**Hasan Khan ¬** [**hasan.k.khan5196@gmail.com**](mailto:hasan.k.khan5196@gmail.com) **¬ +44 07399211196**

**Data Analysis Insights Report Project:**

**Vehicles Sales Data [Large Dataset]**

Contents

[Abstract/Summary 2](#_Toc186290938)

[Section 1 – Exploratory Data Analysis 3](#_Toc186290939)

[1.1 – Dataset Details 3](#_Toc186290940)

[1.2 – Missing or Anomalous Values 3](#_Toc186290941)

[1.2.1 Via Excel 3](#_Toc186290942)

[1.2.2 Via SQL 4](#_Toc186290943)

[1.3 Distribution of Numerical Fields with Insights 6](#_Toc186290944)

[1.3.1 Via Excel 6](#_Toc186290945)

[1.3.2 Via SQL 8](#_Toc186290946)

[Section 2 – Pricing and Market Analysis 9](#_Toc186290947)

[2.1 -How does Selling Prices Compare to MMR Values 9](#_Toc186290948)

[2.1.2 – Via SQL 9](#_Toc186290949)

[2.1.1 – Via Excel 10](#_Toc186290950)

[2.2 – Which makes/models tend to have the highest resale value? 12](#_Toc186290951)

[2.2.1 – Via SQL 12](#_Toc186290952)

[2.3 - How does vehicle condition rating influence final selling price? 12](#_Toc186290953)

[2.3.2 – Via SQL 12](#_Toc186290954)

[2.3.1 – Via Excel 13](#_Toc186290955)

[Section 3 – Geographic and Regional Analysis 15](#_Toc186290956)

[3.1 - Which states have the highest volume of sales and the highest average selling price? 15](#_Toc186290957)

[3.1.2 – Via SQL 16](#_Toc186290958)

[3.1.1 – Via Excel – Number of Sales by State 16](#_Toc186290959)

[Section 4 – Vehicle Attributes and Customer Preferences 18](#_Toc186290960)

[4.1 - Which body types (e.g. SUV, Sedan or Truck) are most popular? 18](#_Toc186290961)

[4.2 - Transmission type vs. Price 19](#_Toc186290962)

[Section 5 – Interactive PowerBI Dashboard 20](#_Toc186290963)

# Abstract/Summary

**Dataset:**

Vehicle Sales Data

**About Dataset**

The "Vehicle Sales and Market Trends Dataset" provides a comprehensive collection of information pertaining to the sales transactions of various vehicles. This dataset encompasses details such as the year, make, model, trim, body type, transmission type, VIN (Vehicle Identification Number), state of registration, condition rating, odometer reading, exterior and interior colours, seller information, Manheim Market Report (MMR) values, selling prices, and sale dates.

**About Sections**

I will explore many different questions to provide actionable insights.

# Section 1 – Exploratory Data Analysis

## 1.1 – Dataset Details

The dataset size is 52.3 MB. Via Excel, this dataset consists of 16 columns and 558,838 rows. Each column shows:

**Year** – Manufacturing year of the vehicle.

**Make** – Vehicle brand.

**Model** – the specific model of the vehicle.

**Trim** – Vehicle special features and configurations (packages included).

**Body** – Body type (SUV or Sedan etc).

**Transmission** – automatic or manual (semi-automatic not included).

**Vin** – Vehicle Identification Number (the unique code for each vehicle).

**State** – the American state that it is registered too.

**Condition** – vehicle condition rated on a scale of 1-49 (49 being near perfect)

**Odometer** – The mileage (distance driven)

**Colour** – Exterior colour of the vehicle

**Interior** – Interior colour of the vehicle

**Seller** – The business selling the vehicle

Mmr – ‘Manheim Market Report’, estimated market value

**Selling Price** – Price at which the vehicle was sold

**Saledate** – Date and Time when the vehicle was sold

## 1.2 – Missing or Anomalous Values

### 1.2.1 Via Excel

**Eliminating and Viewing Null Values**

**Steps**: selected all data, selected data on the ribbon, filters, using the drop-down feature I unchecked all ‘(blanks)’ which represents rows with empty data. This eliminated all rows with empty data that are not useful to us. I then copied and pasted this no-null dataset onto another sheet.

