Data Analysis Portfolio

BY NISHTHA MEHAR

"We are surrounded by data but starve for insights"

Professional Background

Currently in my pre- final year pursuing B.Tech-CSE. I have secured 9.02 CGPA (till 5th sem) and have several skills including ,Data Analysis, Python, Data Visualization.

I am working as an Intern in Twilearn as Data Analyst and have experienced virtual internship from various courses like Data Visualization: Empowering Business with Effective Insights ,Accenture North America Data Analytics and Visualization course from scratch and managed different teams. I also cleared NPTEL examination in Cloud Computing making place among top 1% students

As I am a fresher it would be great to experience the real challenges of the corporate world and understand how things work. Being a fresher, I think I am very flexible and adaptive to learn new things. I have theoretical knowledge. But I am waiting to use my theoretical knowledge in a practical way. And I believe by putting significant efforts I will learn.

Table of Contents

TOPIC	PAGE NO
Professional Background	1
Table of Contents	2-5
Data Analytics Process • Problem • Examples • Conclusion	7-8
Instagram User Analytics • Problem • Design • Insights • Conclusion	9-11
Operation & Metric Analytics • Problem • Design • Insights • Conclusion	12-15
Hiring Process Analytics • Problem • Design • Insights • Conclusion	16-19
IMDB Movie Analysis • Problem • Design • Insights • Conclusion	20-23
Bank Loan Case Study • Problem • Design	24-33

• Insights • Conclusion	
Impact of Car Features • Problem • Design • Insights • Conclusion	34-40
ABC Call Volume Trend • Problem • Design • Assumptions • Insights • Conclusion	41-43
Conclusion	44
Appendix	45

Data Analytics Process

Problem Statement

Task is to give the example(s) of such a real-life situation where we use Data Analytics and link it with the data analytics process.

Examples

EXAMPLE 1

WHERE TO EAT

PLAN: First we need to decide which type of food we want to eat like Chinese,Italian etc.

PREPARE: What is our budget limit

PROCESS: Which restaurants nearby our within our budget limit

ANALYZE: Now for eg we want Chinese so in that what we preferably want to eat Manchurian, noodles etc and analyzing which restaurant has best rating and reviews.

SHARE: Now we will take opinion of our friends or family which restaurant majority of us wants to prefer

ACT: We will go that restaurant

EXAMPLE 2

TRAVELLING

PLAN: To decide where we wish to go.

PREPARE: What is our budget limit

PROCESS: Open different apps like ola ,uber and see the distance and ride options available.

ANALYZE: Now analyze from the apps that where the ride is easily available and the fair is low and estimated time of arrival.

SHARE: If you're traveling with someone, communicate the options to them. Discuss which service provides the best balance between price, comfort, and time.

ACT: Book the ride

Conclusion

The data analytics process, as illustrated through everyday examples, highlights its integration into daily decision-making. For instance, deciding where to eat or travel involves sequential steps: planning (defining objectives), preparing (setting constraints like budget), processing (identifying options), analyzing (evaluating options based on preferences), sharing (gathering feedback or opinions), and acting (making the final choice). These examples underscore the practicality and accessibility of data analytics, showcasing how structured analysis supports informed decisions in routine activities.

Instagram User Analytics

Problem Statement

Imagine you're a data analyst working with the product team at Instagram. Your role involves analyzing user interactions and engagement with the Instagram app to provide valuable insights that can help the business grow.
User analysis involves tracking how users engage with a digital product, such as a software application or a mobile app. The insights derived from this analysis can be used by various teams within the business. For example, the marketing team might use these insights to launch a new campaign, the product team might use them to decide on new features to build, and the development team might use them to improve the overall user experience.

Design

Steps taken to clean the data:

- First importing the datasets provided and merging them. Then removing the duplicates and the blank cells.
- Improving the headers of each column with proper values.
- Finding and replacing the data with correct values.
- Tools used for visualization: SQL

Insights

Loyal User Reward: The marketing team wants to reward the most loyal users, i.e., those who have been using the platform for the longest time.

These are oldest user of Instagram:

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

Inactive User Engagement: The team wants to encourage inactive users to start posting by

sending them promotional emails.

Kenton_Kirlin	
Andre_Purdy85	
Harley_Lind18	
Arely_Bogan63	
Aniya_Hackett	
Travon.Waters	
Kasandra_Homenick	
Tabitha_Schamberger11	
Gus93	
Presley_McClure	
Justina.Gaylord27	
Dereck65	
Alexandro35	
Jaclyn81	
Billy52	
Annalise.McKenzie16	
Norbert_Carroll35	
Odessa2	
Hailee26	
Delpha.Kihn	
Rocio33	

Contest Winner Declaration: The team has organized a contest where the user with the most likes on a single photo wins.

Mike.Auer39	257
Rocio33	257
Maxwell.Halvorson	257
Julien_Schmidt	257
Duane60	257
Mckenna17	257
Ollie_Ledner37	257
Annalise.McKenzie16	103
Keenan.Schamberger60	98
Karley_Bosco	97
Adelle96	96
Andre_Purdy85	94
Josianne.Friesen	94
Arely_Bogan63	93
Alexandro35	93
Rick29	92
Lennie_Hartmann40	92
Willie_Leuschke	91
Frederik_Rice	91
Kenneth64	91
Irwin.Larson	91

Hashtag Research: A partner brand wants to know the most popular hashtags to use in their posts to reach the most people.

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

Ad Campaign Launch: The team wants to know the best day of the week to launch ads.

day_of_week	cnt
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

User Engagement: Calculate the average number of posts per user on Instagram

avg
2.5700

Bots & Fake Accounts: Investors want to know if the platform is crowded with fake and dummy accounts.

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

Conclusion

Working on this project has provided me valuable hands-on experience with MySQL Workbench for real-world data analysis. Each task yielded insights that can directly influence Instagram's growth strategies, improve user engagement, and support brand partnerships. This analysis also helped develop a deeper understanding of how data-driven decisions can shape the platform's future, from feature prioritization to marketing tactics, while providing transparent metrics for investors.

The ability to derive such detailed insights through SQL has highlighted the importance of data analysis in strategic planning, ensuring that Instagram's product, marketing, and development teams can make well-informed, impactful decisions.

Operation & Metric Analytics

Problem Statement

Operational Analytics involves analyzing a company's end-to-end operations to identify areas for improvement. As a Data Analyst, you collaborate with teams like operations, support, and marketing to extract insights from collected data. A critical task is investigating metric spikes, such as dips in user engagement or sales, to understand and explain sudden changes. In this project, acting as a Lead Data Analyst at a company like Microsoft, you'll use advanced SQL to analyze datasets, answer departmental queries, and provide actionable insights to enhance operations and address key metric fluctuations.

Design

Steps taken to clean the data:

- First importing the datasets provided and merging them. Then removing the duplicates and the blank cells.
- Improving the headers of each column with proper values.
- Finding and replacing the data with correct values.
- Tools used for visualization: SQL

Insights

Jobs Reviewed Over Time:

Objective: Calculate the number of jobs reviewed per hour for each day in November 2020.

ds	job_review
2020-11-25	80.0000
2020-11-26	64.2857
2020-11-27	34.6154
2020-11-28	218.1818
2020-11-29	180.0000
2020-11-30	180.0000

Throughput Analysis

Objective: Calculate the 7-day rolling average of throughput (events per second).

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

Weekly Throughput
0.03

Language Share Analysis

Objective: Calculate the percentage share of each language over the last 30 days.

language	Percentage	total
English	12.50	8
Arabic	12.50	8
Persian	37.50	8
Hindi	12.50	8
French	12.50	8
Italian	12.50	8

Duplicate Rows Detection

Objective: Identify duplicate rows in the job_data table.

actor_id	duplicate_count
1003	2

job_id	duplicate_count
23	3

Weekly User Engagement

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

week_number	active_user
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

User Growth Analysis

Objective: Analyze the growth of users over time for a product.

year	week_num	num_users	cum_users
2013	0	23	23
2013	1	30	53
2013	2	48	101
2013	3	36	137
2013	4	30	167
2013	5	48	215
2013	6	38	253
2013	7	42	295
2013	8	34	329

Weekly Retention Analysis

Objective: Analyze the weekly retention of users based on their sign-up cohort.

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

Weekly Engagement Per Device

Objective: Measure user activeness on a weekly basis per device.

week_number	dell_inspiron_notebook	iphone_5	iphone_4s	iphone_5s	ipad_air	windows_surface	macbook_air	macbook_pro	ipad_mini	kindle_fire	amazon_fire_phone	ne
17	46	65	21	42	27	10	54	143	19	6	4	40
18	77	113	46	73	52	10	121	252	30	27	9	73
19	83	115	44	79	55	16	112	266	36	21	12	87
20	84	125	55	79	59	21	119	256	32	23	11	103
21	80	137	45	74	51	17	110	247	23	30	5	91
22	92	125	45	71	58	15	145	251	34	21	5	96
23	103	152	53	79	41	14	124	266	33	25	16	88
24	99	142	53	79	57	22	152	255	39	25	11	87
25	105	137	40	78	57	22	121	275	30	24	13	89
26	89	152	50	94	56	21	134	269	43	26	13	87
27	89	163	67	83	55	33	142	302	35	25	10	84
28	103	151	61	93	54	33	148	295	35	31	6	85
29	113	144	60	90	52	28	148	295	34	37	12	77
30	127	152	65	103	70	19	159	322	35	25	12	84
31	113	135	56	71	55	19	147	321	27	14	14	69
32	104	119	34	67	48	10	125	307	30	12	12	67

Email Engagement Analysis

Objective: Analyze how users engage with the email service.

email_open_rate	email_clicked_rate
33.58339	14.78989

Conclusion

Working on this project provided valuable hands-on experience in using SQL for operational analytics and metric investigation. Each task offered insights that can help improve company operations, understand sudden metric changes, and drive product growth and user engagement. This analysis has contributed to a clearer understanding of how data-driven strategies can enhance decision-making across various departments, from operations to marketing, ultimately supporting the company's objectives.

The ability to extract such insights through advanced SQL queries emphasized the role of data analytics in strategic planning. It ensures that teams across the company, such as product, support, and marketing, are equipped with actionable metrics to make informed and impactful decisions that drive growth and efficiency.

Hiring Process Analytics

Problem Statement

As a data analyst at a multinational company like Google, your task is to analyze hiring process data to provide insights for improvement. You'll handle missing data, combine similar categories, detect and manage outliers, and clean the dataset for analysis. Summarize findings through statistical measures like averages and medians, and create visualizations to uncover trends. Key focus areas include analyzing rejections, interviews, job types, and vacancies. Your insights aim to enhance the company's hiring process and decision-making.

Design

Steps taken to clean the data:

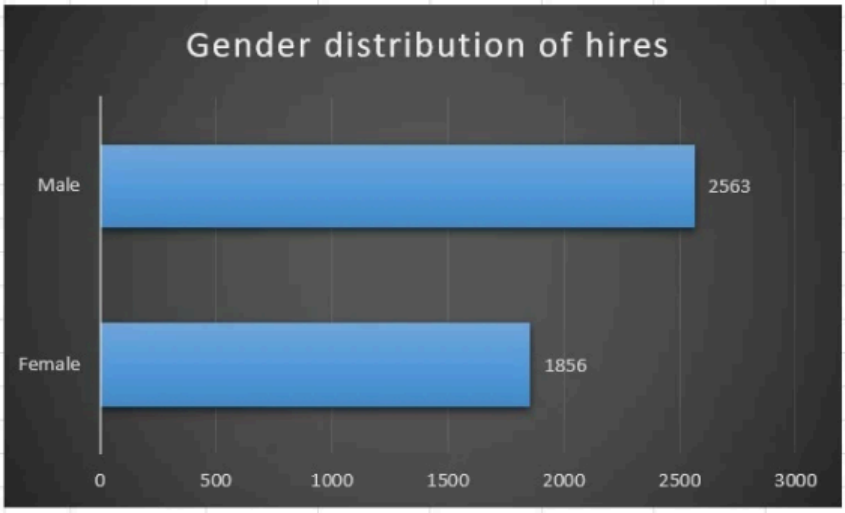
- Handled missing values
- Removed unwanted characters
- Deleted duplicate rows
- Tool Used for Visualization: Microsoft Excel 365

Insights

A. Hiring Analysis

Your Task: Determine the gender distribution of hires by calculating the number of males and females hired.

