**Comparative Trends in Used Car Market**

Gozde Karamanoglu

Maham Rasheda

Sumaiya Gulshan

CIS 3120 MUA

**Abstract**

The present study compares the different car models in the used car market, specifically in the US. As a group, we will write a thorough paper with supporting evidence demonstrating the most common color, popular model, odometer readings, price ranges, and condition of a used car. Data information will be presented and cleaned from a Craigslist dataset. This paper will also go into great detail about our procedure for producing and cleaning our data.

***Keywords:***  color, model, value, COVID, condition, comparative, trend

**Review of Literature:**

The following paper includes different sources and data. Each source was thoroughly analyzed to see what best fit the paper. Some of the keywords we picked up from the various literature are: used, data, odometer, conditions, buyers.

**Intro:**

Since the late 1800s, used cars have been the preference of many Americans due to the reliable prices. Most often used cars are brought into outlets by independent car sellers or dealerships. Since there is a plethora of used cars being listed there can be confusion among buyers. Buyers can be overwhelmed in deciding what car is worth buying. As a result, this study is being carried out in order to properly analyze used data in the hopes of uncovering a comparative pattern that will show what used car colors, model production year, odometer readings, price ranges, and quantities of "good" and "bad" conditions were most popular over the years.

**General Background**

When sellers/dealers sell cars, they look at the car market at the current time. After thoroughly analyzing the market, sellers should understand which models are popular and when to sell them. For a used car, the seller must tell you about the car's condition, history, fuel type, number of miles driven, and model. When examining the mileage of a used car, a buyer will typically inquire with the seller as to how many miles it has been driven. This can be determined from the car's odometer. The odometer is a component of the car that informs the owner of the total distance that the vehicle has ever covered. It can be measured in either kilometers or miles [1]. A used car shouldn’t read zero on the odometer, or else it would be listed as a new car. Most of the time, a used car has an odometer reading of over 200 miles. It's been shown that vehicles with less mileage will even sell for higher prices. When selling a car, the condition is important. The used car must be in good condition for consumers to buy it or believe the cost is reasonable. Additionally, the model of used cars should be displayed in the event that customers have a preference. The model of the car can determine which type of drive transmission the car has. There are three types of car transmissions; four-wheel drive, forward wheel drive, and rear-wheel drive. With all of these key factors in mind, there is a benefit to buying a used car.

**Reasons to buy a used car:**

As previously mentioned, a growing number of Americans prefer used cars to brand-new ones. Buying a used car may be a better investment, especially for people who have never owned a car before. There are several reasons why used cars are better to buy, such as they are cost efficient, have less depreciation, less insurance rates, and there's more availability in different car models. Cars have become a growing need for consumers. And consumers will buy something they need without trying to hurt their pocket, that's where used cars come in as the market offers used cars for values of $3,000-$20,000.Although, some are cautious of buying a used car currently because the prices at the moment have increased. According to J.P. Morgan Research, prices will drop by roughly 2.5% to 5% for new cars and by 10% to 20% for used cars in 2023[2]. Also, used cars don’t lose their value as fast as a new car does. It's exaggerated to claim a new automobile loses one-third of its value the moment you drive it off the lot, but the sharpest decrease comes in the first few years of ownership. Buying a secondhand automobile avoids the large depreciation [3] . Consumers on a monthly basis are also saving money since car insurance for a used car costs less than a new car. Furthermore, used cars are also more available in the car markets, as manufacturers are having difficulty producing new cars. Used-car sales outnumber new-car sales three to one in the US. The coefficient of variation of used-car pricing is almost five times that of new-car costs, demonstrating that secondary markets widen the range of goods available to consumers. [4]. Thus, it is further demonstrated that a used car is a better investment.

**Methodology:**

For this project, we've decided to use the Kaggle dataset that scrapes Craigslist for car information from 2021. We first gathered the dataset, and then we downloaded the dataset, which was a csv file but we viewed it as an excel file. Then we uploaded it as a CSV file into a Jupyter notebook. Followed by importing the library's Seaborn, Pandas, and Numpys to clean, analyze, and visualize the imported dataset.

**Technical Errors:**

One of our first steps was to attempt to web scrape the data of used cars from Craigslist.com. After much inspection, as a team we couldn’t web scrape craigslist directly because they blocked the developers/bots/third parties from extracting their data. This is due to a 31 million dollar settlement with craigslist and Instamotor. Craigslist stated that the company breached terms of use by web scraping user information for their gain [5]. The website blocked dev tools under the network tab, and they also changed their Html elements and base URL. This change makes it harder for us to extract. Also, since we have a large dataset and tried to clean it up, we had trouble figuring out whether the code was error-free because the Jupitner notebook wouldn't produce the output as efficiently. Furthermore, our technological devices crashed a few times due to the data being really big.

**Database Implementation Process:**

To begin, we wanted to become acquainted with the first five rows of data. This will allow us to see how the data looks after we import it into a Jupyter notebook. We discovered 26 columns in total, including: id, Url, region, price, model, manufacturer, number of cylinders, size, type, color, imageurl, country, state, latitude, longitude, and hosting date. As soon as we had finished our final observations, we could finish Part 1 of the project, which involved importing the vehicles.csv data into the Jupyter notebook. As previously stated, because this was a large dataset, we couldn't see all of the columns in the Jupitner notebook. As a result, we double-checked the data on Excel throughout the process.

Part 2 of our project was motivated by our agreement that the data needed to be cleaned before we could better understand it. During the data cleaning phase, we attempted to remove some columns that we were not using, such as Id, region, region url, and county. In addition, we limited the display options to a maximum of 500 columns in python. We observed on Excel that many columns lacked data, so we removed rows with 19-22 columns or more that were empty. After checking the dataset information for null columns, we looked at each useful column that had a high number of null values. The region and price columns were the only ones among the 26 that lacked null values. To analyze relevant data, we filtered each column to determine its composition. The VIN number, image URL, latitude, longitude, and posting date columns were eliminated after analysis.

In order to clean explanatory variables and deal with missing data, we moved on to part 3 of our project. There were 17 columns and 426812 rows, according to our examination of the dataset's shape. Once the unnecessary columns had been removed, we examined the remaining columns. Given the size of the dataset, we chose to discard the values that are not numbers (NaN) rather than replacing them. Our dataset had 79195 rows after removing these. We checked the dataset to make sure that the values had really been removed. This showed that 22 columns and 34585 rows had been cleaned.

We proceeded with the data aggregation portion of our project after observing that there were no longer any empty rows or missing values. By looking for anomalies in the odometer readings that a used car should have, we were able to accomplish this. First we checked the minimum value of the odometer, which came out to be zero. Usually any car under 200 miles is considered brand new which is why we don’t want any zero values. We checked the maximum odometer reading. If an odometer reaches 999,999, it will either remain at that value or go to zeros only. The datasets that were larger than 999,999 were eventually removed. Returning to our minimum value, we wanted to see how many rows had zero as their minimum value. It turned out that there were 121 rows. Our total row count decreased to 78260 from 79045 after removing the rows with values below 200. Considering that we are analyzing Craigslist, we wanted to make sure that no one was intentionally entering random numbers. We were looking for numbers like (1,111),(11,111),(111,111), (1,234),(12,345), and (123,456) as odometer readings. The numbers were taken out after we made sure they were there. Moving past the odometer readings, we chose to examine the years to ensure that we didn't have any float values. In the end, we got zero float values.

Then we went on to fix the price column. Zeros for the prices were something we wanted to avoid. Checking the columns with zero prices yielded 3,277 rows of free products, which made no sense. After removing those rows, we were left with a total of 74814 rows. We wanted to go with the range of minimum of $5000-$500,000 maximum of car price value. Then we wanted to check the minimum price for cars, there were cars listed for $1. We decided to clean the data because the maximum price for cars was over $3 billion. After doing so, the new maximum price was revealed to be $195,000. Once the unusual data had been removed, we chose to examine the different makes and models of cars that had been included in the dataset. Following that, the model year was used to determine the car's condition. Any vehicle manufactured prior to 2000 was classified as "fair," with the remainder classified as "good."

Finally, we could start the data visualization part of our project, which required displaying charts of comparison data. We created Boxplot to check for odometer outliers . To find the most popular models, we made a manufacturer bar graph. We visualized the top-selling automobile brand. Another box plot was made to determine if the most recent years are 2000–2020.

**Data implementation discussion and results:**

According to our research, sedans are the most popular kind of vehicle. SUVs were the second most popular vehicle, followed by trucks. Based on our data, the most popular car color was white, followed by black, and then silver. In addition, cleaning our data revealed that 4687 of the vehicles were in fair condition and 53063 were in good condition. According to our charts, the most popular car manufacturers were Ford, Chevrolet, Toyota, and Honda. And most models were found within the years 2000–2020.

