BY KARAN PANCHAL

DATA ANALYSIS PORTFOLIO

TABLE OF CONTENT

PROFESSIONAL BACKGROUND

Karan, Aspiring Data Analyst

Driven by a passion for data-driven solutions, I'm a dedicated data analyst with a strong foundation in computer science. Currently pursuing a Bachelor's degree, I've honed my analytical skills through rigorous coursework and practical projects. My academic journey has equipped me with a deep understanding of programming principles and data analysis methodologies.

Beyond the classroom, I've gained hands-on experience with software tools like Excel, SQL, Tableau, and Power BI. Working as a data analyst trainee at Trainty further solidified my grasp of the data analysis process, from data collection and cleaning to insightful reporting.

I'm eager to apply my skills in a corporate setting, tackling real-world challenges and contributing to meaningful projects. My strong theoretical background, coupled with my practical experience, positions me well to quickly learn and adapt to new environments.

I'm seeking internship opportunities to further develop my skills and contribute to data-driven initiatives. I'm excited to collaborate with experienced professionals and apply my knowledge to real-world problems.

PROJECT 1:DATA ANALYTICS PROCESS

Planning a vacation can be an exciting yet complex task, and utilizing data analytics can streamline the decision-making process, enhance experiences, and ensure a memorable trip. Here's a step-by-step overview of how the data analysis process can be applied to vacation planning:

Title: Using Data Analytics to Plan a Vacation

Plan:

Determine the purpose of the vacation (e.g., relaxation, adventure,

cultural exploration).

Prepare:

Data Gathering: Collect information on potential destinations, flight

prices, accommodation options, weather conditions, and local

attractions.

Process

Data Cleaning: Organize the data into categories such as costs, weather patterns, and reviews.

Analyze

Analysis Techniques: Use statistical methods and visualization tools to compare destinations and Identify the most cost-effective and enjoyable options.

Share

Communication: Present the findings to your travel companions through charts, graphs,and summaries.

Act

Decision Making: Choose the best destination based on the analysis.

PROJECT 2:INSTAGRAM USER ANALYTICS

DESCRIPTION:

As a data analyst working with the Instagram product team, your role focuses on analyzing user interactions and engagement within the app to generate insights that drive business growth. This user analysis involves tracking how users engage with the app, providing valuable data for various teams. The marketing team may leverage these insights for new campaigns, the product team for feature development, and the development team for enhancing user experience.

In this project, you will utilize SQL and MySQL Workbench to analyze Instagram user data and address questions from the management team. Your findings will aid the product manager and team in making informed decisions regarding the app's future direction. Ultimately, the goal is to use your SQL skills to extract meaningful insights that could influence the ongoing development of one of the world's leading social media platforms.

Insights

- Loyal User Reward:

Identified the five oldest users on the platform who have been using it the longest. These users are potential candidates for loyalty rewards.

- Inactive User Engagement:

Found users who have never posted a single photo.These users can be targeted for engagement campaigns to increase activity on the platform.

- Contest Winner Declaration:

Determined the user with the most likes on a single photo, who can be declared the contest winner.Provided user details for prize distribution.

- Hashtag Research:

Identified the top five most commonly used hashtags.These hashtags can be recommended to partner brands for better reach and engagement.

- User Engagement:

Calculated the average number of posts per user and the total number of photos per user.Provided insights into overall user activity and engagement levels on the platform.

- Bots & Fake Accounts:

Identified potential bot accounts that liked every single photo.Highlighted the need for further investigation and potential account moderation.

RESULTS:

```
mysql> SELECT username, created_at
    -> FROM users
    -> ORDER BY created_at ASC
    -> LIMIT 5;
+------------------+---------------------+
| username         | created_at          |
+------------------+---------------------+
| Darby_Herzog     | 2016-05-06 00:14:21 |
| Emilio_Bernier52 | 2016-05-06 13:04:29 |
| Elenor88         | 2016-05-08 01:30:40 |
| Nicole71         | 2016-05-09 17:30:22 |
| Jordyn.Jacobson2 | 2016-05-14 07:56:25 |
+------------------+---------------------+
```

```
mysql> SELECT u.username, u.created_at
    -> FROM users u
    -> LEFT JOIN photos p ON u.id = p.user_id
    -> WHERE p.id IS NULL;
+--------------------+---------------------+
| username           | created_at          |
+--------------------+---------------------+
| Aniya_Hackett      | 2016-12-07 01:04:39 |
| Kasandra_Homenick  | 2016-12-12 06:50:07 |
| Jaclyn81           | 2017-02-06 23:29:16 |
| Rocio33            | 2017-01-23 11:51:15 |
| Maxwell.Halvorson  | 2017-04-18 02:32:43 |
| Tierra.Trantow     | 2016-10-03 12:49:20 |
| Pearl7             | 2016-07-08 21:42:00 |
| Ollie_Ledner37     | 2016-08-04 15:42:20 |
| Mckenna17          | 2016-07-17 17:25:44 |
| David.Osinski47    | 2017-02-05 21:23:37 |
| Morgan.Kassulke    | 2016-10-30 12:42:31 |
```

```
mysql> SELECT p.id AS photo_id, p.image_url, p.user_id, u.username, COUNT(l.user_id) AS like_count
    -> FROM photos p
    -> JOIN likes l ON p.id = l.photo_id
    -> JOIN users u ON p.user_id = u.id
    -> GROUP BY p.id, p.image_url, p.user_id, u.username
    -> ORDER BY like_count DESC
    -> LIMIT 1;
+----------+--------------------+---------+---------------+------------+
| photo_id | image_url          | user_id | username      | like_count |
+----------+--------------------+---------+---------------+------------+
|      145 | https://jarret.name |      52 | Zack_Kemmer93 |         48 |
+----------+--------------------+---------+---------------+------------+
```

```
mysql> SELECT t.tag_name, COUNT(pt.photo_id) AS usage_count
    -> FROM tags t
    -> JOIN photo_tags pt ON t.id = pt.tag_id
    -> GROUP BY t.tag_name
    -> ORDER BY usage_count DESC
    -> LIMIT 5;
+----------+-------------+
| tag_name | usage_count |
+----------+-------------+
| smile    |          59 |
| beach    |          42 |
| party    |          39 |
| fun      |          38 |
| concert  |          24 |
+----------+-------------+
```

```
mysql> SELECT DAYNAME(created_at) AS day_of_week, COUNT(*) AS registrations
    -> FROM users
    -> GROUP BY DAYNAME(created_at)
    -> ORDER BY registrations DESC
    -> LIMIT 1;
+-------------+---------------+
| day_of_week | registrations |
+-------------+---------------+
| Thursday    |            16 |
+-------------+---------------+
```

PROJECT 3:OPERATION & METRIC ANALYTICS

Project Description

The objective of this project is to analyze user engagement and retention for a product using SQL. The goal is to understand user behavior and retention patterns, which can provide valuable insights for improving user engagement and the overall success of the product.

Insights

Weekly User Engagement: We identified that user engagement tends to peak during certain weeks, which can inform marketing and engagement strategies.

Retention Rates: Analyzing weekly retention rates helped identify the drop-off points where users are most likely to disengage.

Email Engagement: By calculating email open and click rates, we gained insights into the effectiveness of email campaigns, which can guide future communication strategies.

