“USED BIKE PRICES - FEATURE ENGINEERING AND EDA”

A Virtual Internship Project Report On Used Bike Prices-Feature Engineering and EDA

Submitted to:

UNIFIED MENTOR

YOUR SKILL, SUCCESS & JOURNEY

Submitted by:

Kanika Chaudhary
UMID: UMID3003252671
BATCH-01/04 to 01/08/2025

Introduction: An Analytical Perspective on Used Bike Pricing in India

The automotive sector in India, particularly the two-wheeler segment, has witnessed remarkable growth, establishing itself as a pivotal contributor to the nation's economy. Within this dynamic landscape, the used bike market has emerged as a significant and complex domain, driven by factors such as affordability, accessibility, and evolving consumer preferences. Understanding the intricate dynamics of pricing in this secondary market is not merely an academic exercise but a critical endeavour for stakeholders ranging from individual consumers and dealers to financial institutions and manufacturers.

This project, titled "Used Bike Prices - Feature Engineering and EDA," delves into a comprehensive analysis of the factors influencing the valuation of pre-owned motorcycles in the Indian context. Leveraging advanced data science methodologies, the project aims to unravel the multifaceted relationships between various bike attributes and their corresponding market prices. The core objective is to develop a robust understanding of price determinants and, ultimately, to construct predictive models that can accurately forecast used bike prices. This endeavor is particularly relevant for finance analysts, enabling them to assess asset valuations, mitigate risks in lending, and inform investment strategies within the automotive resale sector.

The foundation of this project lies in a meticulously curated dataset encompassing key characteristics of used bikes in India. These features, crucial for a granular analysis, include:

Model Name and Manufacturing Year: These attributes provide insights into the technological advancements, design trends, and depreciation patterns associated with specific bike models over time.

Kilometres driven (kms_driven): As a primary indicator of usage and wear, this metric directly correlates with the physical condition and remaining lifespan of a bike, significantly impacting its resale value.

Owner Type: The number of previous owners (first, second, or subsequent) reflects the bike's ownership history, often influencing buyer perception regarding maintenance and reliability.

Location: Geographical variations in demand, supply, regional preferences, and local economic conditions play a substantial role in determining pricing disparities across different Indian cities and states.

Mileage (kmpl): Fuel efficiency is a paramount concern for Indian consumers, making mileage a critical determinant of a bike's economic viability and thus its market price.

Power (Bhp) and Engine Capacity (cc): These specifications directly relate to a bike's performance capabilities, influencing its appeal to various buyer segments (e.g., performance enthusiasts vs. daily commuters) and consequen3tly its price.

Brand: The reputation, reliability, after-sales service, and market positioning of a manufacturer significantly impact the residual value and desirability of its used bikes.

Key Insights

- **Data Quality is Paramount:** Emphasizes thorough cleaning, handling missing values and duplicates.
- **Feature Engineering Boosts Models:** Creates new features (e.g., "Age of Bike," "Power-to-Weight Ratio") and processes existing ones for better model performance.
- **EDA Informs Understanding**: Uses statistical summaries and visualizations to reveal data distributions and relationships between variables.
- **Model Selection & Evaluation**: Discusses various regression models and metrics (MAE, MSE) for assessing their predictive accuracy.
- **Identify Price Drivers**: Aims to determine the most influential factors on used bike prices.
- **Iterative Improvement:** Recommends continuous refinement through more data, advanced features, and complex models. Impact and Applications

This project demonstrates how data analysis can provide actionable insights into personal finance. The findings can be used to

- Recommend savings strategies
- Detect anomalies in transaction behaviour
- Forecast monthly expenses and optimize cash flow
- Prepare individuals for financial planning and goal-setting

Tools & technologies used

- Python (pandas, matplotlib, seaborn, numpy)

Dataset details

File name: (“bike_prices.csv”)

File Path: https://drive.google.com/file/d/1Nbq7uIvVdj3iF8b7VWfIfhuEV5jCzsXH/view

Analytical Approach

- **Data Collection and Preparation:** This initial phase focuses on loading and inspecting the dataset to understand its structure and identify inconsistencies
- **Exploratory Data Analysis (EDA):** The goal of EDA is to gain insights into the data's characteristics and relationships.
- **Data Visualization:** Beyond EDA, this stage involves creating various plots for a deeper understanding of relationships and trends.
- **Model Building:** This is the core machine learning phase aimed at predicting bike prices.
- **Conclusion and Recommendations:** The final stage involves summarizing key insights, particularly which factors significantly influence bike prices