Taking a closer look at the car model data chart we can conclude that ford sold the most cars between the years of 2018-2019. However, Chevrolet had a decline in 2019. The most prevalent pattern we discovered was that all of the manufacturers experienced a comparable decline between 2018 and 2011. In terms of popular car manufacturers, we discovered that Nissan had the most listings between $16500 and $17000. Which was then followed by Ford,Chevrolet,Toyota, then Honda.

Our box plot of the odometer showed that the average was 108,660, the lowest was 205, and the highest was 999,999. The data points are 61649 units away from the mean, which is what the standard deviation shows. $68,000 is less than 25%, $105,000 is less than 50%, and $143,000 is less than 75%.

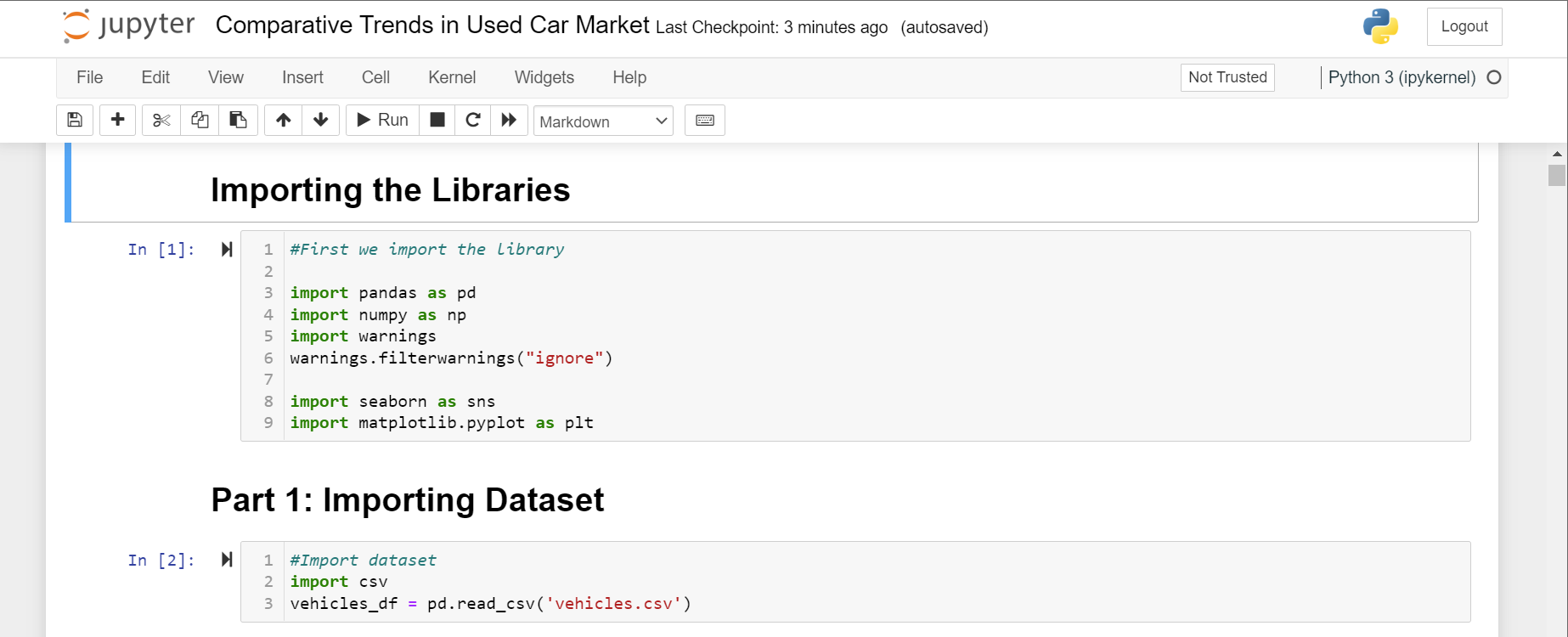
The box plot's outliers are that only three cars are listed for more than $150,000. And more cars were listed below $125,000. Furthermore, the box plot of prices showed that the average price was $16849, the lowest price was $5000, and the highest price was $195,000. There were 12411 standard deviations from the mean among the data points. 25% of the vehicles are priced under $7999, 50% are priced under $12950, and 75% are priced under $21598.

The most produced cars were in 2009, with the standard deviation being 10. The minimum year of production was 1900, and the maximum year of production was 2021. 25% of the cars were below 2007, 50% below 2012, and 75% were below 2015. With a count of 57750, the Ford F-150 was the most popular car model on the list. Ford F-Series trucks have been the most popular selling vehicle in America for the past 38 years [6].

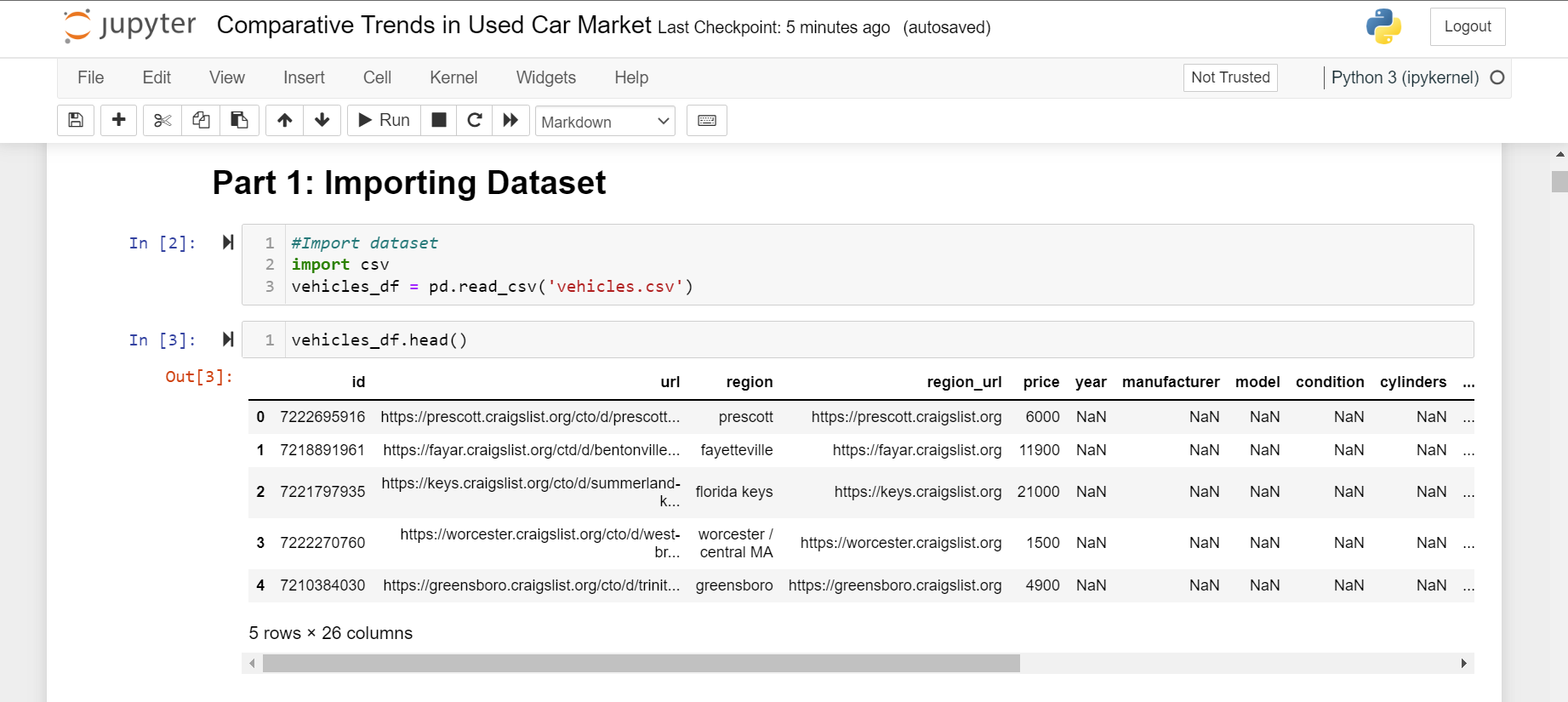
**Future Work**

For future work, we will be doing more regression and predictive analysis and using data mining to check the accuracy of the results since our data had so many inaccuracies.

