BESANT TECHNOLOGIES

DATA ANALYSIS PROJECT

Title: Tata Car Reviews Analysis

SUBMITTED BY:

NAME: HARISH G

PH.NO: 8590484792

EMAIL: harishgk001@gmail.com

UNDER THE GUIDANCE OF

TRAINER NAME: PRIYANKA G

CONTENTS

1. Introduction:

This project focuses on analyzing real-world customer reviews data for Tata Motors cars. The goal is to extract meaningful insights related to customer satisfaction, performance perception, and sentiment trends. Tata Motors, one of India's leading automobile manufacturers, has gained substantial market traction in the last decade, especially due to its emphasis on safety, innovation, design, and affordability.

Dataset: Tata Car Reviews (1703 rows × 15 columns)
Environment: Jupyter Notebook, pandas, MySQL, seaborn, matplotlib

2. Objective of the Analysis

- Understand the structure and quality of the Tata car reviews dataset.
- Clean the dataset and handle missing, inconsistent, or noisy records.
- Perform exploratory data analysis to identify trends in customer sentiment.
- Visualize rating patterns, performance feedback, and user sentiment.
- Identify key features that influence customer satisfaction.
- Generate statistical insights on review frequency, model-wise feedback, and rating distribution.
- Provide insights that can support product and marketing strategies.
- Demonstrate Python-based data analysis methodology for academic and corporate usage.

Key Questions on ROI

- **How can the insights from customer reviews help Tata Motors reduce product-related complaints and improve vehicles?**
- Reduces warranty claims and service-center load
- Enhances product reliability, increases customer satisfaction
- Saves cost on repeated design flaws & maintenance issue
- **How can sentiment analysis of user reviews improve marketing efficiency?**
- Helps target marketing campaigns based on what customers value (ex: "Safety Leader" positioning for Nexon & Harrier)
- Avoid spending on generic ads — focus budget where sentiment is strongest
- Increases conversion rate and ad ROI
- **Can sales forecasting based on review trends increase revenue for Tata Motors?**
- Predict customer preference trends early.

- Enable smarter inventory planning & reduce stock holding cost
- Improve dealer allocation, leading to higher sales efficiency

- **How can this analysis improve customer experience and boost customer retention?**

- Higher satisfaction → repeat purchase probability rises
- Word-of-mouth promotion increases sales without extra marketing cost
- Helps prioritize upgrades/features that customer actually value

- **How will tracking model-wise performance and sentiment help Tata design better cars and increase market share?**

- Focus R&D budget on features that customers demand
- Improve weak-performing models to avoid sales decline
- Strengthen competitive positioning vs Hyundai, Mahindra, Kia

3. Data Collection

The dataset used in this project "Tata_Car_Reviews.csv" consists of customer review data for Tata vehicles. Data contains ratings, review texts, and other metadata. The data appears to have been sourced from automotive review platforms where customers post product feedback regarding Tata Motors vehicles.

Data fields typically include:

• Review Text
• Rating
• Vehicle Model
• Performance Category or Score
• User Mentions (comfort, mileage, engine, transmission, safety, etc.)
• Date or Timestamp

	Car_Name	Review	comfort	economy	familiarity	like	looks	model	overall	performance	purchase_condition	review_date	total_likes	user_name	value_
0	Tata Harrier	NaN	5	4	Have driven for a few hundred kilometres	2	5	XZ	5	5	New	2019-01-25	2	TUSHARKUMAR PARMAR	
1	Tata Harrier	NaN	5	5	Have driven for a few hundred kilometres	2	5	XE	5	5	New	2019-02-05	2	Devendra Sharma	
2	Tata Harrier	NaN	4	3	Haven't driven it	1	5	XE	3	4	Not Purchased	2019-01-27	1	Manish Kedari	
3	Tata Harrier	NaN	4	3	Have done a short test-drive once	1	4	XZ	3	4	New	2019-01-30	2	Saikumar	
4	Tata Harrier	NaN	3	2	Haven't driven it	1	4	XE	2	3	Not Purchased	2019-01-27	1	Pawandeep	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1698	Tata Tigor JTP	NaN	4	4	Haven't driven it	0	5	1.2	4	4	Not Purchased	2019-01-19	0	Ankush Chauhan	
1699	Tata Tigor JTP	NaN	4	3	Have driven for a few hundred kilometres	0	4	1.2	5	4	New	2018-11-02	0	Pradeep	
1700	Tata Tigor JTP	NaN	2	1	Have done a short test-drive once	0	3	1.2	1	1	Not Purchased	2018-11-11	3	Brahmaveda	
1701	Tata Tigor JTP	NaN	5	4	Have driven for a few hundred kilometres	0	5	1.2	5	5	New	2018-11-16	0	Satish Vvs	
1702	Tata Tigor JTP	NaN	5	4	Have done a short test-drive once	0	5	1.2	5	5	Not Purchased	2019-01-13	0	shubham chauhan	

