

Used Car Brand Analysis

Analysing used cars listed on ebay to identify which car brands are most resistant to price depreciation

Sudhanshu Chib | Nov 11th 2017

# Abstract

The project aims to analyse data for used cars listings to evaluate how value of cars depreciates over time. Multiple car brands are analysed to find out which car brands are most resilient to depreciation with time and mileage. The report talks about steps taken to collect real world data through web-scrapping, data cleansing & evaluation, project assumptions and evaluation methodology.

# Introduction

Cars, being a depreciating asset, loose market value with factors like time and mileage contributing to their decline. This project attempts to understand the effect of price depreciation by analysing quoted price of used car.

Findings of this project can be useful for any individual researching to buy a used car.

The project aims to rank order car brands based on their resistance to depreciation (from high resistance to low resistance)

# Project Methodolgy

The following steps were performed while executing the project:

1. Data Capture: Scrapping eBay to capture sufficient data for analysis
2. Exploratory Data Analysis: Evaluating data to identify logical errors. Cleaning data captured to arrive at a sample population that can be evaluated
3. Data analysis: Segmenting car brands based on their starting price and calculating depreciation over time. Building a dashboard on tableau to report findings

Statistical validation of segmentation and findings

## 

## 

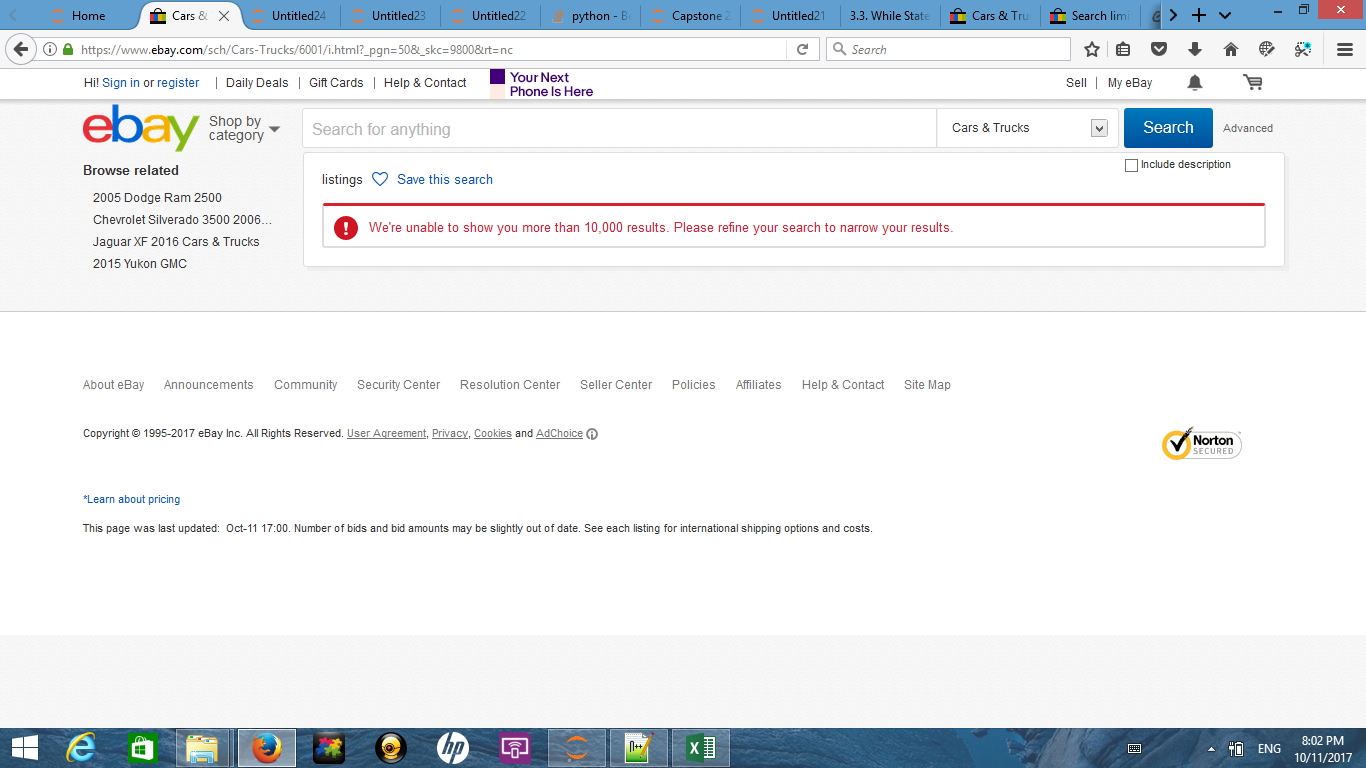
## Data Capture:

Website: eBay

URL: https://www.ebay.com/sch/Cars-Trucks/6001/i.html?LH\_ItemCondition=2%7C0&\_dcat=6001&\_dmpt=US\_Cars\_Trucks

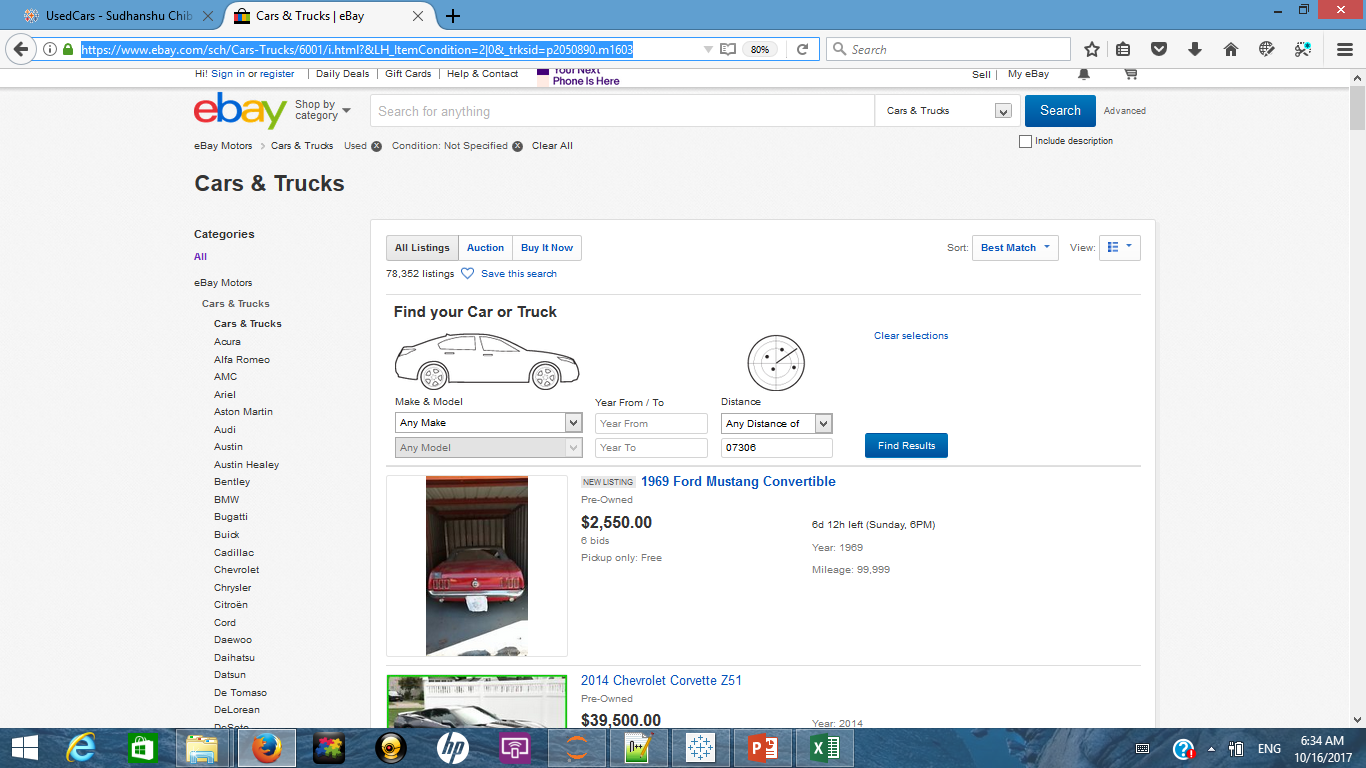
eBay hosts a section on their website where users can put their used cars for sale. This section was used to scrape car listings using a combination of Beautiful Soup and Request libraries in python.

One drawback of the eBay is that a user can browse through or evaluate only 10,000 results at a time. That is, if you search for a product on the website ebay will show that there are X number (say 70,000) of products available but in reality if you navigate through their webpages only 10,000 results will be shown. Snapshot for the same is shared below:



As a result any scrapping attempt would have restricted the data to a size of 10,000 records.