RESULTS:

```
mysql> select avg(t) as 'avg jobs reviewed per day per hour',
    -> avg(p) as 'avg jobs reviewed per day per second'
    -> from
    -> (select
    -> ds,
    -> ((count(job_id)*3600)/sum(time_spent)) as t,
    -> ((count(job_id))/sum(time_spent)) as p
    -> from
    -> job_data
    -> where
    -> month(ds)=11
    -> group by ds) a ;
+-----------------------------------+-----------------------------------+
| avg jobs reviewed per day per hour | avg jobs reviewed per day per second |
+-----------------------------------+-----------------------------------+
|                       126.18048333 |                        0.03505000 |
+-----------------------------------+-----------------------------------+
```

```
mysql> SELECT
    ->     ds AS Dates,
    ->     ROUND(COUNT(event) / SUM(time_spent), 2) AS Daily_Throughput
    -> FROM
    ->     job_data
    -> GROUP BY
    ->     ds
    -> ORDER BY
    ->     ds;
+------------+------------------+
| Dates      | Daily_Throughput |
+------------+------------------+
| 2020-11-25 |             0.02 |
| 2020-11-26 |             0.02 |
| 2020-11-27 |             0.01 |
| 2020-11-28 |             0.06 |
| 2020-11-29 |             0.05 |
| 2020-11-30 |             0.05 |
+------------+------------------+
6 rows in set (0.00 sec)
```

```
mysql> SELECT
    ->     language AS Languages,
    ->     ROUND(100 * COUNT(*) / sub.total, 2) AS Percentage,
    ->     sub.total
    -> FROM
    ->     job_data
    -> CROSS JOIN
    ->     (SELECT COUNT(*) AS total FROM job_data) AS sub
    -> GROUP BY
    ->     language, sub.total;
+-----------+------------+-------+
| Languages | Percentage | total |
+-----------+------------+-------+
| Arabic    |      12.50 |     8 |
| English   |      12.50 |     8 |
| French    |      12.50 |     8 |
| Hindi     |      12.50 |     8 |
| Italian   |      12.50 |     8 |
| Persian   |      37.50 |     8 |
+-----------+------------+-------+
```

```
mysql> SELECT actor id, COUNT(*) AS Duplicates FROM job_data
    -> GROUP BY actor_id HAVING COUNT(*) > 1;
ERROR 1054 (42S22): Unknown column 'actor' in 'field list'
mysql> SELECT actor_id, COUNT(*) AS Duplicates FROM job_data
    -> GROUP BY actor_id HAVING COUNT(*) > 1;
+----------+------------+
| actor_id | Duplicates |
+----------+------------+
|     1003 |          2 |
+----------+------------+
```

Project 4:HIRING PROCESS ANALYTICS

DESCRIPTION:

This project aims to analyze various aspects of employee data for a company, including hiring distribution by gender, average salary, salary distribution, departmental proportions,and position tiers. The dataset was analyzed using Microsoft Excel.

Insights

- Hiring Analysis: The gender distribution of hires wasdetermined, showing the number of males and femaleshired.
- Salary Analysis: The average salary offered by thecompany was calculated.
- Salary Distribution: Class intervals were created tounderstand the distribution of salaries.
- Departmental Analysis: A pie chart was used to show theproportion of employees in different departments.
- Position Tier Analysis: A bar chart displayed the distribution of position tiers within the company.

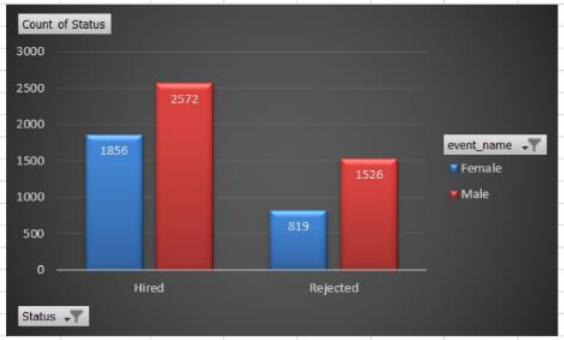
DRIVE LINK:

https://docs.google.com/spreadsheets/d/16RhvtqDperSn8PBpAevmZgX68ZwJL9ak/edit?usp=drivesdk&ouid=114474171974972831050&rtpof=true&sd=true

RESULTS:

Task A:

Count of Status	Column Labels		
Row Labels	Female	Male	Grand Total
Hired	1856	2572	4428
Rejected	819	1526	2345
Grand Total	2675	4098	6773

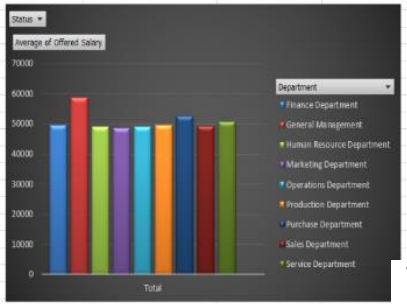


Task B:

Row Labels	Average of Offered Salary
Finance Department	49628.00694
General Management	58722.09302
Human Resource Department	49002.27835
Marketing Department	48489.93538
Operations Department	49151.35438
Production Department	49448.48421
Purchase Department	52564.77477
Sales Department	49261.24832
Service Department	50629.88418
(blank)	
Grand Total	49978.01493

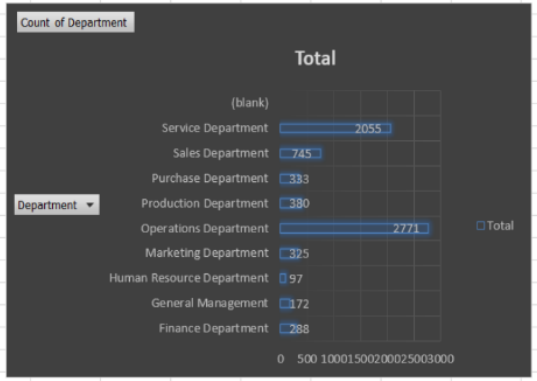
Task C:

Status	(All)								
	Column Labels								
	Finance Department	General Management	Human Resource Department	Marketing Department	Operations Department	Production Department	Purchase Department	Sales Department	Service Department	(bla
Average of Offered Salary	49628.00694	58722.09302	49002.27835	48489.93538	49151.35438	49448.48421	52564.77477	49261.24832	50629.88418	



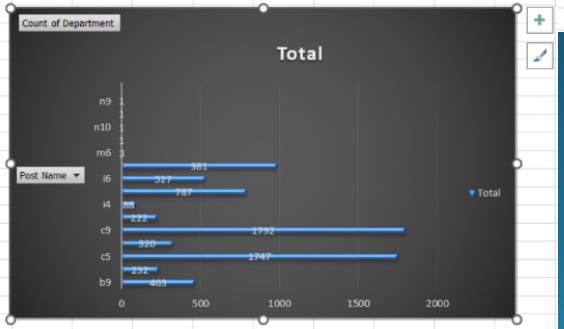
Task D:

Row Labels	Count of Department
Finance Department	288
General Management	172
Human Resource Department	97
Marketing Department	325
Operations Department	2771
Production Department	380
Purchase Department	333
Sales Department	745
Service Department	2055
(blank)	
Grand Total	7166



Task E:

Row Labels	Count of Department
b9	463
c-10	232
c5	1747
c8	320
c9	1792
i1	222
i4	88
i5	787
i6	527
i7	981
m6	3
m7	1
n10	1
n6	1
n9	1
(blank)	
Grand Total	7166



PROJECT 5:IMDB MOVIE ANALYSIS

DESCRIPTION:

This project aims to demystify the factors driving movie performance by conducting a comprehensive analysis of a movie dataset. By examining attributes such as genre, duration, language, director, and budget, we will explore their influence on both critical acclaim (IMDB ratings) and commercial success (gross earnings). Our goal is to uncover hidden patterns and correlations that can provide actionable insights for industry stakeholders, enabling them to make data-driven decisions to optimize film production and distribution strategies.

OBJECTIVE:

- Genre Exploration

Genre Popularity: Identify the most prevalent film genres within the dataset.

Genre Performance: Analyze the correlation between genre and IMDB score to determine which genres consistently garner higher ratings.

Genre Profitability: Explore the financial success of different genres by examining their gross earnings.

- Runtime Impact

Duration Preference: Determine the average and median movie durations to understand audience preferences.

Runtime and Rating: Investigate the relationship between movie length and IMDB score to identify optimal runtimes.