**Appendix:**

****

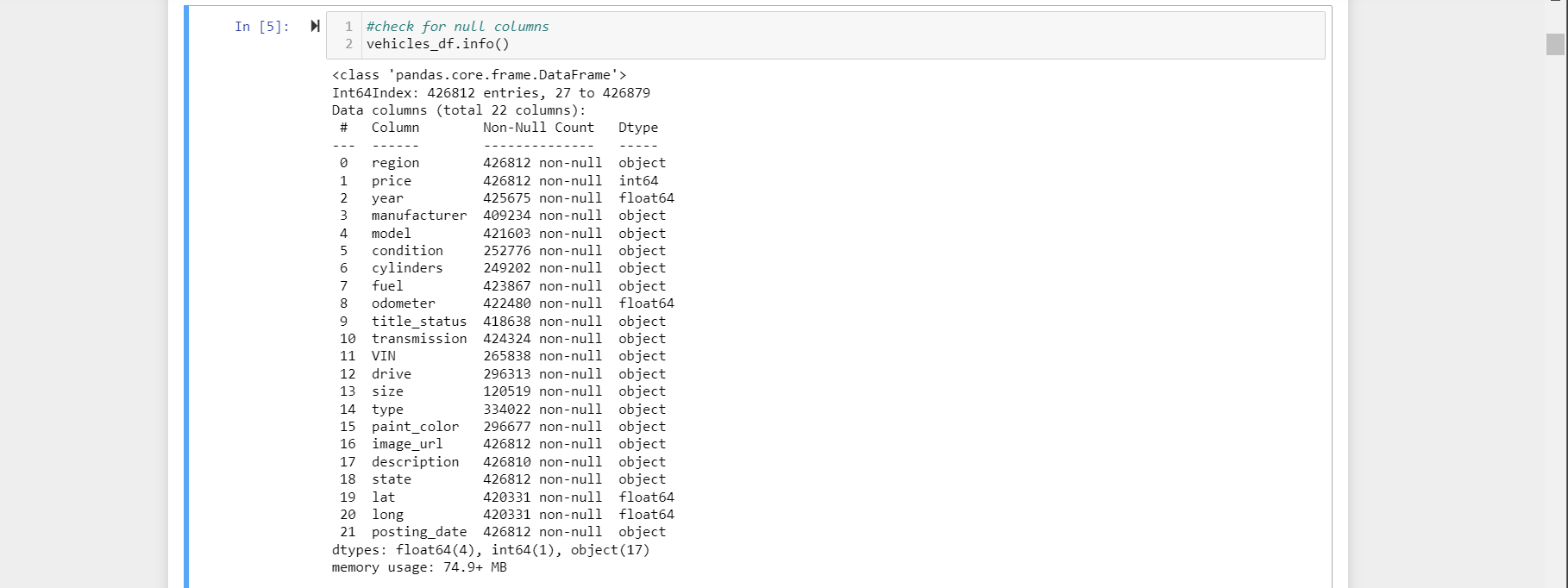
**Figure 1: Imported the libraries**

****

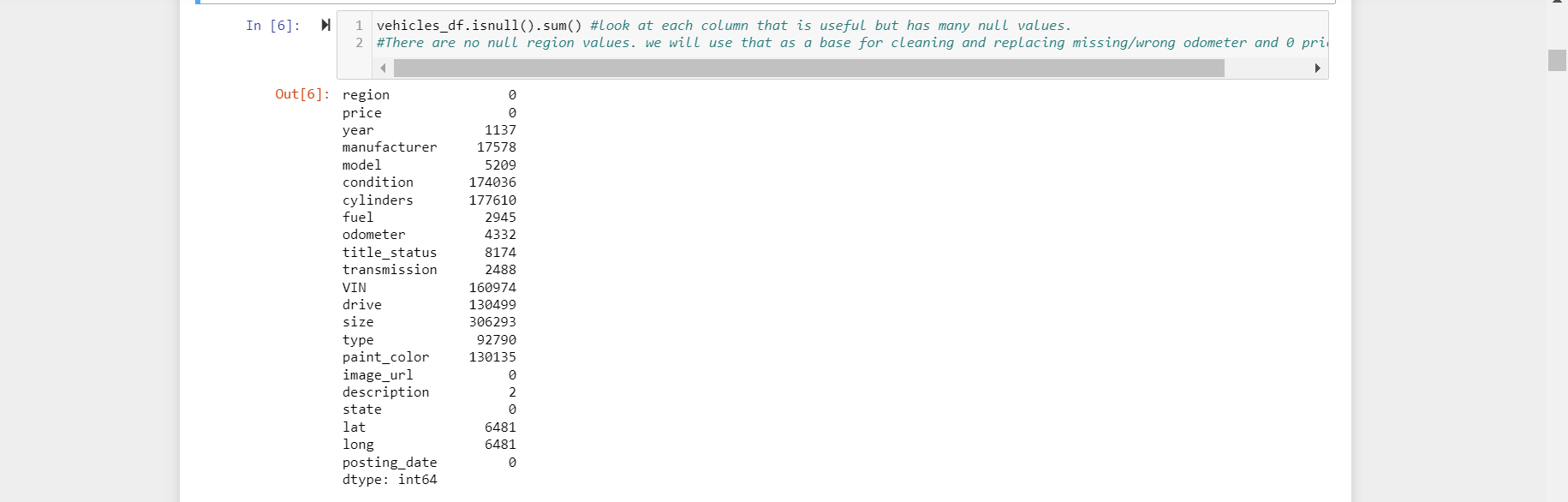
**Figure 2: Opened the vehicles csv file into Jupyter**

****

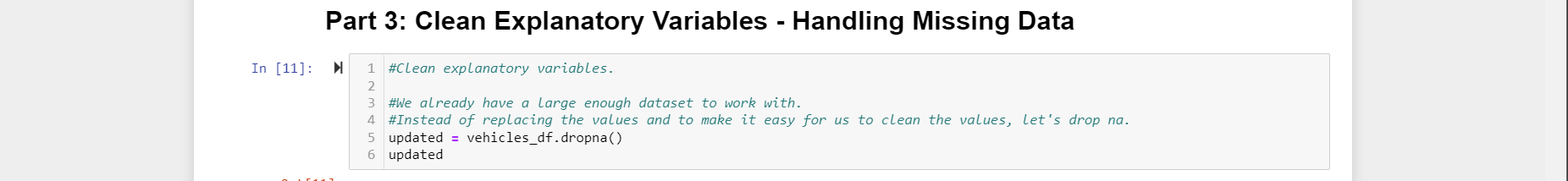
**Figure 3: Removal of irrelevant columns**

****

**Figure 4: Checking for Null Columns**

****

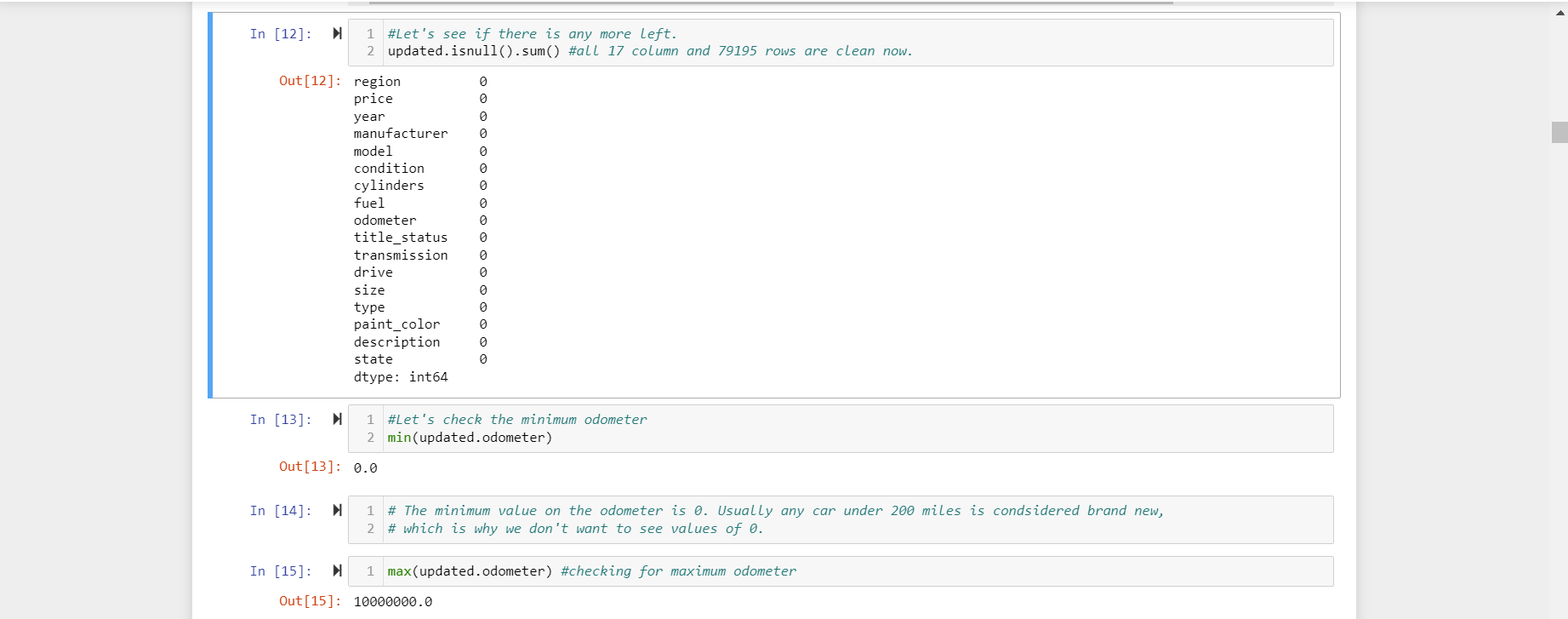
**Figure 5: Checking Columns for Null values**