To avoid this the web scrapping scripts was modified to capture brand categories and their subsequent redirection page links. Brand and their redirection page link was stared in separate lists. These lists were then used to access individual brand webpages and extract data from them.



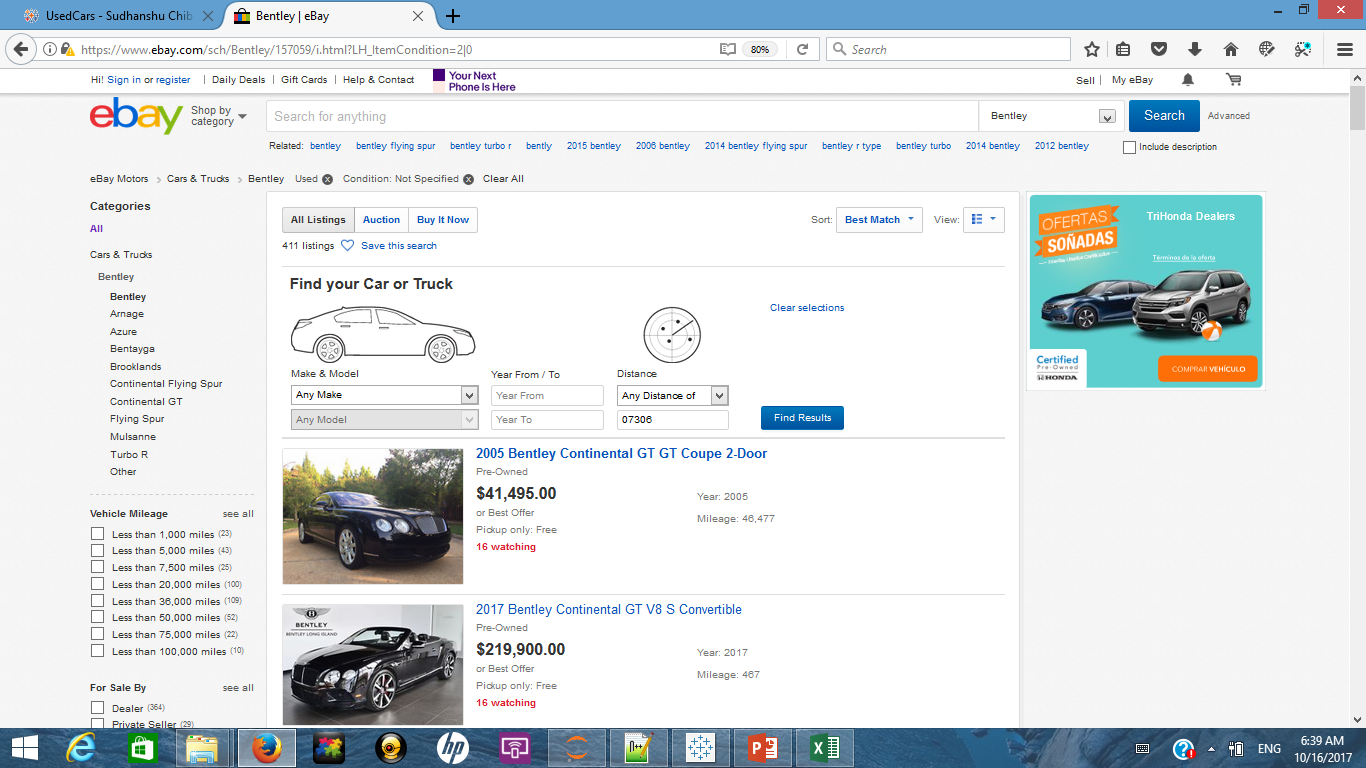
When accessing data from individual brand webpages

To iterate through multiple car listings across pages a URL string was constructed to mimic eBays navigation string.

Navigation URL = Url1 + "&\_pgn=" + str(Page) + "&\_skc=" + str(Count)

Where Url1 is the navigation url for different car brands and variables Page and Count are integers iterated through for loops.

The constructed navigation URL allowed the script to navigate through multiple pages (by posting url requests through python’s request library) on eBay’s website and gather data.

Listing data captured:

A typical car listing on eBay has the following information

Car information string that gives information about the year of manufacture, brand of car and model information.

Quoted price and Mileage on the car

Python script was used to iterate through these information and store it as a data frame.