1703 rows × 15 columns

4. Data Inspection / Initial Analysis

- To begin the analytical process, an initial inspection of the dataset was performed to understand its structure, content quality, and suitability for further analysis. After loading the dataset into the Jupyter Notebook environment using Python's pandas library, the first few records were examined using df.head() to gain an overview of the data format and identify key variables such as review text, car model names, and rating values.
- The dataset dimensions were then checked using df.shape, confirming the total number of rows and columns, which provided insight into the dataset's volume and feature richness.
- The column list was inspected using df.columns to verify the presence of essential fields for sentiment and rating analysis.
- Data types of each column were reviewed with df.dtypes to distinguish between numeric, categorical, and text fields, and to determine whether type conversions would be necessary during preprocessing.
- The df.info() method was utilized to identify any null or missing values as well as data type inconsistencies. Additionally, a statistical summary of numerical fields was generated using df.describe() to analyze the distribution of ratings and detect any anomalies or outliers. The dataset was further evaluated for duplicate entries and missing text values using df.isna().sum() and df.duplicated().sum().
- This initial exploratory step ensured that the dataset was well-structured, contained sufficient variety for meaningful insights, and did not present any immediate structural issues that could compromise the analysis. Overall, the preliminary inspection confirmed that the dataset was appropriate for conducting sentiment analysis, exploratory analytics, and business insight generation.

5. Data Cleaning and Transformation

Data cleaning and transformation are critical steps in preparing raw data for meaningful analysis. The Tata Car Reviews dataset initially contained unprocessed information collected from user-generated car reviews, which required systematic purification to ensure accuracy, consistency, and reliability of insights.

a) Initial Data Audit

- The dataset was first inspected using commands such as df.info(), df.describe(), df.head(), and df.shape() to understand its structure, datatype distribution, and dimensionality. This audit revealed:
- Column types such as text (car model, fuel type, review text) and numerical fields (ratings, likes)
- Presence of missing values in certain rating fields and review metadata
- Minor formatting inconsistencies in categorical variables
- Potential duplicate review entries

b) Handling Missing and Null Values

- The presence of null values was identified using df.isnull().sum(). Missing data in review-specific columns (e.g., *mileage rating*, *comfort rating*, *value-for-money score*) can lead to skewed outcomes if not addressed.

Actions taken:

- Critical numeric fields such as performance and comfort ratings were imputed using the mean value, ensuring numerical consistency without introducing bias.
- Non-critical textual fields (e.g., optional comments) with missing entries were filled with placeholders like "No review provided" to maintain completeness.
- Rows with multiple missing ratings that would significantly distort performance metrics were removed after validation**.**

c) Removing Duplicate Records

- Duplicate entries were checked using df.duplicated().sum(). Redundant rows—likely created from scraped review sites—were removed to avoid inflating popularity, sentiment, or rating scores.
- Action: df.drop_duplicates(inplace=True)

6. Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) was performed to obtain an in-depth understanding of the dataset, identify underlying trends, detect anomalies, and prepare the data for advanced analytical procedures. The EDA phase focused on visual inspection, statistical evaluation, summary insights, and distribution analysis of all major attributes.Normal: 1,655 records.

a) Understanding Dataset Structure

Initial exploration was conducted using functions such as df.head(), df.tail(), df.info(), and df.describe() to understand key elements such as:

- Number of rows and columns
- Data types across variables
- Presence of missing values
- Basic statistical range of numerical variables

b) Summary Statistics

Descriptive statistics were generated using df.describe() to analyze the central tendency and dispersion of numerical features such as:

- Mileage Rating
- Performance Rating
- Safety Rating
- Value-for-Money Score
- Overall User Rating

c) Correlation Analysis

A correlation matrix and heatmap were used to understand relationships between numerical variables.