Status	Hired
Row Labels	**Count of application_id**
Don't want to say	278
Female	1856
Male	2563
Grand Total	**4697**



B. Salary Analysis

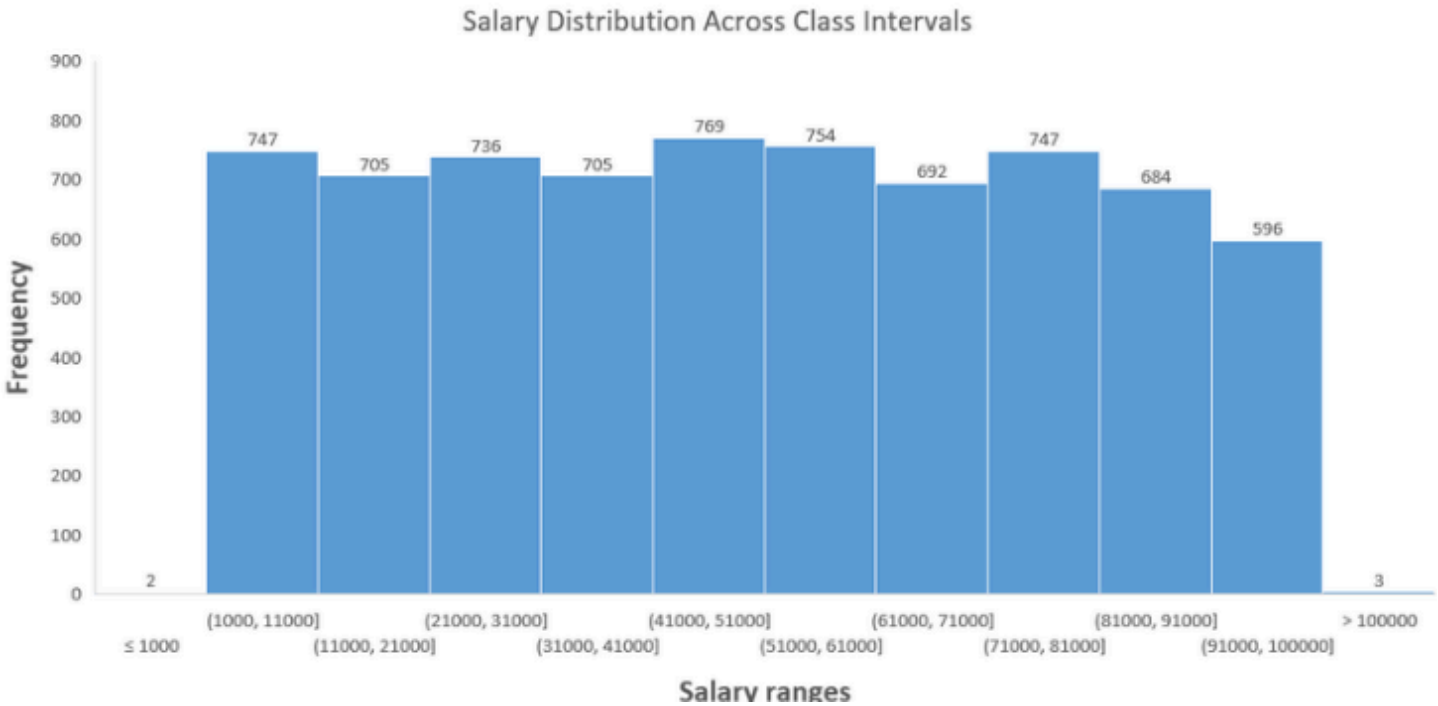
Your Task: Calculate the average salary offered by the company using Excel functions.

```
=AVERAGE(Table1[Offered Salary])
AVERAGE(number1, [number2], ...)
```

AVERAGE OFFERED SALARY
49983.03

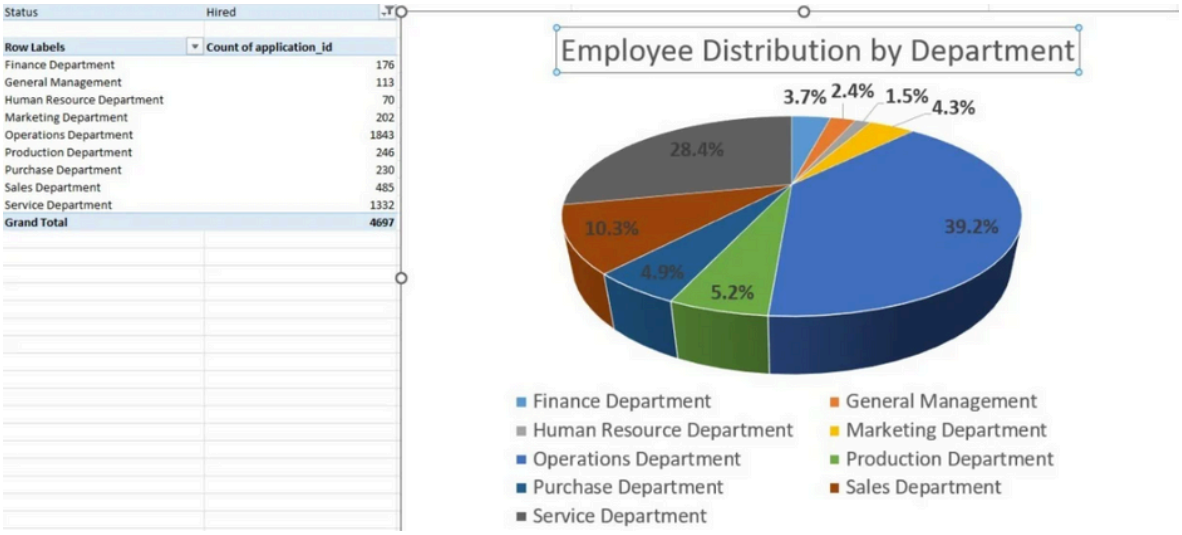
C. Salary Distribution

Your Task: Create class intervals for salaries to analyze the company's salary distribution.



D. Departmental Analysis

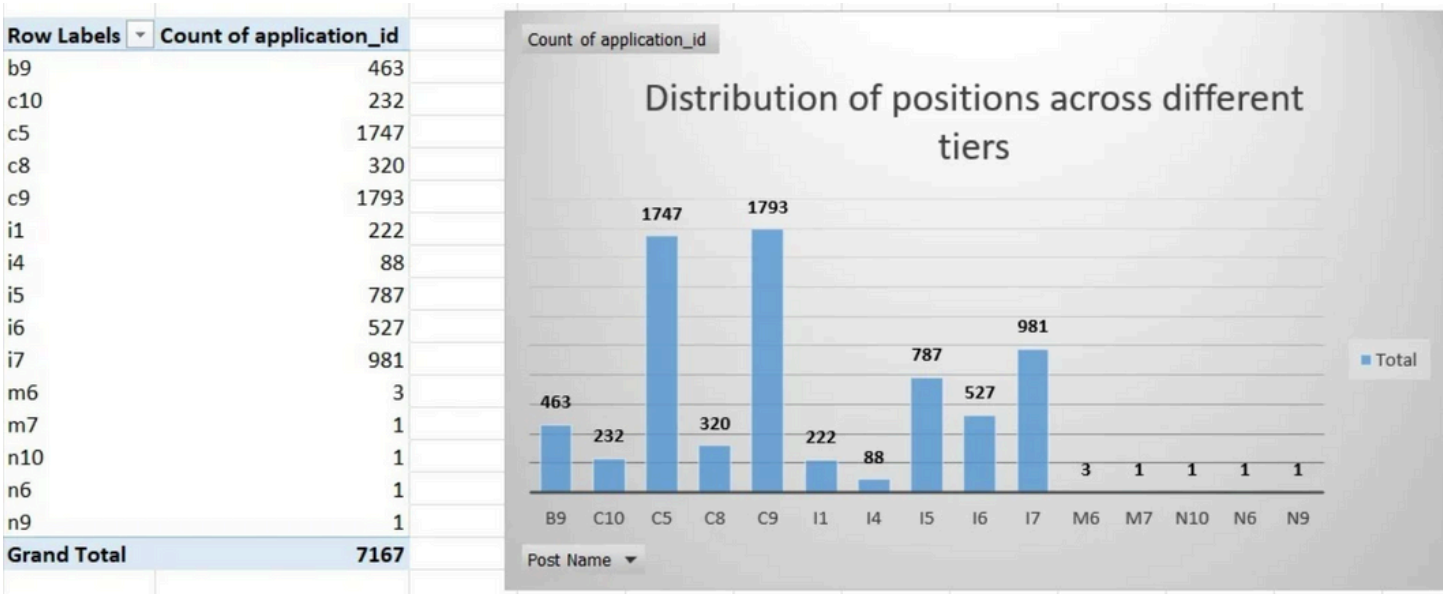
Your Task: Visualize the proportion of employees in different departments using a pie chart, bar graph, or other suitable visualization.



Status	Hired
Row Labels	**Count of application_id**
Finance Department	176
General Management	113
Human Resource Department	70
Marketing Department	202
Operations Department	1843
Production Department	246
Purchase Department	230
Sales Department	485
Service Department	1332
Grand Total	**4697**

E. Position Tier Analysis

Your Task: Represent the distribution of position tiers within the company using charts or graphs.



Row Labels	Count of application_id
b9	463
c10	232
c5	1747
c8	320
c9	1793
i1	222
i4	88
i5	787
i6	527
i7	981
m6	3
m7	1
n10	1
n6	1
n9	1
Grand Total	**7167**

Conclusion

Working on this project provided valuable hands-on experience in using Excel for data analysis and visualization. Each task offered insights that can help improve the company's hiring process, optimize recruitment strategies, and support data-driven decision-making. This analysis has contributed to a clearer understanding of how data can be used to shape workforce diversity, salary fairness, and departmental balance, ultimately supporting the company's human resource objectives.

The ability to extract such insights through Excel’s advanced functions and visualization tools highlighted the importance of data analytics in shaping strategic decisions. It ensures that departments like HR, operations, and management are equipped with actionable metrics to make informed decisions that drive efficiency and enhance the hiring process.

