



DATA ANALYSIS PORTFOLIO

By Mirra. G

PROFESSIONAL BACKGROUND

I am currently in the third year of my B.Tech in Artificial Intelligence and Data Science, where I have immersed myself in the study of cutting-edge technologies and methodologies. Throughout my academic journey, I have cultivated a strong foundation in Machine Learning, Data Analysis, and Data Visualization, which has been complemented by hands-on experience in various university-level projects.

My technical proficiency extends to Python, where I am well-versed in libraries such as Pandas for data manipulation, Matplotlib for data visualization, NLTK for natural language processing, and Keras for deep learning. Additionally, I have a solid grasp of programming languages such as Java and C, which have further broadened my ability to approach problems from multiple angles.

As a fresher, I am open to new experiences and am particularly enthusiastic about collaborating with industry professionals to turn theoretical concepts into tangible outcomes. I am driven by a desire to bridge the gap between academic learning and practical application, and I look forward to contributing to projects that push the boundaries of what is possible in the fields of artificial intelligence and data science.



TABLE OF CONTENTS

TITLE	PG.NO
Data Analytics Process	1
Instagram User Analytics	3
Operation Analytics and Investigating Metric Spike	10
Hiring Process Analytics	18
IMDB Movie Analysis	25
Bank Loan Case Study	33
Analyzing the Impact of Car Features on Price and Profitability	39
ABC Call Volume Trend Analysis	47
What Did I Learn?	54
Appendix	55



DATA ANALYTICS PROCESS

Application in Real Life Scenario Case Study





DESCRIPTION



DESIGN

CONCLUSION



The project outlines the Data Analytics Process, using three practical examples: buying a new car, grocery shopping, and adopting a pet. Each example is used to demonstrate how the steps of the Data Analytics Process—planning, preparing, processing, analyzing, sharing, and acting—can be applied to make informed decisions in everyday situations.



The project is designed to simplify the understanding of the Data Analytics Process by applying it to relatable, real-world activities. Each example is broken down into six steps: Plan, Prepare, Process, Analyze, Share, and Act. This structured approach helps to make complex decision-making processes more accessible and applicable to daily life.



The project concludes by illustrating that the Data Analytics Process is versatile and can be applied beyond technical or professional domains. By using this process in everyday activities, such as buying a car, managing grocery expenses, or caring for a pet, individuals can make more informed and effective decisions, leading to better outcomes.





INSTAGRAM USER ANALYTICS

SQL Fundamentals





DESCRIPTION

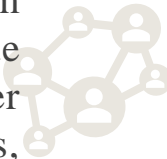


The project, titled "Instagram User Analytics," aims to provide in-depth insights into user behavior, engagement patterns, and platform performance on Instagram. The analysis focuses on various aspects such as user demographics, activity levels, content preferences, and the identification of potential issues like fake accounts or bot activity. The project is designed to support marketing teams, investors, and partner brands in making data-driven decisions.



THE PROBLEM

The key challenges addressed in this project include identifying the most loyal users for rewards, engaging inactive users, determining contest winners, analyzing popular hashtags for marketing, optimizing ad campaign schedules, assessing user engagement levels, and detecting fake accounts or bots that could affect the platform's integrity.



DESIGN



The project utilizes SQL queries to extract and analyze data from Instagram's database. The design involves structuring the tasks to address specific inquiries from stakeholders, using SQL for data extraction and analysis, and interpreting the results to provide actionable insights. The project leverages MySQL Workbench as the primary tool for database management and query execution.







FINDINGS

1. FIVE OLDEST USERS OF INSTAGRAM

The output provides information about the five oldest users on the platform based on their registration dates. It allows you to identify the users who joined the platform earliest. Hence, we have identified 5 oldest users to reward as the most loyal users.

OUTPUT:

Result Grid			 Filter Rows: <input type="text"/>	E
	id	username	created_at	
▶	80	Darby_Herzog	2016-05-06 00:14:21	
	67	Emilio_Bernier52	2016-05-06 13:04:30	
	63	Elenor88	2016-05-08 01:30:41	
	95	Nicole71	2016-05-09 17:30:22	
	38	Jordyn.Jacobson2	2016-05-14 07:56:26	
*	NULL	NULL	NULL	

2. USERS WITHOUT A SINGLE POST

The output provides a list of users who have not posted any photos on the platform. These users may be inactive or may prefer to engage with the platform in other ways, such as liking or commenting on posts. This data can be used to encourage the inactive users to start posting by sending them promotional emails.

OUTPUT:

Result Grid	Filter Rows:	Exports
id	username	
5	Aniya_Hackett	
7	Kassandra_Homenick	
14	Jadyn81	
21	Rodo33	
24	Maxwell.Halvorson	
25	Tierra.Trantow	
34	Pearl7	
36	Ollie_Ledner37	
41	Mckenna17	
45	David.Osinski47	
49	Morgan.Kassulke	
53	Linnea59	
54	Duane60	
57	Julien_Schmidt	
66	Mike.Auer39	
68	Franco_Keebler64	
71	Nia_Haag	
74	Hulda.Macejkovic	
75	Leslie67	
76	Janelle.Nikolaus81	
80	Darby_Herzog	
81	Esther.Zulauf61	
83	Bartholome.Bernhard	
89	Jessyca_West	
90	Esmeralda.Mraz57	
91	Bethany20	





FINDINGS


3. WINNER OF CONTEST

The user with the most likes on a single photo is considered the winner of the contest. This information is crucial for determining which user's content has resonated the most with the audience and has garnered the highest level of engagement. By identifying the user with the highest number of likes on a single photo, the team can declare them as the contest winner.

OUTPUT:

Result Grid



Filter Rows:

Export:


	username	photo_id	image_url	total_likes
▶	Zack_Kemmer93	145	https://jarret.name	48

4. TOP 5 MOST COMMONLY USED HASTAGS

The output reveals the top five hashtags most frequently used on Instagram, offering insights into prevalent themes and user engagement patterns. These hashtags signify popular topics, enabling marketers to align content strategies accordingly for increased visibility and audience engagement. Additionally, they provide opportunities for timely and relevant content creation, allowing brands to capitalize on trending discussions and enhance their competitive edge.

OUTPUT:

Result Grid	Filter Rows:
tag_name	tag_count
▶ smile	59
beach	42
party	39
fun	38
concert	24

FINDINGS

5. DAYS WITH HIGHEST REGISTRATIONS

Based on the analysis of user registration data, it's evident that the distribution of registrations is maximum on Thursdays and Sundays. By using this identification, the team can gain insights into user behavior and preferences, which can inform the scheduling of ad campaigns on Instagram.

OUTPUT:

	registration_day	registration_count
▶	Thursday	16
	Sunday	16
	Friday	15
	Tuesday	14
	Monday	14
	Wednesday	13
	Saturday	12

6. AVERAGE NUMBER OF POSTS PER USER

This query assesses user engagement on Instagram by calculating the average posts per user and the ratio of total photos to total users. A high average suggests active user participation, while discrepancies between averages may indicate varied user activity levels.

OUTPUT:

	avg_posts_per_user
▶	3.4730

	total_photos	total_users	photos_per_user
▶	257	74	3.4730



FINDINGS

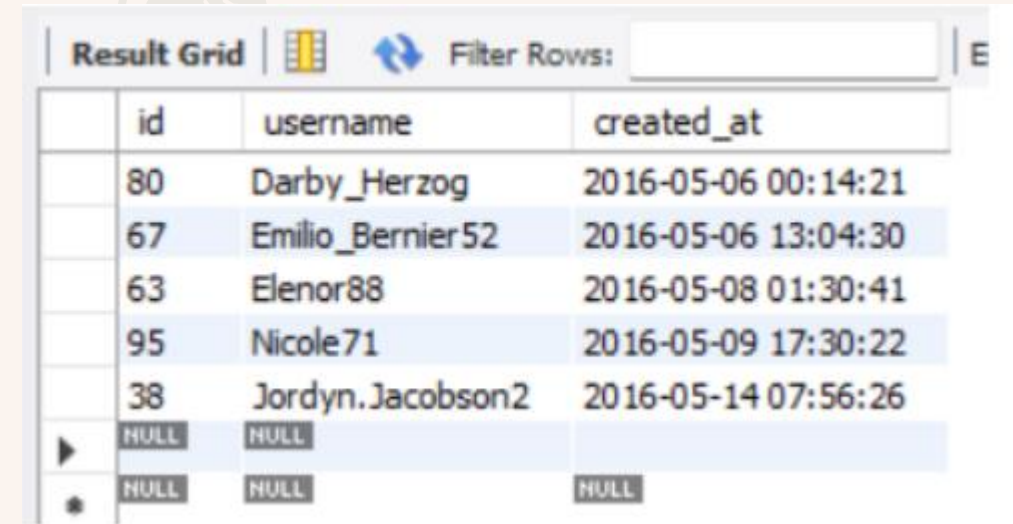


7. POTENTIAL BOTS & FAKE ACCOUNTS

This query identifies potential bots or fake accounts on the platform by finding users who have liked every photo. The presence of such users may indicate automated or fraudulent activity, highlighting the need for further investigation and mitigation strategies to maintain platform integrity and user trust.



OUTPUT:



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





ANALYSIS

The analysis provides insights into user behavior and platform trends. It highlights active user participation and engagement through the average posts per user and the popularity of certain hashtags. The detection of potential bots or fake accounts is critical for maintaining the platform's credibility and user trust. The findings also help in strategic planning for marketing efforts and improving user engagement.

CONCLUSION

The project successfully implements the Instagram User Analytics tasks, providing detailed queries, outputs, and analyses. The insights derived from this project are valuable for stakeholders, enabling them to make informed decisions regarding user engagement, marketing strategies, and platform management. The project underscores the importance of data analytics in understanding user behavior and optimizing platform performance.



OPERATION ANALYTICS AND INVESTIGATING METRIC SPIKE

Advanced SQL





DESCRIPTION



The project "Operation Analytics and Investigating Metric Spike" focuses on analyzing user engagement and growth metrics for a digital product. The goal is to understand user behavior, track growth, evaluate retention, and analyze email engagement by using SQL queries on data from user databases, event logs, and email service logs.



THE PROBLEM



The project addresses key challenges in understanding and optimizing user engagement, growth, and retention for a digital product. It seeks to identify trends, detect anomalies such as metric spikes, and provide actionable insights to improve user satisfaction and platform performance.



DESIGN

The project design involves collecting data from various sources, preprocessing it, and then using SQL to query and analyze key metrics. These metrics include weekly user engagement, user growth, retention rates, and email engagement. The insights are visualized through charts, graphs, and dashboards to support decision-making.



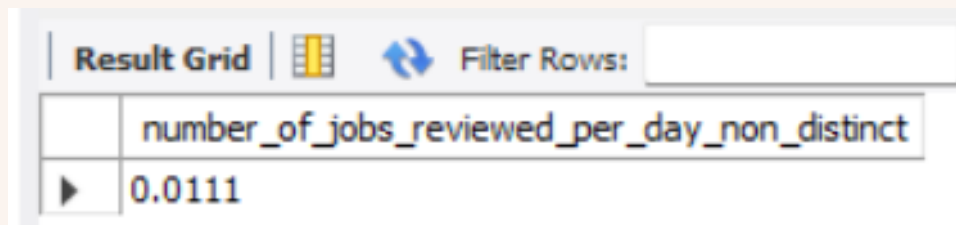


FINDINGS

1. NO. OF JOBS REVIEWED PER DAY

The output of the SQL query provides the average number of jobs reviewed per day over a 30-day period. It offers insights into the daily workload of reviewers, aiding in resource planning and performance evaluation. Comparing this metric over time or across teams can highlight trends and areas for improvement.