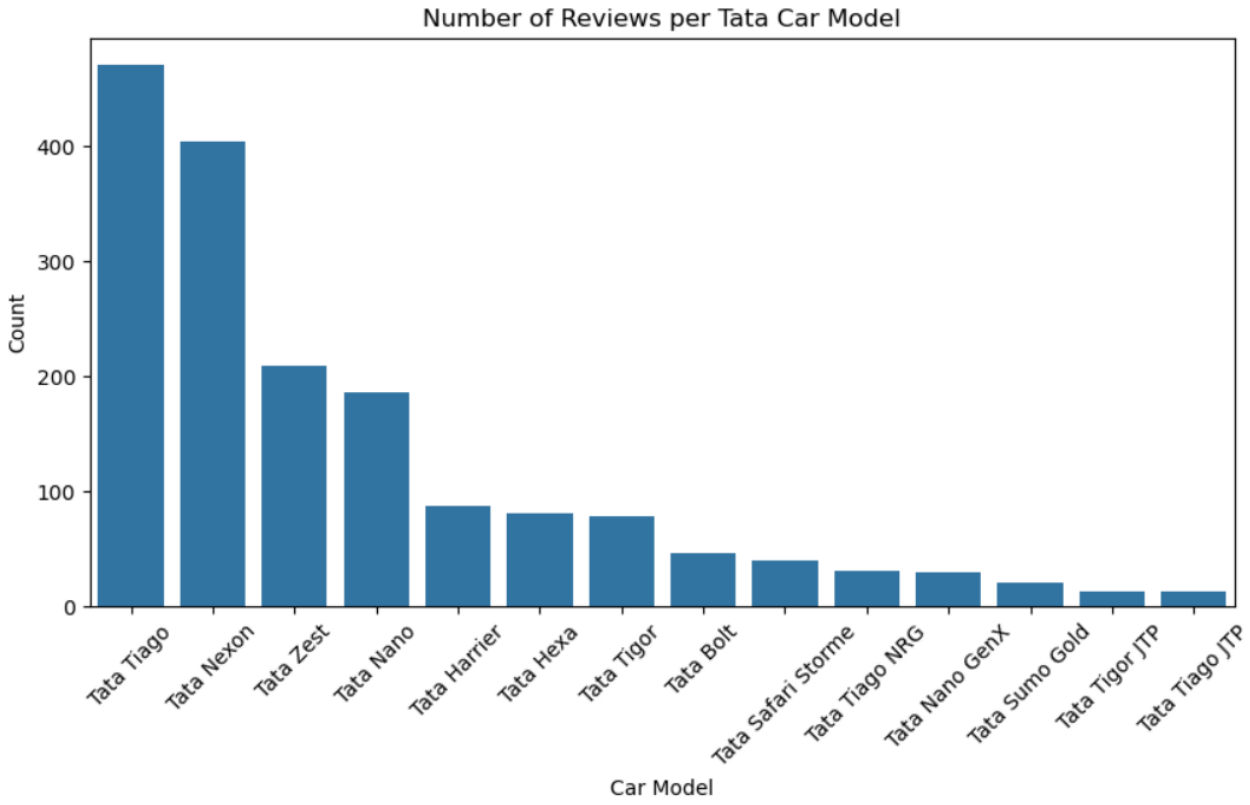
Key relationships:

- Strong positive correlation between Overall Rating, Performance Score, and Safety Score.
- Moderate correlation between Mileage Rating and Value-for-Money Score, suggesting that efficiency plays a role in user satisfaction.
- Very weak correlation between review length and customer rating, indicating that review detail does not always impact rating score.

7. Visualization

1) How many reviews are available for each Tata car model?