**Additional**: It is possible to view only rows with empty data using filters and you can then fill in data or assess why its empty.

Some columns didn’t contain any null values, but the rows were still eliminated. This is to ensure that all the data I am processing for the future can have an extensive search for other analysis. The new sheet now has 472,327 rows.

**Pivot Table to Spot Anomalies**

**Steps**: Selected all the data, created a pivot table based on all the data, I then selected values I wanted the pivot table to showcase.

A screenshot of a computer

Description automatically generated

Using this pivot table I can now inspect or use other commands to help me assess whether there are anomalies or not. There also multiple other benefits of using a pivot table such as summarizing the data, drilling down the details and grouping/sorting the data.

**Other**: It is possible to use another tactic known as conditional formatting to highlight and visually represent the anomalies, so they stand out. This can be helpful when wanting to spot unusual trends or patterns in a dataset.

### 1.2.2 Via SQL

I have successfully converted the Excel file to Microsoft SQL (a more efficient way of converting a dataset over other options such as MySQL Workbench and PhpMyAdmin).

A screenshot of a computer

Description automatically generated

**Eliminating and Viewing Null Values**

A screenshot of a computer

Description automatically generated

I have successfully removed all rows where there are null values in any of the columns.

**Anomalous Values**

I have written a short code where I can see if there are any duplicate vin numbers (Vehicle Identification Numbers are meant to be unique to every vehicle manufactured)

A screenshot of a computer

Description automatically generated

As you can see there are over 6400 rows with duplicate vin numbers. I will now remove duplicates

A screenshot of a computer

Description automatically generated

The query is successful, this dataset has been optimized. There are now 459,347 rows.

## 1.3 Distribution of Numerical Fields with Insights

### 1.3.1 Via Excel

**Histogram – Odometer Sales Analysis**

Firstly, I filtered and got rid of all non-numeric values as there will be errors if there is non-numeric data. Using the add-in ‘data analysis’ I will now create the histogram.

A screenshot of a computer

Description automatically generated

**Output:**

A graph on a white background

Description automatically generated

As you can see, the data has been sorted into a histogram. For easy analysis, I can edit this and make it look much more readable:

A graph of a vehicle

Description automatically generated

**From a Data Analyst Viewpoint:**

This histogram reveals a concentration of odometer readings between 20,000 and 60,000 miles, with a sharp decline beyond 70,000 miles. Notable spikes at 30,000-40,000 miles suggest key ranges for vehicle sales and use. Outliers are above 200,000 miles so this may be cause for investigation on the dataset.

**Questions I might ask:**

1. Why are most vehicles clustered in this range?
2. Would you find it easier to sell a vehicle above 200,000 miles compared to above 140,000 miles?
3. Is 40,000 miles the optimal mileage?

**Answers:**

1. Likely due to resale patterns or leasing mileage limits.
2. These are outliers, it could be solely based on a very cheap price to customers with little or no knowledge of mileage
3. It could be the right price to mileage ratio where most customers are happy with spending ‘x’ amount of money for ‘y’ amount of mileage.

### 1.3.2 Via SQL

**Years – Oldest, Newest, Most Common and Average**

Using SQL, I will find sales data showing the oldest year, the newest year, the most common year and the average year that the vehicles have been sold.

A screenshot of a computer

Description automatically generated

**Data Analyst Insights**

I decided to merge all these questions into one rather than separately addressing each question. The oldest recorded sale year is 1990 while the newest is 2015. 2013 is the most frequent year of sales and on average, vehicles were sold around early 2010.

This distribution may show a relationship between vehicles produced and a market availably in that 2013 may have been a year of excess inventory or significant market shifts, such as lease returns or increased production of popular models. This leads to a decreased price in these models and thus, customers get more value for money with these models.

# Section 2 – Pricing and Market Analysis

## 2.1 -How does Selling Prices Compare to MMR Values

### 2.1.2 – Via SQL

**Are vehicles typically sold above or below mmr estimates?**

A screenshot of a computer

Description automatically generated

As you can see, I can view all rows and I have created 2 new tables showing the price difference and whether it is below, above or equal to MMR. For small datasets, it is common to use this method to visually inspect what vehicles make the most and least profit.