USED BIKE MARKET

OPPORTUNITIES AND FORECAST, 2021 - 2031

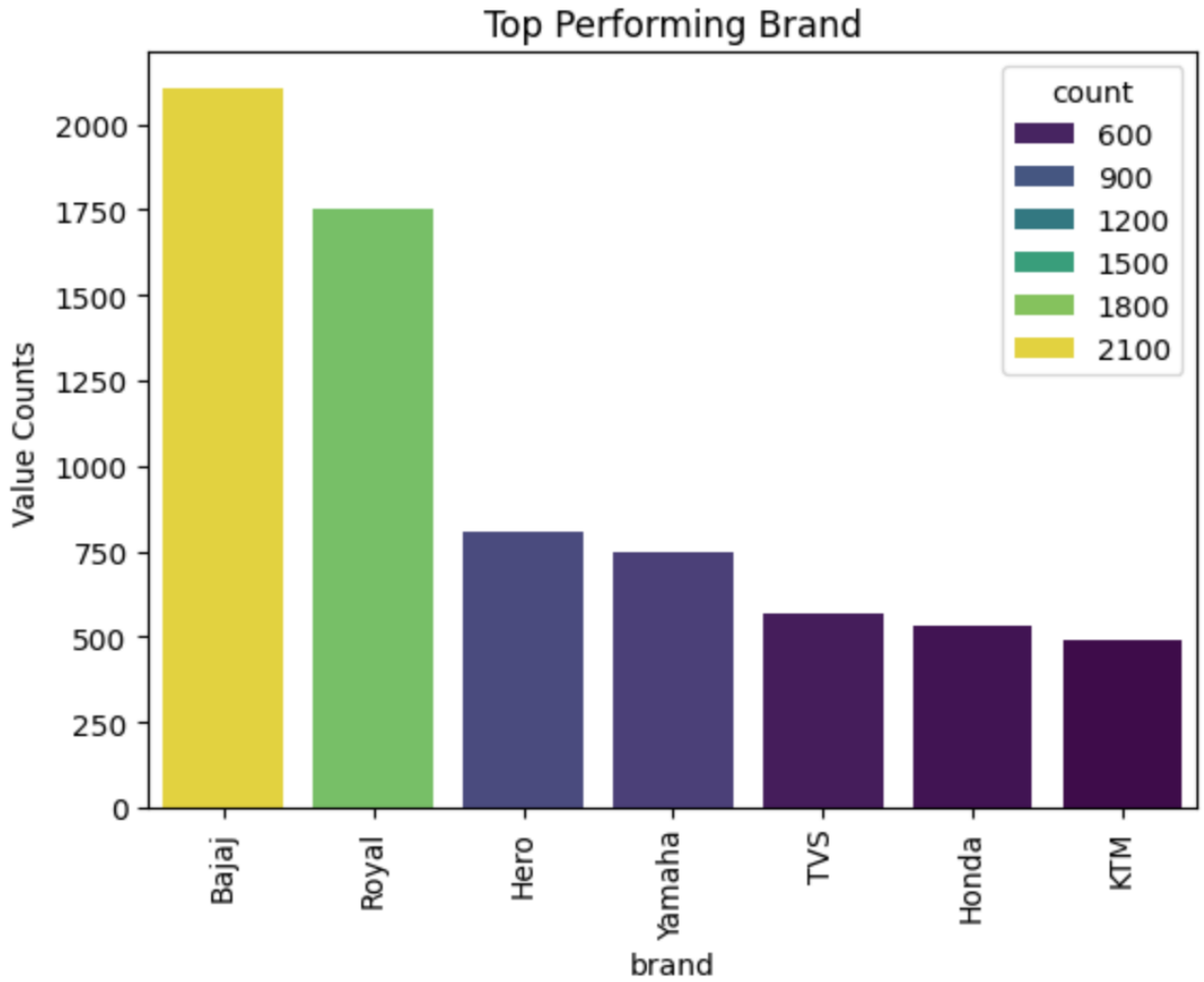
Used bike market is expected to reach $66.2 Billion in 2031

Growing at a CAGR of 5.6% (2022-2031)

Report Code: A09641, www.alliedmarketresearch.com

```
brand_VC = df.brand.value_counts().iloc[:7]

sns.barplot(y=brand_VC.values,x=brand_VC.index ,hue=brand_VC,palette='viridis')
plt.xticks(rotation=90)
plt.title('Top Performing Brand')
plt.ylabel('Value Counts')
plt.show()
```



The graph represents the "Top Performing Brand" based on "Value Counts." The x-axis lists various brands, while the y-axis represents the "Value Counts. and this graph could represent sales data, customer interactions, or another metric relevant to the brands listed.

