**Data** : <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

This data is scraped from Craigslist. Craigslist is the world's largest collection of used vehicles for sale.

**Goal**: Predict the price.

**Columns**:

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| City | Nominal | City |
| URL | Nominal | URL |
| City URL | Nominal | City URL |
| Price | Numerical | Price of the vehicle |
| Year | Numerical | Year of the model may be? Numerical Not sure about this |
| Manufacturer | Categorical/Nominal | Brand |
| Make | Categorical/Nominal | Model |
| Condition | Categorical | Excellent/Good/Like New/Fair/New/Salvage |
| Cylinders | Numerical | Number of Cylinders |
| Title\_Status | Categorical | Clean/Rebuilt/Salvage/Linen/Missing/Parts only |
| Transmission | Categorical | Automatic/Manual/Other |
| Drive | Categorical | 4wd/fwd/rwd |
| Size | Categorical | Full/Mid/Compact/Sub Compact |
| Type | Categorical | Type of the car(Sedan/SUV etc) |
| Paint Color | Categorical | Car Color |
| Odometer | Numerical | Number of miles run by the car |
| Desc | String | Description |
| Latitude |  |  |
| Longitude |  |  |
| Fuel | Categorical | Gas/Diesel/hybrid/Electric etc |
| VIN | String | Unique Identification Number |
| Image URL |  |  |

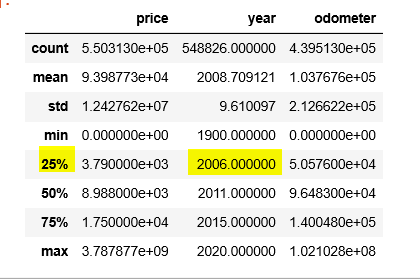
**Size of the Data:** (550313, 22)

**Data Cleaning Steps:**

* I initially created a column ‘missing\_values’ which counts the number of missing values(column) in each row. Then, I observed that 75% of the data has less than 4 missing values for each row, so I deleted those rows which have more than 4 missing values in them.
* **Dropped these columns :** City URL, URL, Latitude, Longitude, Description, Image URL

**Year:**

**Criteria:** Year > 2005. Because there is only 25% of data which has year < 2006, decided to set the limit as 2005.



PS: Did not fill the missing values in year. We need to see if the year has any effect on Price and then decide how can we impute them. (EDA helps here)

**Price:**

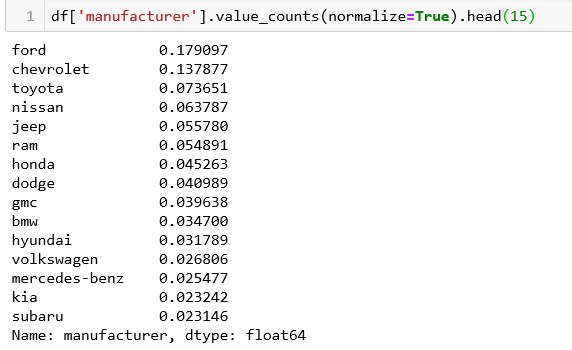
**Criteria:** Price > 500 and Price < 100000

**Odometer:**

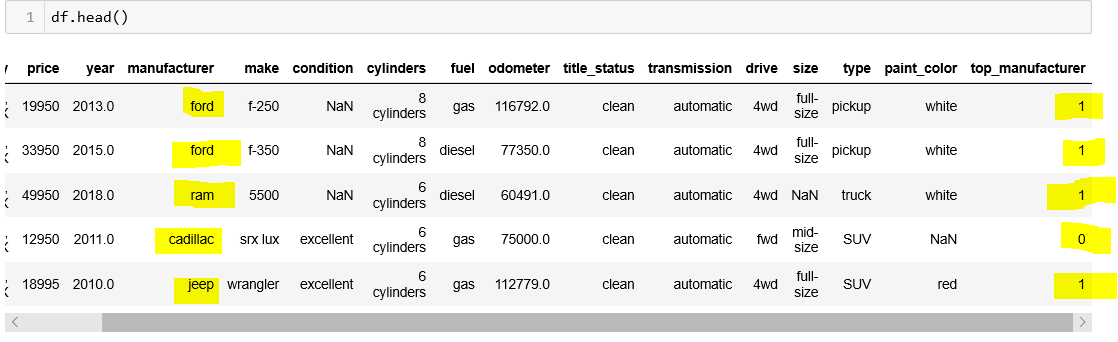
**Criteria:** Odometer>12000 and Odometer < 120000

**Manufacturer:**

There are 40 unique manufacturers in the data set. Below are the top 15 manufacturers.

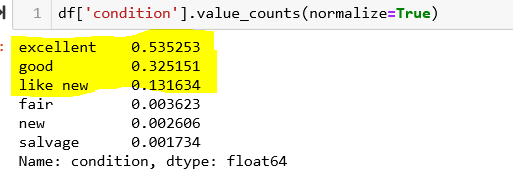


As seen in the above screenshot, the top 6 brans(ford, Chevrolet, Toyota, Nissan, Jeep and Ram) constitute to more than 5% of the total data. So I created a new field which says **‘top\_manufacturer’** which has values of 0 or 1. )(1-Belongs to top 6 brands,0-Doesn’t belongs to Top 6 Brands)

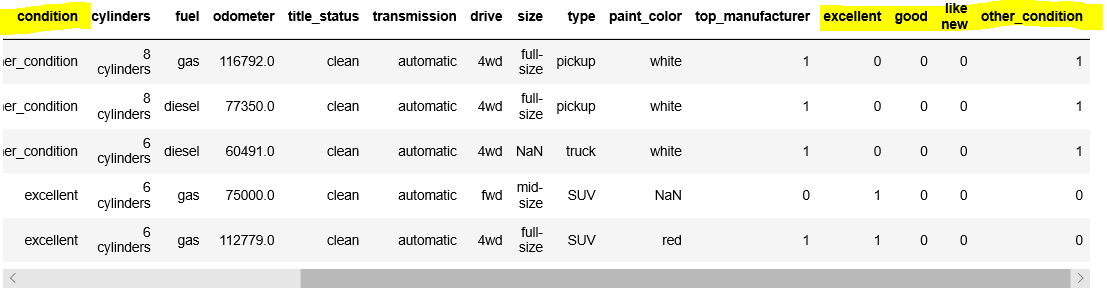


PS: After the insights we get from EDA, we can decide on creating dummy variables for this column.

**Condition:**



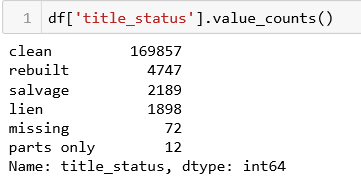
Here, The conditions(Excellent, GOOD and Like new) has more number of records of records in the data. So I grouped the other conditions(fair,new,salvage) as ‘Other Condition’ and created 4 dummy variables.