IMDB Movie Analysis

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?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Design

Steps taken to clean the data:

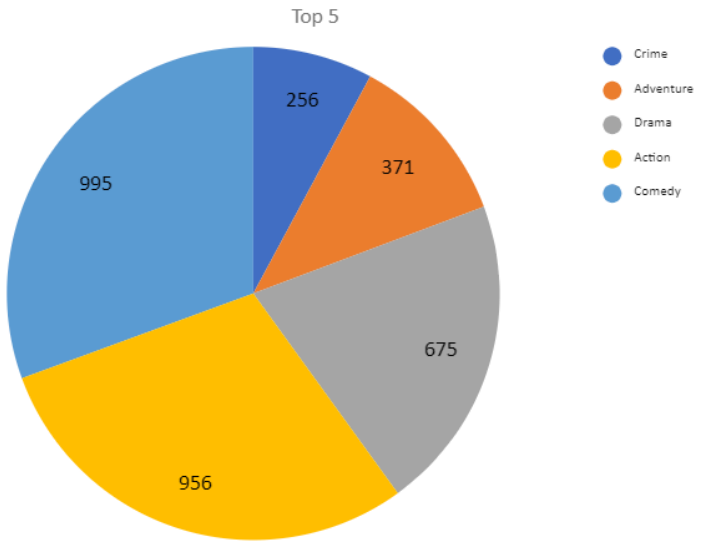
- Handled missing values
- Removed unwanted characters
- Deleted duplicate rows
- Tool Used for Visualization: Microsoft Excel 365

Insights

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

- **Task:** Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

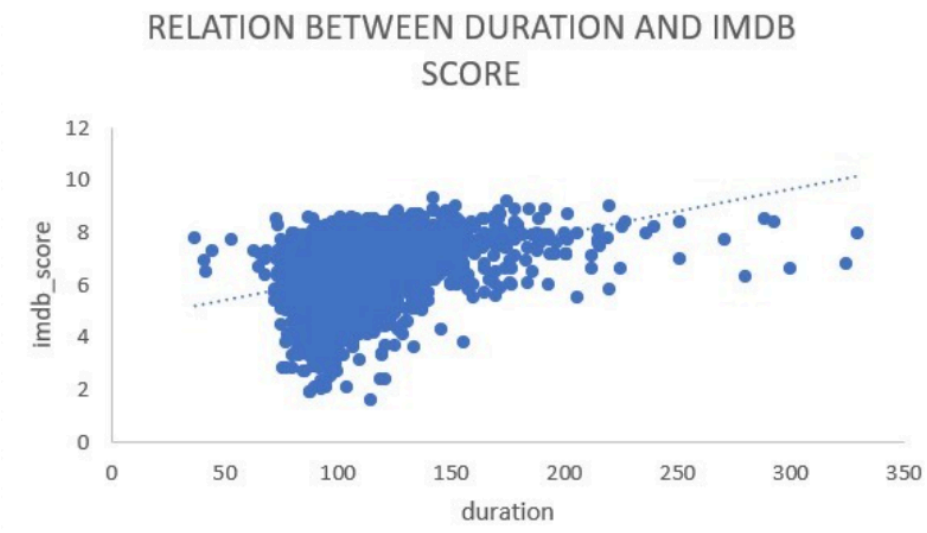
Genre Manipulation	
Median	675
Max	995
Min	256
Mode	#N/A
Var	111590.3
Std	334.0513434



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

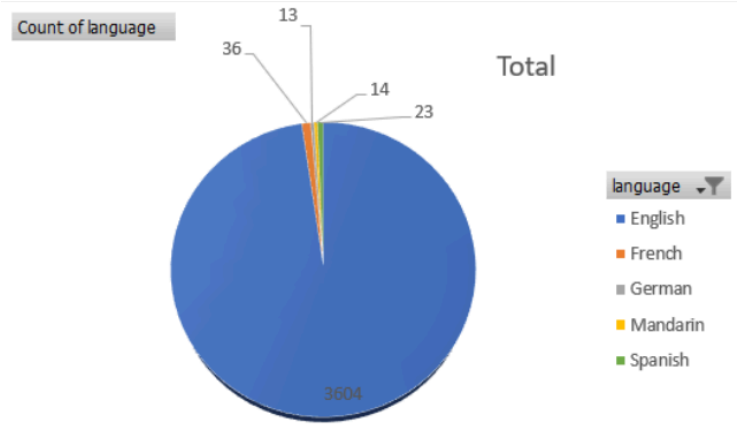
- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

DURATION STATISTICS	
Mean	110.1837437
Median	106
Std	22.77240051



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

- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

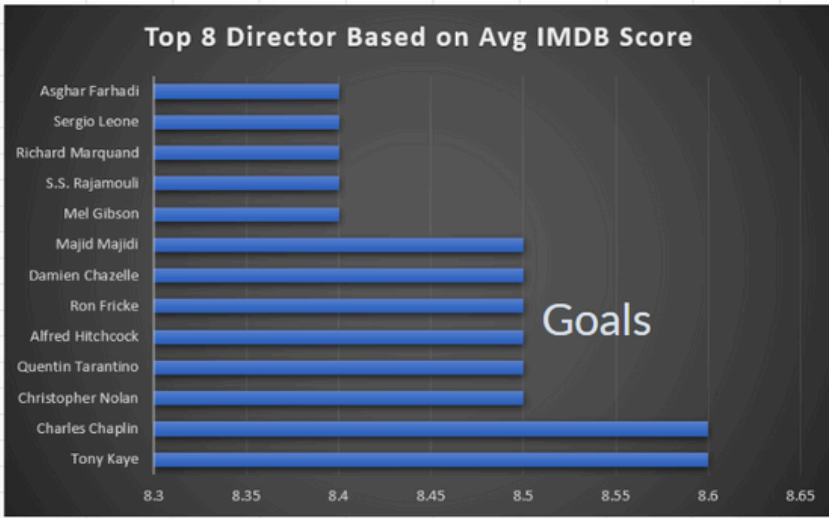


D. Director Analysis: Influence of directors on movie ratings.

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

director_name	imdb_score	avg
Tony Kaye	8.6	8.6
Charles Chaplin	8.6	8.6
Christopher Nolan	8.5	8.5
Quentin Tarantino	8.5	8.5
Alfred Hitchcock	8.5	8.5
Ron Fricke	8.5	8.5
Damien Chazelle	8.5	8.5
Majid Majidi	8.5	8.5
Mel Gibson	8.4	8.4
S.S. Rajamouli	8.4	8.4
Richard Marquand	8.4	8.4
Sergio Leone	8.4	8.4
Asghar Farhadi	8.4	8.4

90 percentile
7.5



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

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

gross	budget	Profit Margin
760505847	237000000	523505847
652177271	150000000	502177271
658672302	200000000	458672302
460935665	11000000	449935665
434949459	10500000	424449459
623279547	220000000	403279547
422783777	45000000	377783777
474544677	115000000	359544677
533316061	185000000	348316061
407999255	78000000	329999255
363024263	58000000	305024263
424645577	130000000	294645577
356784000	63000000	293784000
368049635	76000000	292049635
350123553	58800000	291323553
380838870	94000000	286838870
436471036	150000000	286471036
377019252	94000000	283019252
309125409	32500000	276625409
329691196	55000000	274691196
290158751	18000000	272158751
285761243	18000000	267761243
380262555	113000000	267262555
403706375	139000000	264706375
336029560	74000000	262029560

highest profit margin
52350584

correlation function
0.09515869

Conclusion

Working on this analysis allowed me to dive deep into various factors influencing the success of a movie on IMDB. Each task provided valuable insights into how genres, duration, language, directors, and budgets affect movie ratings. The project helped me understand the complexities of the movie industry and the importance of leveraging data for decision-making.

The ability to analyze and visualize data in Excel gave me hands-on experience with real-world data, improving my skills in data manipulation, statistical analysis, and data storytelling. By using descriptive statistics, correlation analysis, and visualization techniques, I was able to uncover patterns and trends that provide actionable insights for movie producers, directors, and investors.

This analysis not only shed light on what factors contribute to a movie's success but also demonstrated the power of data-driven decision-making. The insights gained from this project can assist stakeholders in making informed choices about future movie productions, improving their chances of success in a highly competitive industry.

Bank Loan Case Study

Problem Statement

The finance company faces challenges in accurately assessing loan applicants, leading to potential financial risks. Approving applicants who cannot repay results in losses, while rejecting capable applicants causes missed business opportunities. Using Exploratory Data Analysis (EDA), your task is to analyze patterns in loan application data to improve decision-making. The dataset includes loan scenarios like payment difficulties and timely payments, with outcomes such as approved, cancelled, refused, or unused offers. Your analysis aims to identify trends and ensure fair and accurate loan approvals, minimizing risks and maximizing profitability.

Design

Steps taken to clean the data:

- Handled missing values
- Removed unwanted characters
- Deleted duplicate rows
- Removed Outlier
- Tool Used for Visualization: Microsoft Excel 365

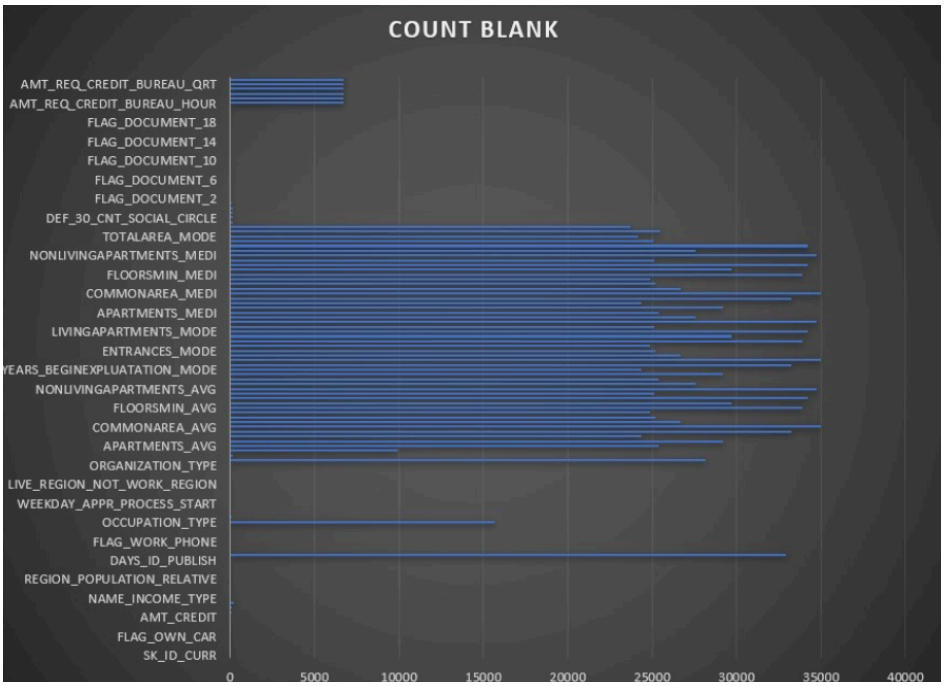
Insights

A. **Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

- **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.