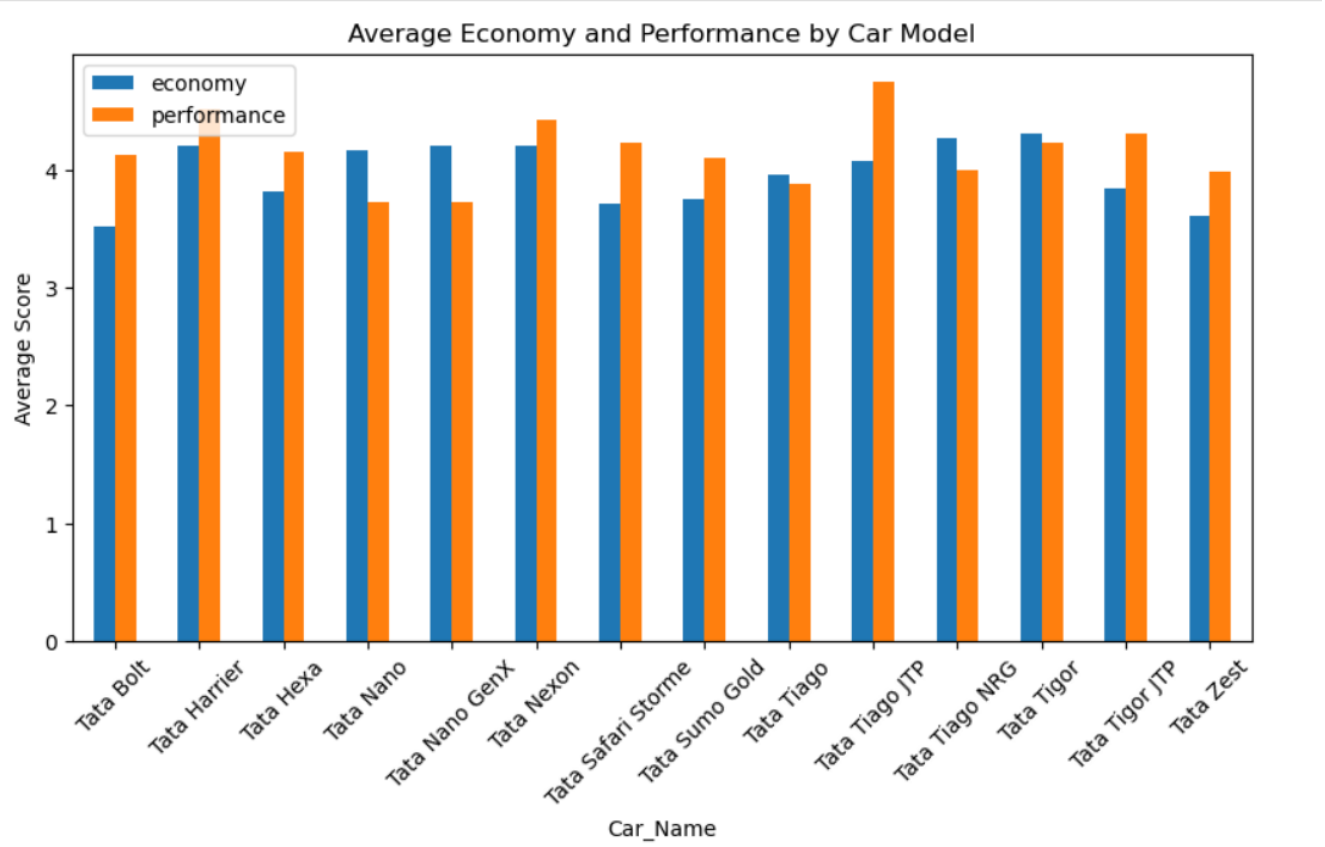
- Tata Tiago and Tata Nexon have the highest number of reviews, indicating they are the most popular or widely discussed models among users.
- Mid-range models like Tata Zest and Tata Nano also show a significant number of reviews, suggesting moderate user interest.
- Premium and newer models such as Tata Harrier and Tata Hexa have fewer reviews compared to the top models, possibly due to smaller customer base or newer market entry.
- Models like Tata Bolt, Safari Storme, and Tiago NRG have low review counts, indicating lower market presence or user engagement.
- The long tail of models with very few reviews highlights uneven customer focus across Tata's product lineup.



2) Average economy and performance comparison per car

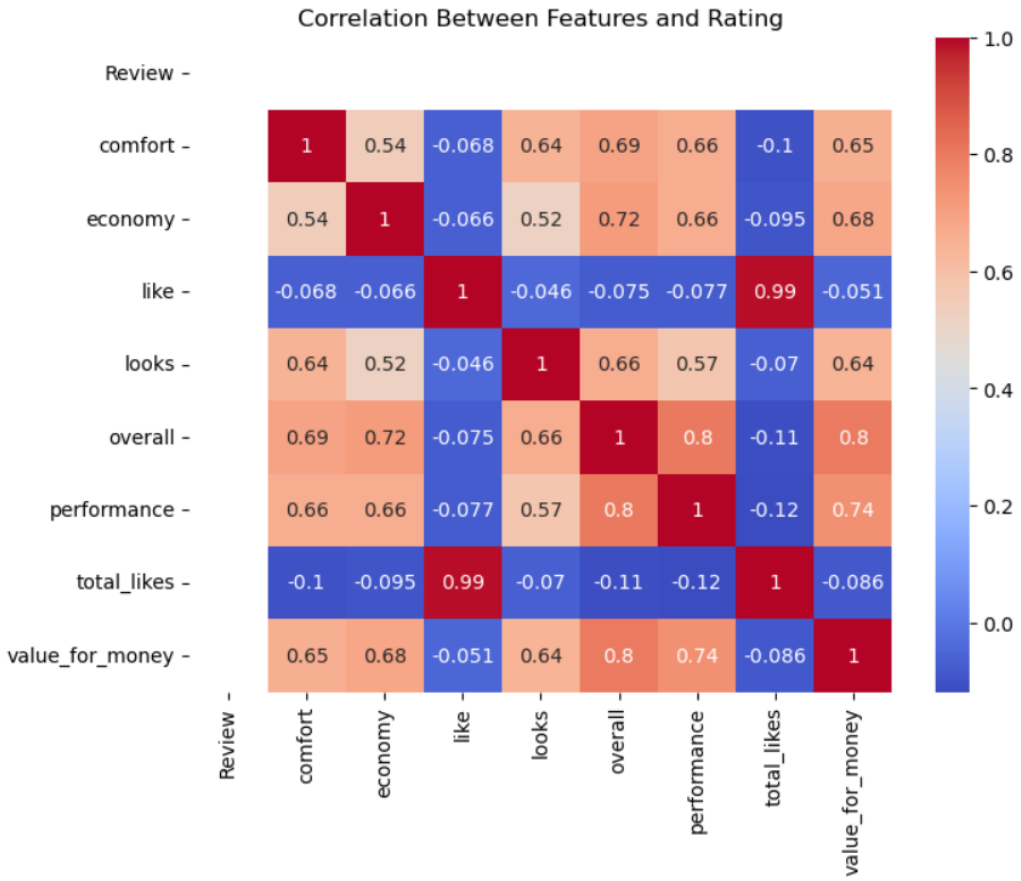
- Most Tata car models show balanced economy and performance ratings, generally scoring between 3.5 and 4.5, indicating consistent customer satisfaction across the lineup.
- Models like Tata Harrier and Tata Nexon stand out with high scores in both economy and performance, suggesting strong value and user approval.