- **Dominant Leaders:** Bajaj is the clear leader with the highest value count, significantly outperforming all other brands. Royal (likely Royal Enfield) is a strong second, also showing substantial performance. These two brands collectively hold a significant portion of the total "value counts" represented.
- **Mid-Tier Performers:** Hero and Yamaha occupy the middle ground, with respectable but considerably lower counts than Bajaj and Royal. Their performance is relatively similar to each other.
- **Lower-Tier Performers:** TVS, Honda, and KTM are in the lower tier, with the lowest value counts among the brands shown. KTM has the lowest performance in this specific metric.
- **Performance Gap:** There's a noticeable performance gap between the top two brands (Bajaj and Royal) and the rest, indicating a highly concentrated market leadership.
- **Consistency:** The overall trend is a decreasing "Value Count" as you move from left to right (from Bajaj to KTM), indicating a clear hierarchy in performance.

Important Objectives:

- **Market Share Analysis:** The graph provides insights into the market share of different brands. Bajaj and Royal Enfield dominate the market, while Hero and Yamaha are significant players. TVS, Honda, and KTM have a relatively smaller market share.
- **Competitive Positioning:** The graph illustrates the competitive positioning of each brand. Companies can use this data to benchmark their performance against competitors and identify areas for improvement.
- **Brand Valuation:** The graph can be used to estimate brand valuation. Brands with higher value counts, such as Bajaj and Royal Enfield, are likely to have higher brand equity and valuation.
- **Investment Opportunities:** The graph can help investors identify attractive investment opportunities. Consistently high-performing brands like Bajaj and Royal

Enfield may be considered more attractive due to their stable revenue streams and strong market position.

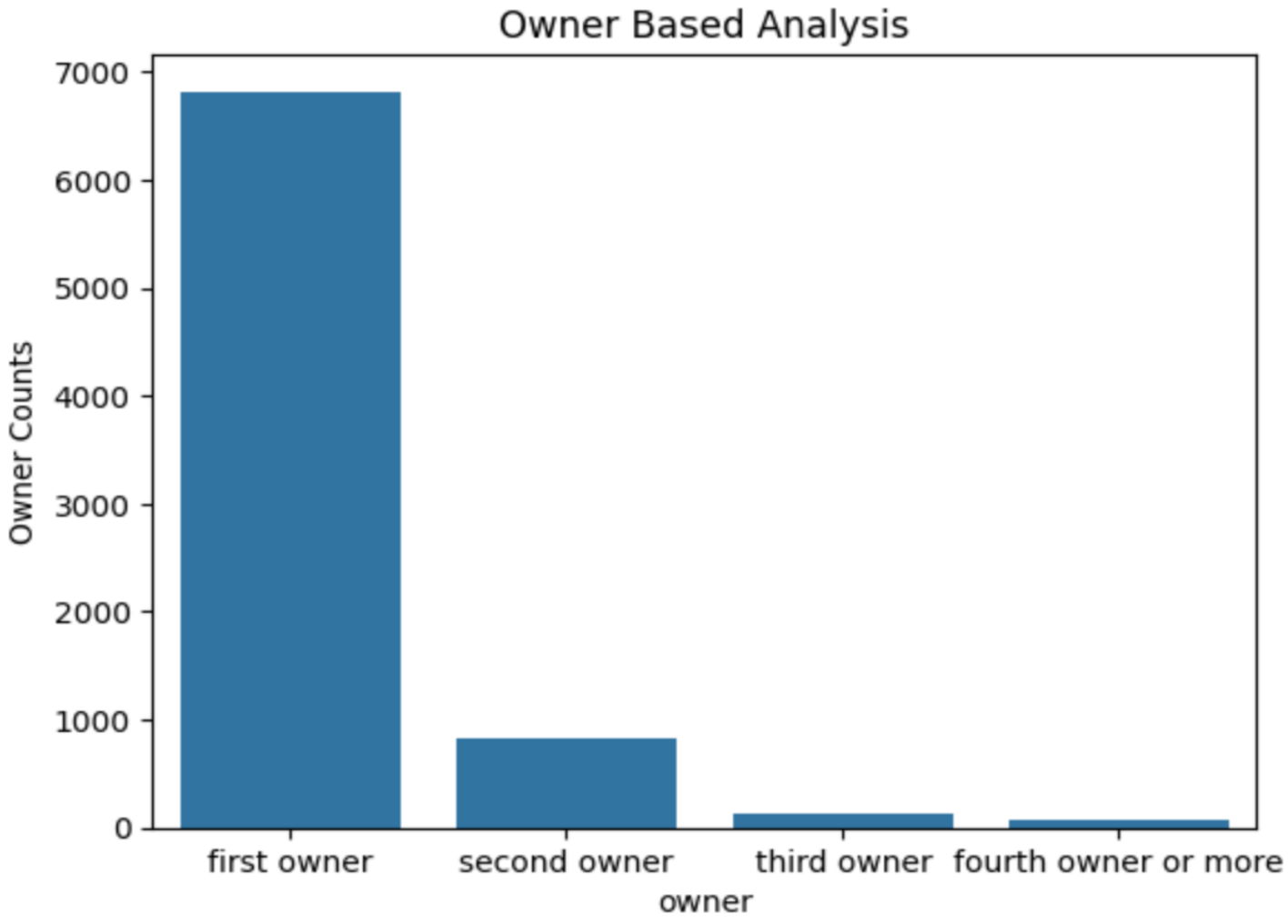
- **Strategic Decision-Making:** The graph can inform strategic decisions regarding pricing, marketing, and product development. Companies can use this data to identify opportunities to improve their performance and gain a competitive edge.