Duration and Earnings: Explore how movie duration affects box office performance.

- Linguistic Influence

Language Dominance: Identify the most common languages in the dataset.

Language and Rating: Analyze the impact of different languages on IMDB scores.

Language and Earnings: Explore whether language influences a film's global commercial success.

- Directorial Impact

Director Performance: Rank directors based on their average IMDB scores.

Director Consistency: Analyze the consistency of director performance across different genres and budgets.

Director Profitability: Assess the financial success of top-performing directors.

- Financial Analysis

Budget-Revenue Correlation: Determine the relationship between production budget and gross earnings.

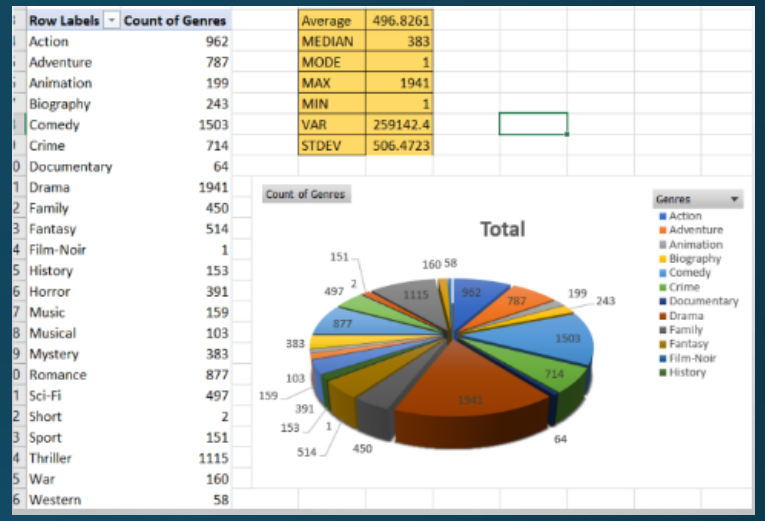
Profit Margins: Identify movies with the highest profit margins relative to their budgets.

Budget Efficiency: Analyze the impact of budget on IMDB scores to assess the efficiency of spending.

RESULTS:

TASK A: GENRE ANALYSIS

GOT TO KNOW WHICH GENRES HAS THE MOST NUMBER OF MOVIES AND ALSO CALCULATED THE MEAN. AS RESULT "ACTION" AND "THRILLER" HAS MOST NUMBER OF MOVIES.

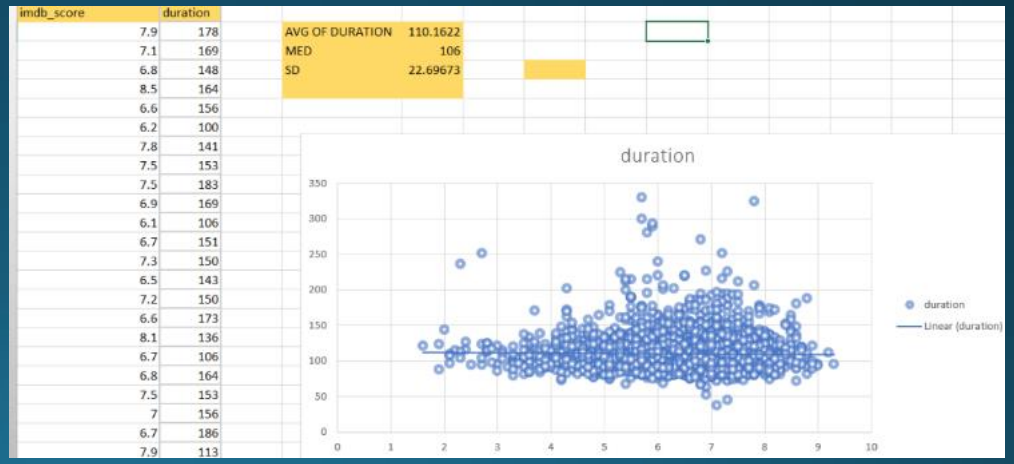


Row Labels	Count of Genres
Action	962
Adventure	787
Animation	199
Biography	243
Comedy	1503
Crime	714
Documentary	64
Drama	1941
Family	450
Fantasy	514
Film-Noir	1
History	153
Horror	391
Music	159
Musical	103
Mystery	383
Romance	877
Sci-Fi	497
Short	2
Sport	151
Thriller	1115
War	160
Western	58

Average	496.8261
MEDIAN	383
MODE	1
MAX	1941
MIN	1
VAR	259142.4
STDEV	506.4723

TASK B:DURATION ANALYSIS

WE JUST GOT TO KNOW THAT DURATION OF THE MOVIE AND IMDB SCORE IS NOT DEPENDEND ON EACHOTHER. IT MEANS THAT MOVIE WITH MORE DURATION WILL HAVE THE HIGHEST IMDB SCORE AND THAT'SA FACT . EVEN A SMALL DURATION MOVIES CAN HAVE HIGHEST IMDB SCORE, ITS MORE ABOUT THE CONTENT AND QUALITY OF THE MOVIE.



imdb_score	duration
7.9	178
7.1	169
6.8	148
8.5	164
6.6	156
6.2	100
7.8	141
7.5	153
7.5	183
6.9	169
6.1	106
6.7	151
7.3	150
6.5	143
7.2	150
6.6	173
8.1	136
6.7	106
6.8	164
7.5	153
7	156
6.7	186
7.9	113

AVG OF DURATION	110.1622
MED	106
SD	22.69673

TASK C: LANGUAGE ANALYSIS:

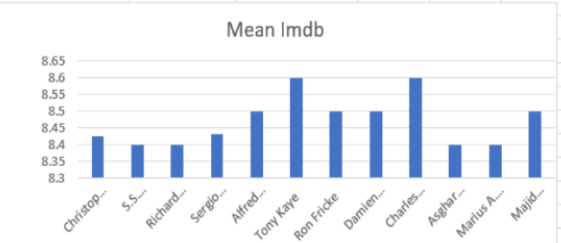
• HERE WE JUST FOUND OUT THE UNIQUE LANGUAGES IN WHICH MOVIES ARE MADE. THEN WE JUST CALCULATED THE NUMBER OF MOVIES FOR EACH LANGUAGE. AFTER DOING THAT I GOT AN INTERESTING RESULT THAT 'ENGLISH' LANGUAGE HAS MORE NUMBER OF MOVIES AD IT IS A GLOBAL LANGUAGE AND MORE NUMBER OF MOVIE ARE DUBBED IN 'ENGLISH'.

language	NO. of movies	Mean	Median	Standard dev
English	3669	6.423549	6.5	1.048835543
Mandarin	14	7.021429	7.25	0.765786244
Aboriginal	2	6.95	6.95	0.777817459
Spanish	26	7.05	7.15	0.826196103
French	37	7.286486	7.2	0.561328861

TASK D: DIRECTOR ANALYSIS:

• HERE I FOUND OUT THE UNIQUE DIRECTORS AND THEN I CALCULATED THE AVERAGE IMDB SCORE FOR RESPECTIVE DIRECTORS. THEN JUST CALCULATED THE 25TH PERCENTILE , 50TH PERCENTILE(MEDIAN) AND75TH PERCENTILE. DIRECTORS WITH AVERAGE SCORES ABOVE THE 75TH PERCENTILE ARE PERFORMING BETTER THAN THE MAJORITY OF MOVIES. DIRECTORS WITH SCORES BETWEEN THE 25TH AND 75TH PERCENTILES ARE PERFORMING BETTER THAN SOME BUT NOT ALL. DIRECTORS WITH SCORES BELOW THE25TH PERCENTILE ARE PERFORMING LOWER THAN THE MAJORITY.