Identified missing data using CountBlank function

Row	COUNT BLANK
SK_ID_CURR	0
TARGET	0
NAME_CONTRACT	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REAL	0
CNT_CHILDREN	0
AMT_INCOME_TC	0
AMT_CREDIT	0
AMT_ANNUITY	1
AMT_GOODS_PRI	38
NAME_TYPE_SUIT	192
NAME_INCOME_T	0
NAME_EDUCATIO	0
NAME_FAMILY_ST	0
NAME_HOUSING_	0
REGION_POPULAT	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
DAYS_REGISTRATI	0
DAYS_ID_PUBLISH	0
OWN_CAR_AGE	32950
FLAG_MOBIL	0
FLAG_EMP_PHON	0
FLAG_WORK_PHC	0
FLAG_CONT_MOB	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYP	15654
CNT_FAM_MEMBI	1



Created relevant columns

R	S	T	U	V	W	X	Y
DAYS_BIRTH	AGE	DAYS_EMPLOYED	YRS_EMPLOYED	DAYS_REGISTRATION	YRS_REG	DAYS_ID_PUBLISH	YRS_PUBLISHED
-9461	26.28	-637	1.77	-3648	10.13	-2120	5.89
-16765	46.57	-1188	3.30	-1186	3.29	-291	0.81
-19046	52.91	-225	0.63	-4260	11.83	-2531	7.03
-19005	52.79	-3039	8.44	-9833	27.31	-2437	6.77
-19932	55.37	-3038	8.44	-4311	11.98	-3458	9.61
-16941	47.06	-1588	4.41	-4970	13.81	-477	1.33
-13778	38.27	-3130	8.69	-1213	3.37	-619	1.72
-18850	52.36	-449	1.25	-4597	12.77	-2379	6.61
-20099	55.83	365243	1014.56	-7427	20.63	-3514	9.76
-14469	40.19	-2019	5.61	-14437	40.10	-3992	11.09
-10197	28.33	-679	1.89	-4427	12.30	-738	2.05
-20417	56.71	365243	1014.56	-5246	14.57	-2512	6.98
-13439	37.33	-2717	7.55	-311	0.86	-3227	8.96
-14086	39.13	-3028	8.41	-643	1.79	-4911	13.64
-14583	40.51	-203	0.56	-615	1.71	-2056	5.71
-8728	24.24	-1157	3.21	-3494	9.71	-1368	3.80
-12931	35.92	-1317	3.66	-6392	17.76	-3866	10.74
-9776	27.16	-191	0.53	-4143	11.51	-2427	6.74
-17718	49.22	-7804	21.68	-8751	24.31	-1259	3.50
-11348	31.52	-2038	5.66	-1021	2.84	-3964	11.01
-18252	50.70	-4286	11.91	-298	0.83	-1800	5.00
-14815	41.15	-1652	4.59	-2299	6.39	-2299	6.39
-11146	30.96	-4306	11.96	-114	0.32	-2518	6.99
-24827	68.96	365243	1014.56	-9012	25.03	-3684	10.23
-11286	31.35	-746	2.07	-108	0.30	-3729	10.36
-19334	53.71	-3494	9.71	-2419	6.72	-2893	8.04

Replaced blanks of numerical column with mode and categorical column with maximum category used using Pivot Table

Row Labels	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
(blank)	
Grand Total	**49807**

B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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

COLUMNS	Q1	Q3	IQR	UB	LB
CNT_CHILDREN	0	1	1	2.5	-1.5
AMT_INCOME_TOTAL	112500	202500	90000	337500	-22500
AMT_CREDIT	270000	808650	538650	1616625	-538975
AMT_ANNUITY	16457	34596	18139.5	61705.25	-10652.75
AMT_GOODS_PRICE	238500	679500	441000	1341000	-423000
REGION_POPULATION_RELATIVE	0.01	0.0287	0.018657	0.056648	-0.018979
YRS	34.383	54.567	19.7334	84.1667	5.2333
YRS_REG	5.55	20.73	15.18	43.5	-17.22
YRS_EMPLOYED	2.5917	15.883	13.29166	35.82083	-17.34583