****

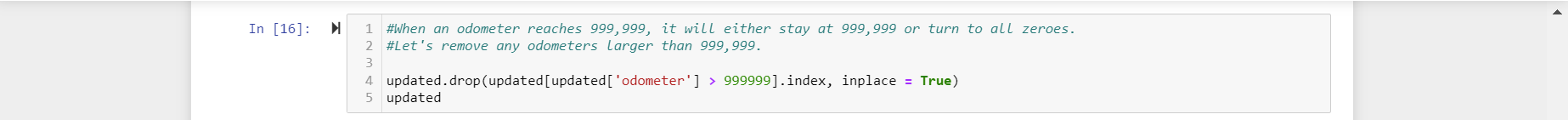
**Figure 6(a): Cleaning missing data**

****

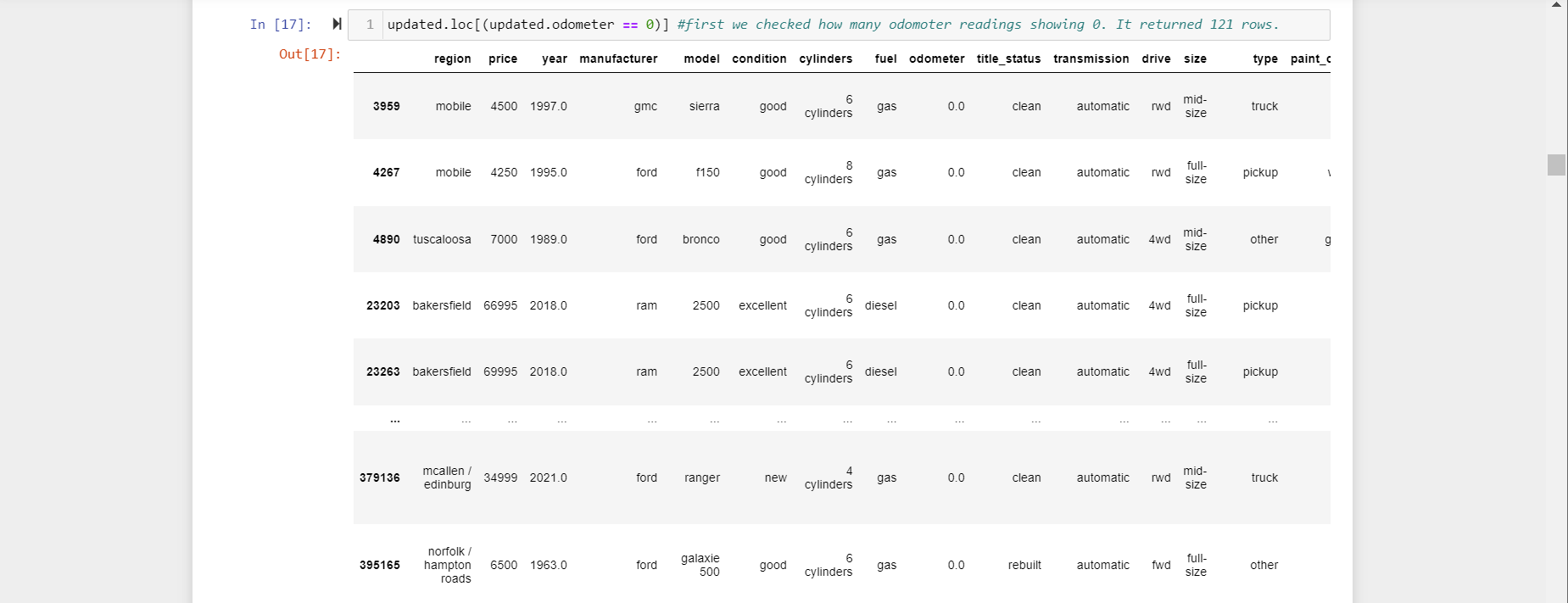
**Figure Figure 6(b): Cleaning Missing Data Results**

****

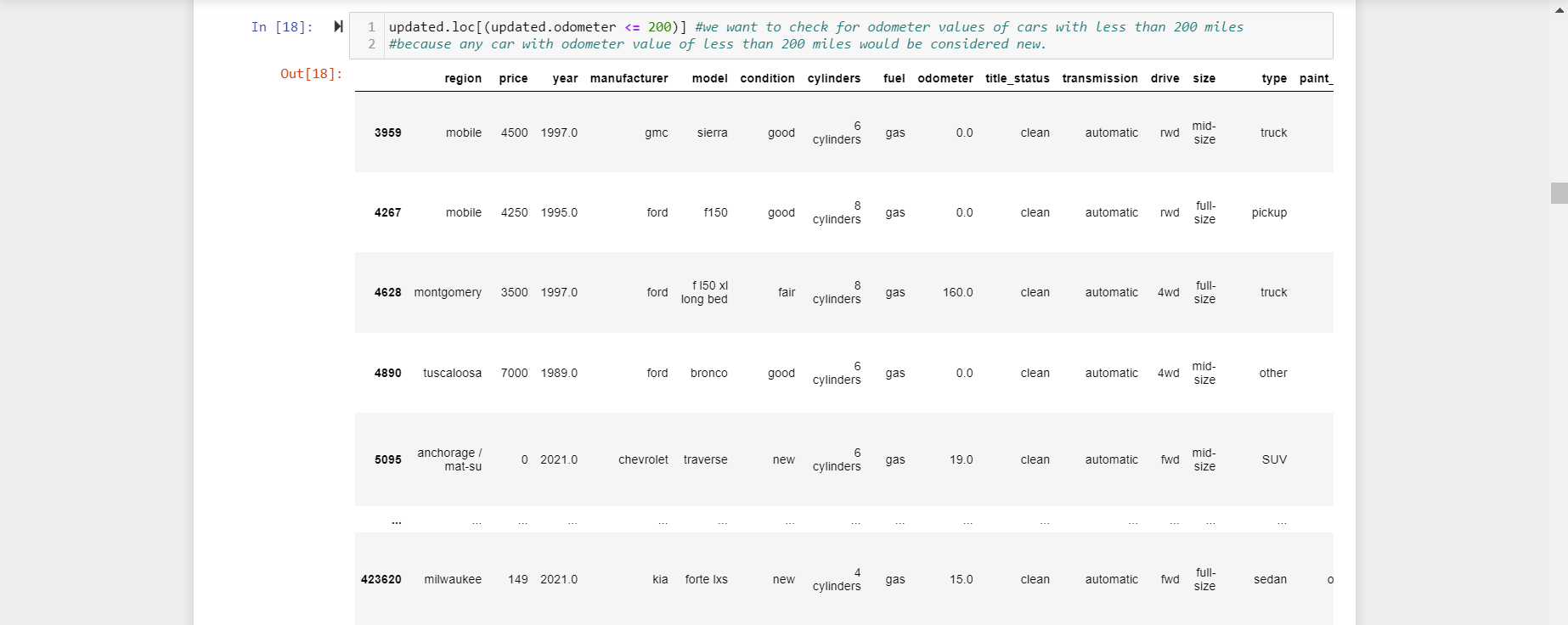
**Figure 7: Minimum and Maximum Odometer Readings**