OUTPUT:

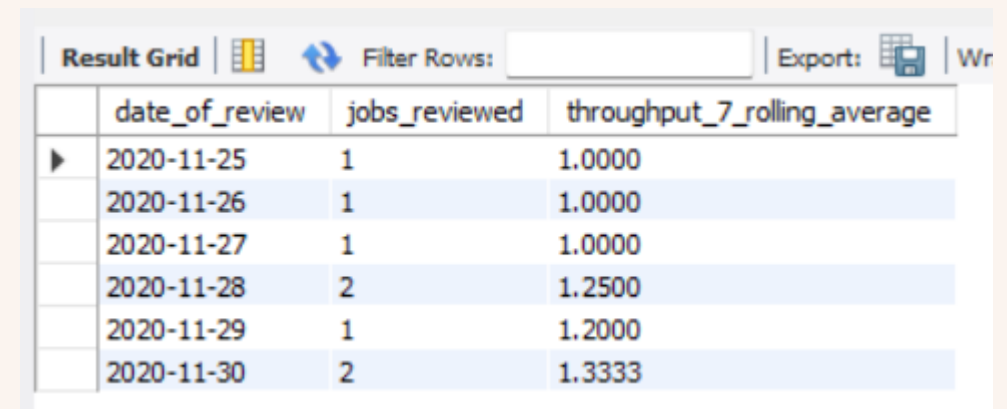


Result Grid	Filter Rows:
number_of_jobs_reviewed_per_day_non_distinct	
0.0111	

2. 7-DAY ROLLING AVG. OF THROUGHPUT

The output shows daily job reviews, their counts, and a 7-day rolling average. Comparing daily counts with the rolling average helps identify trends and deviations in reviewer activity. For immediate operational decisions, daily metrics are useful. For longer-term trends, the 7-day rolling average is preferred.

OUTPUT:



Result Grid	Filter Rows:	Export:	Wr
date_of_review	jobs_reviewed	throughput_7_rolling_average	
2020-11-25	1	1.0000	
2020-11-26	1	1.0000	
2020-11-27	1	1.0000	
2020-11-28	2	1.2500	
2020-11-29	1	1.2000	
2020-11-30	2	1.3333	

FINDINGS

3. PERCENTAGE SHARE OF LANGUAGES

The output provides insights into the distribution of job reviews across languages, presenting the total count and percentage share of each language in the dataset. It helps in understanding the prevalence of different languages in the review process, guiding resource allocation and content management strategies to align with user needs and preferences.

OUTPUT:

	job_id	language	total_of_each_language	percentage_share_of_each_language
▶	21	English	1	12.5000
	22	Arabic	1	12.5000
	23	Persian	3	37.5000
	25	Hindi	1	12.5000
	11	French	1	12.5000
	20	Italian	1	12.5000

4. DUPLICATE ROWS FROM THE DATA

The output of the SQL query highlights duplicate rows in the `job_data` table based on the `job_id` column. It achieves this by assigning row numbers to each row within partitions defined by `job_id`, then filtering out rows with row numbers greater than 1. This concise analysis provides valuable insights for data quality assessment and corrective actions to ensure dataset accuracy.

OUTPUT:

	ds	job_id	actor_id	event	language	time_spent	org	row_num
▶	2020-11-28	23	1005	transfer	Persian	22	D	2
	2020-11-26	23	1004	skip	Persian	56	A	3



FINDINGS

5. WEEKLY USER ENGAGEMENT

The output of the query provides a weekly count of distinct active users, revealing trends in user engagement over time. By examining the number of active users each week, businesses can identify patterns such as increases due to successful marketing campaigns or decreases that may indicate retention issues. This data is crucial for understanding user behavior, planning resources, and making informed decisions to enhance user satisfaction and engagement.

OUTPUT:

	week_number	number_of_users
▶	20	2346
	21	259
	24	2116
	25	1970
	28	1953
	29	2712
	33	2674
	34	1056



6. USER GROWTH FOR PRODUCT



The query output provides weekly and cumulative counts of active users by year and week. The `num_active_users` column shows the weekly activations, while `cum_active_users` accumulates these totals. This data is crucial for tracking user engagement, identifying growth patterns, and spotting trends in user activation. It also helps highlight successful onboarding or marketing efforts, aiding in strategic planning and improving user retention.

OUTPUT:

	year_num	week_num	num_active_users	cum_active_users
▶	2001	2	23	23
	2001	7	13	36
	2001	11	14	50
	2001	15	29	79
	2001	20	44	123
	2001	24	15	138
	2001	28	46	184
	2001	33	54	238
	2001	37	4	242
	2001	41	16	258
	2001	46	17	275
	2001	50	5	280
	2002	3	42	322

	year_num	week_num	num_active_users	cum_active_users
	2030	20	50	9009
	2030	24	38	9047
	2030	28	56	9103
	2030	33	29	9132
	2030	37	14	9146
	2030	42	15	9161
	2030	46	6	9167
	2030	50	19	9186
	2031	3	29	9215
	2031	11	35	9250
	2031	20	18	9268
	2031	29	55	9323
	2031	33	20	9343
	2031	42	16	9359
	2031	50	22	9381



FINDINGS

7. WEEKLY RETENTION OF USERS

The query output provides insights into user retention by tracking how many users engage within their first week after signing up. The `COUNT(user_id)` column shows total engagement events per user, while `per_week_retention` tracks users who engage within their first week. This data is key for assessing early user engagement and the effectiveness of onboarding processes. High first-week retention suggests a strong onboarding experience, while low retention may indicate areas needing improvement.

OUTPUT:

	user_id	COUNT(user_id)	per_week_retention
▶	11768	1	0
	11770	1	0
	11775	1	0
	11778	2	0
	11779	1	0
	11780	1	0
	11785	1	0
	11787	1	0
	11791	1	0
	11793	2	0
	11795	1	0
	11798	2	0
	11799	6	0
	11801	1	0

	user_id	COUNT(user_id)	per_week_retention
	11901	2	0
	11906	6	1
	11908	2	1
	11909	4	1
	11914	4	1
	11919	2	0
	11920	1	0
	11924	1	0
	11926	3	0
	11928	4	0
	11929	1	0
	11931	4	0
	11933	3	0
	11936	2	0

8. WEEKLY ENGAGEMENT PER DEVICE

The query output provides a weekly breakdown of user engagement by device type, showing distinct users per device each week. This reveals trends in device preferences and shifts in user behavior, like a move towards mobile usage. These insights help guide development and marketing strategies, ensuring the platform aligns with user preferences. Notable changes in engagement can also highlight the impact of specific events or updates.

OUTPUT:

	year_num	week_num	device	no_of_users
▶	2001	20	acer aspire desktop	5
	2001	20	acer aspire notebook	10
	2001	20	amazon fire phone	2
	2001	20	asus chromebook	9
	2001	20	dell inspiron desktop	8
	2001	20	dell inspiron notebook	22
	2001	20	hp pavilion desktop	4
	2001	20	htc one	5
	2001	20	ipad air	8
	2001	20	ipad mini	6
	2001	20	iphone 4s	10
	2001	20	iphone 5	22
	2001	20	iphone 5s	18
	2001	20	kindle fire	2
	2001	20	lenovo thinkpad	40

	year_num	week_num	device	no_of_users
	2031	33	iphone 4s	6
	2031	33	iphone 5	2
	2031	33	iphone 5s	3
	2031	33	kindle fire	3
	2031	33	lenovo thinkpad	16
	2031	33	mac mini	2
	2031	33	macbook air	10
	2031	33	macbook pro	17
	2031	33	nexus 10	2
	2031	33	nexus 5	4
	2031	33	nexus 7	2
	2031	33	nokia lumia 635	2
	2031	33	samsung galaxy note	1
	2031	33	samsung galaxy s4	6
	2031	33	windows surface	3



FINDINGS



9. EMAIL ENGAGEMENT METRICS

The query output calculates key email campaign metrics: the `email_opening_rate`, showing the percentage of sent emails that were opened, and the `email_clicking_rate`, indicating the percentage that led to a click-through. These rates are derived by dividing the number of `email_opened` and `email_clicked` actions by `email_sent` actions. High rates suggest effective subject lines and content, while low rates may highlight the need for improved strategies, such as better targeting or more compelling calls to action.



OUTPUT:

Result Grid			
Filter Rows: <input type="text"/>			
	id	username	created_at
	80	Darby_Herzog	2016-05-06 00:14:21
	67	Emilio_Bernier52	2016-05-06 13:04:30
	63	Elenor88	2016-05-08 01:30:41
	95	Nicole71	2016-05-09 17:30:22
	38	Jordyn.Jacobson2	2016-05-14 07:56:26
	NULL	NULL	
	NULL	NULL	NULL





ANALYSIS

The analysis shows trends in user activity, engagement, and growth, helping to understand how users interact with the product over time. For instance, rolling averages are used to smooth out data fluctuations and provide clearer insights into user behavior. The analysis also helps in identifying successful engagement strategies and areas needing improvement.

CONCLUSION

The project concludes by effectively implementing SQL-based analyses to derive actionable insights. These insights are critical for optimizing user engagement strategies, enhancing retention, and improving overall product performance. The project demonstrates the importance of data-driven decision-making in managing and improving digital products.



HIRING PROCESS ANALYTICS

Statistics





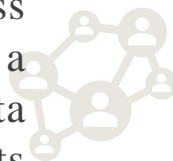
DESCRIPTION



The project "Hiring Process Analytics" aims to analyze a company's workforce data using Excel to gain insights into hiring patterns, salary distribution, departmental composition, and position tiers. The analysis provides a comprehensive view of the workforce structure and helps identify trends in hiring and compensation.



THE PROBLEM



The project addresses the need for a detailed understanding of workforce composition, including gender distribution in hiring, average salary trends, departmental staffing levels, and the hierarchical distribution of positions within the company. This analysis is crucial for making informed HR and management decisions.



DESIGN

The project utilizes Microsoft Excel for data analysis, employing pivot tables, charts, and functions to extract and visualize insights. Key analyses include gender distribution of hires, average salary calculation, salary distribution through class intervals, and visual representations of departmental and positional data.





FINDINGS

1. GENDER DISTRIBUTION OF HIRES

The analysis of gender distribution in hiring revealed that the company has hired 1,856 female employees and 2,563 male employees. This indicates a higher number of male hires compared to female hires, suggesting a potential gender imbalance in the hiring process. The data was summarized using a pivot table, which provided a clear count of hires by gender.

OUTPUT:

Gender	Count of hired
Female	1856
Male	2563

2. AVERAGE SALARY



The project calculated the average salary offered by the company to be approximately \$49,983. This was done after removing outliers using the Interquartile Range (IQR) method to ensure that extreme values did not skew the results. Salaries outside these bounds were excluded from the average calculation, resulting in a more accurate representation of typical employee compensation.