I followed the similar technique with the following columns: **Type and Paint Color(**grouping the categories with less number of records as ‘Other’ and created dummy variables)

Without any grouping like above, only dummy variables are created for these variables **- Fuel, Transmission, Size, Drive.**

**Title \_Status**



Major proportion of the data set are clean cars. So I created a new column ‘title\_clean’ with values 1/0. 1- Clean Car, 0-Not clean car.

After cleaning the data, we are left with **180509** rows of data.

**DATA VISUALIZATION THROUGH TABLEAU**

I could use sum, count, median, average and percentile functions on the outcome variable.

**Sum- Total Price on that group**

**Count- No of records on the group**

**Percentage- Total percent of all on the group**

**Percentile 10 – only 10% of data lies below that value**

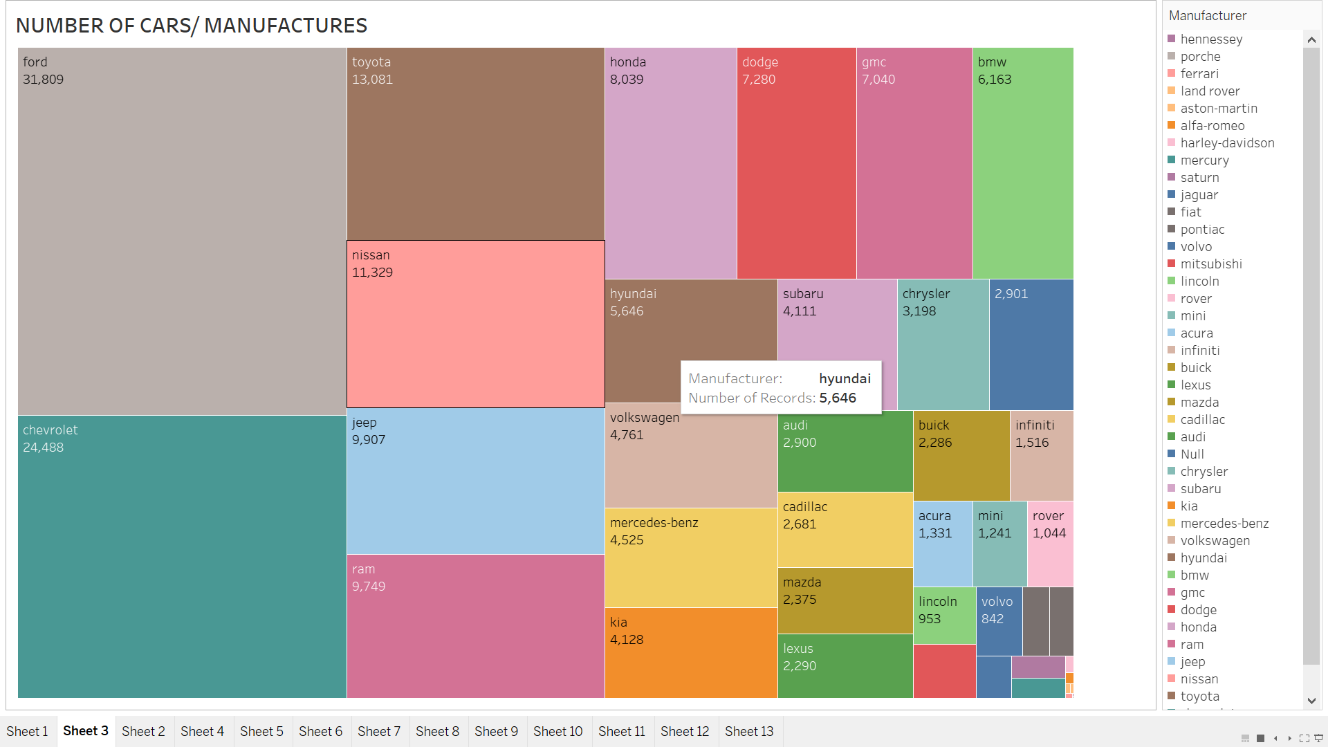
**Percentile 90- only 10% of data lies above that value**

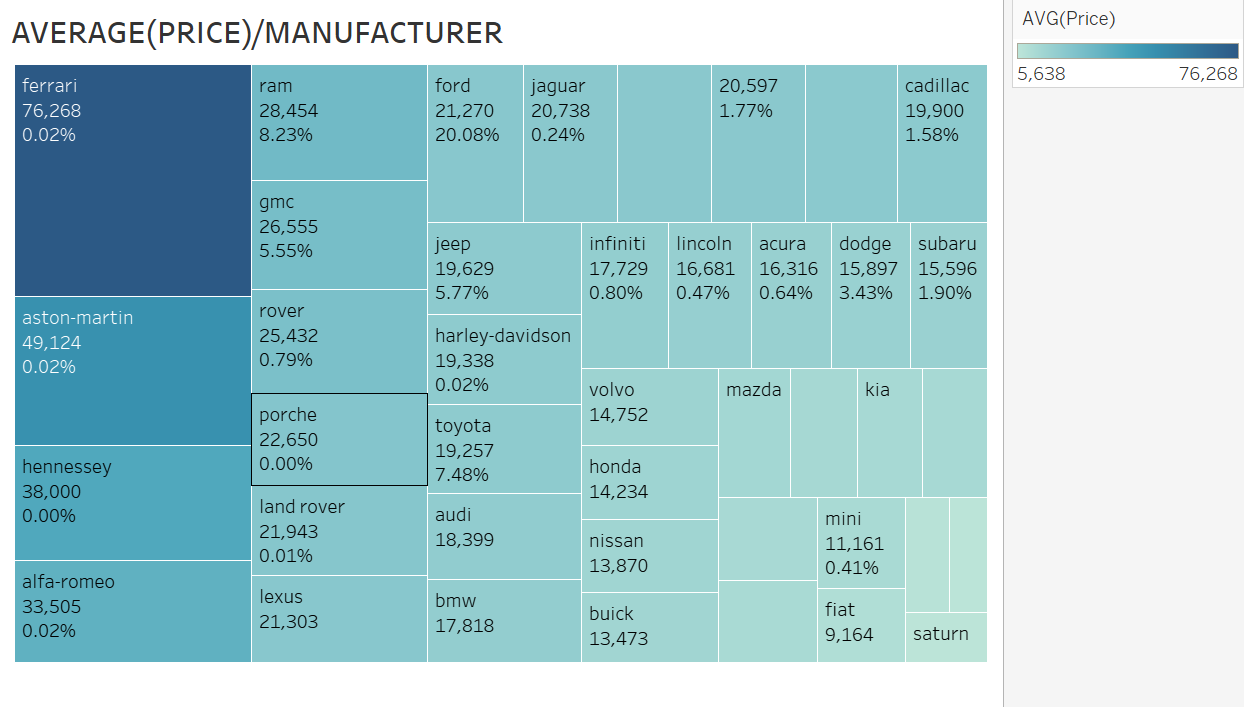
**Average- Mean of price**

**Median- Median of price in that group**

**Average (Price)/Manufacturer-**

Ford, Chevrolet, Toyota, Nissan, jeep, ram, Honda, Hyundai, Volkswagen are top manufactures. Does that define that they are sell the highest priced cars?