****

**Figure 8: Cleaning Odometer Readings**

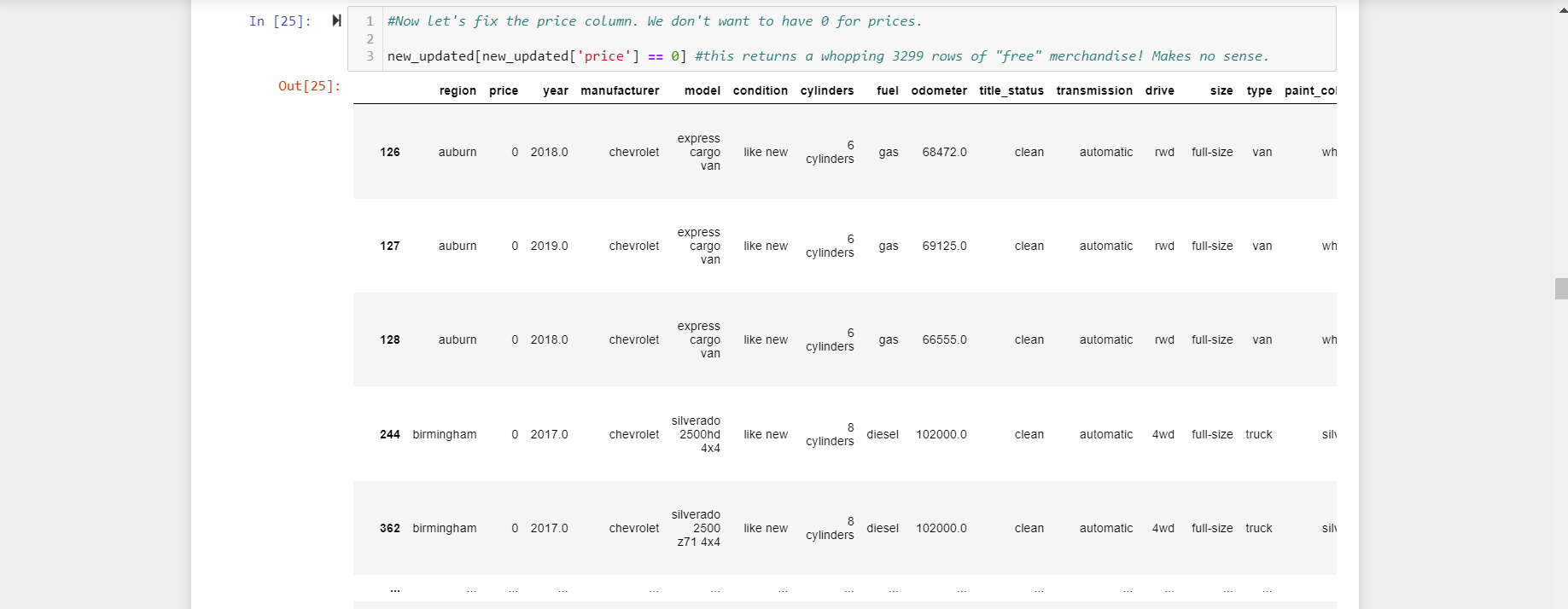
****

**Figure 9: Checking for Odometer Readings of 0**

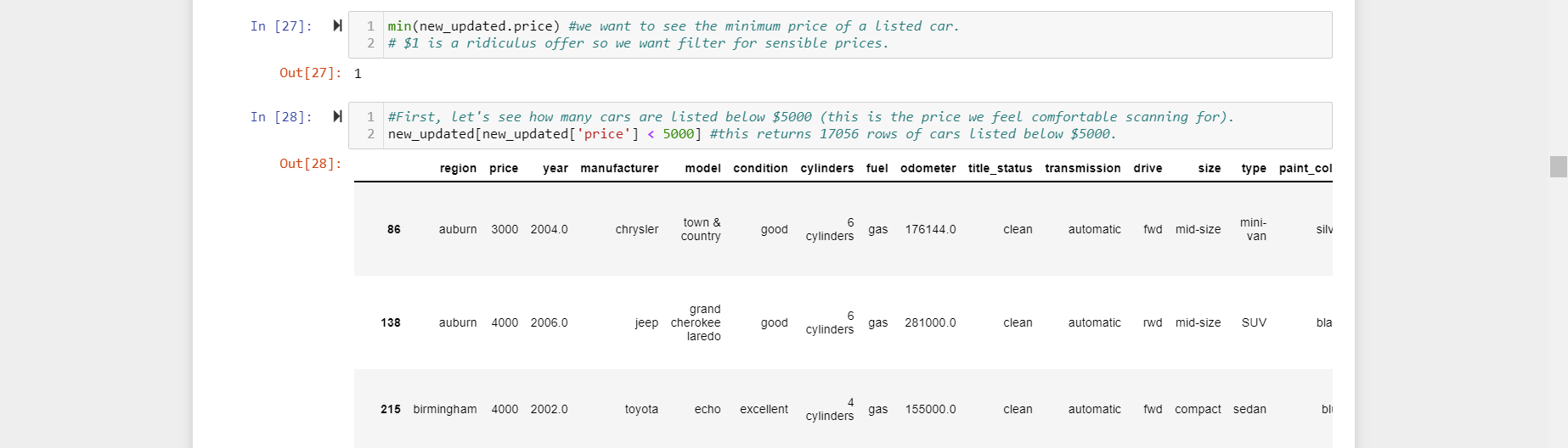
****

**Figure 10: Checking for Odometer Values less than 200 miles**

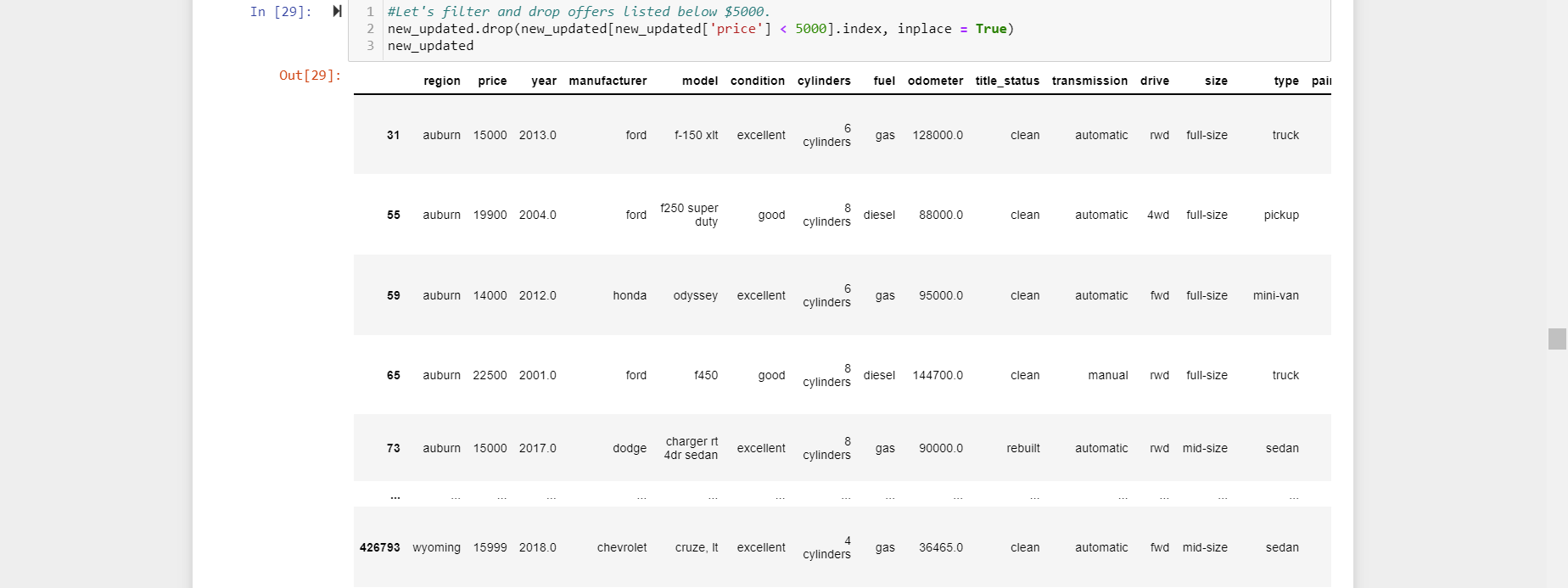
****

**Figure 11: Dropping Odometer Values less than 200 miles  
  
**

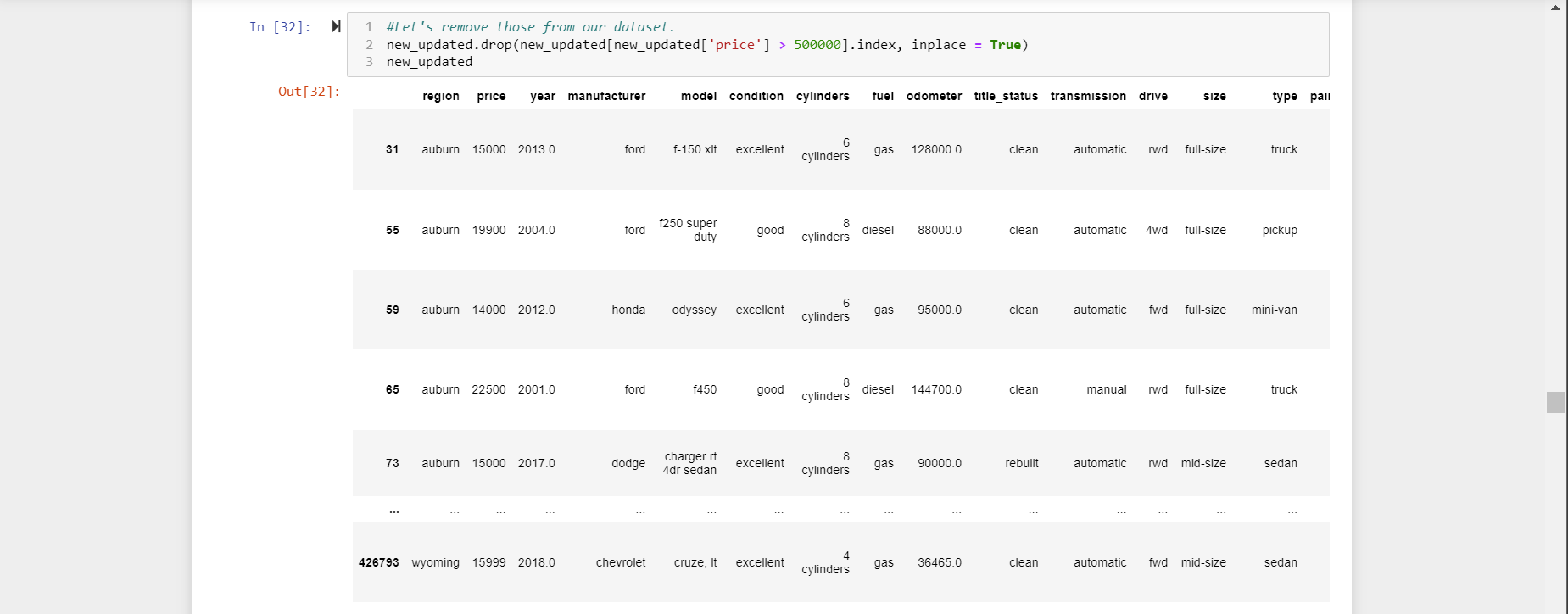