Data frame was exported as a csv file and is available on the github link as ebayRevised1.

**Data cleaning as part of scraping data:**

A common feature of data captured from eBay was presence of newline “\n” and tab “\t” characters. So for every data element captured, as a blanket rule newline and tab characters were removed from the elements.

Tag information for car model and make, price and year is shared below:

Name=soup.find\_all("a",{"class":"vip"})

Price=soup.find\_all("li",{"class":"lvprice prc"})

Year=soup.find\_all("ul",{"class":"lvdetails left space-zero full-width"})

This information was captured using the find all function supported by BeautifulSoup.

Car’s mileage and year of manufacture information was captured using regular expressions shown below:

year=(re.search(r'Year: \d\d\d\d',year))

mileage=(re.search(r'Mileage: \d+',mileage))

Car’s name sometimes had text like “New Listing” appended to the name. Such text was removed by using the replace option in python.

Any missing field in data was imputed by the value “NotFound“.

To avoid getting blocked by eBay a default sleep timer was added to the script which added a delay of 5 seconds between subsequent url requests.

This resulted in 237, 186 different car listing data to be tracked and stored.

[Link for web scraping script](https://github.com/SudhanshuChib/Car-Brand-Analysis/blob/master/eBay_ScrapingScript.ipynb)

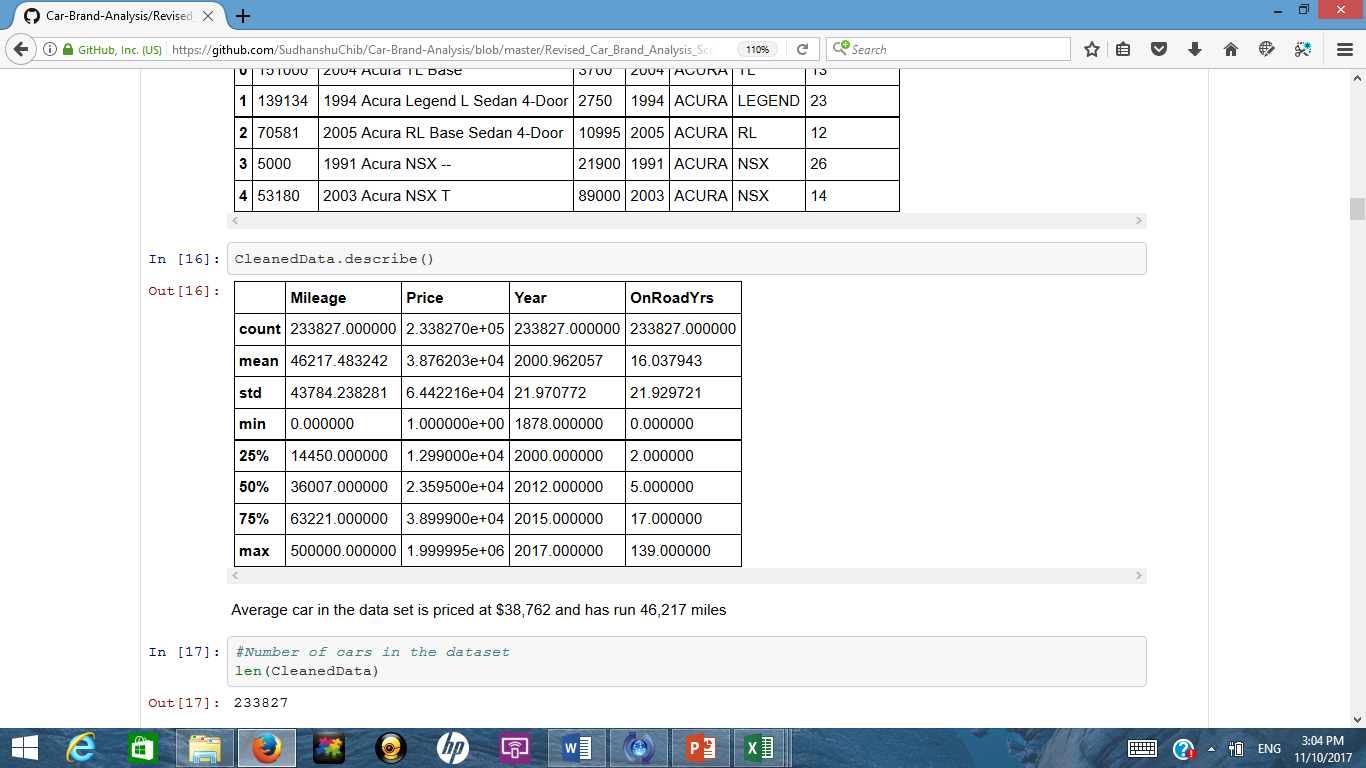
## Data cleaning and Exploratory Data Analysis:

Following are the steps performed to clean and explore captured data-

1. All missing values for price, mileage and car brand/ model names were imputed by string “NotFound” in the web scraping script.