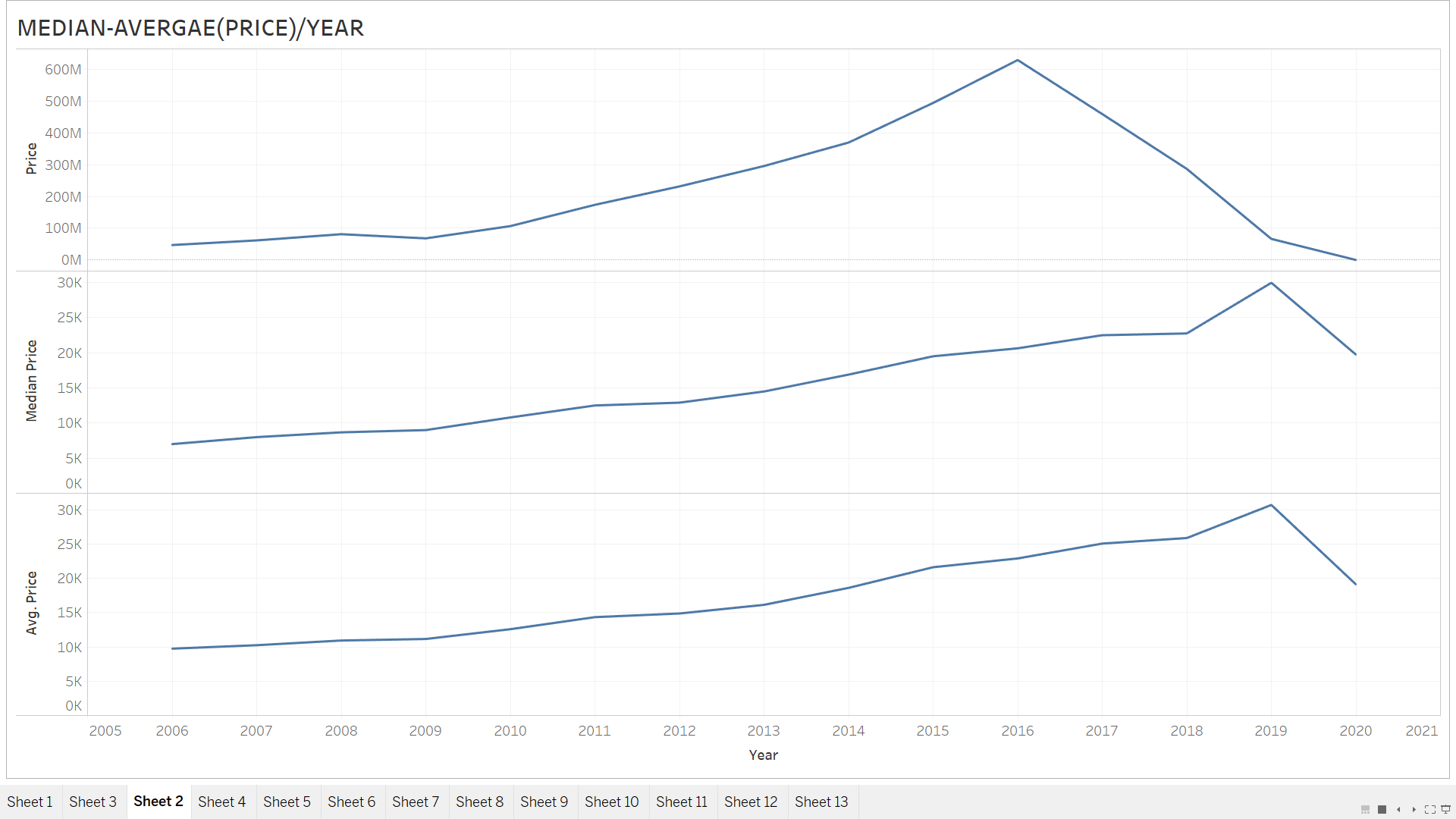


**PRICE/YEAR:**

Year plays an important role as we can notice the increase and decrease in prices w.r.t year.

