled by this project. As a data analyst, I used this dataset to gain insights into various aspects of the automotive industry, such as:

● Exploring trends in-car features and pricing over time

● Comparing the fuel efficiency of different types of cars

● Investigating the relationship between a car's features and its popularity

● Predicting the price of a car based on its features and market category

Data cleaning and Approach:

I started by cleaning the data to ensure its accuracy and reliability before conducting the analysis. This involved managing missing data, removing duplicates, standardizing data formats, and addressing any discrepancies in the dataset. During this process, I found 102 empty rows and 715 duplicate values, which I then removed from the dataset. For this project I used Microsoft® Excel® 2021 MSO (Version 2404 Build 16.0.17531.20152) 64-bit.

Analytical method:

▪ Descriptive statistics were used to summarize the dataset's salient aspects. These numbers provide information on the distribution and average values of several characteristics, such as MPG, MSRP, and Engine output.

▪ Visualization techniques, such as scatter plots, line graphs, and clustered column charts, were used to visually analyze the dataset's relationships and patterns. This facilitated the identification of trends, anomalies, and potential connections among disparate data sets.

▪ The application of analytical techniques was directed by the project's goal, which included developing effective models for price analysis and gaining insights into many aspects of the automotive industry.

Modeling techniques:

▪ Regression analysis, a modeling technique, was used to forecast car prices. more exact linear regression with feature selection and market category considerations. This project's objective was to establish relationships between independent variables (car pricing) and dependent variables (market category, engine power, and MPG) to fit a linear model to the data and conduct predictive analysis.

FINDINGS:

Tasks: Analysis

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

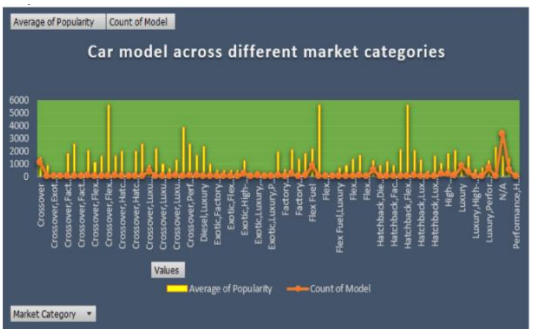
● **Task 1. A**: Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

Output:

Market category	Average of Popul	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	2073.75	64
Crossover,Flex Fuel,Luxury	1173.2	10
Crossover,Flex Fuel,Luxury,Performance	1624	6
Crossover,Flex Fuel,Performance	1657	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	630.3166667	24
Crossover,Luxury,Performance	1343.069286	112
Crossover,Luxury,Performance,Hybrid	3916	2
Crossover,Performance	2585.956522	69
Diesel	1730.904762	84
Diesel,Luxury	2416.106383	47
Exotic,Factory Tuner,High-Performance	1046.380952	21
Exotic,Factory Tuner,Luxury,High-Performance	523.0196078	51
Exotic,Factory Tuner,Luxury,Performance	520	3
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	520	10
Exotic,Flex Fuel,Luxury,High-Performance	520	11
Exotic,High-Performance	1270.326531	245
Exotic,Luxury	112.6666667	12
Exotic,Luxury,High-Performance	473.025974	77
Exotic,Luxury,High-Performance,Hybrid	204	1
Exotic,Luxury,Performance	217.0277778	36
Factory Tuner,High-Performance	1966.442308	104
Factory Tuner,Luxury	617	2
Factory Tuner,Luxury,High-Performance	2133.367442	215
Factory Tuner,Luxury,Performance	1413.419355	31
Factory Tuner,Performance	1818.043383	81
Flex Fuel	2225.71345	855
Flex Fuel,Diesel	1657	16
Flex Fuel,Factory Tuner,Luxury,High-Performance	259	1
Flex Fuel,Hybrid	155	2
Flex Fuel,Luxury	746.5384615	39

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

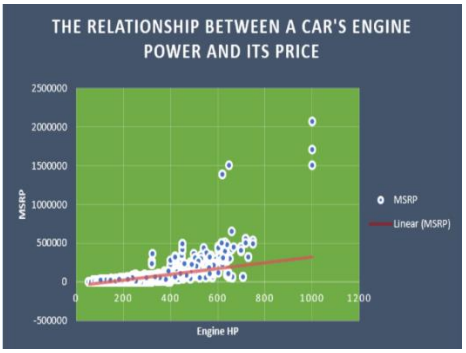
Output:



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

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

Output:



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

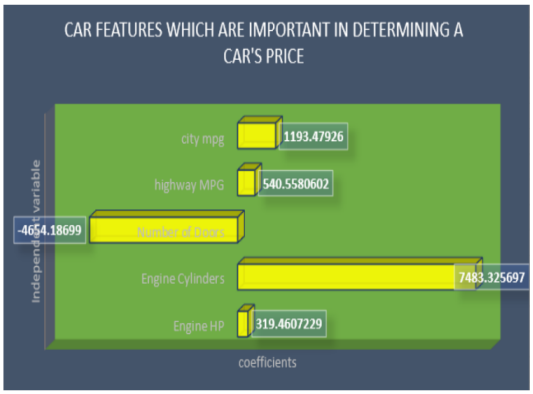
Task 3: Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance**.**

Summary:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.678347							
R Square	0.460155							
Adjusted R	0.459912							
Standard E	45366.26							
Observatio	11097							

ANOVA								
	df	SS	MS	F	Significance F			
Regression	5	1.94568E+13	3.89137E+12	1890.759	0			
Residual	11091	2.28264E+13	2058097531					
Total	11096	4.22832E+13						

	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	-102835	3893.782161	-26.4099912	4E-149	-110467	-95202.2	-110467	-95202.2
Engine HP	319.4607	6.417508778	49.77955371	0	306.8813	332.0402	306.8813	332.0402
Engine Cyli	7483.326	464.1300157	16.12333925	7.93E-58	6573.548	8393.103	6573.548	8393.103
Number of	-4654.19	498.8130353	-9.33052399	1.25E-20	-5631.95	-3676.42	-5631.95	-3676.42
highway M	540.5581	109.9295838	4.917311986	8.9E-07	325.0765	756.0396	325.0765	756.0396
city mpg	1193.479	126.3627723	9.444864481	4.27E-21	945.7858	1441.173	945.7858	1441.173



4. Insight Required: How does the average price of a car vary across different manufacturers?

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

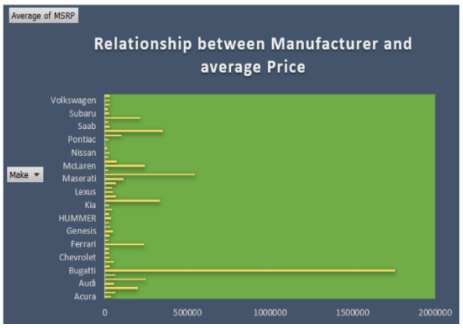
manufacturer.

Output:

Make	Average of MSRP
Acura	35067.4678
Alfa Romeo	61600
Aston Martin	198123.4615
Audi	54574.1215
Bentley	247169.3243
BMW	62162.55864
Bugatti	1757223.667
Buick	29034.18347
Cadillac	56366.26515
Chevrolet	29000.2214
Chrysler	26722.36257
Dodge	24857.04537
Ferrari	237383.8235
FIAT	22206.06695
Ford	28522.86207
Genesis	46616.66667
GMC	32444.08506
Honda	26608.68339
HUMMER	36464.41176
Hyundai	24926.26255
Infiniti	42640.27134
Kia	25316.75
Lamborghini	331567.3077
Land Rover	68067.08633
Lexus	47543.06931
Lincoln	43560.01316
Lotus	68377.14286
Maserati	113684.4309
Maybach	546221.875
Mazda	20106.55612
McLaren	233605
Mercedes-Benz	72135.02647
Mitsubishi	21316.35122
Nissan	28856.42329
Oldsmobile	12843.79545
Plymouth	3296.673239
Pontiac	19800.0442
Porsche	101622.3371
Rolls-Royce	351130.6452
Saab	27879.80734
Scion	19332.5
Spyker	214390
Subaru	24240.67364
Suzuki	18021.0531
Toyota	28758.76676
Volkswagen	28347.36879
Volvo	29724.68421
Grand Total	41901.119

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

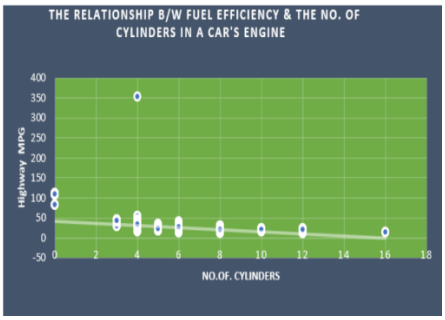
Output:



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

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

Output:



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

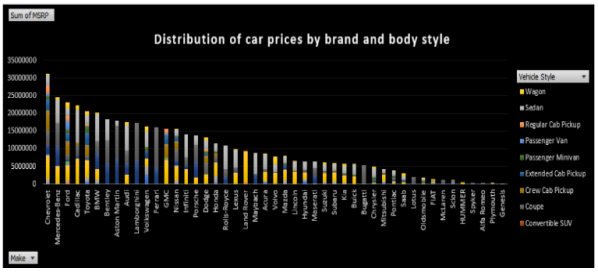
Output:

correlation coefficient	-0.6147

Building the Dashboard

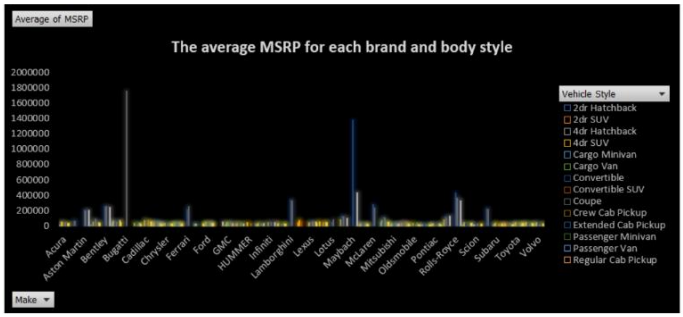
Task 1: How does the distribution of car prices vary by brand and body style?

Output:



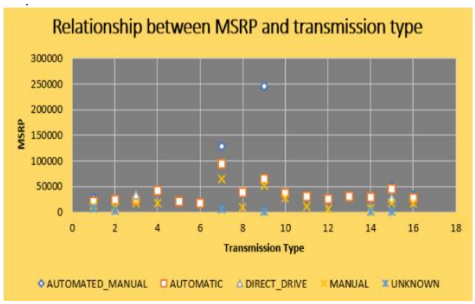
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

Output:



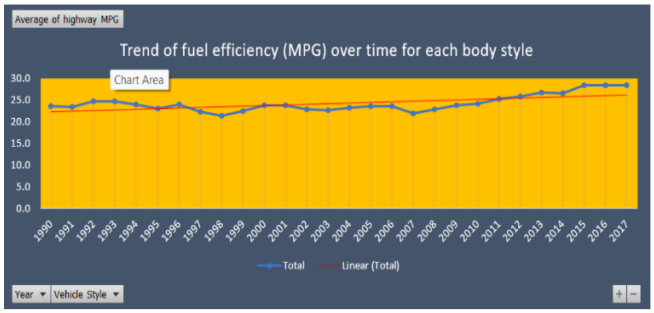
Task 3: How do the different features such as transmission type affect the MSRP, and how does this vary by body style?

Output:



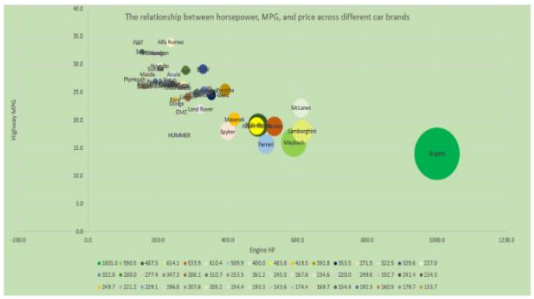
Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

Output:



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

Output:



Analysis & Insights:

We found several key insights, such as the significant impact of engine power on car prices, the varying popularity of car models across market categories, and the relationship between fuel efficiency and engine specifications.

- The market categories with the highest popularity score (5657) **are "crossover, flexfuel, performance" "flexfuel, diesel", and "hatchback, flexfuel".** The market category with the most models is **N/A (3362),** followed by **crossover** (1068).
- The price of an automobile is directly correlated with the engine power of that vehicle. **"engine cylinders"** is the variable that has the strongest relationship with car pricing. This suggests that a car's price is primarily influenced by its cylinder count, with the number of doors having the least bearing on the price.
- By analyzing this data, I discovered that **Bugatti** has the highest average price of any brand, at $175723.7, followed by **Maybach** at $546221.9.
- According to this statistical indicator, highway MPG tends to decline with an increasing cylinder count, quantifying a stronger and negative association.

- Upon conducting investigation, I discovered that, regarding overall prices, **Chevrolet** ranks first with $31175238, followed by **Mercedes Benz** with $24525909. **'Automated-manual transmission'** is the most expensive type of transmission, while **'UNKNOWN transmission'** is the least expensive.