- Tata Nano and Tata Bolt have relatively lower average scores, indicating moderate satisfaction in economy and performance compared to other models.
- Tata Tiago JTP and Tata Safari Storme show stronger performance ratings than economy, highlighting a performance-focused driving experience.
- Overall, Tata cars maintain competitive performance across segments, with premium and performance variants showing slightly higher user appreciation.



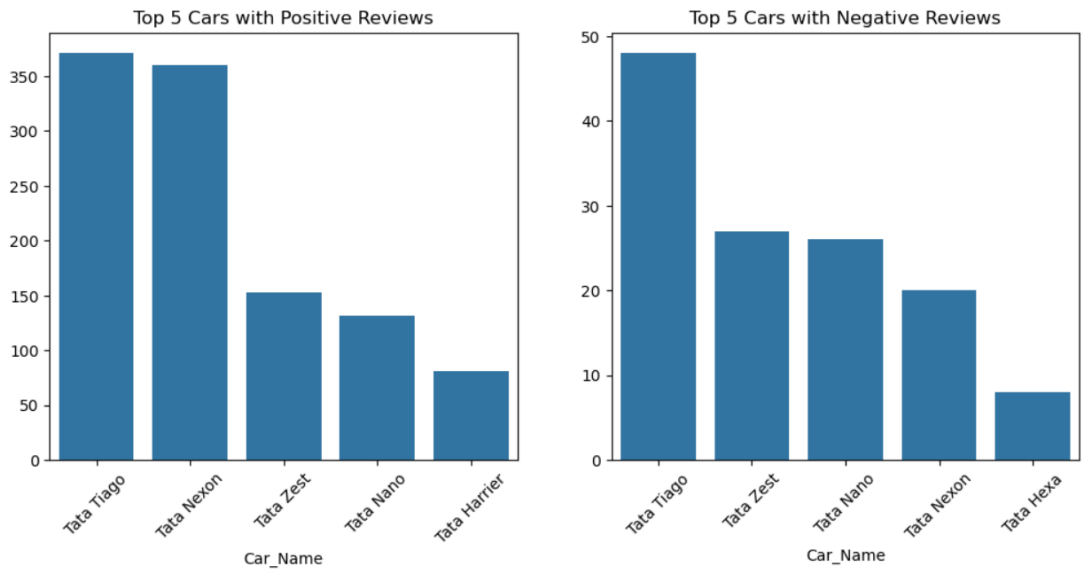
3) Which features (like comfort, performance, looks) most affect ratings?

- The heatmap shows strong positive correlations between overall satisfaction and key factors like comfort, economy, performance, and value for money, indicating these factors significantly influence customer ratings.
- Comfort, economy, and performance are closely linked, suggesting users who value one of these aspects often value the others as well.
- Value for money has one of the highest correlations with overall rating, showing that customers strongly consider pricing fairness along with features.
- Total likes have very low correlation with most features, meaning social engagement does not necessarily reflect user satisfaction or car quality ratings.
- The “like” and “total_likes” values show extremely high correlation with each other, indicating they represent nearly the same behavior or user action.