Sum(price)- Highestin year 2016 and then decreases

Median(price) and average(price)-Highest in 2019 and then decreases

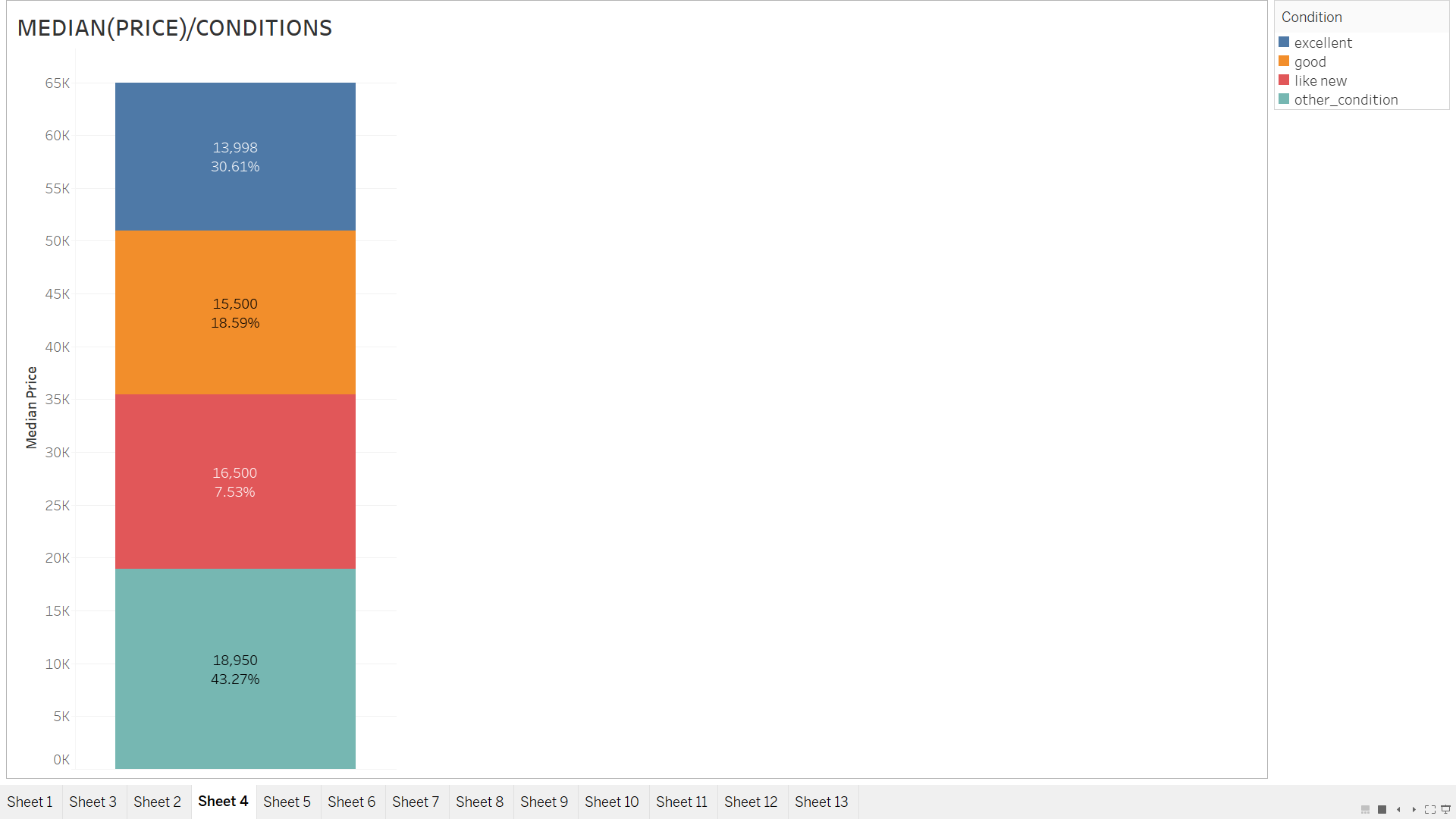


**Median(price)/Conditions:**

Could be excellent, good , fair, and others(new, fair, salvage, nulls)

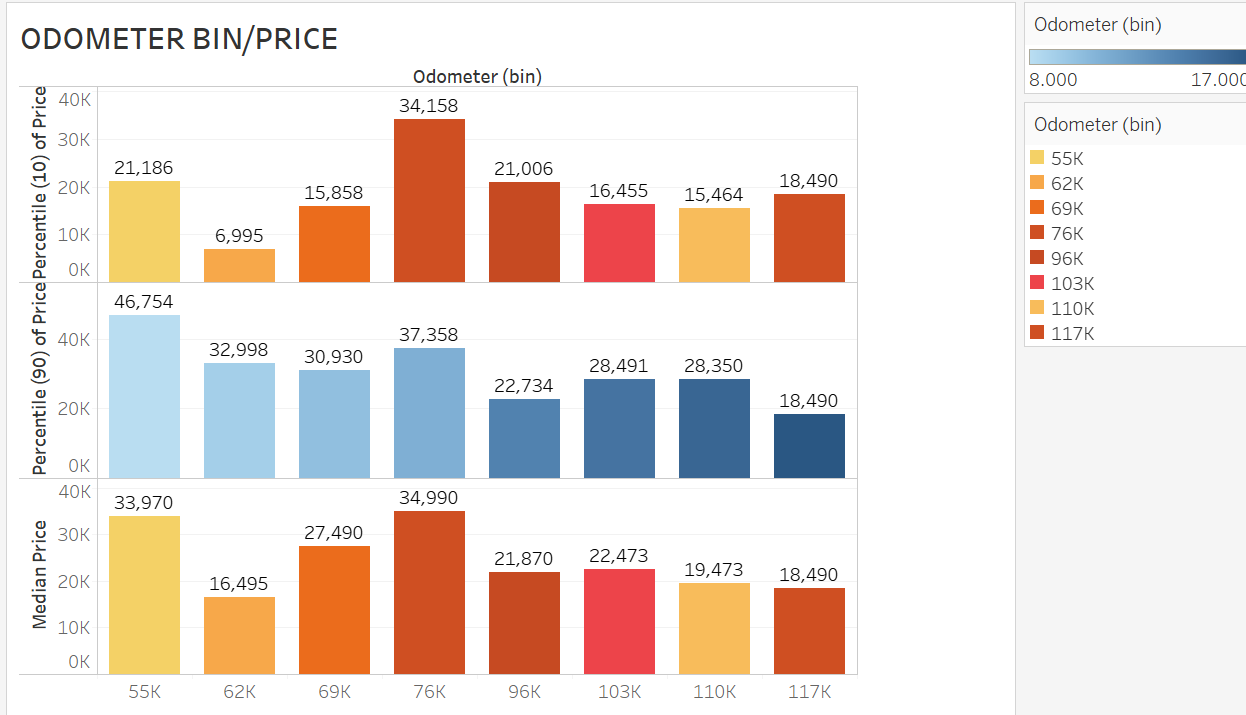
I think others shouldn’t be grouped together, as new cars will have higher prices than excellent and fair.

Excellent<good<like new<others(could be because of new cars)