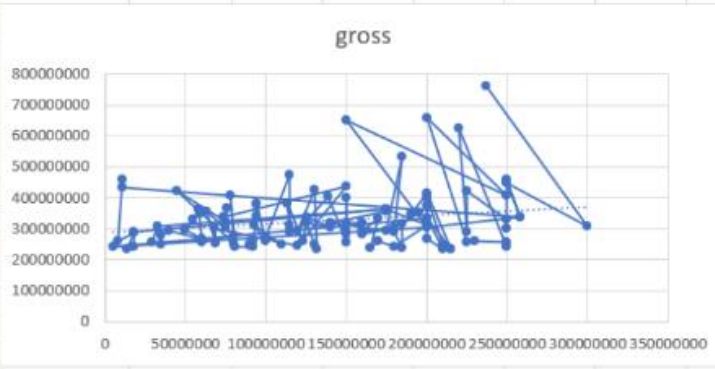
director_names	Mean Imdb	25th percentile	50th percentile	75th percentile
Christopher Nolan	8.425			
S.S. Rajamouli	8.4			
Richard Marquand	8.4			
Sergio Leone	8.433333333			
Alfred Hitchcock	8.5			
Tony Kaye	8.6			
Ron Fricke	8.5			
Damien Chazelle	8.5			
Charles Chaplin	8.6			
Asghar Farhadi	8.4			
Marius A. Markevicius	8.4			
Majid Majidi	8.5			



TASK E: FINANCIAL ANALYSIS

• FIRST I HAVE CALCULATED THE CORRELATION. THE COEFFICIENT OF 0.1 INDICATES THAT THERE IS ONLY A SLIGHT TENDENCY FOR MOVIES WITH HIGHER BUDGETS TO HAVE HIGHER GROSS EARNINGS.THEN I CALCULATED THE PROFIT (GROSS-BUDGET).AFTER THAT I JUST CALCULATED A MOVIE WITH THE HIGHTEST PROFIT AND RESULT WAS 'AVATAR'.

budget	gross	Corelation factor	Profiit	Movie_name	Maximum
237000000	760505847	0.100821908	523505847	AvatarÂ	523505847
300000000	309404152		9404152		
250000000	448130642		198130642		
258000000	336530303		78530303		
250000000	458991599		208991599		
250000000	301956980		51956980		
250000000	330249062		80249062		
225000000	423032628		198032628		
225000000	291021565		66021565		
220000000	623279547		403279547		
250000000	241063875		-8936125		
250000000	255108370		5108370		
230000000	262030663		32030663		
225000000	258355354		33355354		
200000000	658672302		458672302		
250000000	407197282		157197282		
150000000	652177271		502177271		
200000000	304360277		104360277		



DRIVE LINK:

HTTPS://DOCS.GOOGLE.COM/SPREADSHEETS/D/1PARP2FV1ES5LPKDBHA0FX8LSD0JHIKE-/EDIT?USP=DRIVESDK&OUID=114474171974972831050&RTPOF=TRUE&SD=TRUE

PROJECT 6:BANK LOAN CASE STUDY

Overview:

The project aims to analyze loan application data to predict whether a loana pplicant is likely to default on their loan. This analysis will help the company make more informed decisions about loan approvals, reducing financial risk andimproving business outcomes. The dataset contains information about loan applicants, including their demographics, financial history, and loan details.

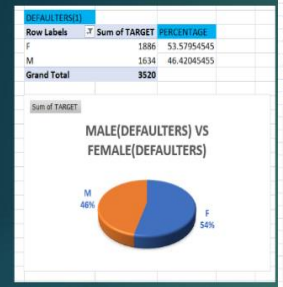
Objectives:

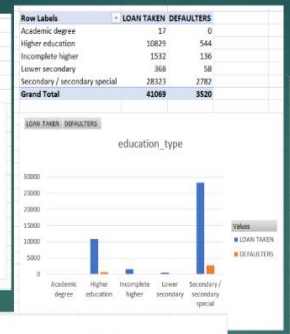
- ❑ Identify missing data and handle it appropriately.
- ❑ Detect and manage outliers in the dataset.
- ❑ Analyze data imbalance and understand its impact on model performance.
- ❑ Conduct univariate, segmented univariate, and bivariate analyses to explore patterns and relationships.
- ❑ Determine top correlations for different scenarios to identify key indicators of loan default.

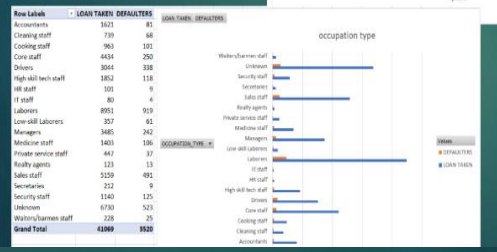
RESULTS:

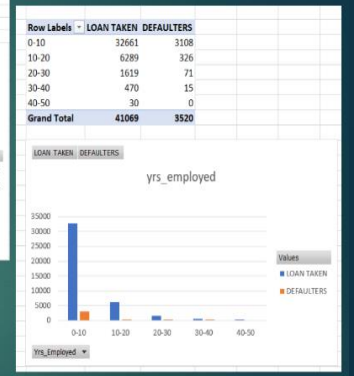
Outliers: We identified unusual data points that might affect our analysis and decided how to handle them.

Univariate Analysis: Provided insights into the distribution of individual variables.





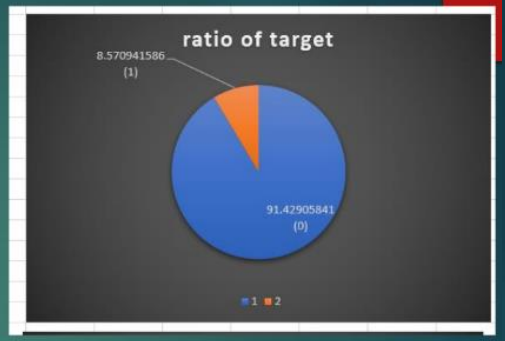




Data Imbalance:

Row Labels	Count of TARGET	ratio of target
0	37549	91.42905841
1	3520	8.570941586
(blank)		
Grand Total	**41069**	

Here the proportion of the defaulter is lesser than the Non defaulters.

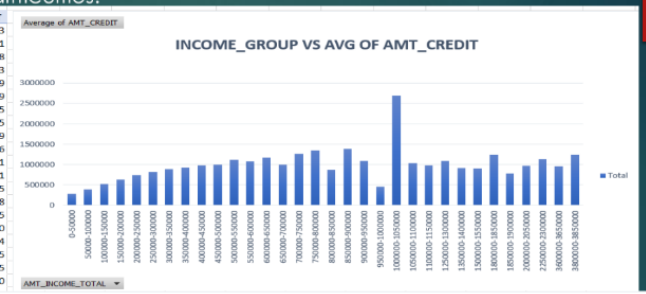


Row Labels	Count of NAME_CONTRACT_TYPE	PERCENTAGE OF CLIENTS
Cash loans	36892	89.82931165
Revolving loans	4177	10.17068835
(blank)		
Grand Total	**41069**	

Here more cash loans are given than revolving loans.

Bivariate Analysis: Explored relationships between variables and loan defaults, identifying potential predictors of loan repayment difficulties.