OUTPUT:

q1	25460.5
q3	74438
q3-q1	48977.5
lower	-48005.8
upper	147904.3

average salary	49983.02902
----------------	-------------



FINDINGS

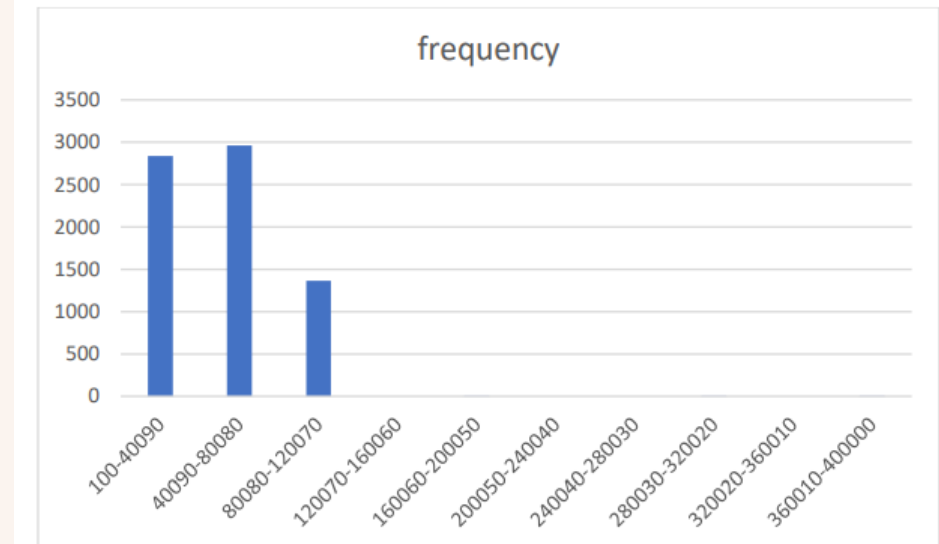


3. SALARY DISTRIBUTION

The salary distribution was broken down into 10 class intervals, ranging from \$100 to \$400,000. The majority of employees fall within the lower salary ranges, with the first interval (\$100 - \$40,090) containing 2,837 employees and the second interval (\$40,090 - \$80,080) containing 2,961 employees. This indicates that a significant portion of the workforce earns between \$100 and \$80,080. Only a few employees fall into higher salary brackets, suggesting that higher salaries are less common within the company. The class intervals were visualized using a bar graph to help in understanding the distribution of salaries across different ranges.

OUTPUT:

		lower limit	upper limit	class interval	frequency
		100	40090	100-40090	2837
		40090	80080	40090-80080	2961
		80080	120070	80080-120070	1365
		120070	160060	120070-160060	0
		160060	200050	160060-200050	1
min	100	200050	240040	200050-240040	0
max	400000	240040	280030	240040-280030	0
no. of intervals	10	280030	320020	280030-320020	1
class width	39990	320020	360010	320020-360010	0
		360010	400000	360010-400000	1





FINDINGS



4. DEPARTMENTAL COMPOSITION

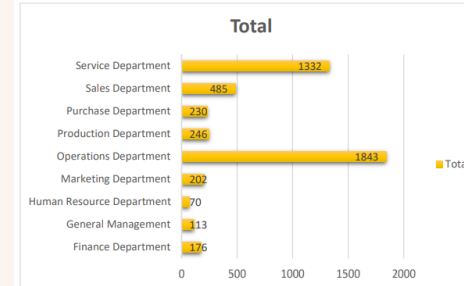
The departmental analysis revealed that the Operations Department has the largest number of employees, with 1,843 staff members, accounting for approximately 39% of the workforce. This is followed by the Service Department, with 1,332 employees (28%), and the Sales Department, with 485 employees (10%). Other departments like Marketing, Finance, and General Management have significantly fewer employees. The distribution of employees across departments was visualized using pie charts and bar graphs, providing a clear picture of how the workforce is spread across different functional areas.

OUTPUT:

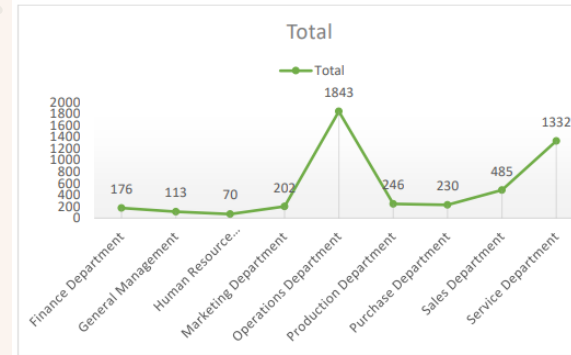
PIVOT TABLE

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

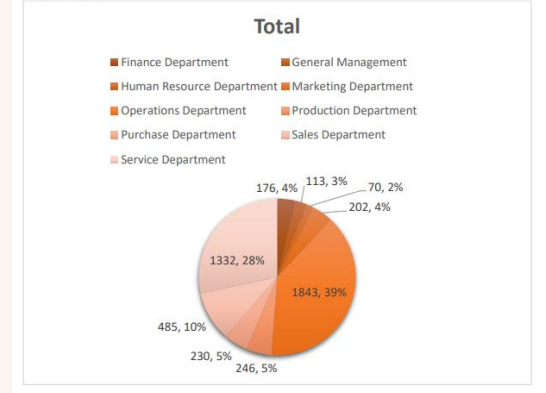
BAR CHART



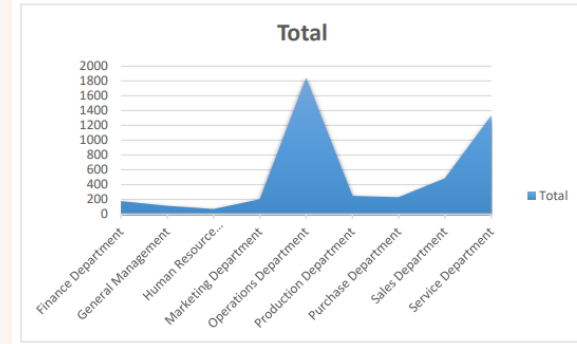
LINE CHART



PIE CHART



AREA CHART





FINDINGS



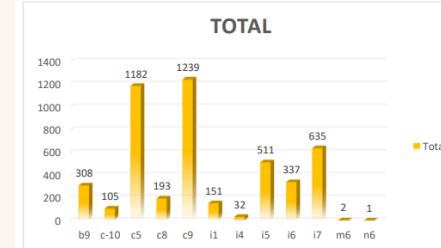
5. POSITION TIER DISTRIBUTION

The analysis of position tiers within the company showed that the most common tiers are 'c5' with 1,182 employees (25%) and 'c9' with 1,239 employees (26%). These tiers represent a significant portion of the workforce, indicating that many employees are concentrated within these specific hierarchical levels. Other tiers, such as 'b9' and 'i7', also have a notable number of employees, with 308 and 635 employees respectively. However, some tiers like 'm6' and 'n6' have very few employees, suggesting a more limited presence at those levels. The distribution of positions was visualized through various charts, helping to understand the hierarchical structure of the company.

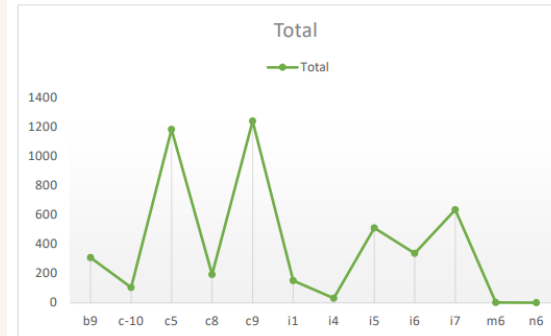
OUTPUT:

PIVOT TABLE	
Row Labels	Count of application_id
b9	308
c-10	105
c5	1182
c8	193
c9	1239
i1	151
i4	32
i5	511
i6	337
i7	635
m6	2
n6	1

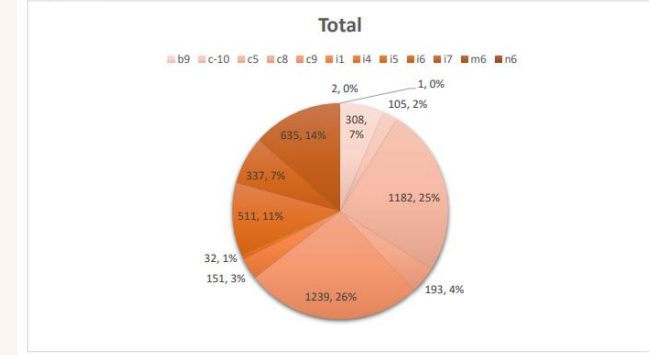
BAR CHART



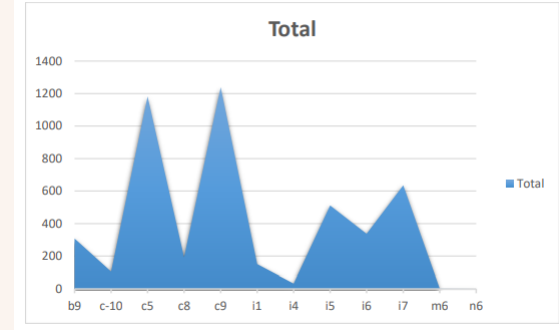
LINE CHART



PIE CHART



AREA CHART





ANALYSIS

The analysis shows patterns in hiring and compensation, identifying key areas where gender balance might be improved, and highlighting salary trends across different departments and positions. It provides a clear picture of how the workforce is distributed and compensated, which is valuable for HR planning and organizational development.

CONCLUSION

The project successfully implements a detailed analysis of the company's hiring process and workforce distribution using Excel. The insights generated can be used to optimize hiring strategies, ensure equitable salary distribution, and better understand the organizational structure, ultimately supporting more informed decision-making in HR and management.



IMDB MOVIE ANALYSIS

Final Project - 1





DESCRIPTION

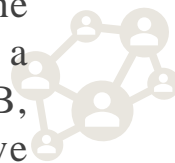


This project delves into the factors that contribute to a movie's success on IMDB, utilizing a comprehensive dataset that includes various attributes such as genre, duration, language, director, and budget. The analysis aims to identify which elements are most influential in determining a movie's success, as reflected in its IMDB score. Excel functions and tools are employed to clean, analyze, and visualize the data to extract meaningful insights.



THE PROBLEM

The goal of this project is to pinpoint the key determinants of a movie's success on IMDB. By examining different dimensions of the dataset, such as genre, duration, language, director, and budget, the project seeks to understand which factors have the greatest impact on a movie's rating. The underlying problem is to discern patterns and correlations that can explain why certain movies receive higher ratings than others.



DESIGN

The analytical process begins with cleaning the dataset to ensure its accuracy and reliability. This involves removing irrelevant columns, handling missing values, and eliminating duplicate records. Descriptive statistics are then applied to summarize the data and provide an overview of each factor. Visualization tools, including charts, graphs, and scatter plots, are utilized to illustrate trends and relationships between variables. These visual aids help in interpreting the data and supporting the findings with clear, graphical evidence.