Financial ideas:-

- **Revenue Generation Potential:** Higher "Value Counts" for a brand generally indicate a larger volume of sales or a greater number of transactions. In a business context, higher sales volume directly translates to higher potential revenue, assuming a stable pricing strategy. Bajaj, with the highest count, likely generates the most revenue among these brands.
- **Market Share:** The "Value Counts" can be interpreted for market share. The brand with the highest count holds the largest share of the market represented in this data. A larger market share often leads to economies of scale, stronger bargaining power with suppliers, and greater brand recognition.
- **Brand Equity and Recognition:** Brands with consistently high performance (like Bajaj and Royal) tend to build stronger brand equity. This equity translates into customer loyalty, willingness to pay a premium, and easier access to financing or partnerships
- **Competitive Landscape and Benchmarking:** The graph clearly illustrates the competitive positioning of each brand. Companies can use this data to benchmark their performance against competitors, identify leaders, and understand where they stand in the market. This informs strategic decisions regarding pricing, marketing, and product development.
- **Investment Opportunity:** From an investor's perspective, consistently high-performing brands might be considered more attractive investment opportunities due to their stable revenue streams, strong market position, and potential for future growth.

- **Efficiency and Profitability:** While not directly shown, a higher volume of sales (implied by "Value Counts") can allow companies to achieve greater operational efficiency through economies of scale, potentially leading to higher profit margins. However, this depends on cost structures, pricing, and operational management.

```
oc =df.owner.value_counts()
sns.barplot(y= oc.values,x=oc.index)
plt.ylabel('Owner Counts')
plt.title('Owner Based Analysis')
plt.show()
```



The graph, titled "Owner Based Analysis," presents a bar chart illustrating the distribution of owner counts across different ownership categories. The x-axis lists four categories: "first owner," "second owner," "third owner," and "fourth owner or more." The y-axis represents the "Owner Counts" and ranges.

Overall Performance (Owner-Wise):

- **First Owner:** This category shows overwhelmingly dominant performance, this indicates that the vast majority of items or assets in this analysis are in the hands of their original purchasers or initial entities. From a financial standpoint, this suggests strong primary market sales, high product stickiness, or long holding periods by the initial owners
- **Second Owner:** There's for second owners. This signifies a significantly smaller secondary market. Financially, this could mean that assets lose substantial value after the first sale, become less desirable it also suggests a faster depreciation curve after the initial ownership period.
- **Third Owner:** This indicates an extremely niche or limited market for items that have already had two prior owners. Financially, such items might have reached the end of their economically viable lifespan, require significant investment to maintain, or have very little residual value.
- **Fourth Owner or More:** This category is almost negligible, with counts just above zero. This reinforces the notion that items rarely change hands more than twice or thrice. Financially, this segment likely represents assets with virtually no resale market, indicating near-complete depreciation or obsolescence.