Visualization

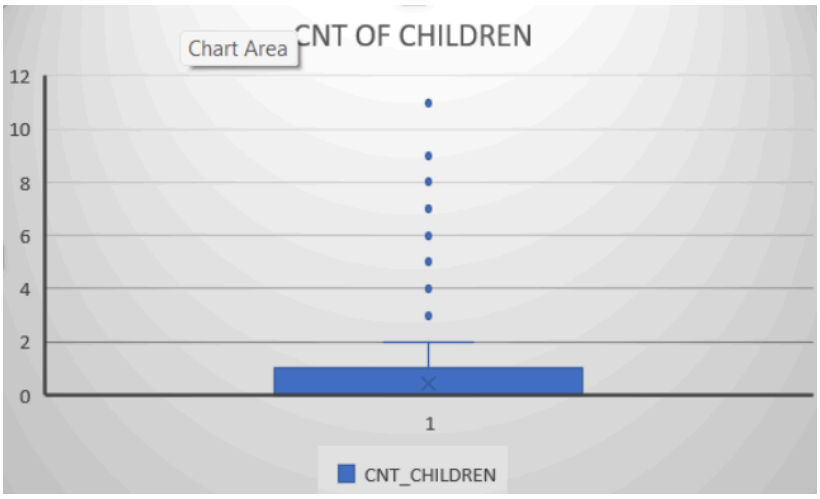


Fig 1

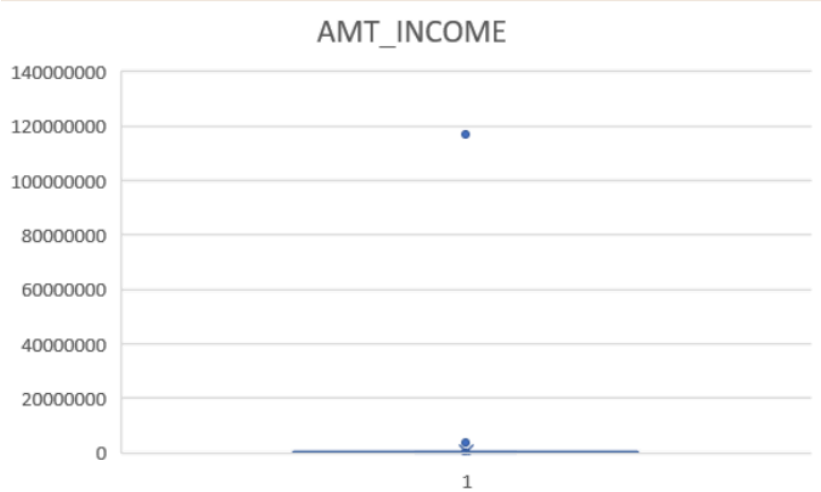


Fig 2

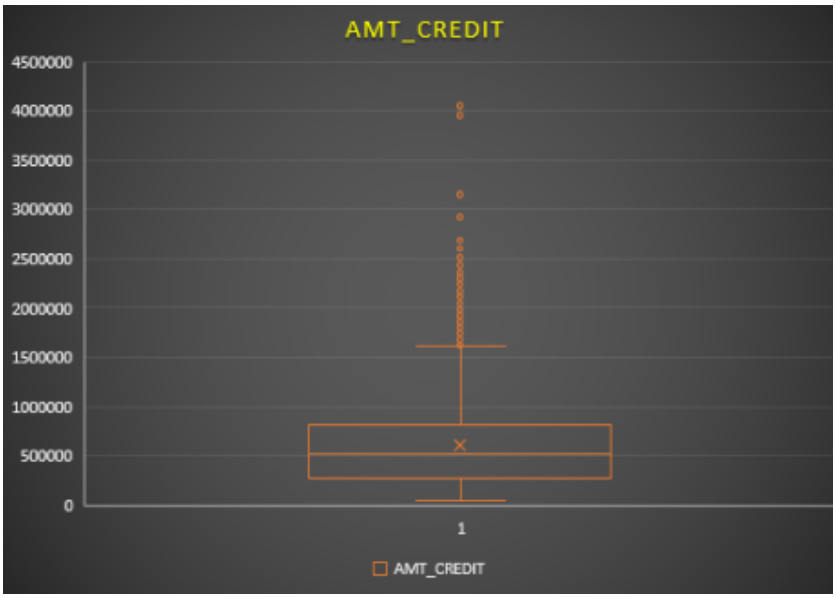


Fig 3

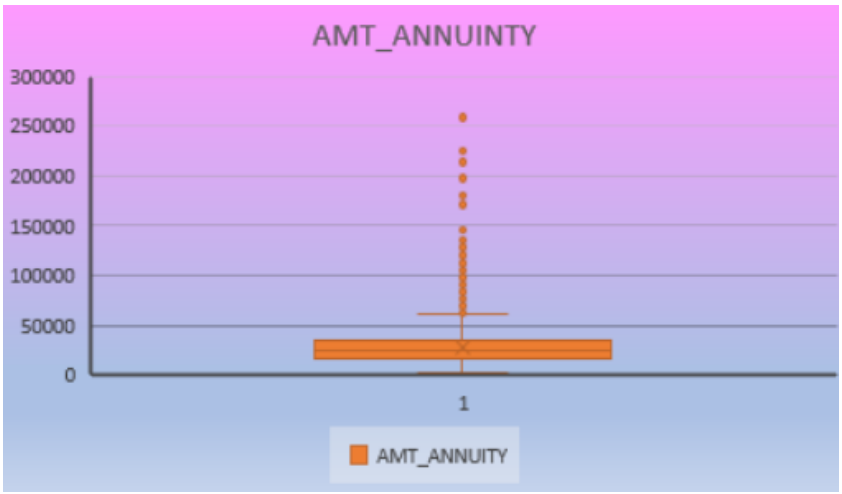


Fig 4

Fig 5

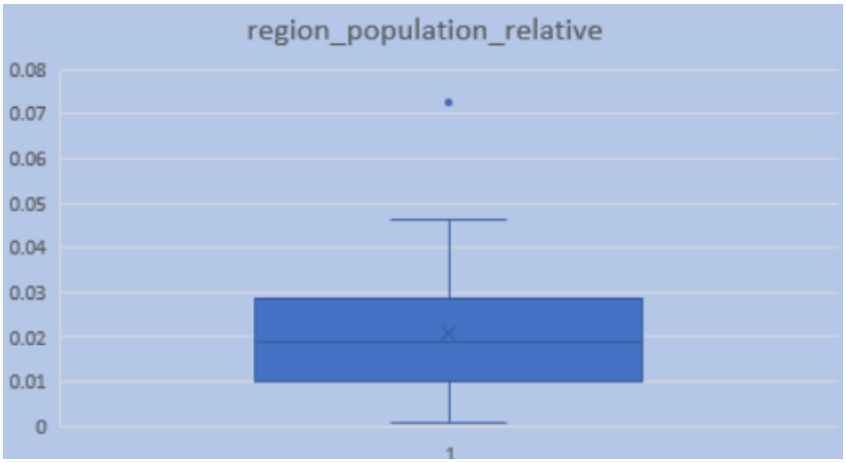


Fig 6

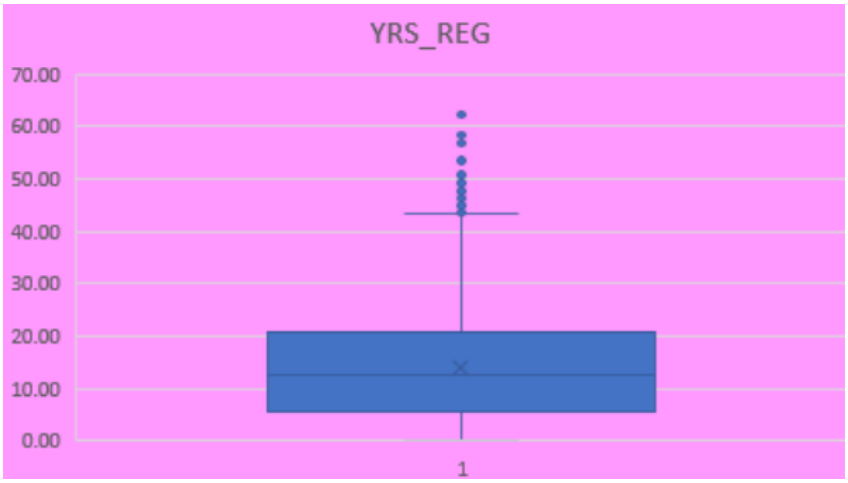


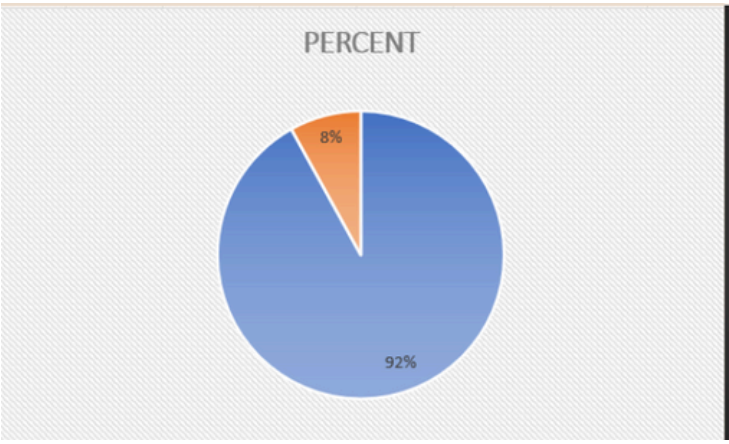
Fig 7

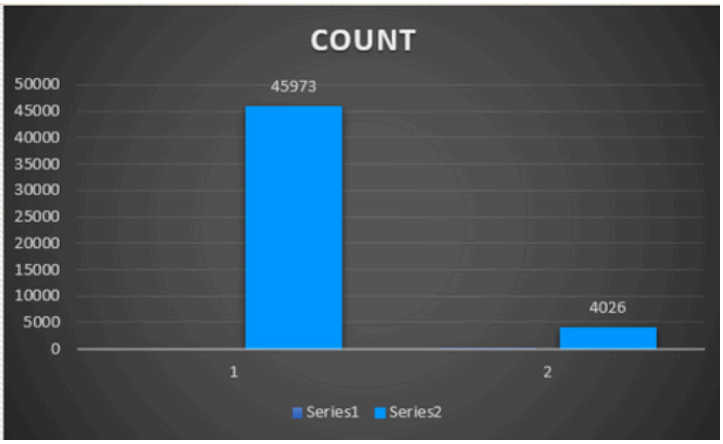
Fig 8

C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

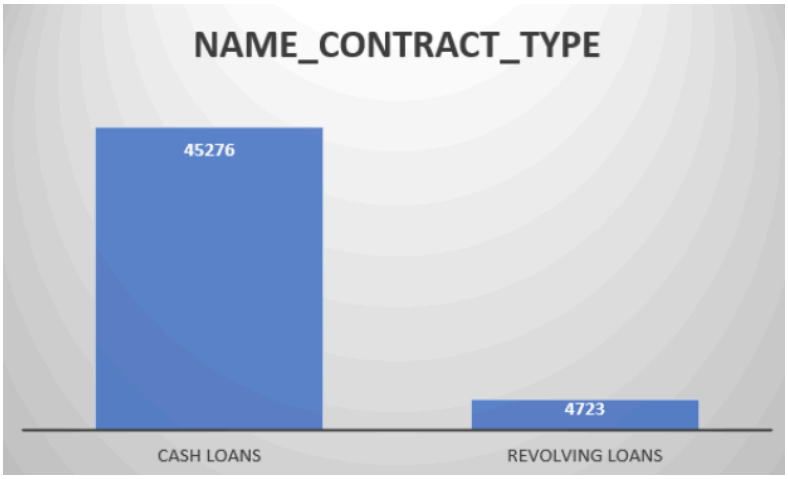
- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

IMBALANCE BETWEEN DEFAULTERS AND NON-DEFAULTERS





IMBALANCE BETWEEN CASH LOAN AND REVOLVING LOAN



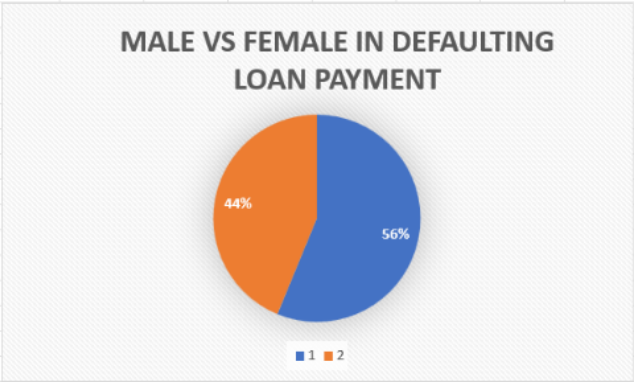
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

- **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.

<u>UNIVARIATE</u>

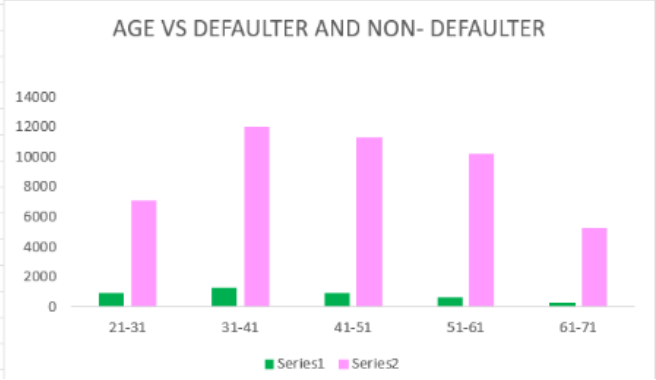
- GENDER

	NON -DEFAULTERS		DEFAULTERS
F	30559	F	2264
M	15412	M	1762



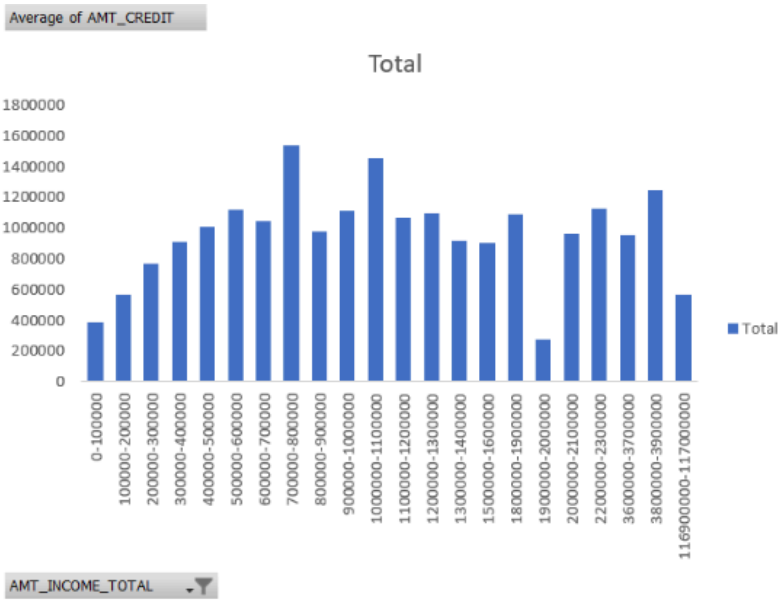
- AGE

AGE	DEFAULTER	NON-DEFAULTER
21-31	910	7114
31-41	1264	12046
41-51	919	11291
51-61	651	10211
61-71	282	5311

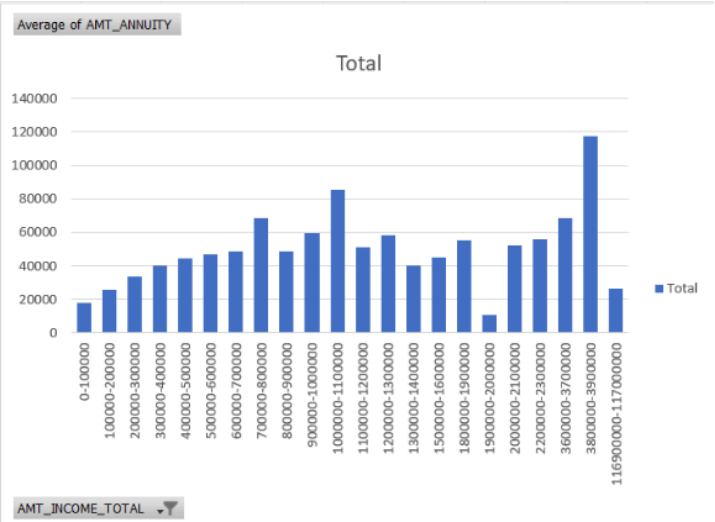


BIVARIATE

- **Avg of AMT_CREDIT VS AMT_INCOME_TOTAL**

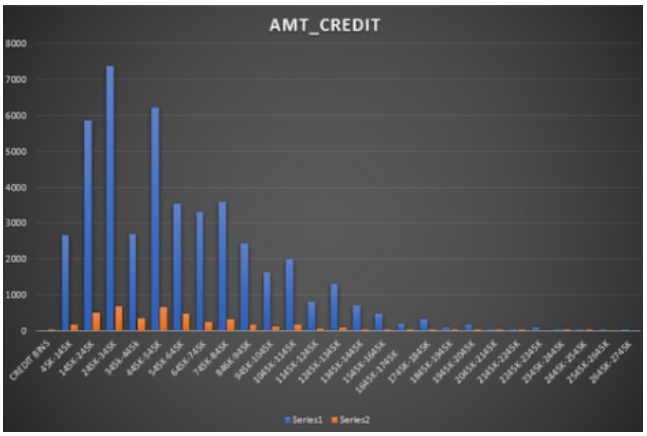


- **Avg of AMT_ANNUITY VS AMT_INCOME_TOTAL**

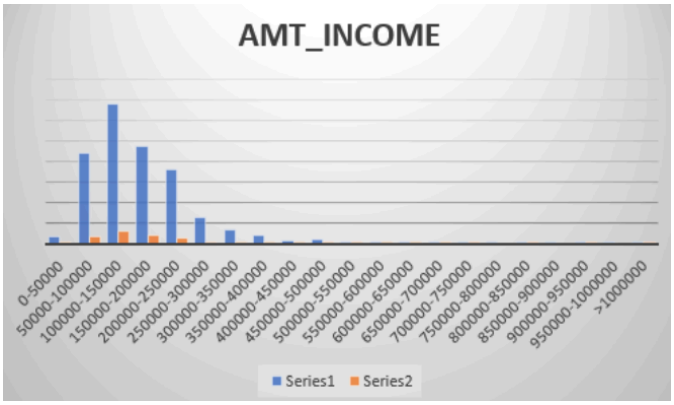


SEGMENTED UNIVARIATE

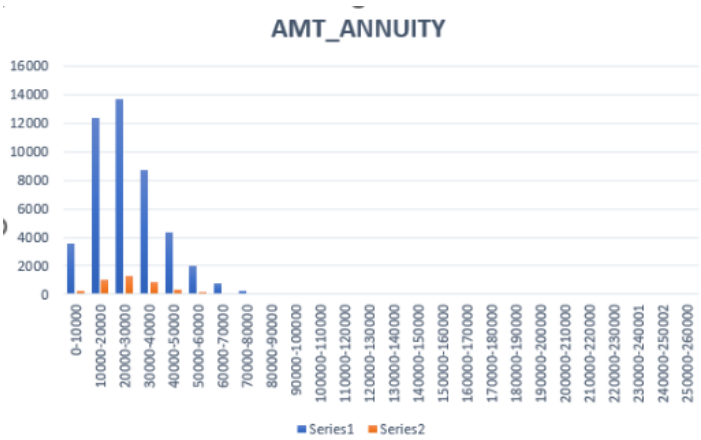
- **AMT_CREDIT**



AMT_INCOME



- **AMT_ANNUITY**



E. **Identify Top Correlations for Different Scenarios:** Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

- **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.