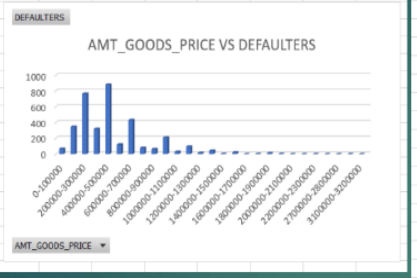
Row Labels	Average of AMT_CREDIT
0-50000	277298.0533
50000-100000	393338.1971
100000-150000	519688.9078
150000-200000	630878.7663
200000-250000	740833.6079
250000-300000	821826.3659
300000-350000	884090.215
350000-400000	920791.1045
400000-450000	985704.8809
450000-500000	997944.6226
500000-550000	1112433.211
550000-600000	1074844.071
600000-650000	1171325.885
650000-700000	1000031.838
700000-750000	1259983.35
750000-800000	1344940
800000-850000	876760.0714
850000-900000	1388400.75
900000-950000	1093675.655
950000-1000000	450000



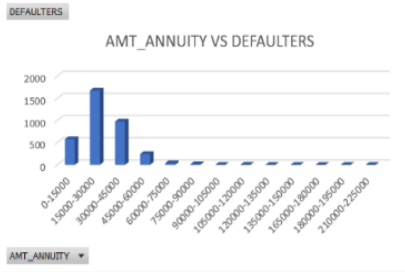
Row Labels	Average of AMT_ANNUITY
0-50000	13640.21326
50000-100000	18704.50089
100000-150000	24008.57266
150000-200000	28654.87312
200000-250000	33007.88849
250000-300000	36100.31341
300000-350000	39301.62898
350000-400000	40810.4589
400000-450000	44097.50693
450000-500000	44784.76415
500000-550000	45984.45868
550000-600000	46788.96429
600000-650000	49857.69231
650000-700000	47379.40541
700000-750000	53482.275
750000-800000	58803
800000-850000	47799.85714
850000-900000	47631.375
900000-950000	58976.22414
950000-1000000	30073.5
1000000-1050000	74254.5
1050000-1100000	89250

Segmented Analysis: Revealed differences in variable distributions between scenarios.

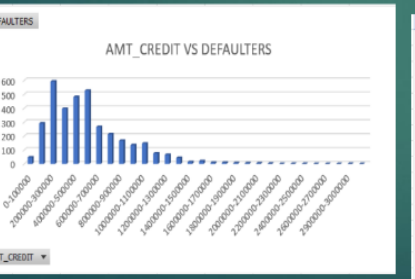
Row Labels	DEFAULTERS
0-100000	68
100000-200000	348
200000-300000	766
300000-400000	321
400000-500000	882
500000-600000	122
600000-700000	432
700000-800000	79
800000-900000	64
900000-1000000	211
1000000-1100000	29
1100000-1200000	96
1200000-1300000	15
1300000-1400000	44
1400000-1500000	3



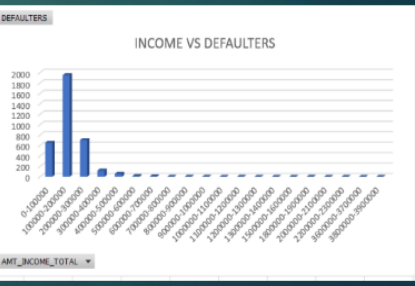
Row Labels	DEFAULTERS
0-15000	578
15000-30000	1676
30000-45000	976
45000-60000	243
60000-75000	37
75000-90000	9
90000-105000	0
105000-120000	1
120000-135000	0
135000-150000	0
165000-180000	0
180000-195000	0
210000-225000	0
Grand Total	3520



Row Labels	DEFAULTERS
0-100000	47
100000-200000	291
200000-300000	593
300000-400000	397
400000-500000	481
500000-600000	527
600000-700000	266
700000-800000	213
800000-900000	167
900000-1000000	135
1000000-1100000	147
1100000-1200000	76
1200000-1300000	65
1300000-1400000	43



Row Labels	DEFAULTERS
0-100000	654
100000-200000	1950
200000-300000	707
300000-400000	121
400000-500000	59
500000-600000	14
600000-700000	8
700000-800000	1
800000-900000	2
900000-1000000	2
1000000-1100000	0
1100000-1200000	1
1200000-1300000	0
1300000-1400000	0
1500000-1600000	0
1800000-1900000	1



Correlations:

Nondefaulter:

	CNT_CHILD	AMT_INCO	AMT_CRED	AMT_ANN	AMT_GOO	REGION_PO	Client_Age	Yrs_Emplo	Yrs_Regist	Yrs_Id_pu
CNT_CHILDREN	1	-0.038026	0.004405	0.007292	-0.003005	-0.013768	-0.169515	-0.01238	-0.13199	0.111081
AMT_INCOME_TOTAL	-0.038026	1	0.312174	0.371245	0.313727	0.096759	0.08763	0.022601	-0.00298	0.037533
AMT_CREDIT	0.004405	0.312174	1	0.745132	0.981928	0.055598	0.194438	0.10511	0.042506	0.054409
AMT_ANNUITY	0.007292	0.371245	0.745132	1	0.746422	0.065867	0.086228	0.054427	-0.01256	0.050271
AMT_GOODS_PRICE	-0.003005	0.313727	0.981928	0.746422	1					
REGION_POPULATION_REL	-0.013768	0.096759	0.055598	0.065867	0.061151	1				
Client_Age	-0.169515	0.08763	0.194438	0.086228	0.188108	0.013409	1			
Yrs_Employed	-0.01238	0.022601	0.10511	0.054427	0.11307	-0.001641	0.305742	1		
Yrs_Registration	-0.131994	-0.002984	0.042506	-0.012559	0.0411	0.046967	0.239861	0.150695	1	
Yrs_Id_publish	0.111081	0.037533	0.054409	0.050271	0.058967	0.008006	0.125405	0.099253	0.04376	1
FLAG_MOBIL	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_EMP_PHONE	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
FLAG_WORK_PHONE	0.025412	-0.099297	-0.053987	-0.056408	-0.026161	-0.027466	-0.059626	0.010873	-0.02151	0.011799
FLAG_CONT_MOBILE	0.007494	-0.055294	0.030352	0.039523	0.027004	0.002007	0.025658	0.008829	0.001113	0.006168
FLAG_PHONE	-0.003387	0.007845	0.031813	0.00741	0.046284	0.062164	0.035471	0.070787	0.058137	0.039391
FLAG_EMAIL	0.005525	0.079619	-0.006264	0.098213	-0.004527	0.052206	-0.058973	-0.032382	0.01728	-0.01833
CNT_FAM_MEMBERS	0.8956	-0.028628	0.048692	0.046939	0.044283	-0.016835	-0.106836	0.011258	-0.13216	0.111309
REGION_RATING_CLIENT	0.066584	-0.160226	-0.047989	-0.064618	-0.053865	-0.436699	-0.051304	-0.003614	-0.12676	-0.0284
REGION_RATING_CLIENT_W	0.066185	-0.168886	-0.059358	-0.08532	-0.062286	-0.439271	-0.047189	0.000284	-0.1158	-0.01963
HOUR_APPR_PROCESS_ST	-0.028317	0.060162	0.047723	0.038564	0.060385	0.157801	-0.037058	0.016151	0.065341	0.002562

Defaulter:

	CNT_CHIL	AMT_INCO	AMT_CRE	AMT_ANN	AMT_GOO	REGION_P	Client_Age	Yrs_Emplo
CNT_CHILDREN	1							
AMT_INCOME_TOTAL	-0.00531	1						
AMT_CREDIT	-0.01771	0.360012	1					
AMT_ANNUITY	-0.00413	0.431027	0.760828	1				
AMT_GOODS_PRICE	-0.02171	0.367729	0.986358	0.765202	1			
REGION_POPULATION_	-0.03148	0.188786	0.096542	0.116108	0.099642	1		
Client_Age	-0.24178	0.049536	0.160879	0.099057	0.155143	0.048987	1	
Yrs_Employed	-0.0659	0.036222	0.094943	0.062134	0.095919	-0.00561	0.352389	1
Yrs_Registration	-0.16136	-0.0335	0.024295	0.000852	0.020392	0.064621	0.304734	0.175693
Yrs_Id_publish	0.122249	0.023116	0.044247	0.032222	0.044496	0.004356	0.107692	0.082502
FLAG_MOBIL	0.003376	0.002543	0.004263	0.000701	0.00416	0.00383	0.008816	0.004667
FLAG_EMP_PHONE	0.004775	0.00184	-0.00332	-0.00081	-0.00212	0.003952	-0.00824	0.000482
FLAG_WORK_PHONE	-0.00335	-0.08101	-0.03428	-0.05343	-0.00966	-0.01906	-0.04097	0.013544
FLAG_CONT_MOBILE	0.000875	-0.0163	0.02748	0.026055	0.024836	-0.00617	0.001801	0.001232
FLAG_PHONE	-0.02833	0.011982	0.022919	0.00812	0.039816	0.09176	0.03854	0.05686
FLAG_EMAIL	0.011251	0.078194	0.00477	0.057543	0.005051	0.041496	-0.06739	-0.03088
CNT_FAM_MEMBERS	0.893736	-0.00084	0.036172	0.04283	0.034141	-0.03101	-0.18033	-0.03596