**Figure 12: Removing Price values of 0**

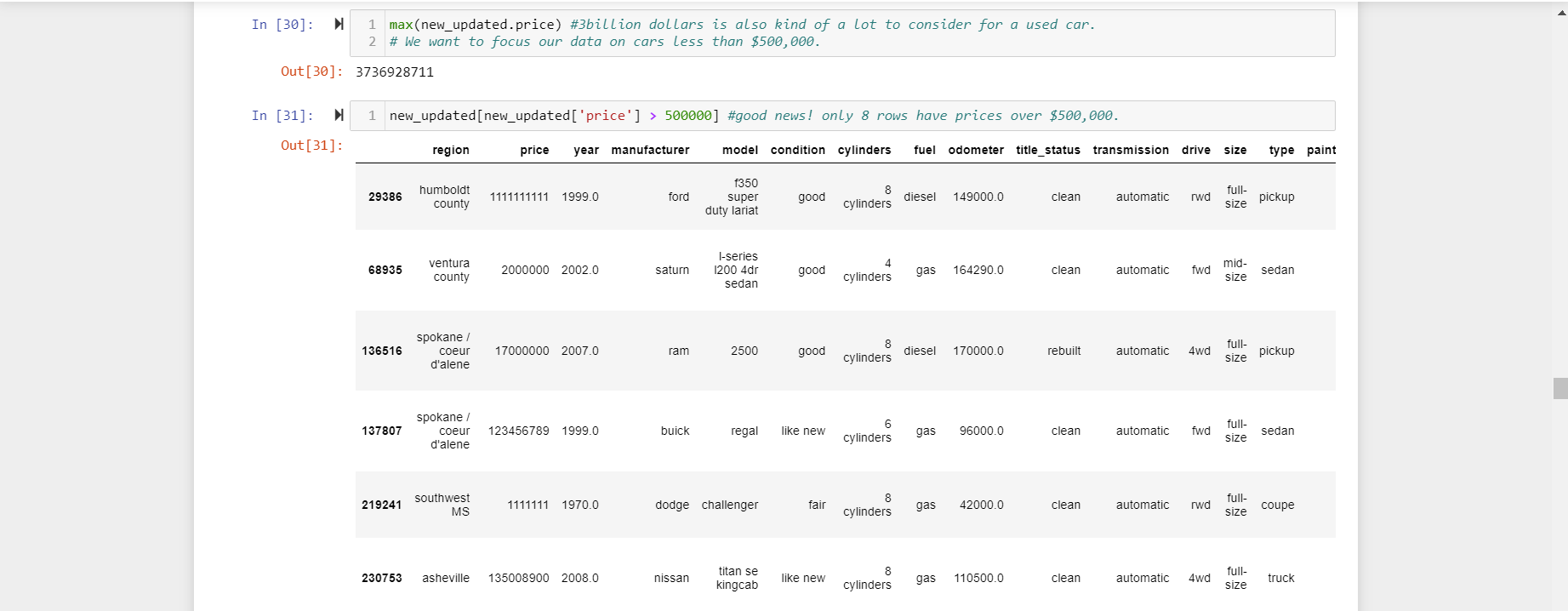
****

**Figure 13: Checking for Minimum Price and Prices less than $5000**



**Figure 14: Dropping Irrelevant Price Values**

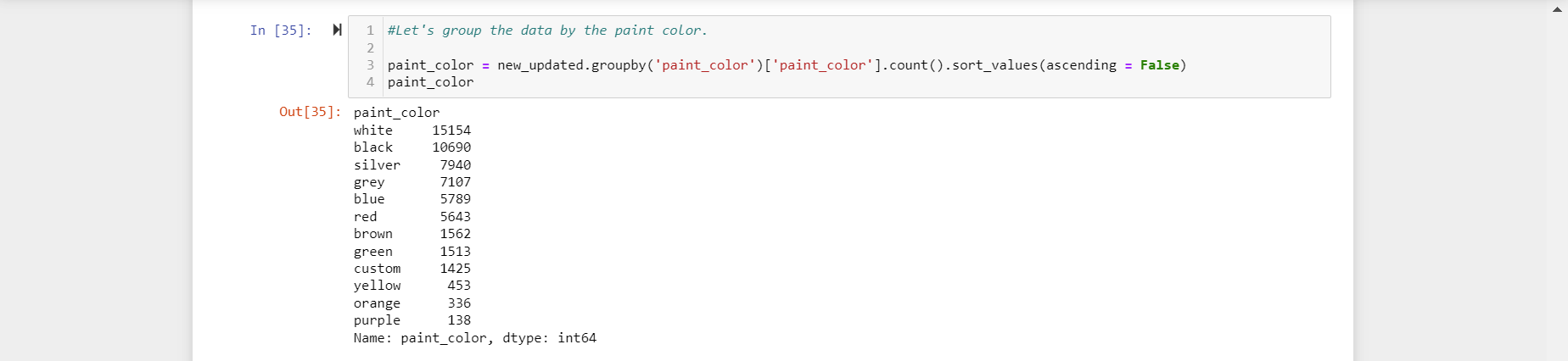


**Figure 15:Checking for Maximum Price and Prices more than $500,000**

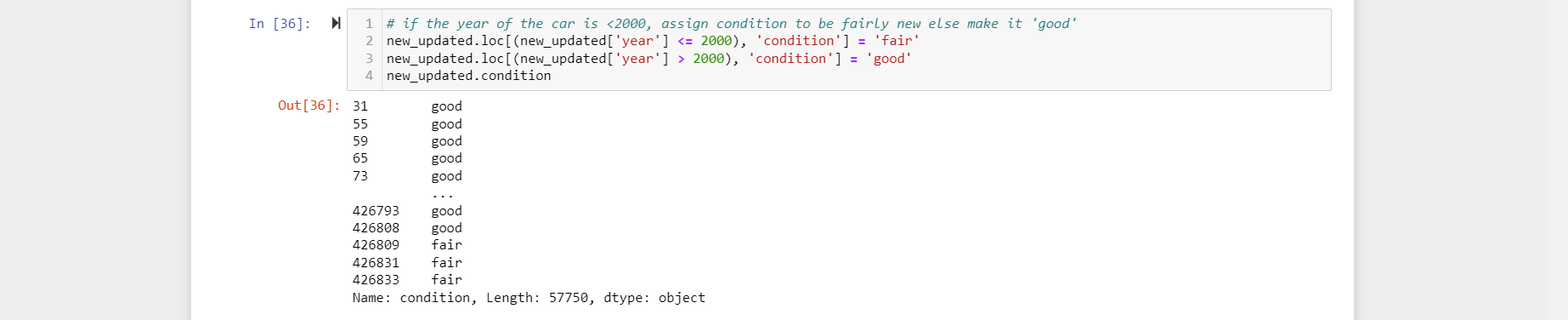


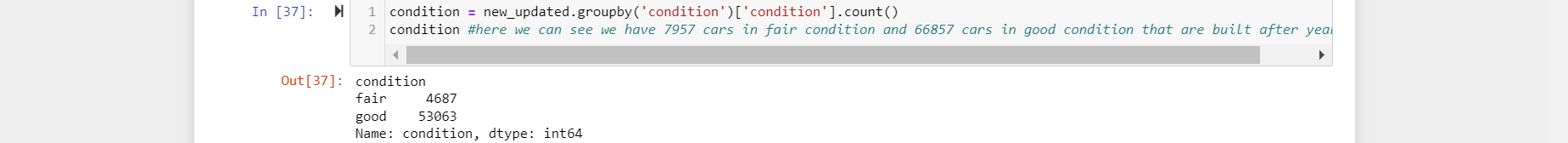
**Figure 16: Dropping Irrelevant Price Values and checking for updated Maximum**