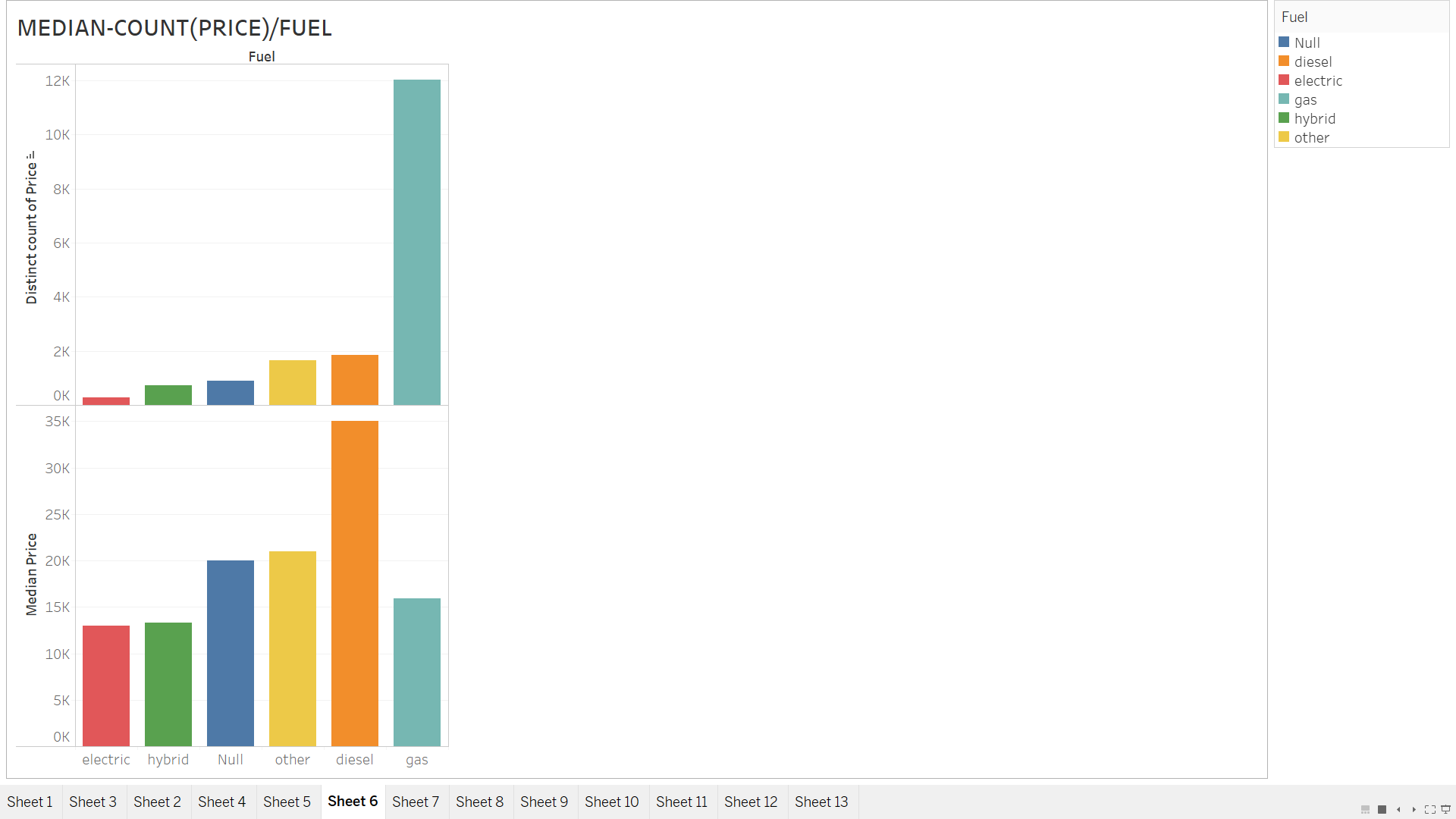
**PRICE/ODOMETER:**

As we have huge bins of odometer, we can create bins. 9 bins. 55k-117k and then check the prices. Percentile 10 is the lowest range and percentile 90 as the upper range. Irregular patterns.