FINDINGS

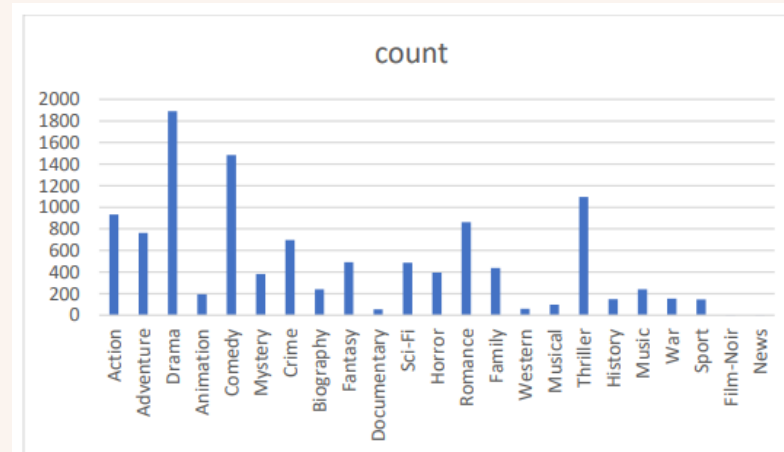


1. GENRE ANALYSIS

The genre distribution reveals that Drama is the most frequently occurring genre in the dataset, with Comedy, Thriller, and Romance following closely. Upon analyzing the average IMDB scores across genres, it is observed that Biographical films tend to receive the highest average ratings. This suggests that movies based on real-life stories might resonate more with audiences or critics. The analysis also highlights that while Drama dominates in number, its average rating does not necessarily exceed those of other genres consistently.

OUTPUT:

unique_genre	count	average_IMDB	median_IMDB	mode_IMDB	max_IMDB	min_IMDB	var_IMDB	stddev_IMDB
Action	933	6.27	6.3	6.6	9	2.1	2.1	1.05
Adventure	762	6.45	6.6	6.6	8.9	2.3	2.3	1.12
Drama	1890	6.79	6.9	6.7	9.3	2.1	2.1	0.9
Animation	195	6.7	6.8	7.3	8.6	2.8	2.8	0.99
Comedy	1484	6.17	6.3	6.7	8.8	1.9	1.9	1.05
Mystery	382	6.46	6.5	6.6	8.6	3.1	3.1	1.03
Crime	698	6.55	6.6	6.6	9.3	2.4	2.4	0.99
Biography	239	7.15	7.2	7	8.9	4.5	4.5	0.7
Fantasy	492	6.28	6.4	6.7	8.9	2.2	2.2	1.14
Documentary	54	7.04	7.4	6.6	8.5	1.6	1.6	1.28
Sci-Fi	487	6.31	6.4	7	8.8	1.9	1.9	1.18
Horror	395	5.84	5.9	6.2	8.6	2.2	2.2	1.03
Romance	862	6.43	6.5	6.5	8.6	2.1	2.1	0.97
Family	438	6.2	6.3	6.1	8.6	1.9	1.9	1.17
Western	60	6.7	6.75	6.8	8.9	3.8	3.8	1.04
Musical	98	6.57	6.7	7.1	8.5	2.1	2.1	1.09
Thriller	1095	6.35	6.4	6.5	9	2.7	2.7	0.99
History	149	7.13	7.2	7.7	8.9	4.4	4.4	0.7
Music	241	6.46	6.6	6.2	8.5	1.6	1.6	1.17
War	154	7.05	7.1	7.1	8.6	4.3	4.3	0.81
Sport	147	6.58	6.8	7.2	8.4	2	2	1.07
Film-Noir	2	7.95	7.95	7.7	8.2	7.7	7.7	0.25
News	1	7.1	7.1	7.1	7.1	7.1	7.1	0





FINDINGS



2. DURATION ANALYSIS

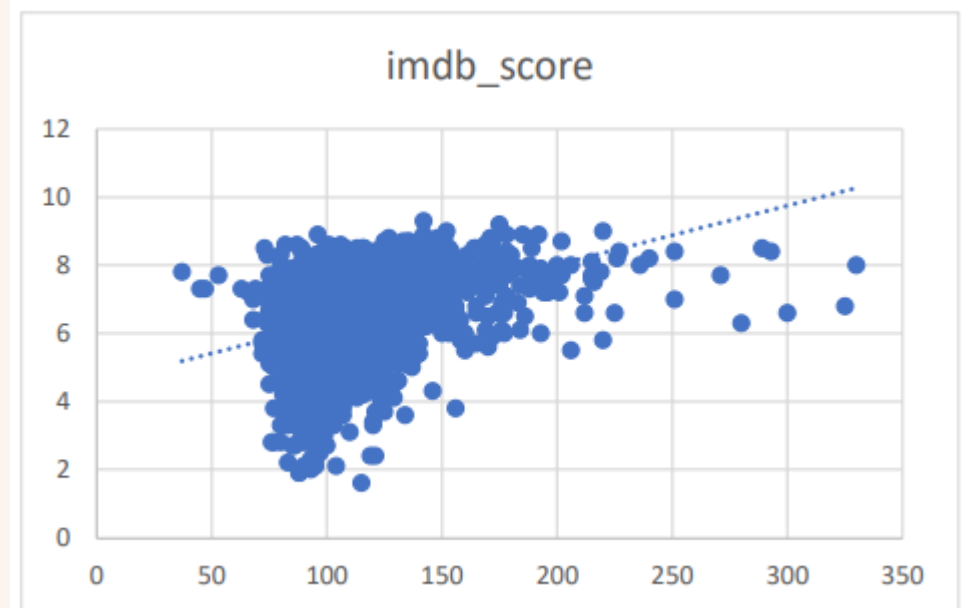


The duration analysis shows that most movies fall within the 50 to 200-minute range. There is a slight tendency for longer movies to achieve higher IMDB scores, although the correlation between duration and rating is not particularly strong. This implies that while longer movies may have a marginally better chance of higher ratings, other factors likely play a more significant role in determining a movie's success.



OUTPUT:

average_duration	109.7407014
mode_duration	101
median_duration	105
stddev_duration	22.64341264
var_duration	512.7241362





FINDINGS

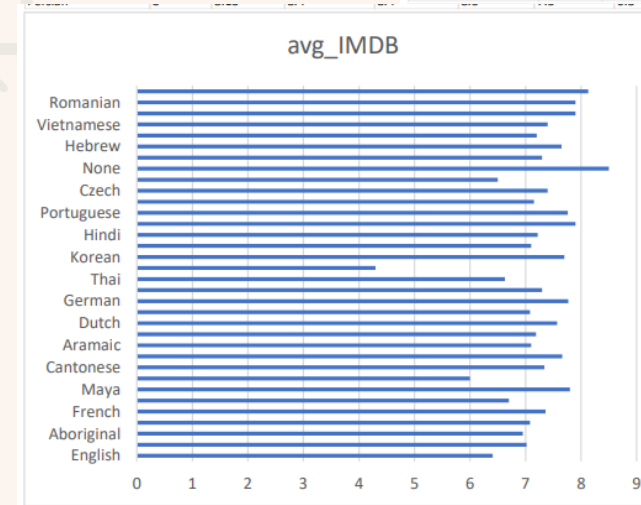


OUTPUT:

3. LANGUAGE ANALYSIS

English emerges as the predominant language for movies in the dataset, with a moderate average IMDB score and a broad variability in ratings. In contrast, languages with fewer movie entries, such as Bosnian and Thai, exhibit extreme ratings, both high and low. These findings suggest that while English-language films are more common and have a relatively consistent rating, films in less common languages might receive more polarized reviews, possibly due to limited audience reach or niche appeal.

unique_languages	count	avg_IMDB	median_IMDB	mode_IMDB	max_IMDB	min_IMDB	var_IMDB	stddev_IMDB
English	3606	6.41	6.5	6.7	9.3	1.6	1.14	1.07
Mandarin	14	7.02	7.25	7.6	7.9	5.6	0.59	0.74
Aboriginal	2	6.95	6.95	6.4	7.5	6.4	0.61	0.55
Spanish	23	7.08	7.2	5.9	8.2	5.2	0.74	0.84
French	34	7.36	7.3	7.2	8.4	5.8	0.27	0.51
Filipino	1	6.7	6.7	6.7	6.7	6.7	0.523	0
Maya	1	7.8	7.8	7.8	7.8	7.8	0.523	0
Kazakh	1	6	6	6	6	6	0.523	0
Cantonese	7	7.34	7.3	7.3	7.8	6.7	0.12	0.32
Japanese	10	7.66	8	7.7	8.7	6	0.98	0.94
Aramaic	1	7.1	7.1	7.1	7.1	7.1	0.523	0
Italian	7	7.19	7	5.3	8.9	5.3	1.33	1.07
Dutch	3	7.57	7.8	7.8	7.8	7.1	0.16	0.33
Dari	16	7.08	7.4	7.6	7.9	5.6	0.54	0.71
German	10	7.77	7.8	7.4	8.5	6.1	0.51	0.68
Mongolian	1	7.3	7.3	7.3	7.3	7.3	0.523	0
Thai	3	6.63	6.6	6.2	7.1	6.2	0.2	0.37
Bosnian	1	4.3	4.3	4.3	4.3	4.3	0.523	0
Korean	5	7.7	7.7	8.1	8.4	7	0.33	0.51
Hungarian	1	7.1	7.1	7.1	7.1	7.1	0.523	0
Hindi	5	7.22	7.4	6	8	6	0.64	0.72
Danish	3	7.9	8.1	7.3	8.3	7.3	0.28	0.43
Portuguese	5	7.76	8	8.1	8.7	6.1	0.96	0.88
Norwegian	4	7.15	7.3	7.6	7.6	6.4	0.33	0.5
Czech	1	7.4	7.4	7.5	7.4	7.4	0.523	0
Russian	1	6.5	6.5	6.5	6.5	6.5	0.523	0
None	1	8.5	8.5	8.5	8.5	8.5	0.523	0
Zulu	1	7.3	7.3	7.3	7.3	7.3	0.523	0
Hebrew	2	7.65	7.65	8	8	7.3	0.25	0.35
Arabic	1	7.2	7.2	7.2	7.2	7.2	0.523	0
Vietnamese	1	7.4	7.4	7.4	7.4	7.4	0.523	0
Indonesian	2	7.9	7.9	7.6	8.2	7.6	0.18	0.3
Romanian	1	7.9	7.9	7.9	7.9	7.9	0.523	0
Persian	3	8.13	8.4	8.4	8.5	7.5	0.3	0.45





FINDINGS



OUTPUT:

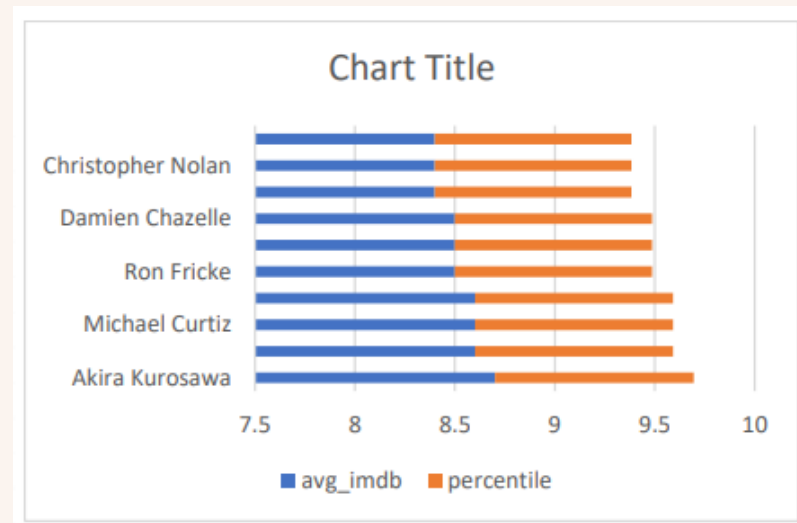
4. DIRECTOR ANALYSIS



The analysis of directors indicates that Akira Kurosawa stands out with the highest average IMDB score, achieving a ranking in the 99.4th percentile. The top 10 directors include a mix of classic icons and modern filmmakers, reflecting a diverse range of directorial styles and their impact on movie ratings. This highlights the significant role that directorial vision and experience play in the success of films, with certain directors consistently achieving high acclaim.