Drive Link:

https://docs.google.com/spreadsheets/d/1mW6mht_ay9me38-szRb7n8HfYBkXHzQ-/edit?usp=drivesdk&ouid=114474171974972831050&rtpof=true&sd=true

PROJECT 7:IMPACT OF CAR FEATURES

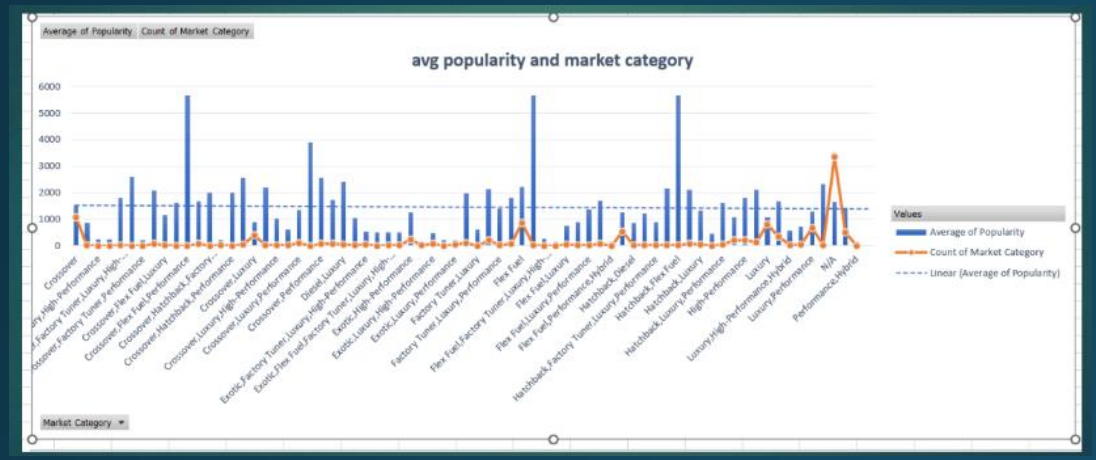
The automotive industry is experiencing significant changes, driven by a heightened focus on fuel efficiency, sustainability, and technological advancements. As competition among manufacturers intensifies and consumer preferences evolve, understanding the factors thatinfluence car demand has become essential. Recently, there has been a marked shift towards electric and hybrid vehicles, along with increased interest in alternative fuel sources like hydrogen and natural gas. Despite this trend, traditional gasoline-powered cars still dominate the market. In this context, the project aims to assist a car manufacturer in optimizing pricing and product development strategies to maximize profitability while addressing consumer needs. By analyzing a comprehensive dataset that includes various car features, market categories, and pricing information, the project will explore the relationships between these variables. Using techniques such as regression analysis and market segmentation, the analysis will identify the most popular and profitable features for consumers. The ultimate goal is to develop a pricing strategy that effectively balances consumer demand with profitability, providing actionable insights for future product development. This strategic approach will enable the manufacturer to enhance its market competitiveness and drive sustainable growth in an evolving automotive landscape.

RESULTS:

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

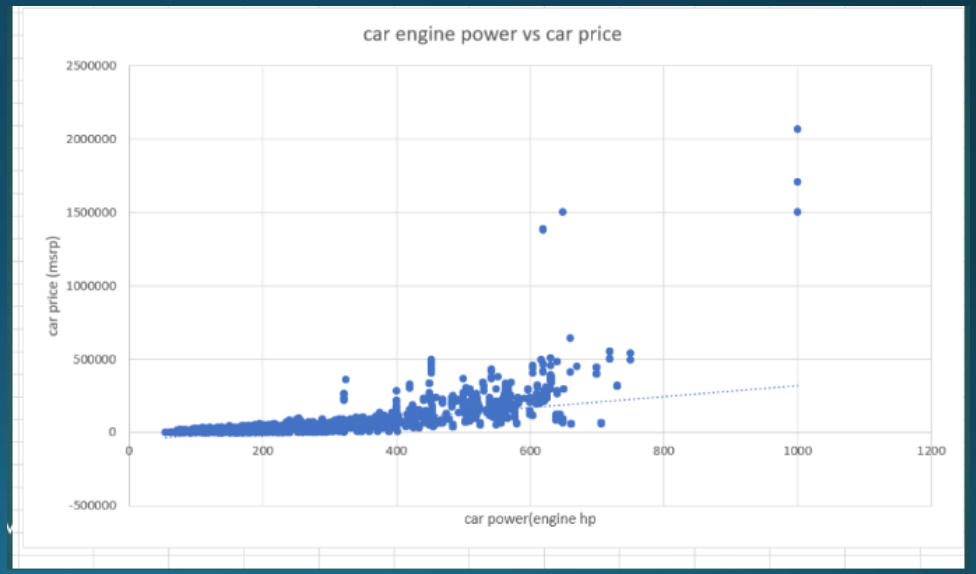
Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

Here the highest avg popularity is of crossover, flex fuel,performance and highest count of market category is not known



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.



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.

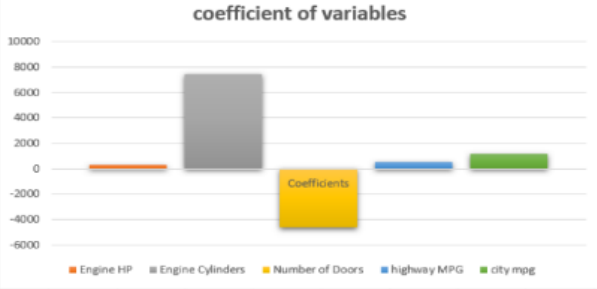
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.67835
R Square	0.46016
Adjusted F	0.45991
Standard E	45366.3
Observati	11097

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	5	1.9E+13	3.9E+12	1890.76	0
Residual	11091	2.3E+13	2.1E+09		
Total	11096	4.2E+13			