CORRELATION	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUIT	AMT_GOOD	REG_POPULAT	AGE	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PH	FLAG_CONT_M	FLAG_PHONE	FLAG_EMAIL
AMT_INCOME_TOTAL	1											
AMT_CREDIT	0.069315897	1										
AMT_ANNUITY	0.083008508	0.76949785	1									
AMT_GOODS_PRICE	0.069884774	0.98660712	0.7740223	1								
REG_POPULATION_RELATIVE	0.029841469	0.095111221	0.115111316	0.0991621	1							
AGE	-0.016002774	0.059342658	-0.00771213	0.0991621	0.032513748	1						
FLAG_MOBIL	0.000376468	0.003568428	0.00036741	0.0991621	0.003300527	0.007637377	1					
FLAG_EMP_PHONE	0.031568121	0.069063079	0.109570201	0.0991621	0.004185769	-0.617702887	-0.002084846	1				
FLAG_WORK_PHONE	-0.009180815	-0.015114949	-0.02096375	0.0991621	-0.016786854	-0.175689744	0.002230968	0.232551933	1			
FLAG_CONT_MOBILE	-0.003245731	0.024481204	0.023110819	0.0991621	-0.005049899	0.01190043	-0.000201207	-0.015157451	0.02244341	1		
FLAG_PHONE	-0.002044185	0.019460051	0.005127012	0.0991621	0.094068351	0.044218509	0.002773195	-0.023557669	0.297934688	0.005022897	1	
FLAG_EMAIL	0.01521548	0.010811876	0.065896194	0.0991621	0.039507678	-0.092149777	0.001085765	0.066934685	-0.008201041	-0.010449537	0.015989278	1

Conclusion

Working on this project provided significant hands-on experience in utilizing Exploratory Data Analysis (EDA) techniques to address challenges in loan approval processes. Each task offered critical insights into how customer and loan attributes influence repayment behavior, enabling the identification of patterns that can predict loan defaults. This analysis contributed to a deeper understanding of risk factors, such as data imbalance, outliers, and correlations, which are vital for shaping data-driven lending strategies.

By leveraging Excel's powerful functions and visualization tools, the project highlighted the importance of data analytics in financial decision-making. It ensured that actionable insights, such as identifying risky applicants or adjusting loan terms, could be derived to mitigate financial losses and enhance business outcomes. This approach underscores the role of data-driven strategies in fostering sustainable growth and improving customer experience in the finance sector.

Impact Of Car Features

Problem Statement

How can a car manufacturer optimize pricing and product development to maximize profitability while meeting consumer demand? The challenge involves analyzing the relationship between a car's features, market category, and pricing to identify the most popular and profitable options. By leveraging data analysis techniques such as regression analysis and market segmentation, the manufacturer aims to develop a strategic pricing model and focus on key features that appeal to consumers, ensuring competitiveness and profitability in a dynamic automotive market.

Design

Steps taken to clean the data:

- Handled missing values
- Removed unwanted characters
- Deleted duplicate rows
- Removed irrelevant columns
- Removed Outlier
- Tool Used for Visualization: Microsoft Excel 365

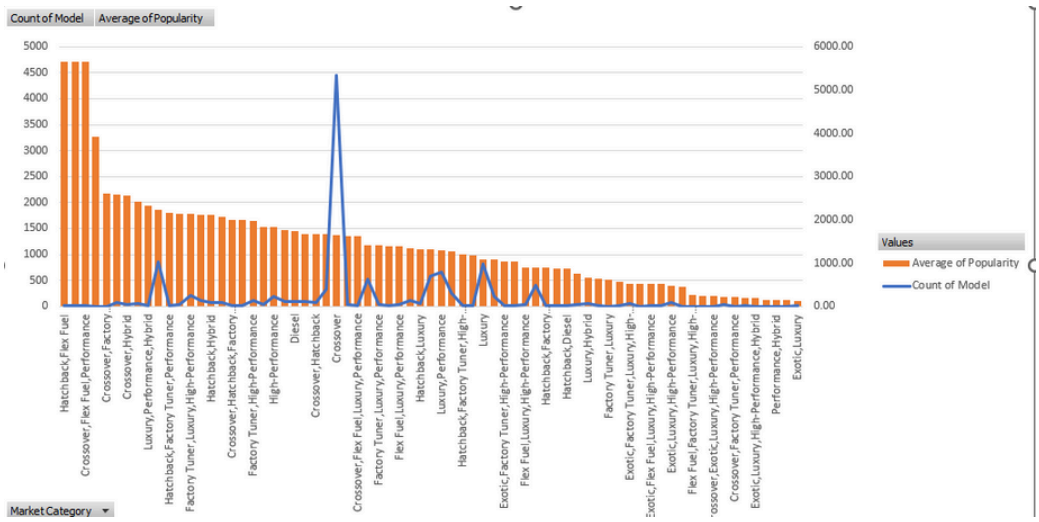
Insights

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.

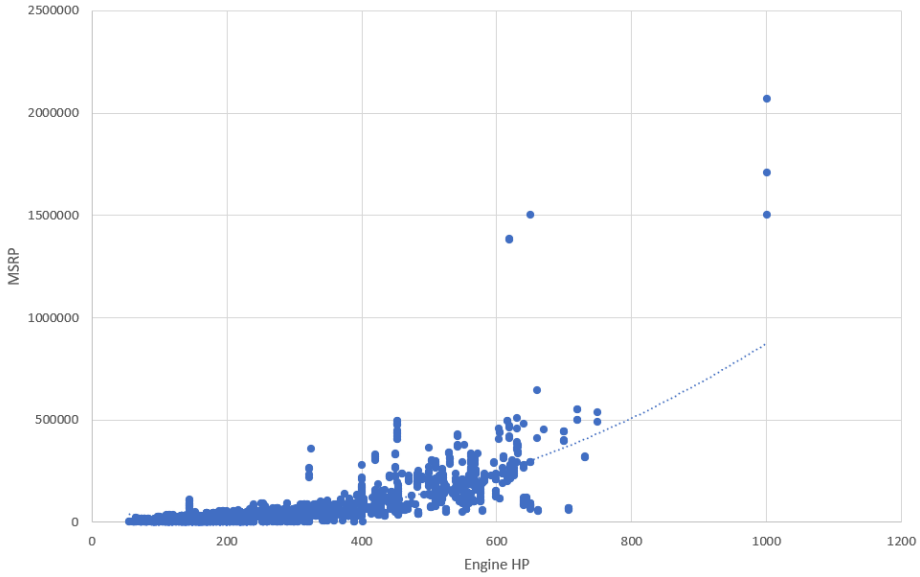
Row Labels	Count of Model	Average of Popularity
Hatchback,Flex Fuel	7	5657.00
Flex Fuel,Diesel	16	5657.00
Crossover,Flex Fuel,Performance	6	5657.00
Crossover,Luxury,Performance,Hybrid	2	3916.00
Crossover,Factory Tuner,Luxury,Performance	5	2607.40
Crossover,Performance	69	2585.96
Crossover,Hybrid	42	2563.38
Diesel,Luxury	47	2416.11
Luxury,Performance,Hybrid	11	2333.18
Flex Fuel	855	2225.71
Hatchback,Factory Tuner,Performance	21	2173.71
Crossover,Luxury,Diesel	34	2149.41
Factory Tuner,Luxury,High-Performance	215	2133.37
Hybrid	121	2116.59
Hatchback,Hybrid	64	2111.16
Crossover,Flex Fuel	64	2073.75
Crossover,Hatchback,Factory Tuner,Performance	6	2009.00
Crossover,Hatchback,Performance	6	2009.00
Factory Tuner,High-Performance	104	1966.44
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.46
High-Performance	198	1823.38
Factory Tuner,Performance	84	1774.05
Diesel	84	1730.90
Flex Fuel,Performance	87	1680.47
Crossover,Hatchback	72	1675.69
Luxury,High-Performance	334	1668.02
Crossover	4451	1638.59
Hatchback,Luxury,Performance	36	1632.25
Crossover,Flex Fuel,Luxury,Performance	6	1624.00
Performance	520	1415.21
Factory Tuner,Luxury,Performance	31	1413.42
Exotic,Performance	6	1391.00
Flex Fuel,Luxury,Performance	28	1380.07
Crossover,Luxury,Performance	112	1349.09

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



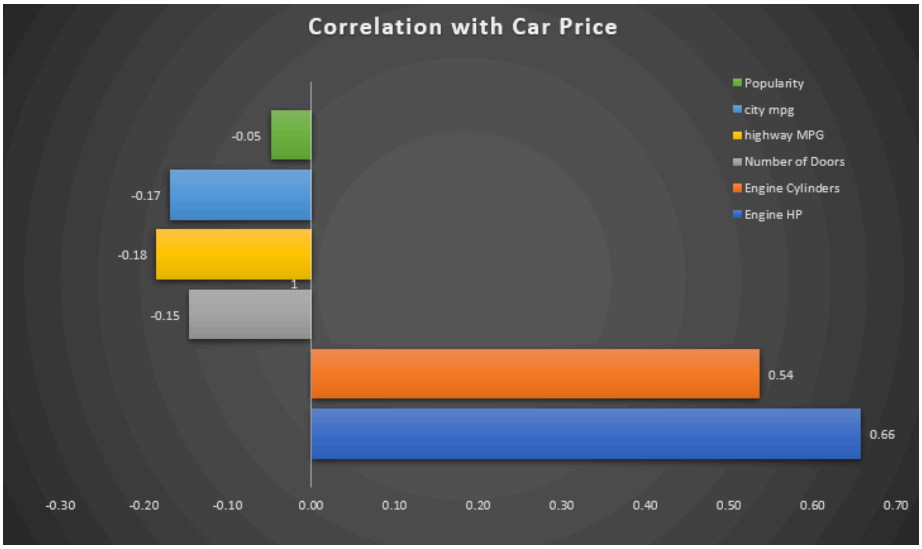
Insight Required: What is the relationship between a car's engine power and its price?

- **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.

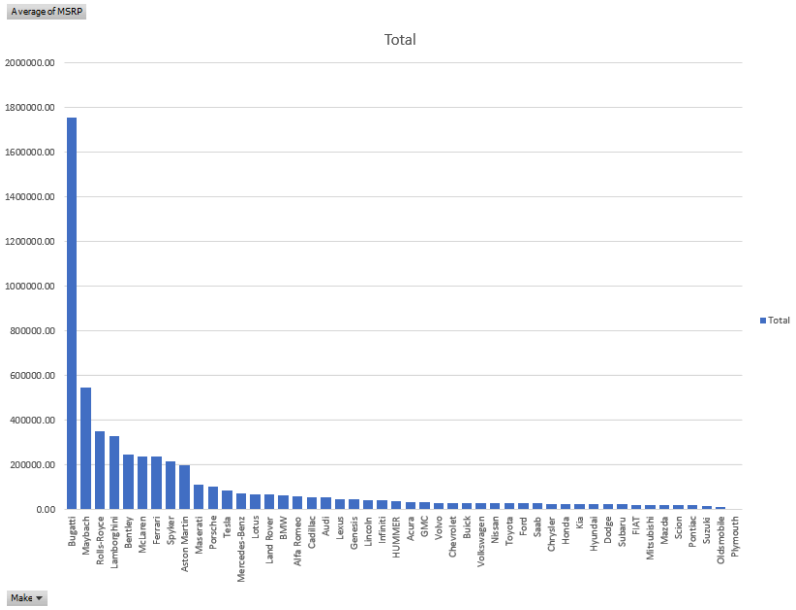


Insight Required: How does the average price of a car vary across different manufacturers?