Financial insights:-

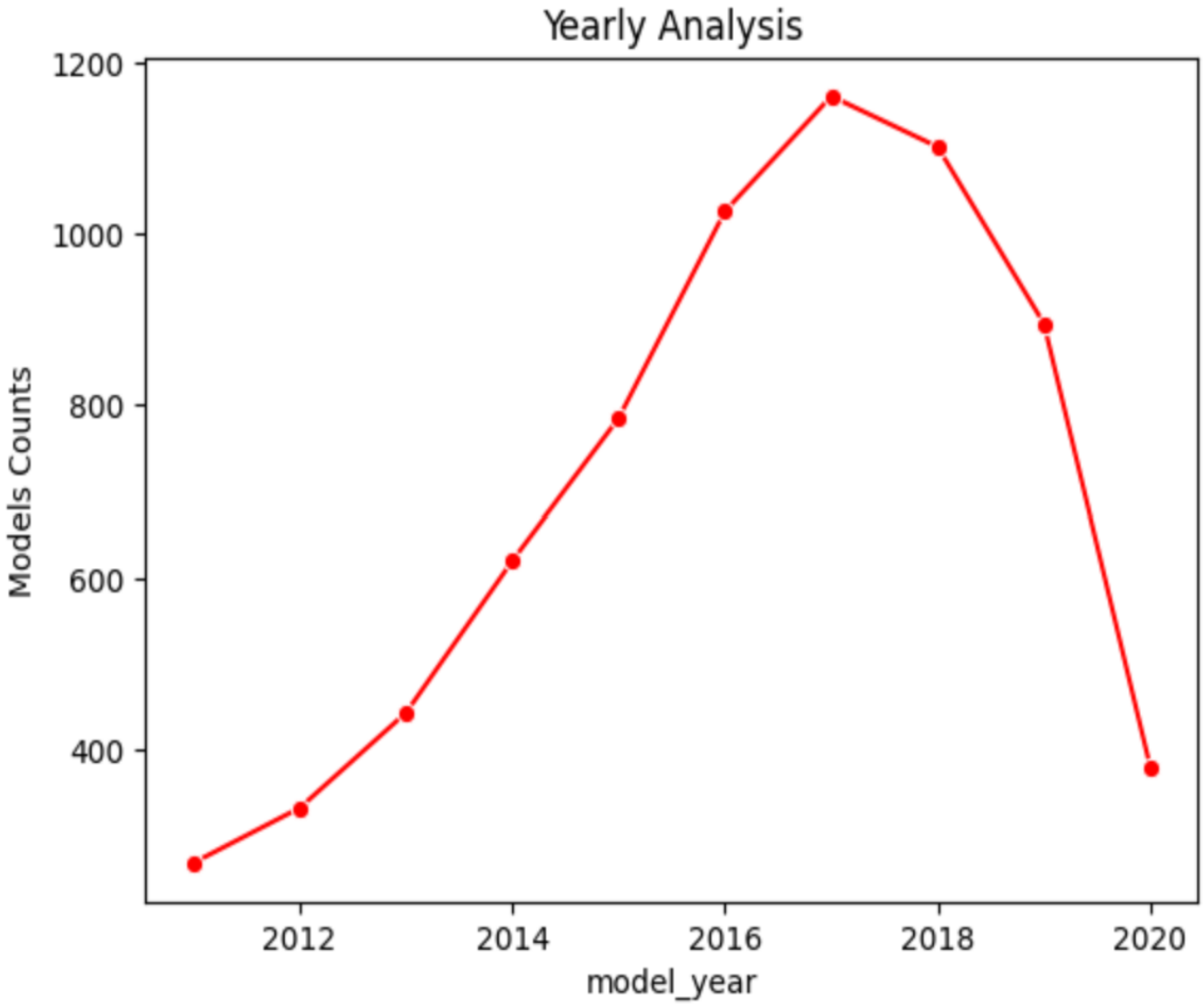
- **Primary Market Dominance:** Highlights the strong prevalence of first-time owners, indicating a robust market for new assets (e.g., new cars, properties).

- **Secondary Market Activity**: Reveals limited activity in the second-hand market, which can suggest high asset retention or lower resale demand/value.
- **Depreciation & Value Retention:** Indirectly informs about asset depreciation trends, where a sharp decline in owner counts after the first owner might imply significant value loss post-initial purchase.
- **Market Saturation & Growth:** Helps assess whether the market is growing (high first-owner numbers) or if there's limited churn in the secondary market.
- **Risk Assessment:** For financial products (e.g., loans), different owner types can indicate varying risk profiles.