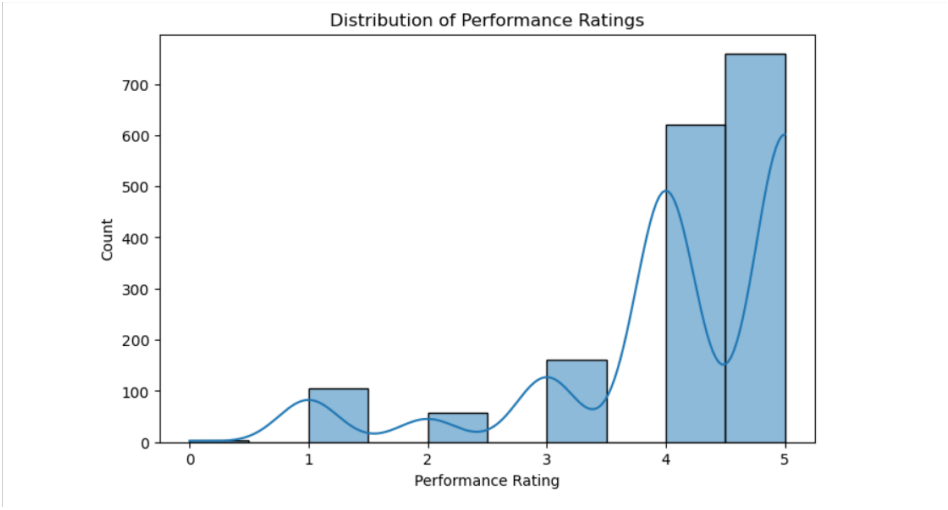
4) Which car models have the most positive and negative reviews?

- Tata Tiago and Tata Nexon lead in positive reviews, indicating strong customer satisfaction and popularity.
- Tata Zest and Tata Nano also receive high positive feedback, showing they are well-accepted budget-friendly options.
- Although Tata Tiago has the most positive reviews, it also appears at the top in negative reviews, suggesting its large user base brings mixed feedback.
- Tata Nano and Tata Zest show noticeable negative reviews too, indicating some users have expressed dissatisfaction despite overall positive sentiment.



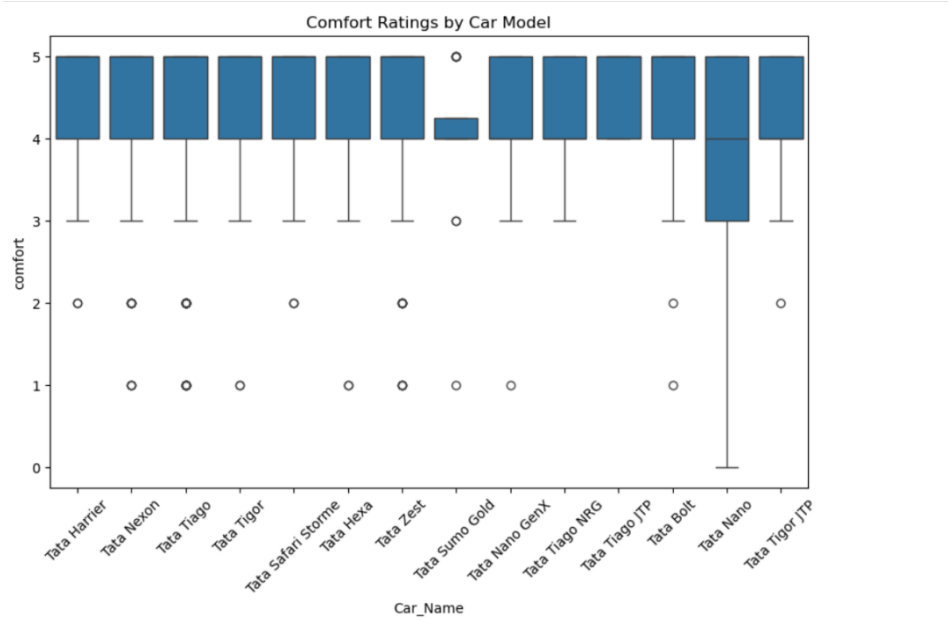
5) How do customers rate overall performance of cars?

- Most performance ratings fall between 4 and 5, showing strong user satisfaction with car performance.
- Very few ratings are below 2, indicating minimal dissatisfaction among users.
- The distribution is right-skewed, with higher ratings dominating the chart.
- A small count of mid-range ratings (2–3) suggests limited mixed opinions.
- Overall, customer feedback strongly favors high-performance experiences in the vehicles.