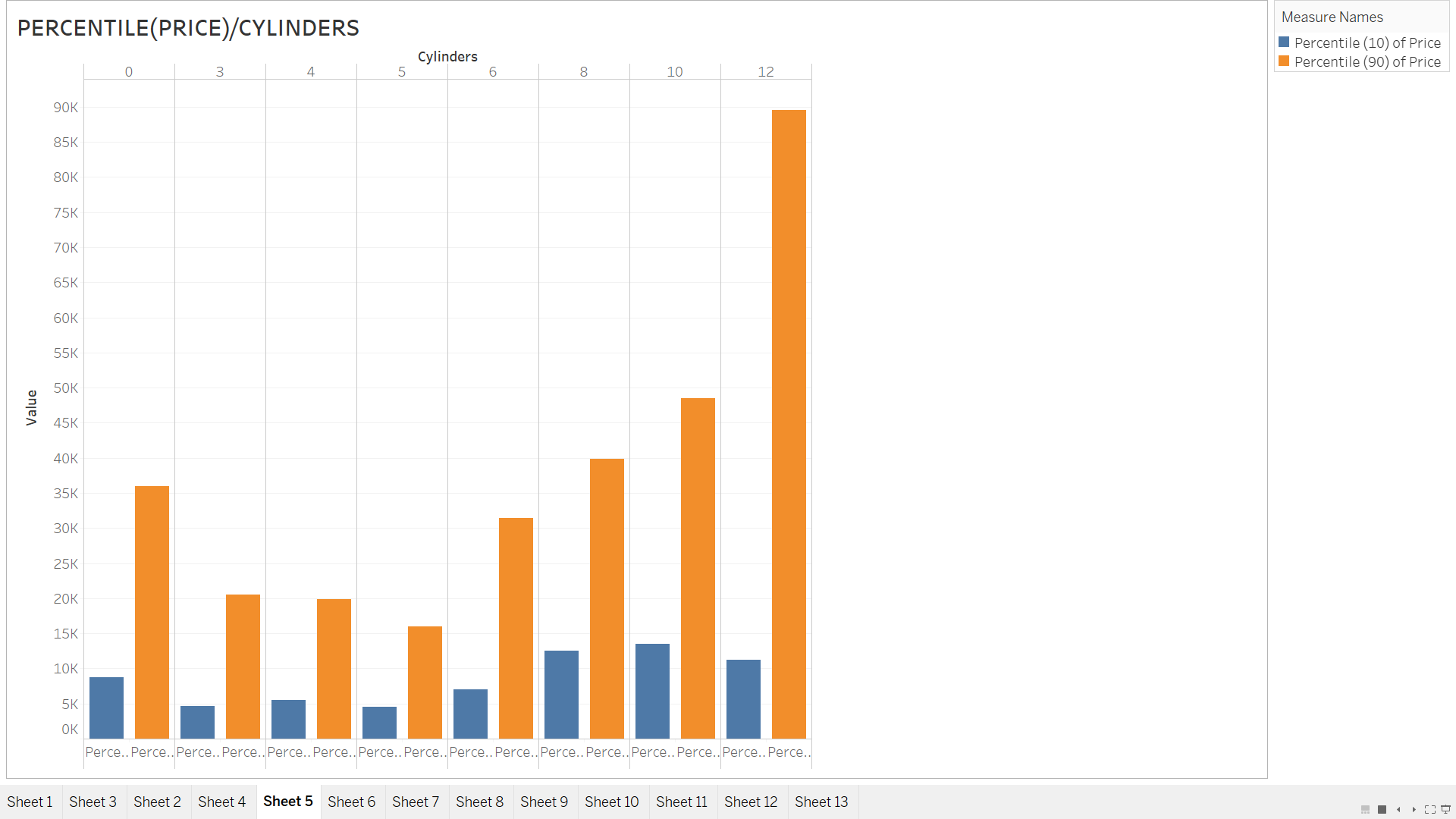
**PRICE/FUEL:**

**Most of the cars are using CNG as the fuel, yet prices are high for diesel cars.**

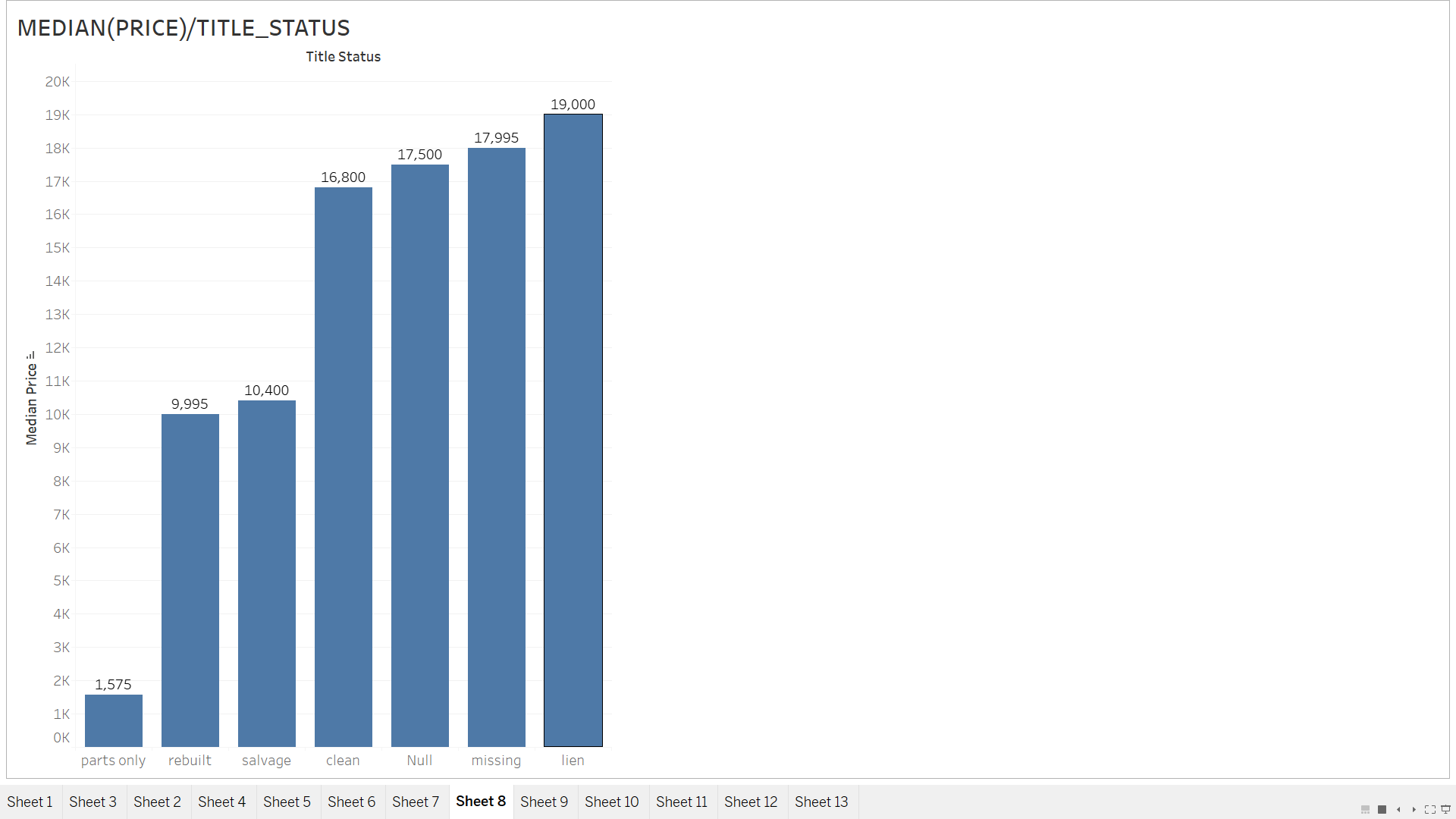


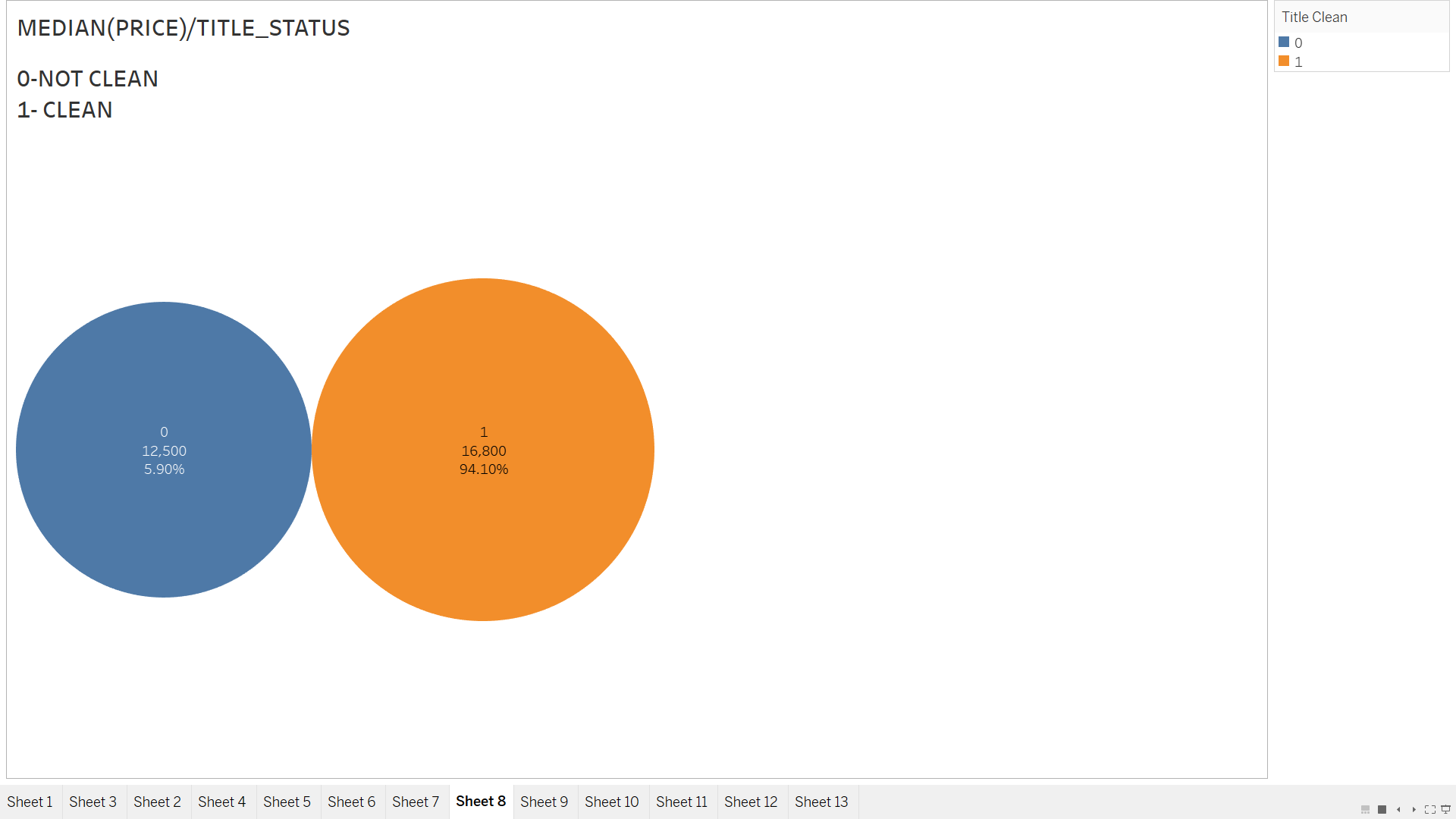
**PRICE/CYLINDERS:**

Converted cylinders into a dimension. Calculated