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

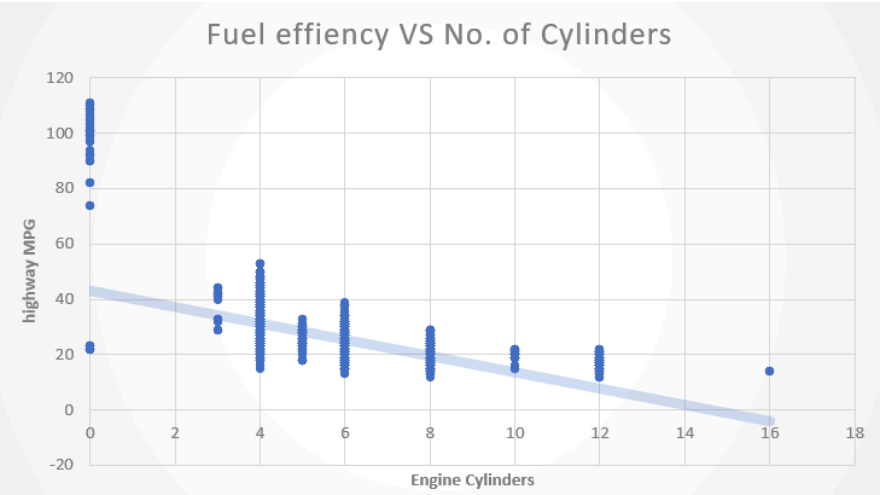
Row Labels	Average of MSRP
Bugatti	1757223.67
Maybach	546221.88
Rolls-Royce	351130.65
Lamborghini	331567.31
Bentley	247169.32
McLaren	239805.00
Ferrari	238218.84
Spyker	214990.00

- **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.

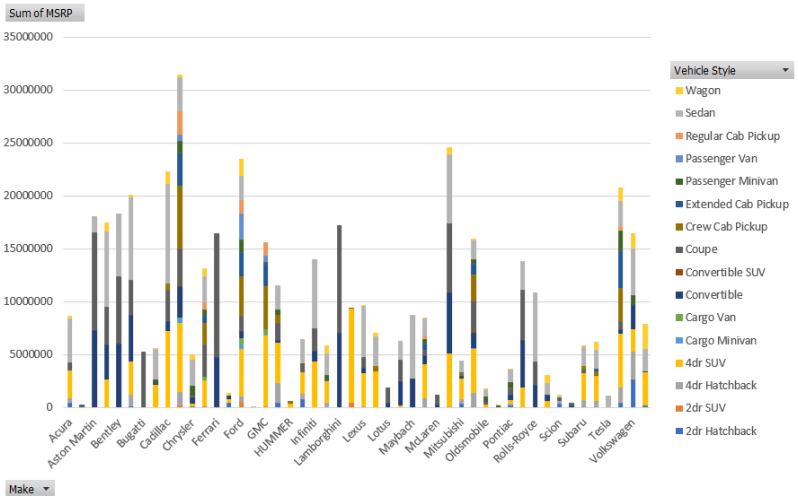


- **Task 5.B:** Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

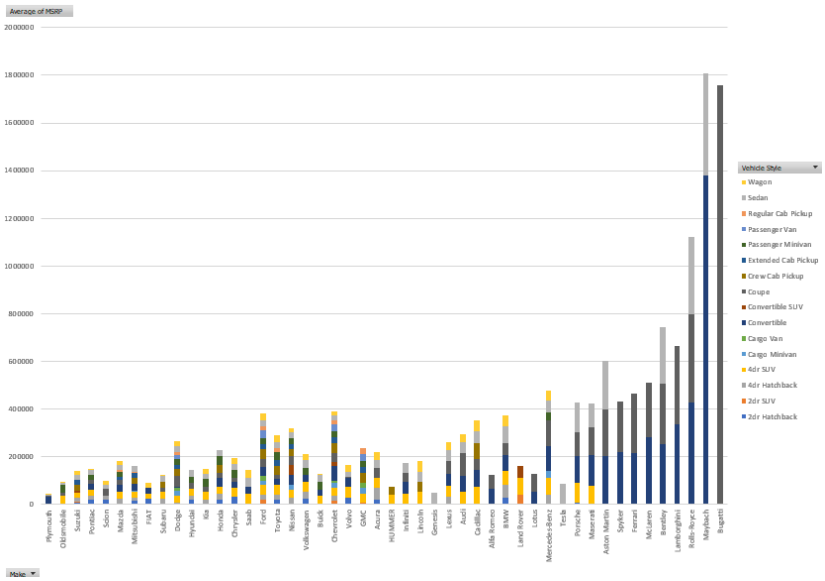
Correlation
-0.64801

Building the Dashboard:

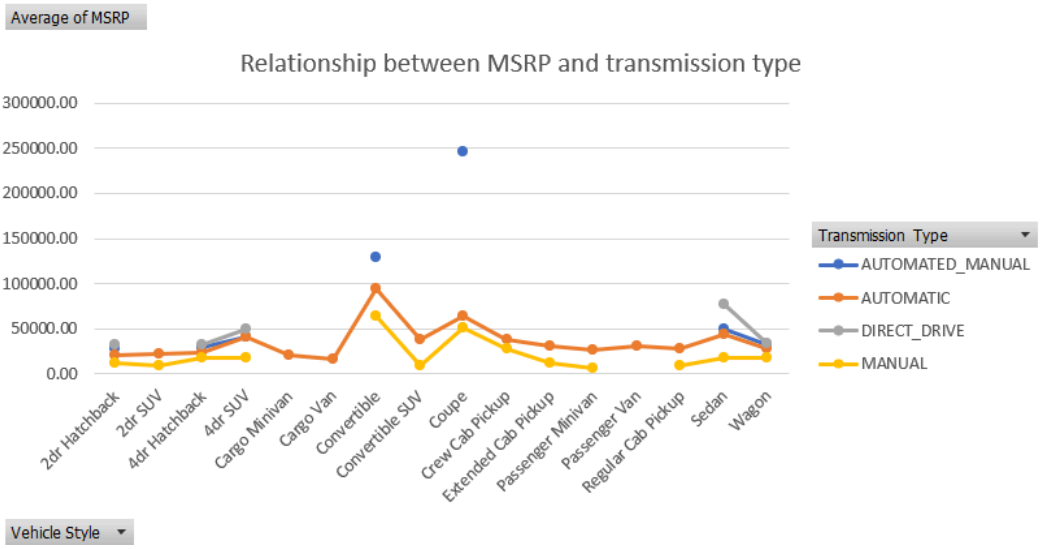
Task 1: How does the distribution of car prices vary by brand and body style?



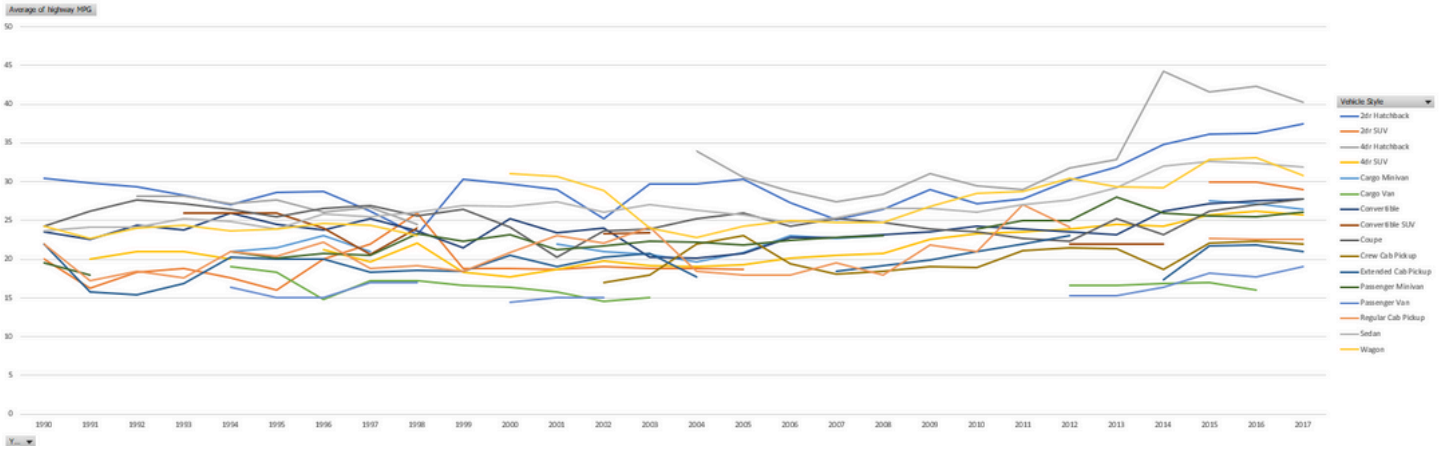
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?



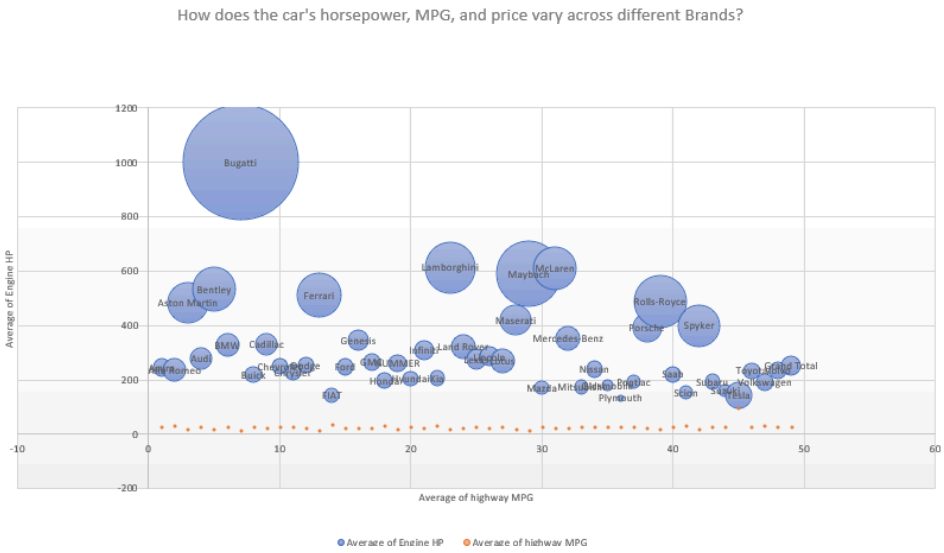
Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?



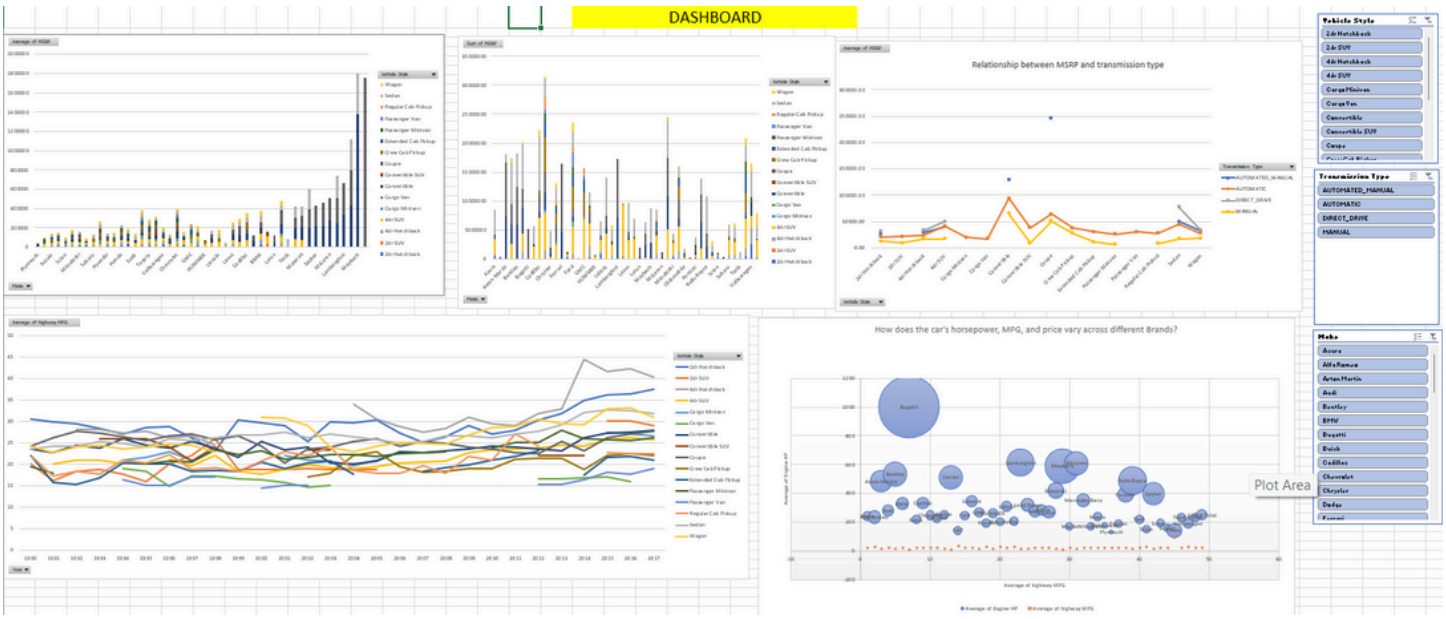
Task 4: How does the fuel efficiency of cars vary across different body styles and model years?