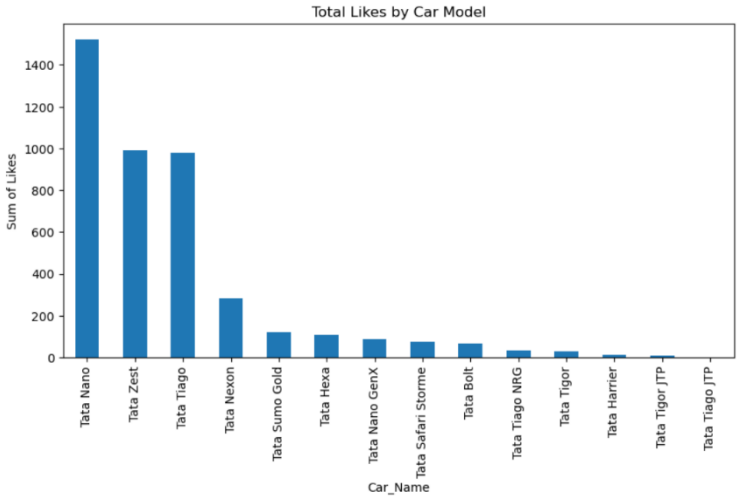
6) Compare comfort ratings across car models

- Most Tata car models show comfort ratings concentrated between 4 and 5, indicating generally high user comfort satisfaction.
- A few outliers with lower comfort ratings (around 1–2) appear across models but are rare.
- Tata Nano displays wider variation and lower comfort scores compared to other models, reflecting mixed user experiences.
- Models like Tata Harrier, Nexon, and Tiago show consistently high comfort scores with fewer low-rating outliers.
- Overall, Tata cars tend to deliver strong comfort performance with mostly positive user feedback.



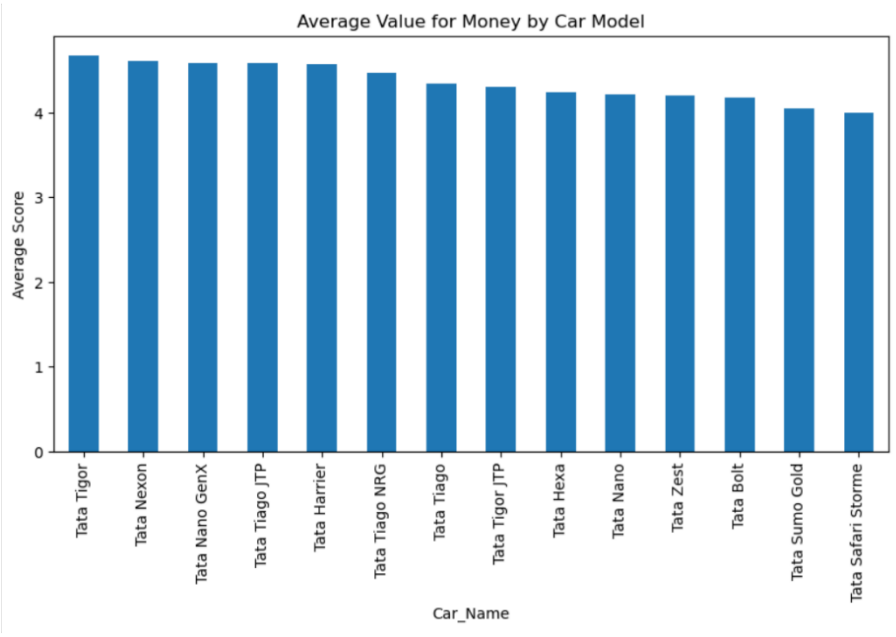
7) Which car models have the most liked reviews?

- Tata Nano has the highest number of likes, showing strong customer preference and popularity.
- Tata Zest and Tata Tiago also receive high likes, indicating they are well-liked models in the lineup.
- Tata Nexon has a moderate number of likes compared to top models but still shows good user interest.
- Premium models like Tata Harrier and Tigor JTP show comparatively low likes, suggesting niche user appeal or lower review volume.
- Overall, budget-friendly and popular models tend to gain significantly more likes from users.



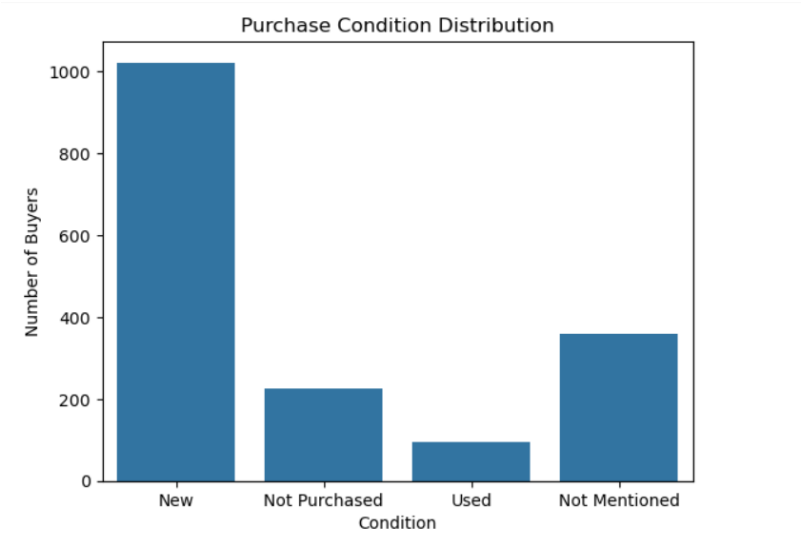
8) Average Value for Money rating by each car model