percentile 10 and 90. We can say that data has some pattern where more no of cylinder corresponds to higher price.



**Median (Price)/Title\_Status and Title\_Clean**





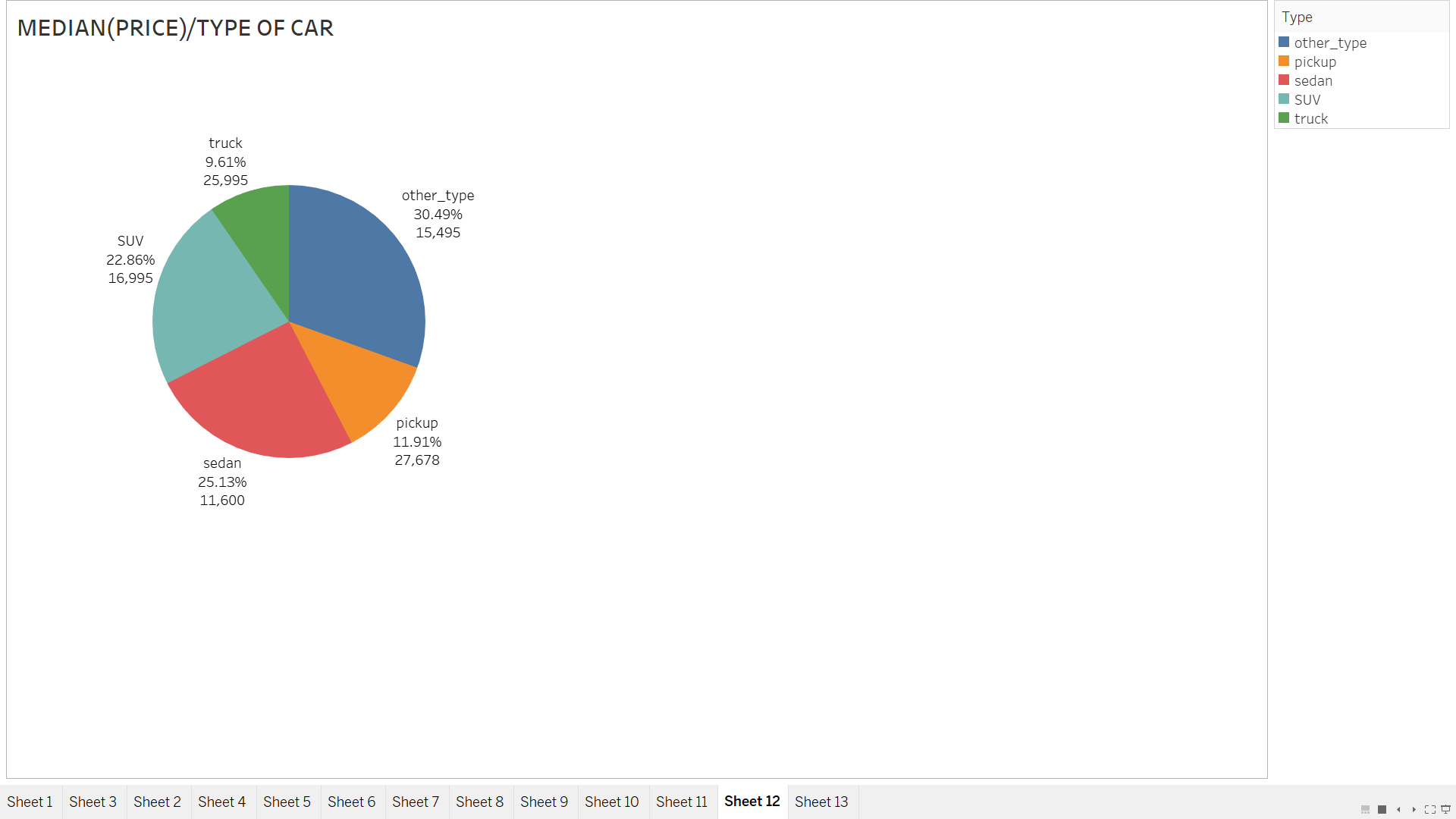
**MEDIAN(PRICE)/COLOR:**

I think we should use all the colors.