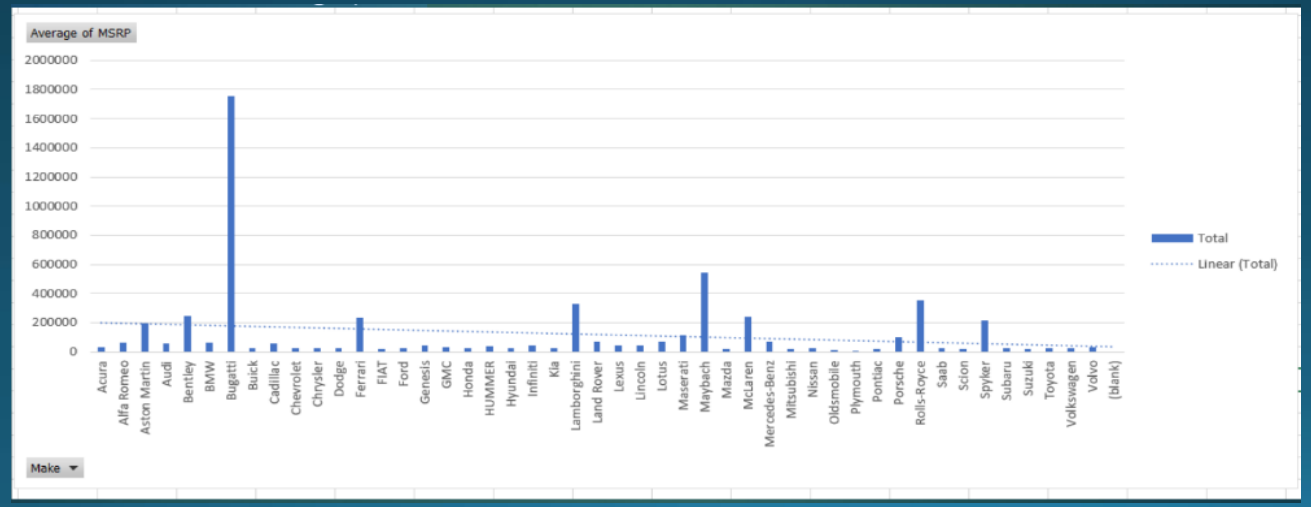
	Coefficient	*andard Err*	*t Stat*	*P-value*	*.ower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-102835	3893.78	-26.41	4E-149	-110467	-95202.2	-110467	-95202.2
Engine HP	319.461	6.41751	49.7796	0	306.881	332.04	306.881	332.04
Engine Cy	7483.33	464.13	16.1233	7.9E-58	6573.55	8393.1	6573.55	8393.1
Number of	-4654.19	498.813	-9.33052	1.3E-20	-5631.95	-3676.42	-5631.95	-3676.42
highway M	540.558	109.93	4.91731	8.9E-07	325.077	756.04	325.077	756.04
city mpg	1193.48	126.363	9.44486	4.3E-21	945.786	1441.17	945.786	1441.17



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.

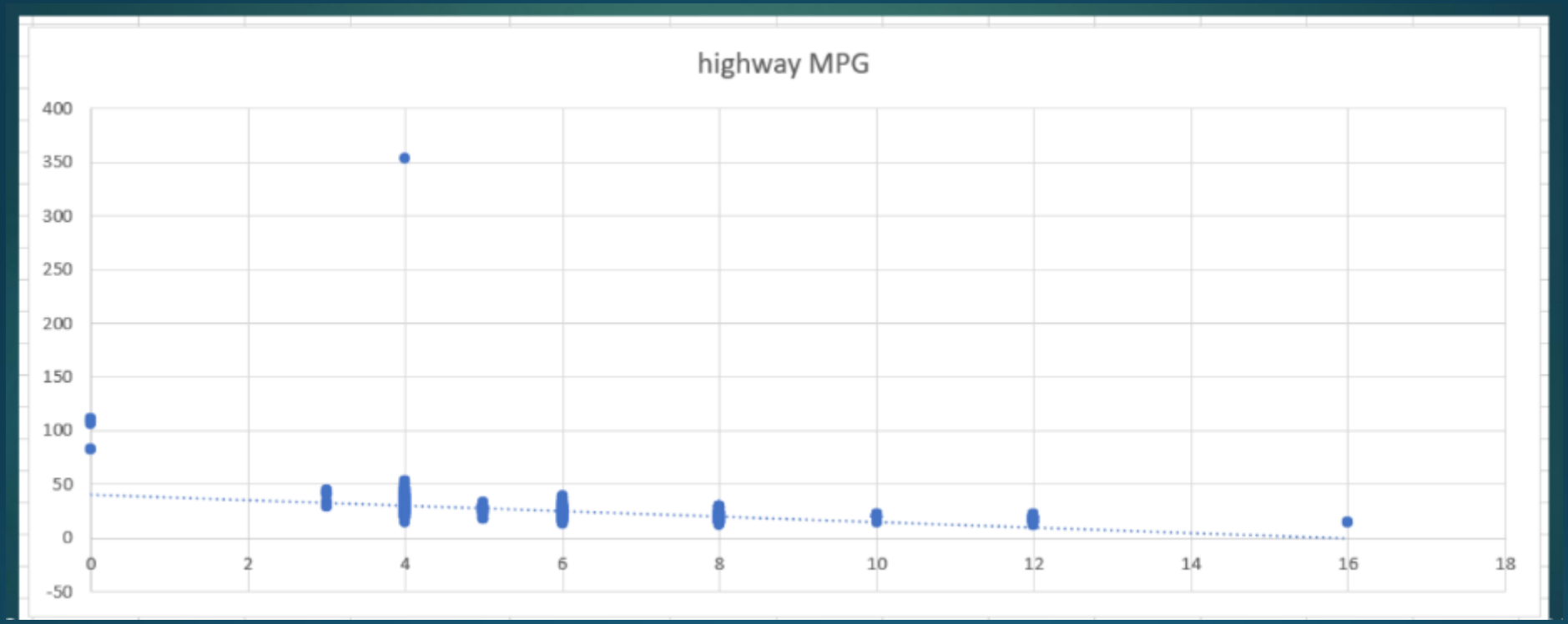
Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



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.

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



DRIVE LINK:

https://docs.google.com/spreadsheets/d/1XoE66OBd28G8QXgG8COpUTJqSXNq_C11/edit?usp=drive_link&ouid=114474171974972831050&rtpof=true&sd=true

PROJECT 8:ABC CALL VOLUME TREND

In this project, you'll be diving into the world of Customer Experience (CX) analytics, specifically focusing on the inbound calling team of a company. You'll be provided with a dataset that spans 23 days and includes various details such as the agent's name and ID, the queue time (how long a customer had to wait before connecting with an agent), the time of the call, the duration of the call, and the call status (whether it was abandoned, answered, or transferred).

A Customer Experience (CX) team plays a crucial role in a company. They analyze customer feedback and data, derive insights from it, and share these insights with the rest of the organization. This team is responsible for a wide range of tasks, including managing customer experience programs, handling internal communications, mapping customer journeys, and managing customer data, among others. In the current era, several AI-powered tools are being used to enhance customer experience. These include Interactive Voice Response (IVR), Robotic Process Automation (RPA), Predictive Analytics, and Intelligent Routing.

One of the key roles in a CX team is that of the customer service representative, also known as a call center agent. These agents handle various types of support, including email, inbound, outbound, and social media support. Inbound customer support, which is the focus of this project, involves handling incoming calls from existing or prospective customers. The goal is to attract, engage, and delight customers, turning them into loyal advocates for the business.

RESULTS:

TASK 1:

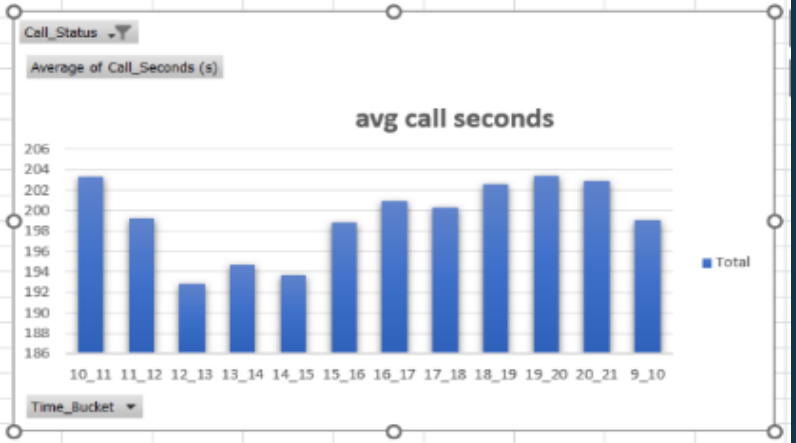
Average Call Duration: Determine the average duration of all incoming calls received by agents. This should be calculated for each time bucket.

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

Insights:

- Maximum call duration is found in Time slot 10 to 11.
- Min is found in timeslot 12 to 13.
- Average call duration is 199 secs.