Some important objectives:-

- **Quantify Market Segmentation:** Clearly show the distribution of ownership across different stages (first, second, third, fourth or more owners). This helps in segmenting the market and understanding the relative size of primary versus secondary markets.
- **Identify Market Dominance and Trends:** Highlight the dominance of first-time ownership, indicating a strong primary market. The rapid drop-off in subsequent ownership reveals trends in asset longevity, depreciation, and potential resale market liquidity.
- **Inform Strategic Planning:** Provide data crucial for developing targeted financial products (e.g., loans, insurance, and warranties), marketing strategies, and sales approaches for each ownership segment.
- **Assess Asset Value Retention and Depreciation:** Indirectly infer insights into how quickly the asset might depreciate or retain value over time, based on how many times it changes hands. A low number of subsequent owners could suggest rapid depreciation or long asset life with initial owners.

```
my=df.model_year.value_counts().iloc[:10]
sns.lineplot(y=my.values,x=my.index, marker='o' ,color='r')
plt.ylabel('Models Counts')
plt.title('Yearly Analysis')
plt.show()
```



This graph represents the "Models Counts" over different "model years", providing a "Yearly Analysis" of how the number of models has changed each year from approximately 2011 to 2020.

Overall Performance of this Graph:

- **Early Growth (2011-2017):**
  The period from 2011 to 2017 shows robust and consistent growth in "Models Counts This indicates a period of significant expansion, possibly driven by successful product development, market demand, or aggressive market penetration strategies. From a financial standpoint, this would likely correspond to increasing revenues, economies of scale, and potentially strong investor confidence.
- **Peak Performance (2017):**
  The year 2017 represents the pinnacle, with the highest number of models. This is the point of maximum "reach" or "output" based on the metric provided. Financially, this would likely be the most profitable or highest-revenue year, assuming profitability per model remained stable or increased.
- **Decline Phase (2017-2020):**
  Following the peak in 2017, there's a noticeable decline which could be a minor correction or a sign of deceleration.
  Financially, this decline phase would likely correspond to decreasing revenues, potentially shrinking market share, and increased pressure on profitability. It could indicate mature products reaching the end of their lifecycle, a shift in market preferences, increased competition, or even internal strategic changes to streamline product offerings.

Important Objectives:-

- **Understanding Growth and Decline Trends:** Identifying periods of significant growth and decline in the number of models. This can indicate market expansion, contraction, or changing business strategies.