top 10 directors	avg_imdb	percentile
Akira Kurosawa	8.7	0.994
Tony Kaye	8.6	0.992
Michael Curtiz	8.6	0.992
Charles Chaplin	8.6	0.992
Ron Fricke	8.5	0.987
Majid Majidi	8.5	0.987
Damien Chazelle	8.5	0.987
Sergio Leone	8.4	0.983
Christopher Nolan	8.4	0.983
Richard Marquand	8.4	0.983





FINDINGS



5. BUDGET ANALYSIS



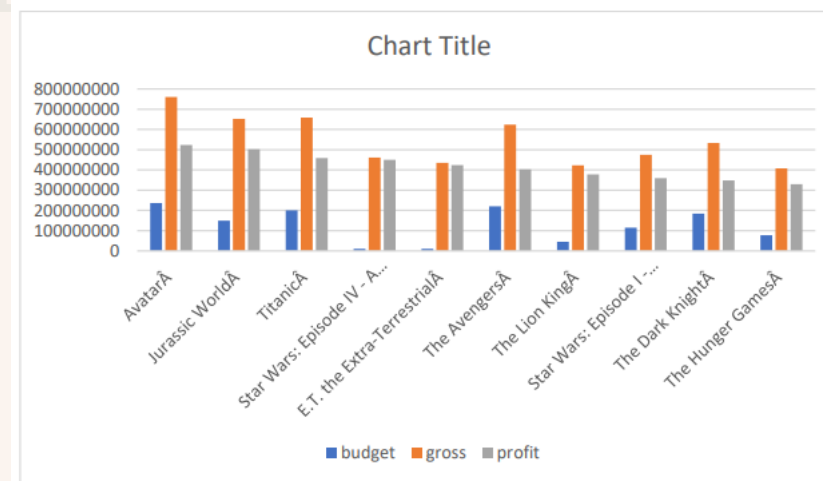
The budget analysis reveals a weak positive correlation between a movie's budget and its gross earnings. While higher budgets can contribute to greater earnings, this relationship is not strong enough to guarantee higher profits. Notably, movies like Avatar and Jurassic World demonstrate high profit margins, suggesting that successful movies can achieve significant returns regardless of their budget. This underscores the importance of factors beyond budget, such as creative content and market appeal, in determining a movie's financial success.



OUTPUT:

correlation_coeff 0.095584

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





ANALYSIS

The project provides a comprehensive look at how various factors contribute to a movie's IMDB rating. It is evident that genre and director have more consistent and significant influences on movie success. The duration of the movie has a moderate impact, while language and budget show varying effects depending on specific contexts. The overall analysis underscores the complexity of movie success, revealing that multiple elements interplay to affect a film's reception and rating.

CONCLUSION

The project concludes that movie success on IMDB is influenced by a range of factors, with some having more consistent impacts than others. Genre and directorial quality emerge as key determinants, while duration and language have more variable effects. Budget, although important, does not guarantee higher profits and highlights the importance of other creative and contextual elements. The findings suggest that a holistic approach, considering multiple factors, is essential for understanding and predicting a movie's success.



BANK LOAN CASE STUDY

Final Project - 2





DESCRIPTION

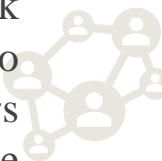


This project analyzes a bank loan application dataset to uncover the key factors influencing loan defaults. The analysis involves data cleaning, outlier detection, handling missing data, and performing correlation analysis. The primary goal is to provide insights into the demographics and financial characteristics that impact loan approval and default rates.



THE PROBLEM

The dataset presents several challenges, including significant missing data, outliers, and potential class imbalances. The core problem revolves around identifying the critical factors that contribute to loan defaults. This involves examining various demographic and financial variables, such as income levels, employment status, and loan types, to understand their impact on the likelihood of a loan default. The project seeks to discern patterns and relationships that can provide actionable insights for improving loan risk assessment and management.



DESIGN



The project started with extensive data cleaning, removing columns like `OWN_CAR_AGE` and `EXT_SOURCE_1` that had over 50% missing values, and imputing others, such as `OCCUPATION_TYPE`, with the most common value. Outliers, particularly in `AMT_ANNUITY`, were managed by capping values above 250,000 to the median. The significant data imbalance between defaulters and non-defaulters was carefully examined. Univariate, segmented univariate, and bivariate analyses were performed to explore variable distributions and relationships with loan default. The analysis concluded with a correlation study, identifying significant predictors like the `EXT_SOURCE_2` score, which showed a strong negative correlation with default rates.





FINDINGS



1. HANDLING MISSING DATA & OUTLIERS

The project identified and removed columns with high levels of missing data, ensuring that the remaining dataset was more manageable and accurate. For example, columns like OWN_CAR_AGE and EXT_SOURCE_1 were removed, and missing values in OCCUPATION_TYPE were imputed with 'Laborers'. Outliers in the AMT_ANNUITY column were capped at the median value to prevent skewed analysis.

2. HANDLING DATA IMBALANCE

The dataset showed a significant imbalance, with 91.95% of the records being non-defaulters and only 8.05% defaulters. Additionally, the analysis revealed that 65% of the applicants were women, and a majority were taking out cash loans rather than revolving loans. This imbalance highlights the importance of considering the distribution of the target variable in predictive modeling.





FINDINGS

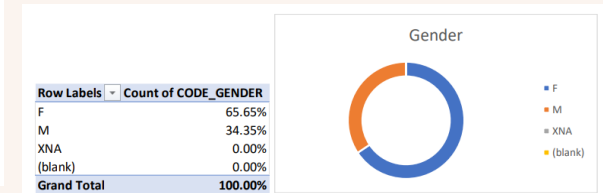
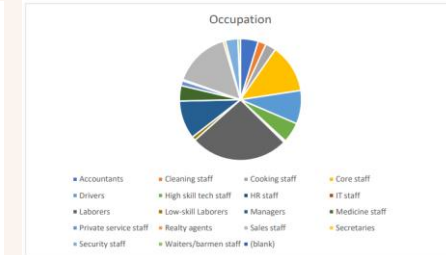
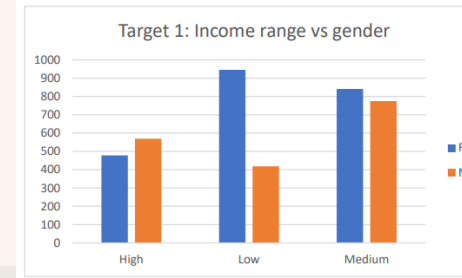


3. UNIVARIATE AND BIVARIATE ANALYSIS

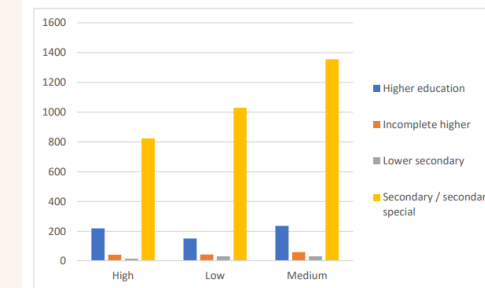
We can see that the percentage of loan defaulters is around 8% and non-defaulters is almost 92%. We can see that the most of the clients are laborers followed by sales staff and core staff. Majority of the clients are women constituting about 65% whereas the remaining 35% are men. We can see that most clients with medium to low income and have family status as married have payment issues. Male clients with low income and have family status as married have payment issues. Clients with medium credit range and an education type of Secondary / Secondary special are those with maximum payment issues.

OUTPUT:

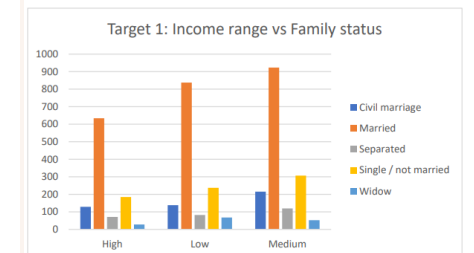
TARGET	1		
Count of CODE_GENDER	Column Labels		
Row Labels	F	M	Grand Total
High	478	569	1047
Low	945	418	1363
Medium	841	775	1616
Grand Total	2264	1762	4026



TARGET	1	3				
Count of NAME_EDUCATION_TYPE	Column Labels					
Row Labels	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special	Grand Total	
High	220	40	14	824	1098	
Low	149	41	30	1030	1250	
Medium	237	57	29	1355	1678	
Grand Total	606	138	73	3209	4026	



TARGET	1					
Count of NAME_FAMILY_STATUS	Column Labels					
Row Labels	Civil marriage	Married	Separated	Single / not married	Widow	Grand Total
High	129	634	71	185	28	1047
Low	138	838	82	237	68	1363
Medium	215	923	119	307	52	1616
Grand Total	482	2395	272	729	148	4026





FINDINGS



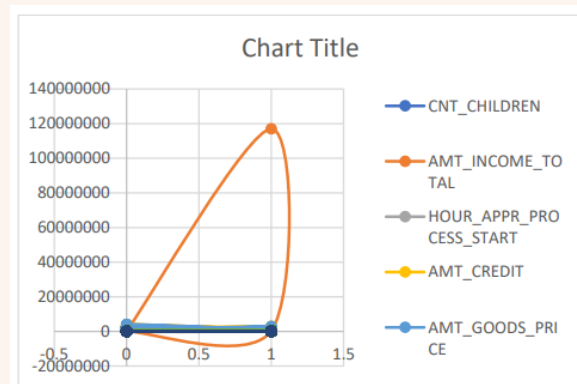
4. CORRELATION INSIGHTS

The correlation analysis identified several key predictors of loan default. For instance, the EXT_SOURCE_2 score had a strong negative correlation with default rates, suggesting that higher scores were associated with lower risk of default. Similarly, longer employment duration (DAYS_EMPLOYED) was linked to a lower likelihood of default, emphasizing the importance of job stability.

OUTPUT:

COLUMN AGAINST TARGET	CORRELATION COEFFICIENT
CNT_CHILDREN	0.026363931
AMT_INCOME_TOTAL	0.010893745
HOUR_APPR_PROCESS_START	-0.032036463
AMT_CREDIT	-0.032428347
DAYS_EMPLOYED	-0.040281269
AMT_GOODS_PRICE	-0.04127611
EXT_SOURCE_2	-0.158424274

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	HOUR_APPR_PROCESS_START	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_EMPLOYED	EXT_SOURCE_2
TARGET	1							
CNT_CHILDREN	0.026363931	1						
AMT_INCOME_TOTAL	0.010893745	0.009588558	1					
HOUR_APPR_PROCESS_START	-0.032036463	-0.006253862	0.01846417	1				
AMT_CREDIT	-0.032428347	0.00497156	0.069315897	0.056676981	1			
AMT_GOODS_PRICE	-0.04127611	0.000232954	0.069891714	0.065891636	0.986704386	1		
DAYS_EMPLOYED	-0.040281269	-0.239673381	-0.031603988	-0.08796449	-0.070416099	-0.067738221	1	
EXT_SOURCE_2	-0.158424274	-0.017641055	0.019517645	0.157147521	0.138125321	0.146862491	-0.026153993	1





ANALYSIS

The analysis highlighted the complex interplay of various factors affecting loan defaults. Missing data and outlier management were crucial in ensuring the accuracy of the results. The significant data imbalance necessitated careful interpretation of the findings, particularly concerning demographic insights. The demographic analysis revealed potential biases in the dataset, such as a higher representation of female applicants. The income and family status findings underscored the importance of socioeconomic factors in predicting loan defaults. The correlation insights provided valuable predictors, such as the EXT_SOURCE_2 score and employment duration, which could be leveraged for more effective risk assessment.