**  
Figure 17: Grouping Car type Data**



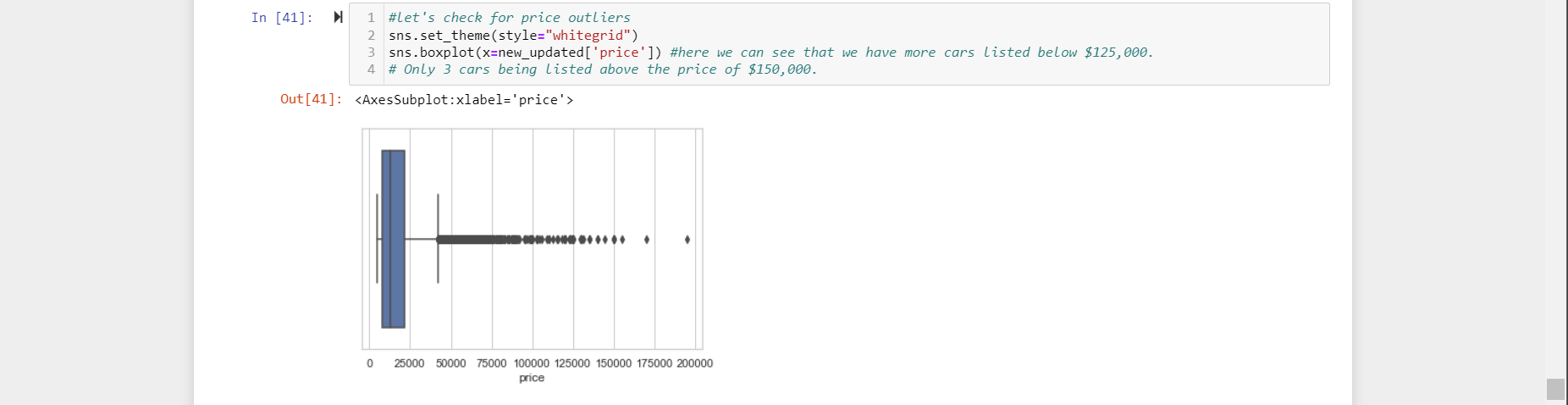
**Figure 18: Grouping by Paint Color**

****

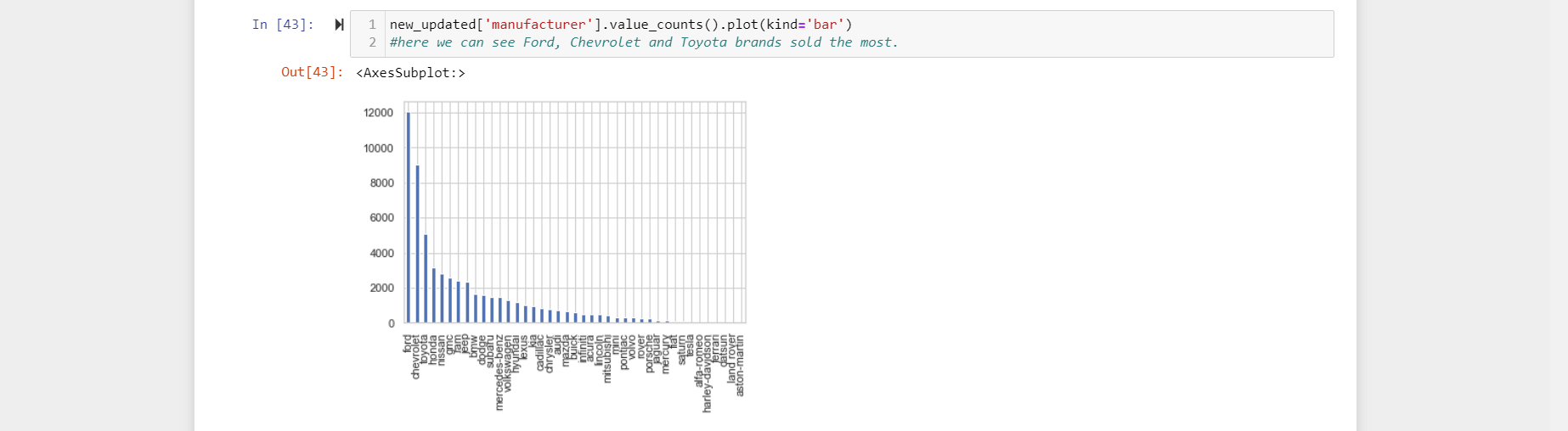
****

**Figure 19: Determining the Condition of the Car**

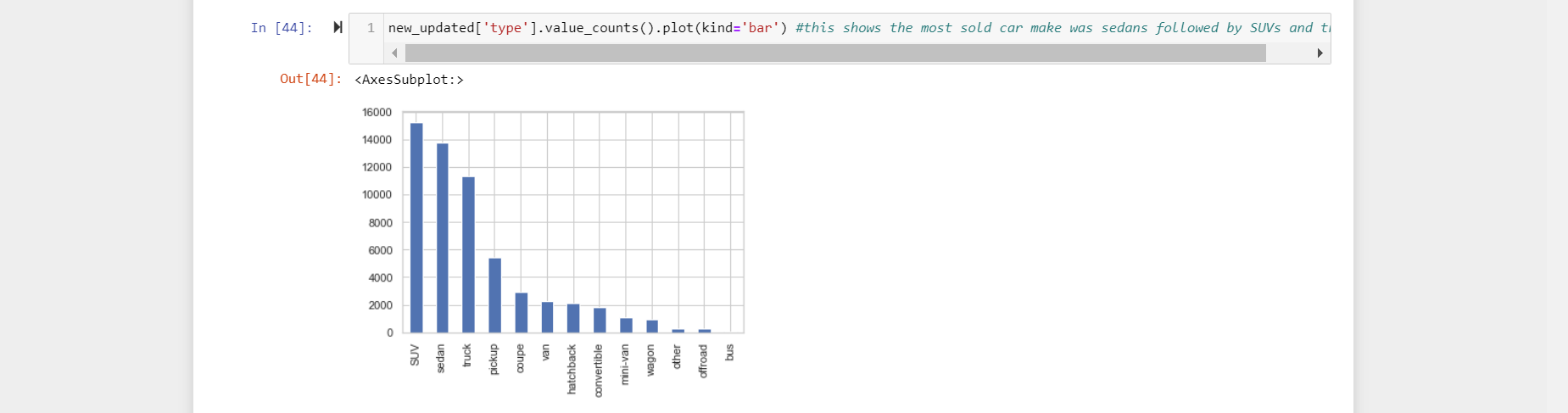
**Figure 20: Odometer Boxplot**

****

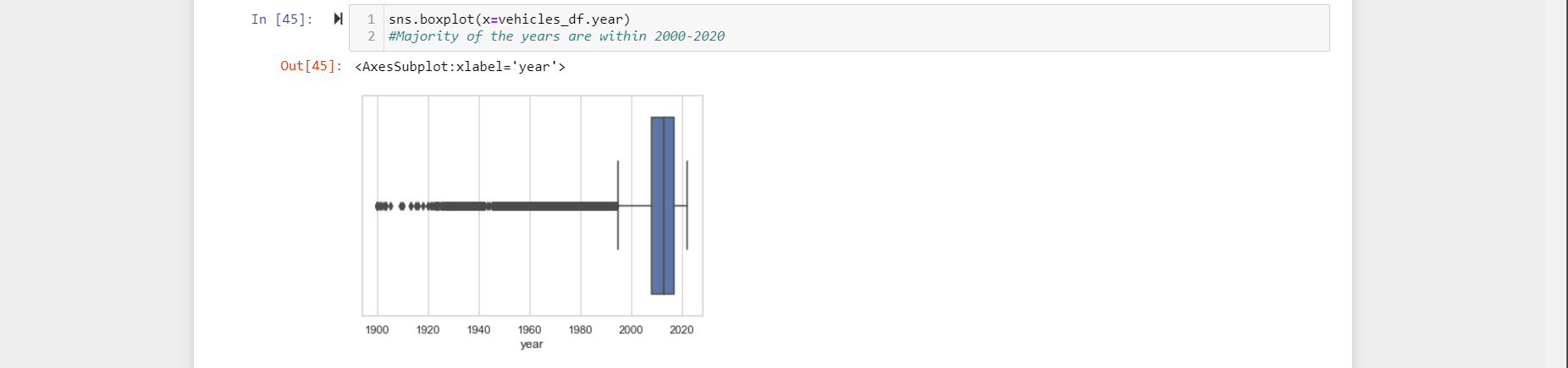
**Figure 21: Price Boxplot**

****

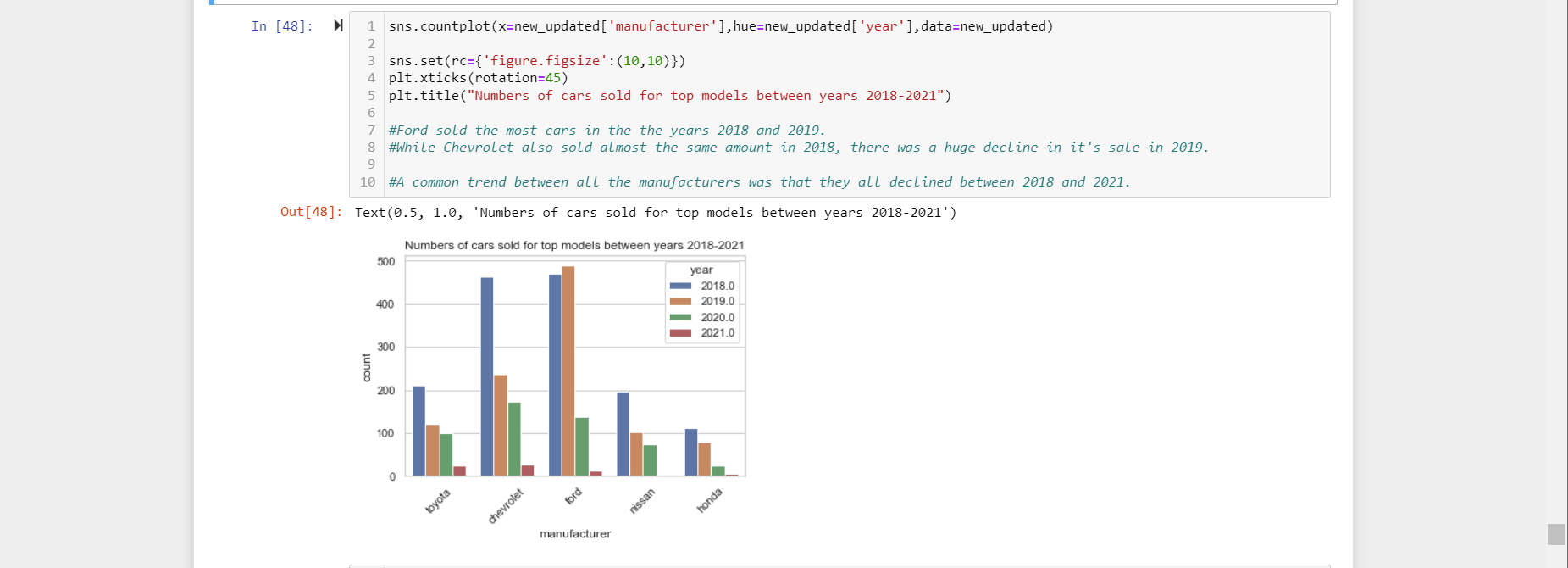
**Figure 21: Manufacturers Bar Chart**

****

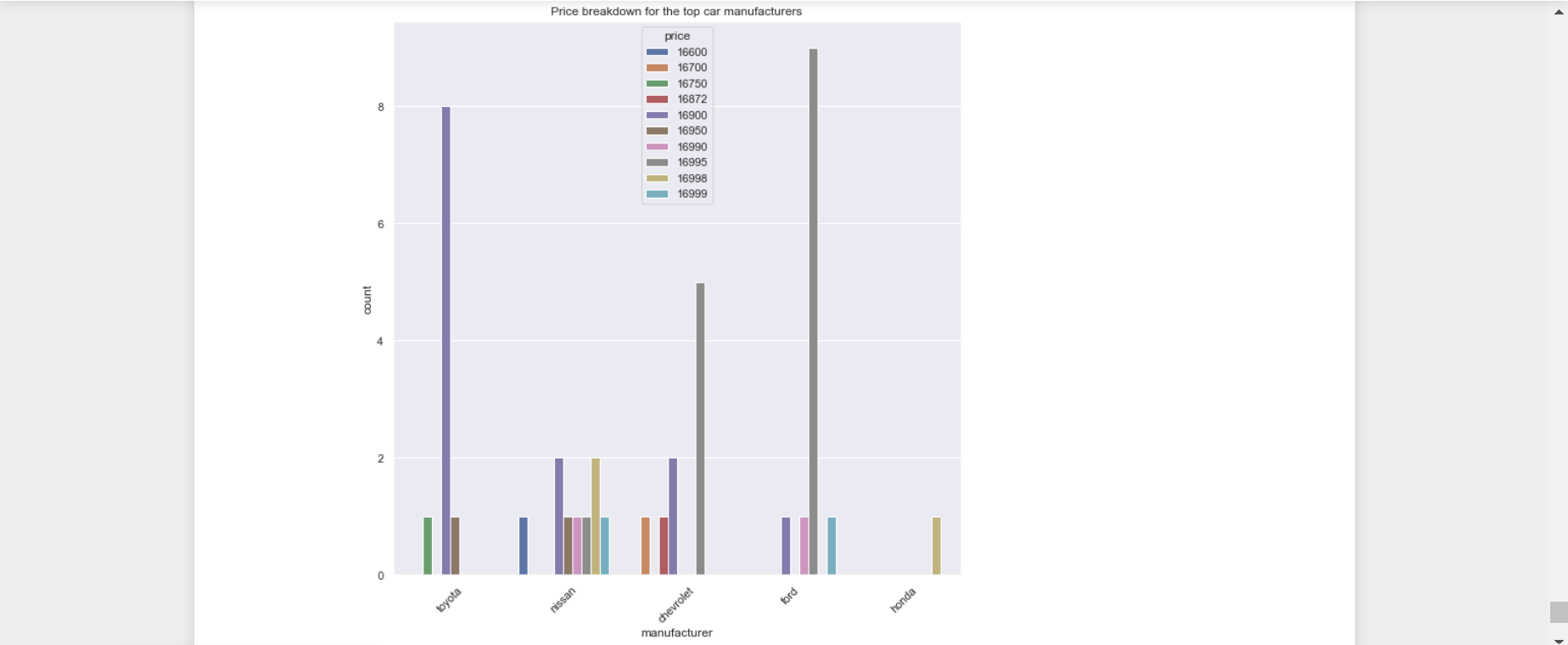
**Figure 22: Car Model Bar Chart**

****

**Figure 23:Boxplot by Year**

****

**Figure 24: No. of Cars Sold for Top Models (2018-2019)**

****

**Figure 25: Price Breakdown for Top Manufacturers by Year**

**References:**

[1]P. Jones, “Do A New Car’s Odometer Say Exactly ZERO Miles? (Checked).” https://motorandwheels.com/new-cars-say-exactly-zero-miles-on-odometer/ (accessed Dec. 04, 2022).

[2]“When Will Car Prices Drop? | J.P. Morgan Research,” *www.jpmorgan.com*. https://www.jpmorgan.com/insights/research/when-will-car-prices-drop

[3]“Buying a Car During COVID-19 Pandemic | Equifax - United States - Evo Prod,” *United States*.