Relevance to Business Problem:

By providing practical information to automakers, helping them prioritize product development, find lucrative market niches, and improve pricing strategies, these insights immediately address the business problem.

Recommendations:

Based on the knowledge acquired, we advise automakers to concentrate on creating fuel-efficient models, strategically determine the pricing of automobiles according to the value of their features, and modify their marketing plans to focus on niche markets.

Conclusion:

▪ In the future, more research may examine dynamic pricing schemes, include current market data, and make use of cutting-edge machine learning methods for predictive modelling.

▪ Furthermore, maintaining competitiveness in the automobile sector requires constant observation of consumer preferences and market developments.

▪ Overall, the knowledge gathered from this analysis offers automakers insightful advice on how to best approach pricing and new product development in a market that is changing quickly.

PROJECT – 7 ABC Call Volume Trend Analysis

Project Description:

Overview - This project will go into the field of Customer Experience (CX) analytics, with a particular emphasis on a company's inbound calling staff. The customer service representative, sometimes referred to as a call center agent, is one of the most important members of a CX team. These agents deal with a variety of support requests, such as email, incoming, outgoing, and social media help.

Problem statement:

The project's main objective is inbound customer assistance, which entails taking inbound calls from current or potential clients. Attracting, involving, and delighting customers is the goal of making them devoted supporters of the company.

Data cleaning & Approach:

I used Microsoft Excel to do the analytical duties listed in the project specifics

• I started by loading the given dataset into Microsoft Excel so that I could analyze it. After that, I managed to comprehend the data and deal with any abnormalities by cleaning it up so that it might be analyzed more effectively.

• In this dataset, 117988 rows with 13 columns are displayed. There are blank cells in the column "Wrapped_By" (47877). This indicates that a blank cell appears for each dropped call.

• #N/A is present in both of the "Agent_Name, Agent_Id" columns. It is a result of missed and ignored calls. so that, because of a shortage of agents, none of them responded to them.

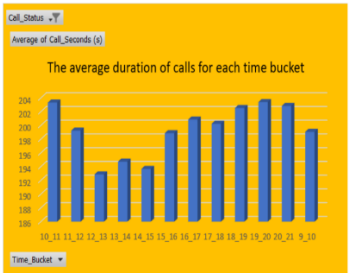
FINDINGS:

Data Analytics Tasks: The dataset contains information about the inbound calls received by a company named ABC, which operates in the insurance sector. By using this data, I have answered the following questions.

1. Average Call Duration: What is the average duration of calls for each time bucket?

Result: The following graph shows the average of calls for each time bucket in seconds

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**



2. Call Volume Analysis: Can you create a chart or graph that shows the number of calls received in each time bucket?

the following pivot table & combo chart shows the number of calls received in each time bucket.

Time_Bucket	Count of Call_Status	Percentage of Time
10_11	13313	8.23%
11_12	14626	9.95%
12_13	12652	9.39%
13_14	11561	9.30%
14_15	10561	9.15%
15_16	9159	8.50%
16_17	8788	8.70%
17_18	8534	8.97%
18_19	7238	8.06%
19_20	6463	7.60%
20_21	5505	6.82%
9_10	9588	5.34%
Grand Total	**117988**	**100.00%**

3. Manpower Planning: What is the minimum number of agents required in each time bucket to reduce the abandon rate to 10%?

Result:

Assumptions:

An agent's total working hours - 9hr	
break -1.5hr	
actual working hours - 7.5hr	
occupancy - 60%	
total working hours - 4.5hr	
total working seconds - 16200sec	
call capacity of any agent/day -	81.6
call capacity of any agent/hour	18

Average no.of call_status	1496	3585	49	5130
call_status in %	29%	70%	1%	100%
here we need to decrease abandon calls from **29% to 10%** and answered calls from **70% to 90%**.				
working hours per agent/seco	16200			
working hours per agent/day	4.5			
Average of call duration in sec	198.62			
hours needed for 90%	255	→ to increase answered calls from 70% to 90%, it needed 255 hr.		
The no.of agents needed is	57	9am-9pm		
call capacity of any agent/day	82			
call capacity of any agent/hour	18			

Time_bucket	abundant/DAY	Answered/DAY	Trnasfer/DAY	Grandtotal/DAY	Abondont %	No.of extra calls to answer	Required Agents
10_11	300	277	1	579	52%	521	244	14
11_12	262	372	2	636	41%	572	200	11
12_13	134	410	6	550	24%	495	85	5
13_14	114	384	5	503	23%	452	69	4
14_15	108	347	5	459	23%	413	67	4
15_16	53	337	8	398	13%	358	21	1
16_17	32	341	8	382	9%	344	2	0
17_18	34	330	7	371	9%	334	3	0
18_19	41	270	5	315	13%	283	14	1
19_20	80	199	2	281	29%	253	54	3
20_21	114	125	0	239	48%	215	91	5
9_10	224	193	0	417	54%	375	183	10
Grand Total	1496	3585	49	5130	29%	4617	1032	57

The total no.of calls required to be answered to decrease the abundant percentage from 29% to 10% is 4617. and the total no.of agents required are 57.