CONCLUSION

The project provided a comprehensive analysis of the factors affecting loan defaults. By addressing issues of missing data, outliers, and class imbalances, the study was able to identify key demographic and financial attributes that influence loan repayment behavior. These findings highlight the need for a nuanced approach in loan risk assessment, taking into account a wide range of factors to better predict and manage loan defaults. The insights gained from this analysis can inform strategies for improving credit risk models and making more informed lending decisions.



ANALYZING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

Final Project – 3

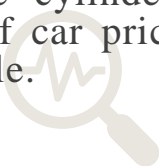




DESCRIPTION



The project aims to create an interactive Excel dashboard to analyze the automotive market. It uses a comprehensive dataset of car models to examine various factors influencing car prices and profitability. Key tasks include analyzing car model popularity across market categories, exploring the relationship between engine power and price, and identifying influential car features on pricing through regression analysis. The project also investigates average price variations across manufacturers, the relationship between fuel efficiency and engine cylinders, and the distribution of car prices by brand and body style.



THE PROBLEM

The primary problem this project addresses is the lack of comprehensive, data-driven insights into the factors that influence car prices and profitability. Automotive stakeholders often rely on fragmented information and intuition when making decisions about car pricing, marketing strategies, and product development. This can lead to suboptimal decisions that do not fully leverage available data. By systematically analyzing a wide range of car features and their impact on price, this project aims to provide a clearer, evidence-based understanding of the market. This understanding can help manufacturers, marketers, and analysts to optimize pricing strategies, identify key market trends, and improve overall profitability.



DESIGN

The project employs Microsoft Excel for creating interactive dashboards with features like filters, slicers, and various chart types. Python libraries (Pandas, StatsModels, and Matplotlib) are used for data processing, statistical analysis, and visualization. Key analyses include pivot tables, scatter plots, combo charts, and regression models to provide insights into the automotive market.





FINDINGS



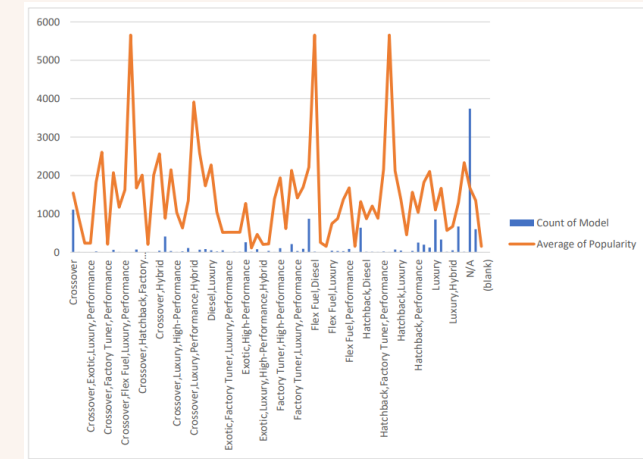
1. POPULARITY BASED ON MARKET CATEGORY

Market analysis reveals that certain categories exhibit exceptionally high popularity scores, indicating consumer preference despite having fewer models. The distribution of models across these categories is uneven, with some categories densely populated and others sparsely represented. This disparity suggests that market popularity is not merely a function of the number of available models but is influenced by additional factors such as brand reputation, features, and prevailing market trends.

OUTPUT:

Row Labels	Count of Model	Average of Popularity
Crossover	1110	1545.263063
Crossover,Diesel	7	873
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.461538
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Factory Tuner,Performance	4	210
Crossover,Flex Fuel	64	2073.75
Crossover,Flex Fuel,Luxury	10	1173.2
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover,Flex Fuel,Performance	6	5657
Crossover,Hatchback	72	1675.694444
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Luxury	7	204
Crossover,Hatchback,Performance	6	2009
Crossover,Hybrid	42	2563.380952
Crossover,Luxury	410	884.5487805
Crossover,Luxury,Diesel	34	2149.411765
Crossover,Luxury,High-Performance	9	1037.222222
Crossover,Luxury,Hybrid	24	630.9166667
Crossover,Luxury,Performance	113	1344.849558
Crossover,Luxury,Performance,Hybrid	2	3916
Crossover,Performance	69	2585.956522
Diesel	84	1730.904762
Diesel,Luxury	51	2275
Exotic,Factory Tuner,High-Performance	21	1046.380952
Exotic,Factory Tuner,Luxury,High-Performance	52	517.5384615
Exotic,Factory Tuner,Luxury,Performance	3	520
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	13	520
Exotic,Flex Fuel,Luxury,High-Performance	11	520
Exotic,High-Performance	261	1271.333333
Exotic,Luxury	12	112.6666667
Exotic,Luxury,High-Performance	79	467.0759494
Exotic,Luxury,High-Performance,Hybrid	1	204
Exotic,Luxury,Performance	36	217.0277778
Exotic,Performance	10	1391
Factory Tuner,High-Performance	106	1941.415094
Factory Tuner,Luxury	2	617
Factory Tuner,Luxury,High-Performance	215	2133.367442
Factory Tuner,Luxury,Performance	31	1413.419355
Factory Tuner,Performance	92	1695.695652
Flex Fuel	872	2217.302752
Flex Fuel,Diesel	16	5657
Flex Fuel,Factory Tuner,Luxury,High-Performance	1	258

Flex Fuel	872	2217.302752
Flex Fuel,Diesel	16	5657
Flex Fuel,Factory Tuner,Luxury,High-Performance	1	258
Flex Fuel,Hybrid	2	155
Flex Fuel,Luxury	39	746.5384615
Flex Fuel,Luxury,High-Performance	33	878.9090909
Flex Fuel,Luxury,Performance	28	1380.071429
Flex Fuel,Performance	87	1680.471264
Flex Fuel,Performance,Hybrid	2	155
Hatchback	641	1318.865835
Hatchback,Diesel	14	873
Hatchback,Factory Tuner,High-Performance	13	1205.153846
Hatchback,Factory Tuner,Luxury,Performance	9	886.8888889
Hatchback,Factory Tuner,Performance	22	2159.045455
Hatchback,Flex Fuel	7	5657
Hatchback,Hybrid	72	2121.25
Hatchback,Luxury	46	1379.5
Hatchback,Luxury,Hybrid	3	454
Hatchback,Luxury,Performance	38	1566.131579
Hatchback,Performance	252	1039.646825
High-Performance	199	1821.447236
Hybrid	123	2105.569106
Luxury	855	1102.65731
Luxury,High-Performance	334	1668.017964
Luxury,High-Performance,Hybrid	12	568.8333333
Luxury,Hybrid	52	673.6346154
Luxury,Performance	673	1292.615156
Luxury,Performance,Hybrid	11	2333.181818
N/A	3742	1676.889364
Performance	601	1348.873544
Performance,Hybrid	1	155
(blank)		
Grand Total	11914	1554.911197





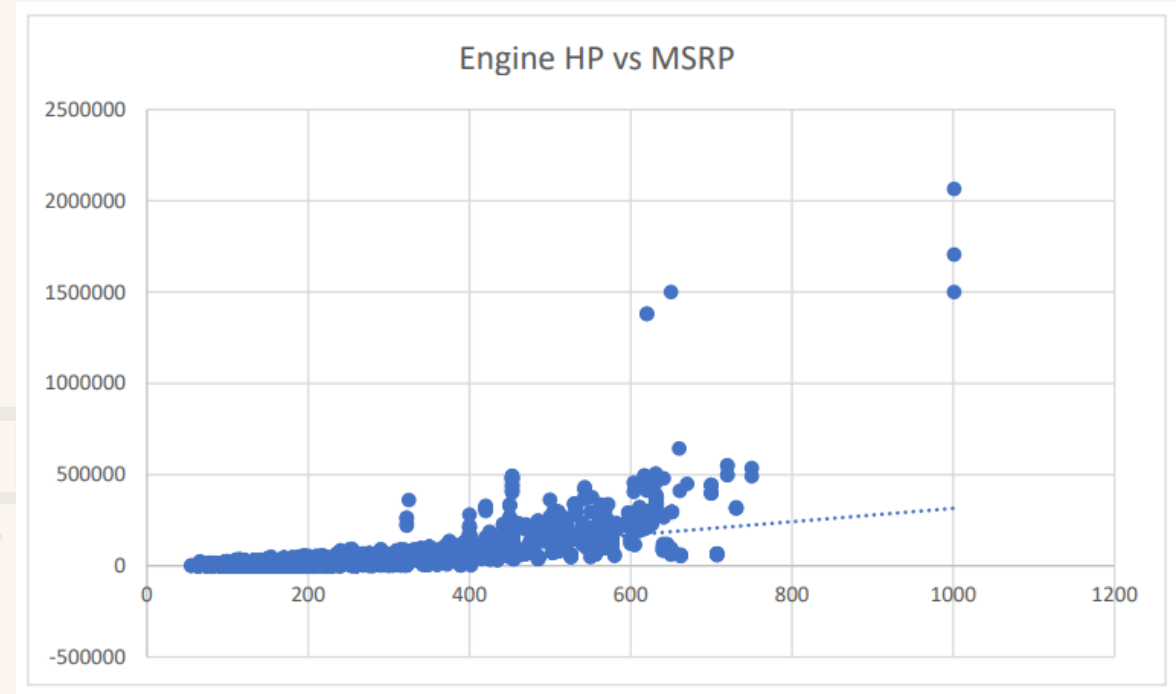
FINDINGS



2. ENGINE HORSEPOWER AND PRICE

OUTPUT:

The analysis shows a positive correlation between engine horsepower (HP) and Manufacturer's Suggested Retail Price (MSRP), indicating that as engine HP increases, the MSRP tends to rise. Most data points are clustered between 100 to 600 HP and \$0 to \$500,000 MSRP, suggesting that the majority of cars fall within this range. The dotted trend line, which has a positive slope, further supports this correlation. However, the relatively flat slope compared to the data spread indicates that while there is a relationship, other factors may also influence the MSRP.