Call_Status	answered
Row Labels	**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**



Task 2:

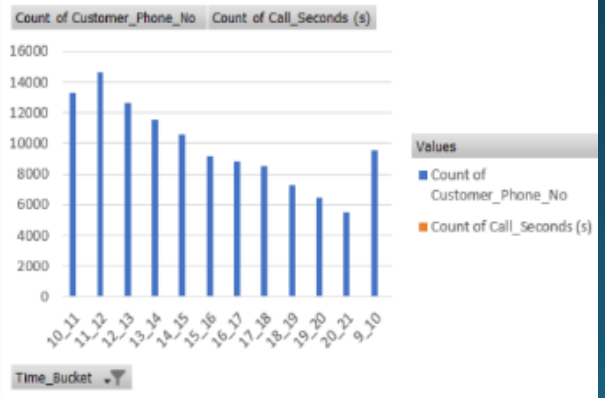
Call Volume Analysis: Visualize the total number of calls received. This should be represented as a graph or chart showing the number of calls against time. Time should be represented in buckets (e.g., 1-2, 2-3, etc.).

Your Task: Can you create a chart or graph that shows the number of calls received in each time bucket?

Insights:

- Maximum call received is between 10 to 11 and 11 to 12.
- Min call received is between 8 to 9.
- Approx 35% of the call are received Between 10 to 13.

Row Labels	Count of Customer_Phone_No	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%**



Task 3:

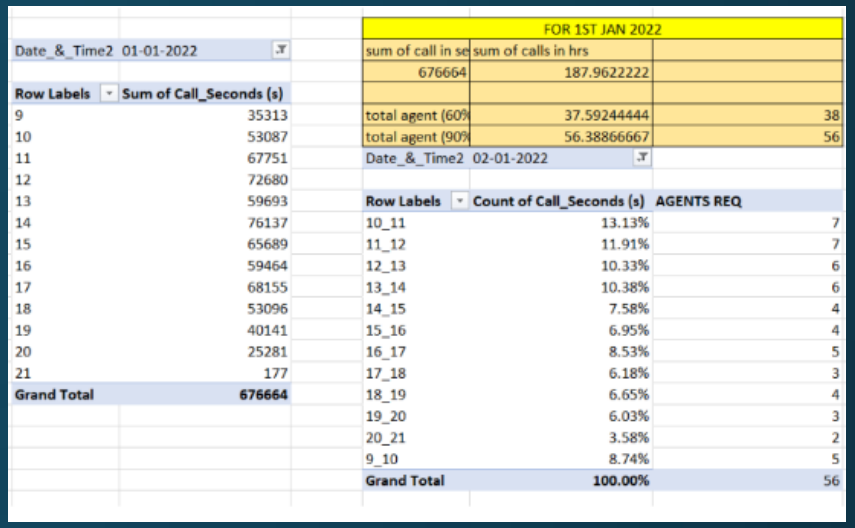
Manpower Planning: The current rate of abandoned calls is approximately 30%. Propose a plan for manpower allocation during each time bucket (from 9 am to 9 pm) to reduce the abandon rate to 10%. In other words, you need to calculate the minimum number of agents required in each time bucket to ensure that at least 90 out of 100 calls are answered.

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

Insights:

- Maximum number of agents are required from 10 to 14.
- Total number of agents required for 1st jan are 56,

For 2nd jan 48..... and so on.



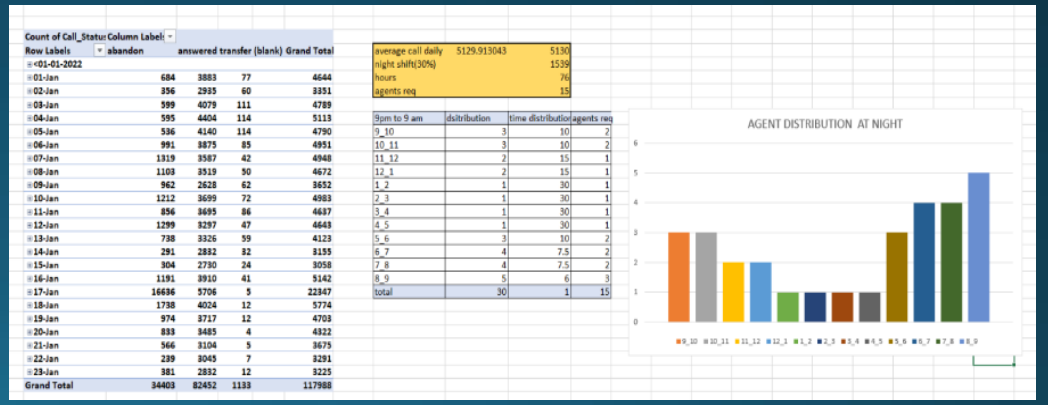
Date_&_Time2	01-01-2022
Row Labels	Sum of Call_Seconds (s)
9	35313
10	53087
11	67751
12	72680
13	59693
14	76137
15	65689
16	59464
17	68155
18	53096
19	40141
20	25281
21	177
Grand Total	676664

FOR 1ST JAN 2022		
sum of call in se	sum of calls in hrs	
676664	187.9622222	
total agent (60%	37.59244444	38
total agent (90%	56.38866667	56
Date_&_Time2	02-01-2022	
Row Labels	Count of Call_Seconds (s)	AGENTS REQ
10_11	13.13%	7
11_12	11.91%	7
12_13	10.33%	6
13_14	10.38%	6
14_15	7.58%	4
15_16	6.95%	4
16_17	8.53%	5
17_18	6.18%	3
18_19	6.65%	4
19_20	6.03%	3
20_21	3.58%	2
9_10	8.74%	5
Grand Total	100.00%	56

Task 4:

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. The distribution of these 30 calls is as follows:

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



Count of Call_Statu	Column Labels			
Row Labels	abandon	answered	transfer	(blank) Grand Total
<01-01-2022				
01-Jan	684	3883	77	4644
02-Jan	856	2935	60	3851
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

average call daily	5129.913043	5130
night shift(30%)		1539
hours		76
agents req		15

9pm to 9 am	dsitribution	time distribution	agents req
9_10	3	10	2
10_11	3	10	2
11_12	2	15	1
12_1	2	15	1
1_2	1	30	1
2_3	1	30	1
3_4	1	30	1
4_5	1	30	1
5_6	3	10	2
6_7	4	7.5	2
7_8	4	7.5	2
8_9	5	6	3
total	30	1	15

Insights:

- Total agent required at night shift are 15.
- Maximum agents are required during 8 to 9.

DRIVE LINK: https://docs.google.com/spreadsheets/d/1CPbq7auimUGhAY29oxRw4PsdLHjrvslZ/edit?usp=drive_link&ouid=114474171974972831050&rtpof=true&sd=true

LEARNINGS:

Data Analytics Foundations:

Data Quality is Critical: Reliable, accurate data is essential for deriving meaningful insights and making sound conclusions.

Visualization is Key: Clear, effective visualizations simplify complex trends, helping communicate findings more effectively.

Data-Driven Decision Making: Analytics empowers organizations to make strategic decisions and optimize operations by providing actionable insights.

Business Applications:

User Behavior Analysis: Gaining insights into customer preferences and behaviors enhances product offerings and tailors marketing efforts.

Operational Efficiency: Analytics identify inefficiencies, enabling process optimization and cost reduction.

Talent Acquisition: Data-driven insights streamline the hiring process, improving candidate selection and reducing the time to hire.

Market Analysis: Analyzing consumer trends and market conditions informs product development and strategic marketing initiatives.

Risk Management: Analytics help assess and mitigate risks, such as evaluating credit risk in financial services.

Analytical Techniques:

Segmentation: Grouping customers into distinct segments to deliver more personalized marketing and product strategies.

Sentiment Analysis: Gauging customer feedback and opinions to refine products and services.

Correlation Analysis: Identifying relationships between variables to uncover hidden patterns and insights.