4. Night Shift Manpower Planning: Propose a manpower plan for each time bucket throughout the day, keeping the maximum abandon rate at 10%.

An agent's total working hours - 9hr	
break -1.5hr	
actual working hours - 7.5hr	
occupancy - 60%	
total working hours - 4.5hr	
total working seconds - 16200sec	
no.of calls at night between 9pm to 9am - 30	

Average no.of call_status	1496	3585	49	5130
call_status in %	29%	70%	1%	100%
Here we need to decrease abandoned calls from **29% to 10%**.				
Agent working hours -	4.5			
Average of call duration in sec-	198.62			
Average no.of calls at night -	1539			
To get 90% call rate at night -	76			
Total no.of agents needed in night shift -	17	9pm-9am		

The total no.of agents required between two slots are :		
9am-9pm	57	90%
9pm-9am	17	90%
Total agents needed for two slots	74	

Analysis & Insights:

- The time interval spanning from 10–11 to 19–20 has the longest average call duration. The shortest is between 12–13-time buckets.
- Upon examining the provided assumptions and approximating, there are 57 minimum number of agents required in each time bucket to reduce the abandon rate from 30% to 10%. the mean number of calls made during the night shift as 1539, I discovered that around 17 agents are needed to achieve a 90% call rate.

Recommendations: I can suggest ABC Insurance Company based on my extensive investigation of this project:

• Determine the most effective means of making expectations clear to every contact center representative. Assign more agents during peak hours and fewer during off-peak hours to lower the percentage of abundant calls.

• Give call centre representatives more weight on attitude than aptitude. That implies that tolerance, empathy, great communication skills, and enthusiasm are all necessary for working in a call centre.

• Lastly, giving both new and experienced agents thorough training is necessary. The abundance call percentage can be decreased with frequent check-ins and retraining sessions. Customer happiness and business improvement are inevitable outcomes of these actions.

Conclusion:

• Through this investigation, I have gained a great deal of knowledge regarding call volume patterns, call length factors, and key performance indicator (KPI) monitoring.

• Comprehending the daily fluctuations in call numbers will facilitate more efficient resource management while responding to client inquiries.

• By using more straightforward techniques and demonstrating my flexibility with analytical procedures, I enhance my planning skills. I've learned a lot from this study that will help to maximize call center productivity and deliver excellent customer service.

ADDENDUM:

1. **Hiring process analytics Excel workbook link**

   https://docs.google.com/spreadsheets/d/1VJdMrvxtUkt_q20-Io24B9mgMYjv6ftu/edit?usp=sharing&ouid=100964008805735709631&rtpof=true&sd=true

2. **IMDB movie analysis Excel workbook link**

   https://docs.google.com/spreadsheets/d/1u4Kpiof9Nq58yhjets2g-JC-Bj9f3man/edit?usp=sharing&ouid=100964008805735709631&rtpof=true&sd=true

Video Presentation -

https://www.loom.com/share/2f7fdc11cfdb49f5ad8e9252db6f068

6?sid=86ba7e0e-68e5-4eb5-9aa1-c20849cc5669

3. Bank loan case study

https://www.loom.com/share/163d01b801a643c6a4e659fe91d4c368?sid=b68c6d15-a57c-40e3-8c32-c42ac176abb4

4. analyzing the impact of car features

https://docs.google.com/spreadsheets/d/16dAuzQMmYj8oj1QrriWOvGFkXFQortnI/edit?usp=sharing&ouid=100964008805735709631&rtpof=true&sd=true

Project video link:

https://www.tella.tv/video/manasas-video-et1x

5. ABC call volume trend analysis

https://docs.google.com/spreadsheets/d/1HcNaUniEURFS5lJpf0siJyN4xTJU1vJ-/edit?usp=sharing&ouid=100964008805735709631&rtpof=true&sd=true

https://www.tella.tv/video/shivaramas-video-dysm