However, this is a large dataset, and I still want to find what vehicle is the most profitable and what is the least profitable:

A screenshot of a computer

Description automatically generated

The most profitable is the GMC Savana LS 1500 with an average profit of $30,475 across all sales. There could be a plethora of reasons for why this is the most profitable such as high demand with limited supply, underestimated MMR values, the trim being favoured across all models or just simply sellers being aware of its popularity and increasing the price.

A screenshot of a computer

Description automatically generated

The least profitable is the Bentley Continental SuperSport’s with the base trim and it yields an average loss of $14,000. This loss does not outweigh the highest profitable vehicle but there could still be a lot of reasons such as a decline in popularity, low demand and high supply or high depreciation.

### 2.1.1 – Via Excel

I am comparing the sum of selling prices, MMR and the sum of the price difference between each body type. I am using a pivot table:

A screenshot of a computer

Description automatically generated

I will now create a pivot chart using this:

A graph with different colored lines

Description automatically generated

Based on the chart, here are some analytical insights:

Body Types with High Positive Price Differences: Certain body types, such as GranTurismo Convertible and Coupe, display significant positive price differences, indicating they are sold well above their MMR values. This suggests strong demand or niche appeal for these vehicles, possibly due to their luxury or performance-oriented nature.

Body Types with Minimal or Negative Price Differences: Body types like Regular Cab and Van appear to have little to no price difference, or even negative differences, suggesting that these vehicles are often sold close to or below their MMR values. This could indicate a saturated market or limited demand for these practical, utilitarian vehicles.

SUV and Sedan Stability: SUVs and Sedans likely have moderate average price differences, reflecting their large market share and consistent demand. These segments tend to show balanced performance between selling price and MMR, making them reliable for volume-driven markets.

High Selling Price Correlation with High MMR: Body types with high average selling prices, such as GranTurismo Convertible, also show high MMR values. This indicates a luxury or premium segment where market estimates align closely with actual sales, suggesting predictability in pricing.

Significant Outliers: A few body types exhibit drastic variations in price differences (positive or negative), which may require further investigation to determine specific market conditions, inventory issues, or geographic demand differences.

Market Dynamics: The variability in price differences across body types highlights how consumer preferences, market demand, and the perceived value of certain vehicles influence their performance against MMR values. Vehicles with niche or luxury appeal often outperform market expectations, while those serving practical purposes tend to meet or fall below estimates.

## 2.2 – Which makes/models tend to have the highest resale value?

### 2.2.1 – Via SQL

I have written some code to give me a list of the top5 highest resale value vehicles. I have included it along with a ratio to roughly show how much on average it is sold ‘x’ amount above the MMR value.

A screenshot of a computer

Description automatically generated

The Chevrolet Silverado 3500 leads with the highest resale value ratio of 3.47 which shows a strong demand for this heavy-duty truck. It is likely that it’s driven by in construction and agriculture sectors. The Suzuki Esteem and Chrysler LHS also have high resale value ratios, and this suggests limited supply and/or niche market demand for economical or luxury vehicles. Luxury brands like the Mercedes-Benz 400-Class maintain high resale values as buyers do have a willingness to pay a lot for brand reputation and features. Lastly, the Isuzu Amigo’s performance portrays the appeal of compact SUVs to the consumer market where it is most likely popular.

## 2.3 - How does vehicle condition rating influence final selling price?

### 2.3.2 – Via SQL

The code below demonstrates the average selling price of vehicles in terms of condition rating. 49 is the highest condition rating in this dataset. Also, I have included total vehicles sold as well as the max and min selling price.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

As you can see, the total number of vehicles sold as well as the selling price increases as condition rating improves. This is because customers do not mind paying more money knowing that the condition of the car is reliable and good – less likely to have faults or break down. Additionally, there is a greater range in lower condition ratings of minimum and maximum price which shows that buyers can negotiate prices more based on the flaws and repair costs.

### 2.3.1 – Via Excel

I will firstly create a pivot table to show me the total amount sold alongside the condition rating. This will show me what condition rating has the best value to price ratio.

A screenshot of a computer

Description automatically generated

Now I will create a pivot chart to better showcase this information.

A graph showing a line

Description automatically generated

**Data Analysis:**

**Selling Price Increases with Condition Rating:**

There is a clear trend showing that higher condition ratings (e.g., 40 and above) are associated with significantly higher total selling prices, indicating that buyers are willing to pay a premium for vehicles in better condition.