FINDINGS



2. ENGINE HORSEPOWER AND PRICE

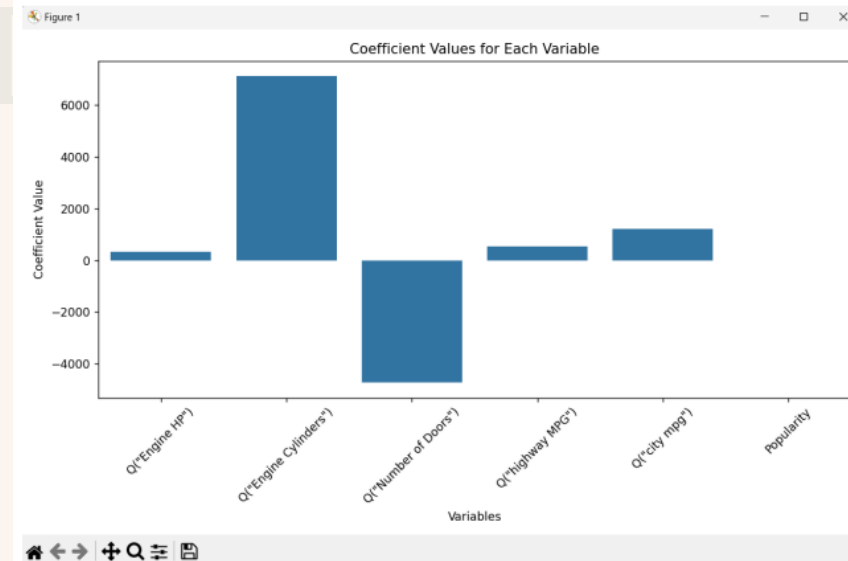
The regression analysis indicates that Engine HP, Engine Cylinders, Number of Doors, highway MPG, city mpg, and Popularity significantly affect MSRP, with Engine HP having the largest impact ($F=2918.36$, $p<0.001$). The model explains approximately 47% of the variability in MSRP ($R^2=0.470$), suggesting a moderate fit. The ANOVA table confirms the significance of all predictors, with p-values close to zero, indicating strong evidence against the null hypothesis for each variable. A bar chart of the regression coefficients reveals that Engine HP and Engine Cylinders are the most influential variables, followed by Popularity, city mpg, Number of Doors, and highway MPG, in descending order of importance.

OUTPUT:

```
PS C:\Courses\TRAINITY\task 7> & "C:/Program Files/Python311/python.exe" "c:/Courses/TRAINITY/task 7/task3.py"

Regression Statistics
Multiple R      0.6853063110272515
R Square       0.46964473993378
Adjusted R Square 0.46937525047236417
Standard Error  688.7236906928455
Observations   11815.0

ANOVA Table
              sum_sq    df      F      PR(>F)
Q("Engine HP")    5.626482e+12    1.0  2918.361271  0.000000e+00
Q("Engine Cylinders") 5.113705e+11    1.0   265.239249  5.428696e-59
Q("Number of Doors") 2.001357e+11    1.0   103.807010  2.810167e-24
Q("highway MPG")    4.967971e+10    1.0    25.768025  3.908767e-07
Q("city mpg")       1.947314e+11    1.0   101.003870  1.142813e-23
Popularity          2.692861e+11    1.0   139.674124  4.753362e-32
Residual           2.276535e+13  11808.0      NaN      NaN
```





FINDINGS



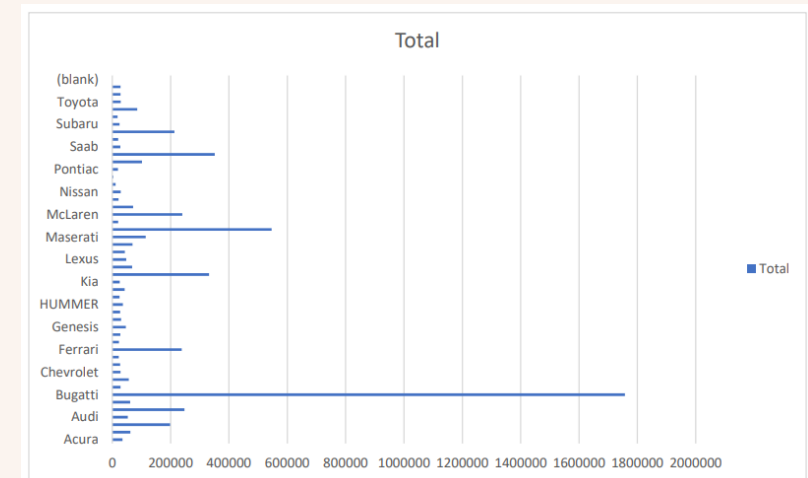
3. AVERAGE PRICE OF MANUFACTURERS

The average price of cars varies significantly across different manufacturers. High-end manufacturers like Alfa Romeo, Aston Martin, and Bentley have much higher average prices, often exceeding \$100,000, reflecting their luxury and performance-focused models. Mid-range brands like BMW, Audi, and Mercedes-Benz also show high average prices, typically in the \$50,000 to \$70,000 range, aligning with their premium market positioning. In contrast, mass-market manufacturers such as Chevrolet, Ford, and Toyota have lower average prices, generally under \$30,000, indicating their focus on affordability and broad market appeal. This variation highlights the diversity in market positioning and target consumer segments among different car manufacturers.

OUTPUT:

Manufacturer	Average Price
Acura	34887.5873
Alfa Romeo	61600
Aston Martin	197910.3763
Audi	53452.1128
Bentley	247169.3243
BMW	61546.76347
Bugatti	1757223.667
Buick	28206.61224
Cadillac	56231.31738
Chevrolet	28350.38557
Chrysler	26722.96257
Dodge	22390.05911
Ferrari	238218.8406
FIAT	22670.24194
Ford	27399.26674
Genesis	46616.66667
GMC	30493.29903
Honda	26674.34076
HUMMER	36464.41176
Hyundai	24597.0363
Infiniti	42394.21212
Kia	25310.17316
Lamborghini	331567.3077
Land Rover	67823.21678
Lexus	47549.06931
Lincoln	42839.82927
Lotus	69188.27586
Maserati	114207.7069
Maybach	546221.875
Mazda	20039.38298
McLaren	239805
Mercedes-Benz	71476.22946
Mitsubishi	21240.53521

Mitsubishi	21240.53521
Nissan	28583.4319
Oldsmobile	11542.54
Plymouth	3122.902439
Pontiac	19321.54839
Porsche	101622.3971
Rolls-Royce	351130.6452
Saab	27413.5045
Scion	19932.5
Spyker	213323.3333
Subaru	24827.50391
Suzuki	17907.20798
Tesla	85255.55556
Toyota	29030.01609
Volkswagen	28102.38072
Volvo	28541.16014
(blank)	
Grand Total	40594.73703





FINDINGS

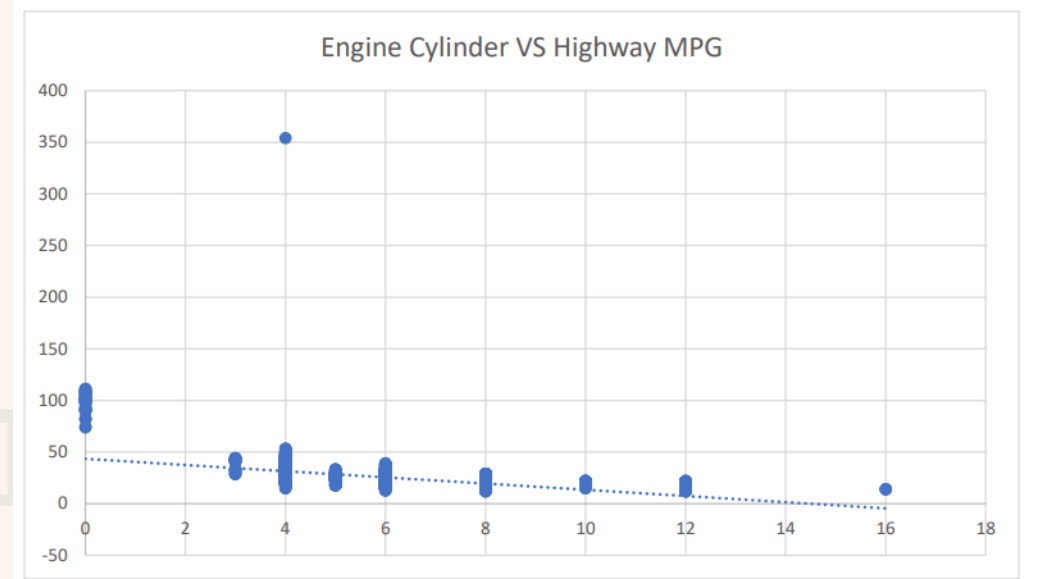


4. FUEL EFFICIENCY AND CYLINDERS



The scatter plot demonstrates a negative correlation between the number of cylinders in a car's engine and its highway miles per gallon (MPG). As the number of cylinders increases, the highway MPG generally decreases, indicating that cars with more cylinders tend to have lower fuel efficiency on the highway. This trend is consistent, although there is a notable outlier with 4 cylinders achieving an exceptionally high MPG.

OUTPUT:



CORRELATION COEFFICIENT	-0.62161
--------------------------------	-----------------



ANALYSIS

The analysis involved several key tasks to uncover insights into the automotive market. Popularity analysis using pivot tables revealed that luxury and high-performance vehicles, though fewer in number, tend to have higher popularity scores, suggesting consumer preference is influenced by factors beyond model availability. A scatter chart analysis showed a general positive correlation between engine horsepower and price, indicating higher engine power typically results in higher prices, though other variables also play a significant role. Multiple regression analysis identified Engine HP and Engine Cylinders as the most significant predictors of MSRP, explaining about 47% of the price variability. Manufacturer price analysis highlighted significant variations, with high-end brands like Alfa Romeo and Bentley having much higher average prices compared to mass-market brands like Ford and Toyota. The final dashboard incorporated various visualizations and interactive elements to effectively communicate these insights and allow dynamic data exploration.

CONCLUSION

The project successfully identifies key factors influencing car prices and provides a user-friendly, interactive dashboard for dynamic data exploration. The insights gained can support automotive stakeholders in making informed decisions regarding pricing strategies and market positioning. The combination of Excel and Python proves effective in handling data analysis and visualization tasks, ensuring a comprehensive understanding of automotive market dynamics. This project highlights the value of integrating data analytics into strategic decision-making processes in the automotive industry, paving the way for more data-driven approaches to understanding market trends and consumer preferences.



ABC CALL VOLUME TREND ANALYSIS

Final Project – 4





DESCRIPTION



This Customer Experience (CX) Analytics project focuses on analyzing inbound call data for ABC Insurance Company. The primary objectives are to enhance customer satisfaction and operational efficiency by addressing various aspects of call handling. Key tasks include calculating the average duration of incoming calls for each time bucket, visualizing call volume trends throughout the day, and proposing a strategic manpower plan to significantly reduce the current 30% call abandonment rate to 10%. Additionally, the project aims to devise a night shift manpower plan to ensure prompt call responses outside regular business hours.

THE PROBLEM

ABC Insurance Company faces a significant challenge in managing its inbound call volume effectively. The high call abandonment rate of approximately 30% indicates that a substantial number of customer calls are not being answered, leading to poor customer satisfaction and potential loss of business. This project aims to analyze call data, identify peak call times, and propose a strategic manpower plan to reduce the call abandonment rate to 10%, thereby improving customer service and operational efficiency.

DESIGN



The project utilizes Microsoft Excel due to its versatility, accessibility, and robust analytical capabilities. Excel's built-in functions and statistical tools enable precise calculations of call durations and abandonment rates. Its powerful visualization features facilitate the creation of detailed charts and graphs to illustrate call volume trends and manpower planning needs, supporting data-driven decision-making.