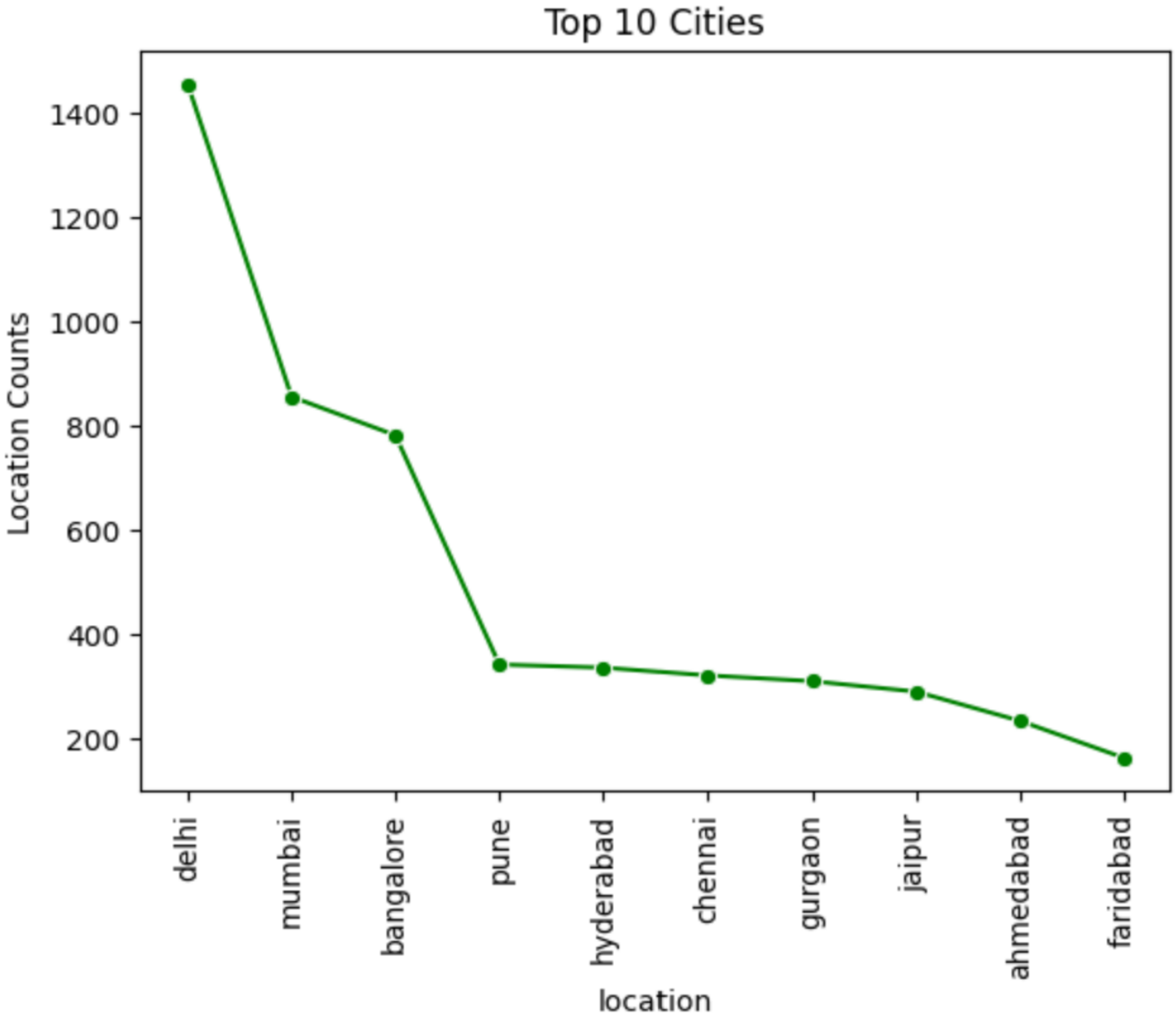
- **Forecasting Future Performance:** While not directly providing financial figures, the trend in model counts can be a leading indicator for revenue, production, or R&D investment.
- **Resource Allocation:** Understanding the peak and trough periods can help in optimizing resource allocation (e.g., manufacturing capacity, marketing spend, R&D budget).
- **Risk Assessment:** A sharp decline, as seen towards the end of the graph, could signal potential market saturation, increased competition, or a shift in consumer preferences, all of which pose financial risks.
- **Strategic Planning:** The data can inform long-term strategic decisions regarding product development, market entry/exit, and investment in innovation.

Financial Ideas and Implications:

- **Revenue and Profitability Impact:** The direct correlation between "Models Counts" and financial performance needs to be established. If each model generates revenue, then the growth phase (2011-2017) likely saw increasing top-line revenue and potentially improved profitability due to scale. Conversely, the sharp decline from 2018 to 2020 would almost certainly imply significant revenue contraction and potentially declining profits, or even losses, if fixed costs remain high.
- **Investment Strategy:**
  - During Growth (2011-2017): This period would have been ideal for investment in R&D, production capacity, and market expansion to capitalize on growth. Companies might have sought equity financing or leveraged debt for expansion.
  - During Decline (2017-2020): This phase necessitates a re-evaluation of investment. Companies might need to pivot their R&D, focus on high-performing models, or consider divesting underperforming assets.
- **Market Share and Competitive Landscape:** The decline could indicate a loss of market share to competitors or a fundamental shift in the industry (e.g., emergence of new technologies making existing models obsolete)

- **Product Lifecycle Management:** This graph clearly demonstrates the importance of product lifecycle management. Businesses need to anticipate the decline phase and proactively introduce new products, innovate existing ones, or diversify to maintain growth. The sharp drop suggests a potential failure to effectively manage this transition or a significant market disruption.
- **Risk Management:** The steep decline highlights significant business risk. This could be operational risk (e.g., supply chain issues, production problems), market risk (e.g., changing consumer tastes, economic downturn), or strategic risk (e.g., poor decision-making).
  - Scenario Planning: For financial forecasting, different scenarios would need to be modelled for the future:
  - Continued Decline: If the trend persists, the business might face severe financial distress.
  - Stabilization: Efforts to stabilize model counts at a lower, but sustainable, level.
  - Rebound: Introduction of new strategies or products leading to a renewed growth phase.

```
py=df.location.value_counts().iloc[:10]
sns.lineplot(y=py.values,x=py.index, marker='o' ,color='g')
plt.ylabel('Location Counts')
plt.title('Top 10 Cities')
plt.xticks(rotation=90)
plt.show()
```



The graph is a line plot showing the distribution of "Location Counts" across the top 10 cities. The cities are listed on the x-axis, and their corresponding counts are on the y-axis. The line connects the data points for each city, illustrating a clear trend.