**Outliers and Anomalies:**

Condition ratings like 4 and 11 stand out with unusually high total selling prices compared to nearby ratings. This could indicate specific vehicle models or types dominating these categories or errors in the data that need further inspection.

**Minimal Activity for Mid-Range Ratings:**

Condition ratings in the mid-range (e.g. 13–17) exhibit very low total selling prices, suggesting limited inventory or low demand for vehicles in this condition range. This could reflect a market preference for either higher-condition vehicles or lower-priced, poorer-condition vehicles.

**Peak Selling Price in Higher Ratings:**

The highest total selling prices are observed in condition ratings 43–46, reflecting a strong demand and market preference for nearly pristine vehicles. This is consistent with expectations that high-quality vehicles command higher market value.

**Opportunity for Low-Rated Vehicles:**

Low condition ratings (e.g., 1 and 2) still contribute to a notable portion of the total selling price. This suggests there is a niche market for these vehicles, likely driven by cost-conscious buyers or buyers looking for project or repair vehicles.

**Market Distribution:**

The data highlights that most of the total selling price value is concentrated in the higher condition ratings, indicating a skewed market toward vehicles in good to excellent condition. Businesses should focus on maintaining high-condition inventory to maximize revenue.

# Section 3 – Geographic and Regional Analysis

## 3.1 - Which states have the highest volume of sales and the highest average selling price?

### 3.1.2 – Via SQL

The code below shows the state with the most amount of sales along with the average selling price of that state:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Florida leads in total sales volume with 73,405 transactions, followed closely by California with 63,579, indicating these states are key markets for vehicle sales. Despite high sales, the average selling price in these states remains moderate, suggesting large-scale sales but not necessarily premium vehicles. On the other hand, Tennessee has a lower sales volume but a higher average selling price of $17,242, highlighting potential demand for higher-value vehicles in that state. The data suggests that focusing marketing efforts on Florida and California can maximize volume, while targeting premium buyers in Tennessee can optimize profitability.

### 3.1.1 – Via Excel – Number of Sales by State

Firstly, I create a pivot table to show each state and then the count of each sale per state.

A screenshot of a computer

Description automatically generated

I can now see an overlook of all states with the count of sales. There is a lot of visual data analysis I can currently do but I want to create a chart to better understand and showcase everything.

A graph with blue lines

Description automatically generated

**Key data analysis insights:**

High Sales Volume States: Florida (FL) and California (CA) dominate with 75,243 and 68,213 sales, respectively. These states represent key markets with high vehicle transaction volumes, likely due to their large populations and strong automotive markets.

Regional Market Leaders: Texas (TX) follows with 41,657 sales, highlighting its importance as a major automotive hub in the southern United States. Georgia (GA) also stands out with 30,939 sales, reflecting strong regional demand.

Low Activity States: States like Alabama (AL) with only 25 sales and New Mexico (NM) with 163 sales show very low market activity. These states may represent underdeveloped markets or lack of data coverage.

Mid-Tier States: States such as Illinois (IL, 21,494) and Ohio (OH, 20,294) exhibit moderate activity levels, suggesting steady but not dominant market presence.

Strategic Opportunities: States with lower sales but notable economic activity, like New York (NY, 11,682) and Washington (WA, 7,012), might indicate potential growth opportunities if the market is strategically targeted.

Overall Market Distribution: The data shows a significant skew toward a few high-volume states (e.g., FL, CA, TX), while most other states have much lower transaction counts. This distribution highlights the need for localized marketing strategies focusing on high-performing states while identifying gaps in lower-performing regions.

# Section 4 – Vehicle Attributes and Customer Preferences

## 4.1 - Which body types (e.g. SUV, Sedan or Truck) are most popular?

I have written code to show me all the body types along with the most popular types.

A screenshot of a computer

Description automatically generated

The Sedan body type dominates sales with over 206,000 transactions which indicates its widespread appeal and practicality for a broad range of buyers. SUVs, with 117,000 sales, are the second most popular, reflecting a growing trend for spacious and versatile vehicles. Niche categories like Hatchbacks and Minivans have a substantial presence but are behind sedans and SUVs, suggesting their appeal is targeted toward specific consumer groups. On the other hand, luxury and specialty vehicles like G Convertible and CTS Coupe have much lower sales, highlighting their limited target audience and specialized market demand.