[4]A. Gavazza, A. Lizzeri, and N. Roketskiy, “A Quantitative Analysis of the Used-Car Market,” *The American Economic Review*, vol. 104, no. 11, pp. 3668–3700, 2014, Accessed: Dec. 04, 2022. [Online]. Available: <http://www.jstor.org/stable/43495350>

[5]J. Neuburger, “Ending Data Scraping Dispute, Craigslist Reaches $31M Settlement with Instamotor,” *New Media and Technology Law Blog*, Aug. 24, 2017.

[6]I. A. Wains, “Exploring and Analyzing Used Car Data Set,” *The Startup*, Dec. 03, 2020. <https://medium.com/swlh/exploring-and-analyzing-used-car-data-set-2e2bf1f24d52>

**Datasets:**

[1]“Used Cars Dataset,” *www.kaggle.com*.<https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data>

[2]T. Shin, “A Machine Learning Project — Predicting Used Car Prices,” *Medium*, May 06, 2020. https://towardsdatascience.com/a-machine-learning-project-predicting-used-car-prices-efbc4d2a4998 (accessed Dec. 02, 2022).

[3]Elladuke, “Elladuke/SCAMP-FINAL-PROJECT-INDI,” *GitHub*, May 20, 2021. https://github.com/Elladuke/SCAMP-FINAL-PROJECT-INDI/blob/master/new\_vehicle.csv (accessed Dec. 02, 2022).