Overall Performance/Analysis:

Dominance of Delhi: Delhi is the clear leader with a "Location Count" of over 1400, significantly higher than any other city. This suggests Delhi has a vastly larger presence or more activity related to whatever "Location Counts" signify.

Strong Second Tier (Mumbai and Bangalore): Mumbai and Bangalore form a strong second tier, with counts around 850 and 780 respectively. While considerably lower than Delhi, they still show substantial activity.

Sharp Drop-off after Bangalore: There's a very steep decline in "Location Counts" after Bangalore. Pune, Hyderabad, Chennai, and Gurgaon all fall into a similar range, between approximately 300 and 350. This indicates a significant drop in scale compared to the top three cities.

Gradual Decline among Middle-Tier Cities: Within this middle tier (Pune to Gurgaon), the decline is more gradual.

Lower Counts for Jaipur, Ahmedabad, and Faridabad: Jaipur, Ahmedabad, and Faridabad show even lower counts, with Faridabad having the lowest among the top 10, just under 200.

Concentration at the Top: The graph clearly demonstrates that the "Location Counts" are highly concentrated in the top few cities, particularly Delhi. The tail of the distribution shows much lower activity.

Important Objectives of this Graph:

- **Identify High-Density Locations:** The primary objective is to visually highlight which cities have the highest "Location Counts." This could represent anything from the number of businesses, residential properties, job opportunities, or even specific types of assets in those cities.
- **Rank Cities by Count:** The graph clearly ranks the top 10 cities based on their respective counts, making it easy to see the leading cities and their relative positions.
- **Show Distribution and Concentration:** It illustrates how the "counts" are distributed across the top cities, indicating where concentration is highest and where it tapers off.
- **Support Strategic Decision-Making:** For businesses or investors, this data can inform decisions related to market entry, resource allocation, expansion, or even where to focus marketing efforts.

Financial Ideas based on the Graph:

- **Investment Hotspots:**
  - Tier 1 (Delhi, Mumbai, Bangalore): These cities represent established, high-activity markets. Financial institutions could focus on large-scale investments, corporate lending, and high-net-worth individual services here. Real estate investments in these cities are likely to be premium but stable.
  - Tier 2 (Pune, Hyderabad, Chennai, Gurgaon): These cities show a moderate level of activity. They might offer growth opportunities, potentially with lower entry costs than Tier 1. Investments could focus on emerging sectors, mid-sized businesses, and residential real estate with good appreciation potential.
- **Market Expansion Strategies:**
  - Aggressive Expansion: Companies looking to maximize market share would prioritize Delhi, Mumbai, and Bangalore for aggressive expansion (e.g., opening new branches, launching major marketing campaigns).
  - Strategic Niche Development: In cities like Pune, Hyderabad, and Chennai, financial services could focus on specific niches or segments where competition might be less intense than in the top three.
  - Emerging Market Exploration: Cities like Jaipur, Ahmedabad, and Faridabad, with lower counts, could be considered for future, long-term expansion, potentially testing new products or services with lower initial investment.
- **Risk Assessment:**
  - Diversification: For financial portfolios, this graph suggests diversifying investments across different tiers of cities rather than concentrating solely on the top few, to mitigate localized risks.
  - Market Saturation: The high counts in Delhi, Mumbai, and Bangalore might indicate higher market saturation and competition, potentially leading to lower profit margins for some businesses.
  - Growth Potential: Cities with lower current counts but strong economic fundamentals might offer higher growth potential for future investments.
- **Targeted Financial Products/Services:**

- Delhi, Mumbai, Bangalore: Products catering to large corporations, high-income individuals, and mature industries (e.g., corporate finance, wealth management, commercial real estate loans).
- Pune, Hyderabad, Chennai, Gurgaon: Products for IT/tech sectors, start-ups, and a growing middle class (e.g., venture capital, personal loans, home mortgages).
- Jaipur, Ahmedabad, Faridabad: Products for local businesses, small and medium enterprises (SMEs), and potentially agricultural or manufacturing sectors depending on the city's primary industries.

- **Resource Allocation:**

- Talent Acquisition: Attracting top talent might be easier in cities with higher "Location Counts" if these counts are correlated with job opportunities.
- Marketing Spend: Advertising and marketing budgets should be allocated proportionally, with higher spending in Delhi, Mumbai, and Bangalore to reach larger potential customer bases.