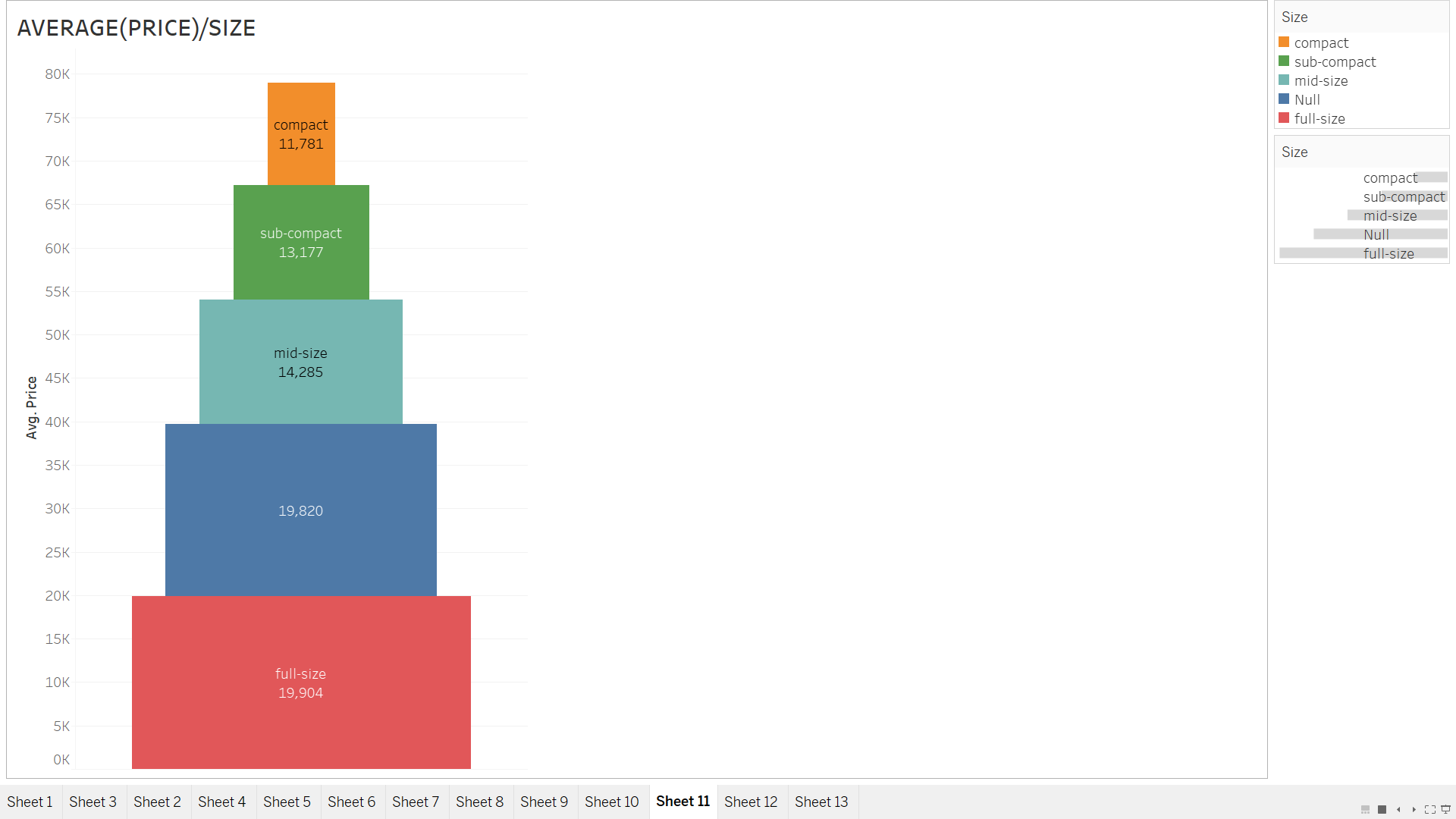
**PRICE/TYPE OF CARS:**

Percentage showing how many percent of cars are there of that type. Median (Price)- pickup vehicles>trucks>other\_types>SUV>Sedan

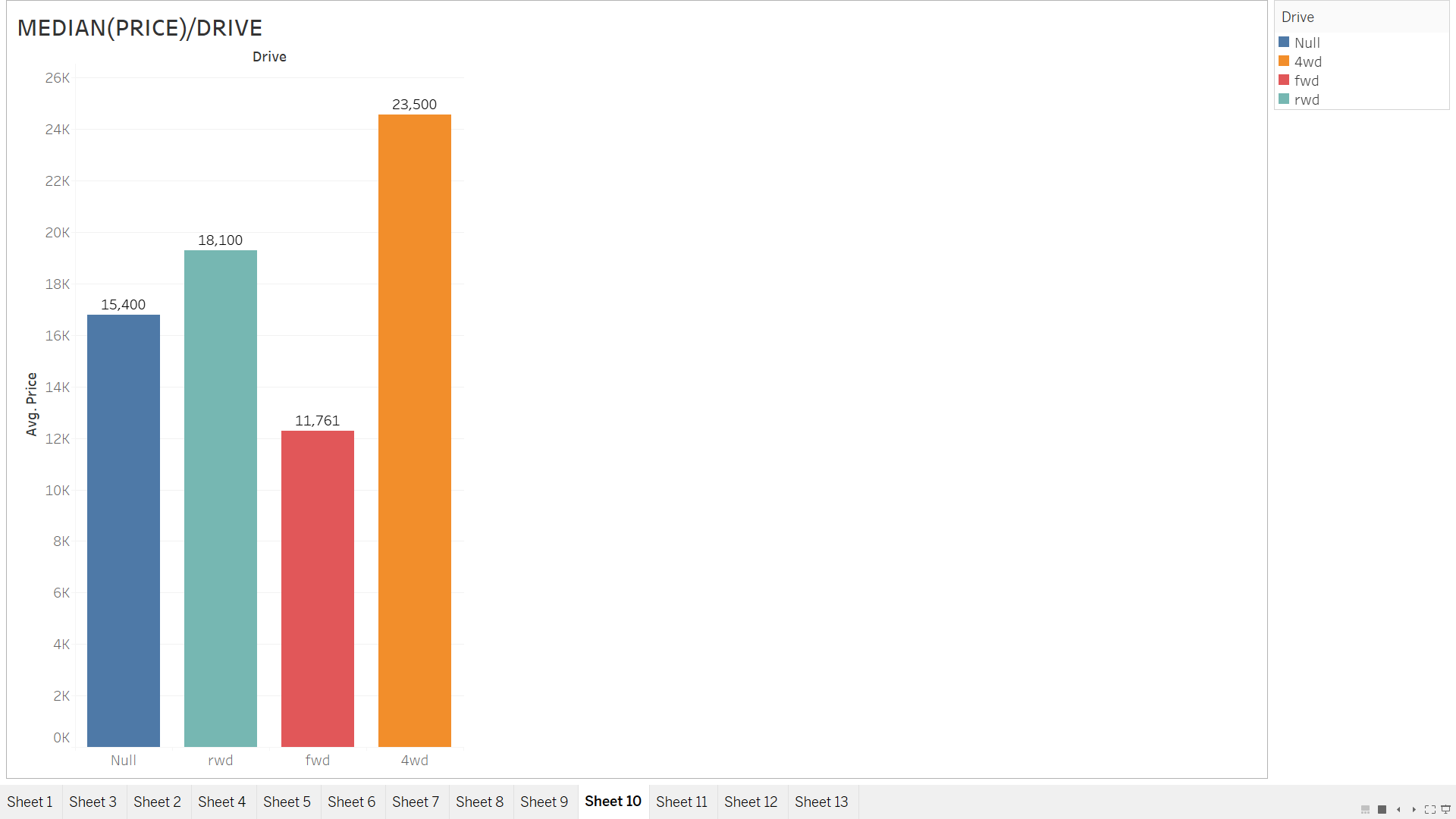


**PRICE/SIZE:**

Full size cars have higher price and null values has second highest price. How to move forward? But the pattern is there. As size increases price increases.



**PRICE/DRIVE:**



**PRICE/TRANSMISSION:**