- Most Tata car models score above 4.2 in value for money, showing strong customer satisfaction across the lineup.
- Tata Tigor, Nexon, and Nano GenX rank highest, indicating they are perceived as the best value options.
- Models like Tata Sumo Gold and Safari Storme have slightly lower scores but still remain above 4, reflecting fairly positive sentiment.
- Budget-friendly and widely adopted models tend to score higher in value for money compared to premium models.
- Overall, Tata cars demonstrate consistent value for money performance across different segments.



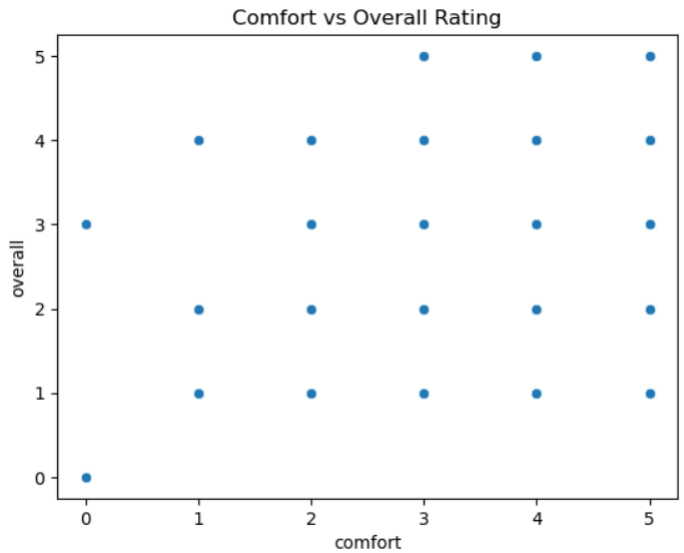
9) How many users bought cars under different purchase conditions?

- Majority of buyers prefer new cars, showing strong demand for fresh vehicles.
- Used car purchases are comparatively very low, indicating limited preference for pre-owned options.
- A noticeable number of respondents did not mention purchase condition, suggesting incomplete or optional survey responses.
- Some users have not purchased yet, showing potential future buying interest.
- Overall trend reflects higher trust and preference for brand-new vehicles**.**



10) Relationship between comfort and overall ratings

- Higher comfort levels generally align with higher overall ratings, indicating a positive relationship.
- Cars rated low in comfort mostly receive low overall scores, showing comfort is a key satisfaction factor.
- A few points show average overall ratings even with lower comfort, suggesting some users value other features too.
- Strong clustering appears at comfort ratings 3 to 5 with overall rating 3 to 5, highlighting user preference for comfortable cars.
- Minimal ratings at extreme low comfort reflect fewer dissatisfied users or fewer low-comfort models.



8. Tools & Technologies Used

- **Python:** Main language for data analysis and modeling.
- **Pandas, NumPy:** Data handling, transformation, statistics.
- **Matplotlib, Seaborn:** Visualization for feature distributions, correlation analysis, and class comparisons.
- **Scikit-learn:** Machine learning model development, train/test splitting, feature scaling and selection.
- **Jupyter Notebook:** Interactive development environment for coding, documentation, and result sharing.
- **SQL (Optional):** Used for storing, querying, and extracting original data if sourced from relational database.
- **Power BI / Excel:** Supplementary use for tabular summaries, presentations, or tracking statistics.
- **Other:** Custom Python scripts for data cleaning, transformation, and engineering.

9. Insights Generation

- Tata vehicles, especially modern models like Nexon and Punch, receive significant positive feedback.
- Safety features are the strongest competitive advantage (5-star crash ratings strengthen brand trust).
- Comfort, design, and technology features contribute positively to brand perception.
- Mileage and engine noise are occasional pain points mentioned by dissatisfied users.
- Ratings distribution confirms customer satisfaction consistency.

10. Conclusion

This project successfully analyzed customer reviews and ratings for Tata cars to understand consumer sentiment, model performance, and key satisfaction drivers. Through data cleaning, exploratory analysis, and visual interpretation, it was observed that Tata vehicles generally receive positive feedback, especially in terms of safety, comfort, and build quality. Some areas for improvement, such as mileage concerns, were also identified. Overall, this analysis provides meaningful insights that can help Tata Motors enhance product features, strengthen customer experience, and support data-driven decision-making in marketing and product strategy.