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



DASHBOARD



Conclusion

The analysis of the "Car Features and MSRP" dataset provides critical insights for optimizing pricing and product development decisions in the automotive industry. By examining popularity trends across market categories, we find that luxury and performance cars have high popularity despite their premium pricing, indicating a demand for premium features. Regression analysis highlights engine power, fuel efficiency, and vehicle size as significant predictors of price, guiding manufacturers on feature prioritization. The scatter plot of engine cylinders versus MPG confirms that higher cylinders correlate with lower fuel efficiency, underscoring the need for innovation in engine technology. Insights into the varying average prices across manufacturers and body styles can inform competitive pricing strategies. Lastly, interactive dashboards combining bubble charts, stacked columns, and trend lines empower stakeholders to make data-driven decisions, ultimately improving market competitiveness and profitability.

ABC Call Volume Trend

Problem Statement

The inbound calling team plays a critical role in shaping customer experience, but the effectiveness of their operations remains unclear. This project aims to analyze a 23-day dataset of call center activities, including agent performance, call queue times, durations, and statuses (answered, abandoned, or transferred). The objective is to identify trends, inefficiencies, and opportunities for improvement in call handling processes. By deriving actionable insights, the goal is to enhance customer satisfaction, optimize resource allocation, and support data-driven decision-making for the CX team. This analysis will also help improve the company's competitive edge in delivering exceptional customer experiences.

Design

Steps taken to clean the data:

- Handled missing values
- Removed unwanted characters
- Deleted duplicate rows
- Removed irrelevant columns
- Removed Outlier
- Tool Used for Visualization: Microsoft Excel 365

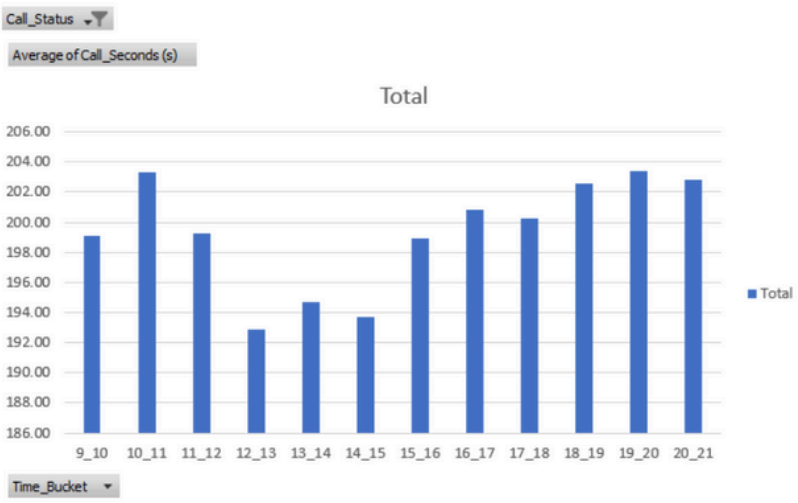
Assumptions

An agent works for 6 days a week; On average, each agent takes 4 unplanned leaves per month; An agent's total working hours are 9 hours, out of which 1.5 hours are spent on lunch and snacks in the office. On average, an agent spends 60% of their total actual working hours (i.e., 60% of 7.5 hours) on calls with customers/users. The total number of days in a month is 30.

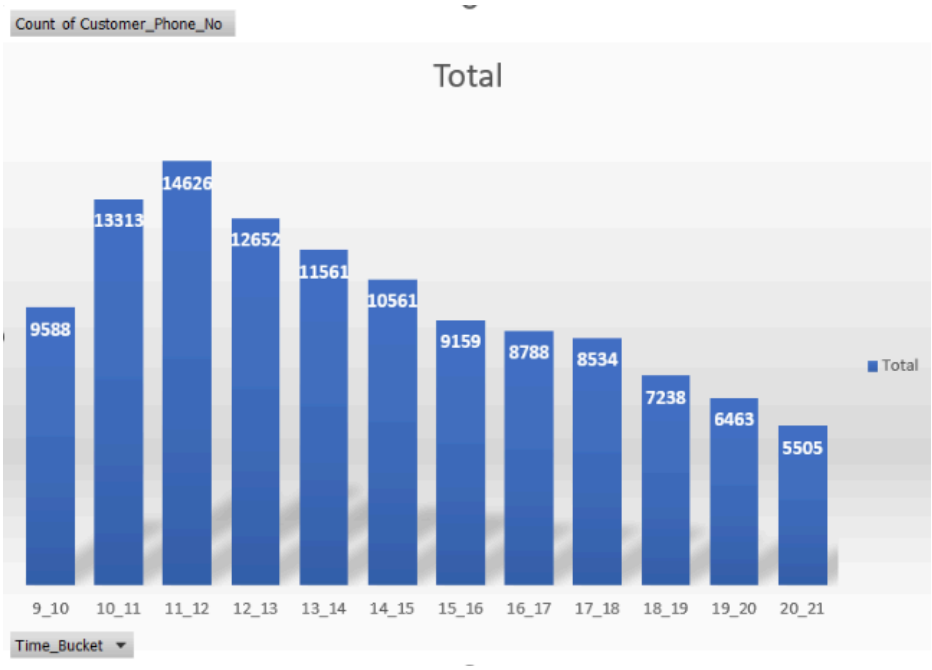
Insights

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

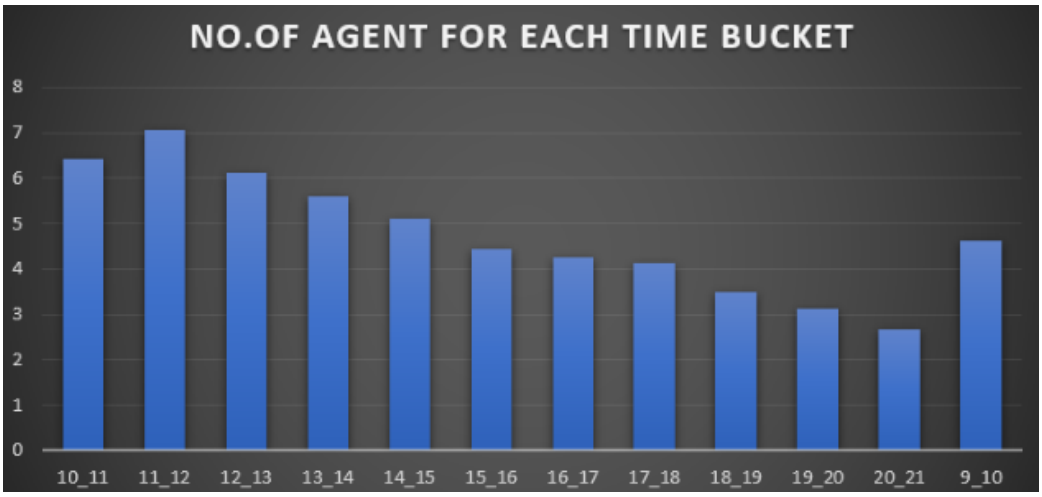
Call_Status	answered
Row Labels	**Average of Call_Seconds (s)**
9_10	199.07
10_11	203.33
11_12	199.26
12_13	192.89
13_14	194.74
14_15	193.68
15_16	198.89
16_17	200.87
17_18	200.25
18_19	202.55
19_20	203.41
20_21	202.85
Grand Total	**198.62**



1. **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.).



1. **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.



1. **Night Shift Manpower Planning:** Customers also call ABC Insurance Company at night but don't get an answer because there are no agents available. This creates a poor customer experience. Assume that for every 100 calls that customers make between 9 am and 9 pm, they also make 30 calls at night between 9 pm and 9 am. The distribution of these 30 calls is as follows:

Conclusion

The analysis of the inbound call dataset provides valuable insights into optimizing customer experience (CX) and improving operational efficiency for the company. By examining call volume trends, average call durations, and abandonment rates across time buckets, key patterns have been identified to better align manpower allocation with peak call periods. Regression analysis highlights the relationship between queue times, agent availability, and call outcomes, offering a framework for reducing abandon rates. Insights from night-shift call patterns underline the need to expand agent availability during off-peak hours to enhance customer satisfaction. Interactive dashboards, combining time-series analysis and graphical representations, empower stakeholders to monitor trends, identify gaps, and make informed, data-driven decisions. These findings enable the company to refine CX strategies, optimize resource utilization, and deliver a seamless and responsive experience for customers.

Conclusion

Through these diverse analytics projects, I have learned how data can be harnessed to solve real-world problems, optimize processes, and drive impactful decisions. Each project provided unique insights, from improving customer engagement on social media platforms to enhancing operational efficiency, refining hiring strategies, and understanding audience behavior. Whether it involved analyzing call center trends, evaluating car features, or assessing financial risks, these experiences demonstrated the value of structured analytics in uncovering actionable insights and solving challenges across various domains.

These learnings underline how analytics is transforming the world. By enabling data-driven decision-making, businesses can improve efficiency, reduce costs, and deliver personalized, customer-centric solutions. Beyond just improving individual organizations, analytics has a broader societal impact, fueling innovation, adapting to market demands, and driving sustainable growth. It highlights the power of turning raw data into meaningful insights that not only solve problems but also create opportunities for progress and betterment in every sphere of life.

APPENDIX

PPT LINK:

Data Analytics Process : 🌐 DATA ANALYTICS.pptx

Instagram User Analytics: 🌐 Instagram User Analytics.pdf

Operation & Metric Analytics: 🌐 OPERATION AND METRIC ANALYTICS.pdf

Hiring Process Analytics: 🌐 Hiring Process Analytic .pdf

IMDB Movie Analysis: 🌐 bankloan.pptx

Bank Loan Case Study: 🌐 BANK LOAN CASE STUDY.pdf

Impact of Car Features: 🌐 IMPACT OF CAR FEATURE ON PRICE.pdf

ABC Call Volume Trend: 🌐 CALL_VOLUME_TREND.pdf

PRESENTATION LINK

IMDB Movie Analysis: 🌐 IMDb Movie Analytics Insights

Bank Loan Case Study: 🌐 Loan Application Data Analysis 📊

Impact of Car Features: 🌐 Analysis of the Automotive Industry

ABC Call Volume Trend:
🌐 Blue and Green Modern Professional Company Profile Presentation - Presentation - 20 D...