FINDINGS



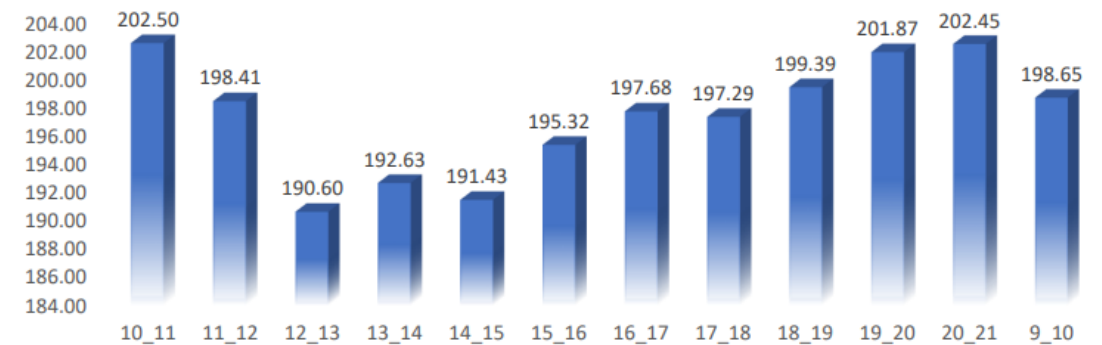
1. AVERAGE CALL DURATION

The analysis of call durations across different time slots revealed that the longest average call durations occur in the 10 AM – 11 AM and 8 PM – 9 PM time buckets, with each call averaging approximately 202.5 seconds. This suggests that customers might have more complex queries or issues requiring more time to resolve during these periods. Understanding these patterns helps in allocating more experienced agents or additional resources during these slots to handle calls more efficiently.

OUTPUT:

Time bucket	Average of Call_Seconds (s)
10_11	202.50
11_12	198.41
12_13	190.60
13_14	192.63
14_15	191.43
15_16	195.32
16_17	197.68
17_18	197.29
18_19	199.39
19_20	201.87
20_21	202.45
9_10	198.65
Grand Total	196.48

AVERAGE DURATION OF CALLS IN EACH TIME BUCKET





FINDINGS



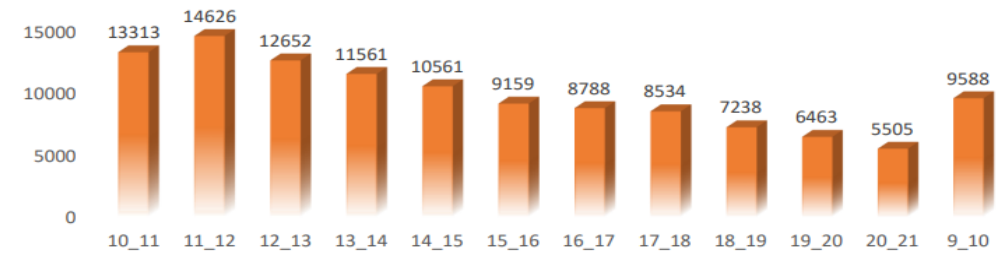
2. CALL VOLUME ANALYSIS

The call volume analysis showed distinct trends in customer call behavior throughout the day. The highest call volumes were observed between 10 AM and 12 PM, peaking in the 11 AM – 12 PM slot. This indicates that customers prefer to contact the company during mid-morning hours, possibly after handling other morning activities. The gradual decline in call volumes towards the evening suggests reduced customer activity as the day progresses. Visualizing these trends through bar charts and pivot tables enables the company to plan staffing needs more effectively, ensuring sufficient coverage during peak times to minimize wait times and abandonment rates.

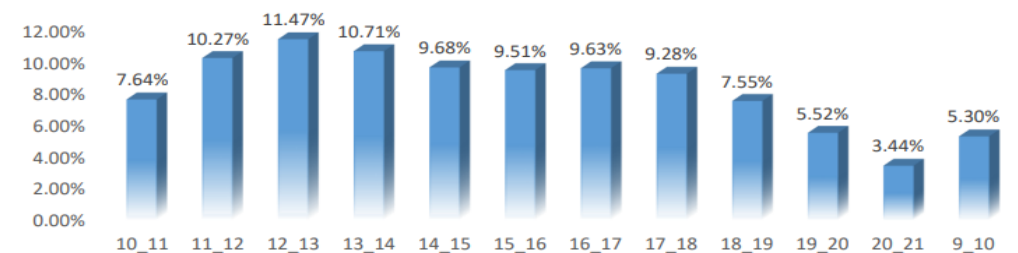
OUTPUT:

Time Bucket	Count of Customer_Phone_No	Time Bucket	Count of Time
10_11	13313	10_11	7.64%
11_12	14626	11_12	10.27%
12_13	12652	12_13	11.47%
13_14	11561	13_14	10.71%
14_15	10561	14_15	9.68%
15_16	9159	15_16	9.51%
16_17	8788	16_17	9.63%
17_18	8534	17_18	9.28%
18_19	7238	18_19	7.55%
19_20	6463	19_20	5.52%
20_21	5505	20_21	3.44%
9_10	9588	9_10	5.30%
Grand Total	117988	Grand Total	100.00%

NUMBER OF CALLS IN EACH TIME BUCKET



% OF CALLS IN EACH TIME BUCKET





FINDINGS

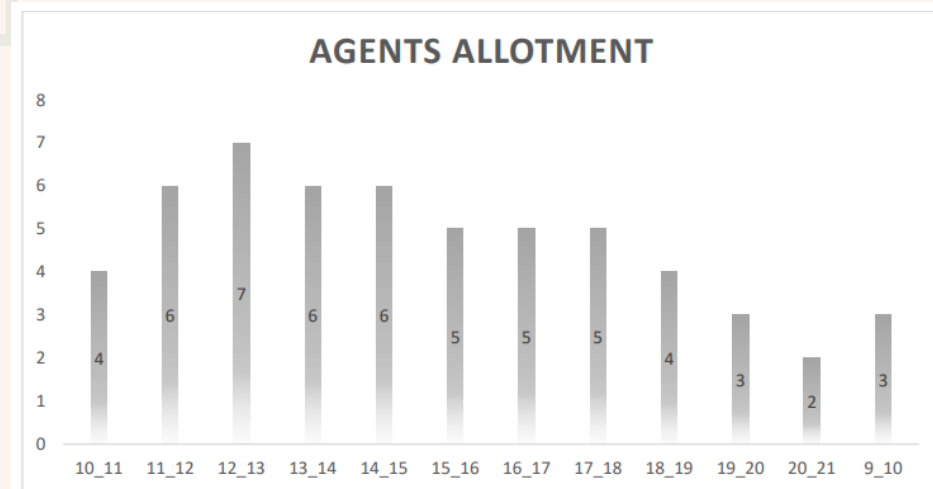


3. MANPOWER PLANNING

To achieve the target of reducing the call abandonment rate to 10%, the project calculated the minimum number of agents required for each time bucket. This involved analyzing the total average number of calls per day, the current answered call rates, and the existing manpower allocation. The findings indicated that increasing the number of agents during peak times is essential. For instance, during the 11 AM – 12 PM slot with the highest call volume, more agents are needed to ensure that at least 90% of calls are answered promptly. The detailed manpower plan includes specific recommendations for each time bucket, optimizing the allocation of agents to handle the predicted call volume efficiently.

OUTPUT:

new total agents working per day		57
AGENT ALLOTMENT PLAN		
Call_Status	answered	
Time Bucket	Count of Customer_Phone_No	Agents allotment
10_11	6368	4
11_12	8560	6
12_13	9432	7
13_14	8829	6
14_15	7974	6
15_16	7760	5
16_17	7852	5
17_18	7601	5
18_19	6200	4
19_20	4578	3
20_21	2870	2
9_10	4428	3
Grand Total	82452	57





FINDINGS



4. NIGHT SHIFT PLANNING

Considering the average of 1539 calls expected during night hours, a comprehensive plan was developed to allocate agents effectively across various night-time slots. The analysis determined that a total of 17 agents are required to maintain a 10% abandon rate during the night shift. This plan involves distributing agents based on the anticipated call volume for each time bucket, ensuring that there are enough agents available to answer calls promptly, even outside regular business hours. By addressing the unique challenges of night-time call handling, this plan aims to improve customer satisfaction by providing reliable service around the clock.

OUTPUT:

AGENT ALLOTMENT PLAN (9 PM - 9 AM)			
time bucket (night)	calls as a part of 30	total hours needed	agents needed
21_22	3	7.64	2
22_23	3	7.64	2
23_00	2	5.09	1
00_01	2	5.09	1
01_02	1	2.55	1
02_03	1	2.55	1
03_04	1	2.55	1
04_05	1	2.55	1
05_06	3	7.64	2
06_07	4	10.19	2
07_08	4	10.19	2
08_09	5	12.74	3
total	30	76.42	19





ANALYSIS

The analysis involved creating pivot tables and bar charts to visualize call duration and volume trends across different time buckets. The average call duration analysis revealed that specific time slots, particularly mid-morning and late evening, require more attention due to longer call durations. The call volume analysis showed peak times during late morning hours, indicating the need for increased staffing during these periods. Manpower planning calculations considered the total average calls per day, current answered rates, and working hours to determine the necessary number of agents required to reduce the abandonment rate from 30% to 10%. For night shifts, a detailed agent allocation plan was devised to ensure adequate coverage and prompt responses to customer calls outside regular business hours.

CONCLUSION

The project successfully identifies critical factors influencing call handling efficiency at ABC Insurance Company. By leveraging Excel for data analysis and visualization, the project provides actionable insights into call durations, volume trends, and optimal manpower allocation. Implementing the proposed manpower plans for both regular and night shifts can significantly reduce the call abandonment rate, enhancing customer satisfaction and operational efficiency. This project underscores the importance of data-driven decision-making in optimizing customer service operations and resource management in the insurance sector.

A series of thin, light brown lines forming an abstract geometric pattern in the top left corner of the page. The lines intersect to create various triangular and polygonal shapes, some of which are nested within others.

WHAT DID I LEARN?

Through my various projects, I have gained extensive experience in data analysis, machine learning, and data visualization. Working on these projects has enhanced my technical skills, especially in using Python, SQL, and Excel for data manipulation and analysis. I have learned how to clean and preprocess data, conduct thorough analyses to uncover insights, and visualize findings to support data-driven decision-making.

Moreover, these projects have taught me the importance of a systematic approach to problem-solving. From defining the problem and planning the analysis to processing data and sharing results, I have developed a structured methodology that can be applied to diverse real-world scenarios. Collaborating on these projects has also improved my ability to communicate complex technical information clearly and effectively, making it accessible to both technical and non-technical stakeholders.

Collaborating on these projects has also improved my ability to communicate complex technical information clearly and effectively. I have learned to create visualizations and reports that make data accessible to both technical and non-technical stakeholders. This skill is crucial for ensuring that insights derived from data analysis can be effectively used to inform decisions and drive improvements. Overall, my projects have provided me with a strong foundation in data analytics and have equipped me with the skills and knowledge needed to tackle a wide range of challenges in the field of artificial intelligence and data science.



APPENDIX

PROJECTS:

<https://drive.google.com/drive/folders/19sPoASO3jRT4QJfBkyIuIM-UTC9pesbH?usp=sharing>

GITHUB:

<https://github.com/mirra-202>

LINKEDIN:

<https://www.linkedin.com/in/mirragavery/>