## 4.2 - Transmission type vs. Price

For this, I created tables to show the difference between Automatic and Manual cars alongside the total sales, average price, the minimum and maximum price of the vehicles. This will show me the key differences between the two transmissions.

A screenshot of a computer

Description automatically generated

As you can see, the data shows that automatic transmissions dominate the market, accounting for 443,636 total sales, far surpassing manual transmissions at 15,711 sales. This suggests a strong customer preference for convenience and ease of use associated with automatic vehicles. This is likely driven by urbanization and heavy traffic scenarios where automatics offer greater comfort and ease of use. Despite the volume disparity, manual transmission vehicles display a higher variability in selling prices, with a maximum price of $105,000 compared to $230,000 for automatics, potentially reflecting their niche appeal in performance or collector markets. The average price for automatics, at $13,774, significantly exceeds the $11,249 for manual vehicles, indicating that automatics are not only more popular but also tend to be valued higher in the market. Interestingly, the minimum price for manual vehicles drops as low as $150, hinting at the possibility of older, less desirable models affecting the average. This data underscores the dominance of automatic transmissions in both volume and value, while manual transmissions maintain relevance in specialized or enthusiast-driven segments.

Interestingly, there is an anomaly of the cheapest automatic car going for $1. This anomaly effects the other values as it is somewhat unreasonable and likely an accounting error.

# Section 5 – Interactive PowerBI Dashboard

**Selling Price, Transmission, Odometer and Condition Averages Across All States**

With Power BI, I have imported the data, and I will create an interactive dashboard showcasing the difference between each American States in terms of automatic and manual cars alongside the averages of selling price and odometer. Furthermore, I want to be able to show average condition of each state too and have a map that you can interact with to show specifics with each state.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

As you can see, the metrics change, and it is fully interactive. Analytical insights:

**Transmission Analysis**

The comparison of **selling price** and **odometer readings** between automatic and manual transmissions reveals key trends:

* Automatic transmissions tend to have higher average selling prices and odometer readings, indicating they might be associated with higher-value vehicles or higher usage rates.
* Manual transmissions, while less expensive on average, might be targeted toward specific buyer segments or represent older vehicles.

Actionable Insight: Customers seeking budget-friendly options may prefer manual transmission vehicles, while those wanting convenience or luxury lean toward automatics.

**Geographic Analysis**

The map visualization of MMR averages by state indicates significant variation in vehicle market values across regions:

* States like Texas show notably higher average MMR values compared to others, possibly due to higher demand, market size, or preference for specific vehicle types (e.g., trucks and SUVs).
* Other regions exhibit lower MMR values, potentially reflecting smaller markets or less demand.

Actionable Insight: For vehicle sellers, targeting states with higher average MMR values could yield higher profits. Buyers can look for opportunities in states with lower MMR values.

**Condition vs. Price Correlation**

Insight: The average condition rating (from 1-49) suggests the data includes vehicles in varying states of wear and tear. This could influence selling price trends:

* States with higher condition scores likely correlate with higher MMR or selling prices.
* Lower condition scores may align with higher odometer readings and reduced market values.

Actionable Insight: Highlight the importance of vehicle condition in pricing, emphasizing the need for sellers to improve condition (detailing, minor repairs) before sale to achieve higher prices.

**Selling Price Trends**

Insight: Selling prices show significant fluctuation between dashboards:

* In some dashboards, the average selling price is 6.48K, while in others, it rises to 13.57K or 13.69K. This could reflect differences in dataset filters (e.g. specific regions, transmission types or vehicle condition).

Actionable Insight: Showcase that sellers and buyers should analyse regional or vehicle-specific trends to optimize pricing and investment decisions.

**Odometer vs. Selling Price**

Insight: Vehicles with higher odometer readings tend to have lower selling prices. However, the difference is less significant for specific transmission types (e.g. automatic vehicles consistently show higher averages even at higher mileage).

Actionable Insight: Buyers should consider odometer readings in their cost-benefit analysis, while sellers might focus on maintenance records to justify higher prices despite higher mileage.