Total number of such cars was 683 which is about 0.2% of the total dataset hence these rows were removed from the data set.

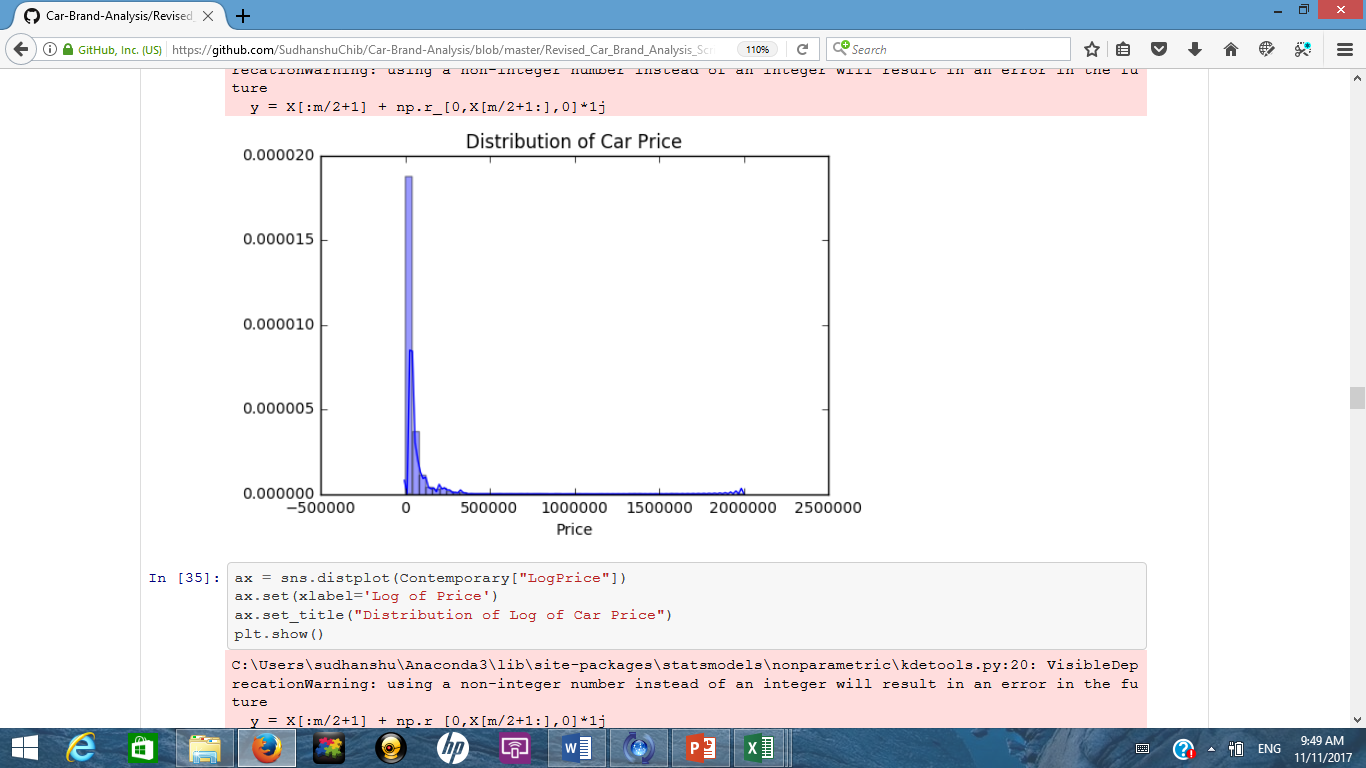
1. For all cars remaining in the data set a variable called “OnRoadYrs” was calculated by subtracting their year of manufacture from 2017. As a result of this some cars had negative value for this variable indicating year of manufacture as 2018 or beyond. Such cars were again removed from the data set. Two additional rules were applied that removed any car listing beyond price of two million or mileage of four hundred thousand miles.



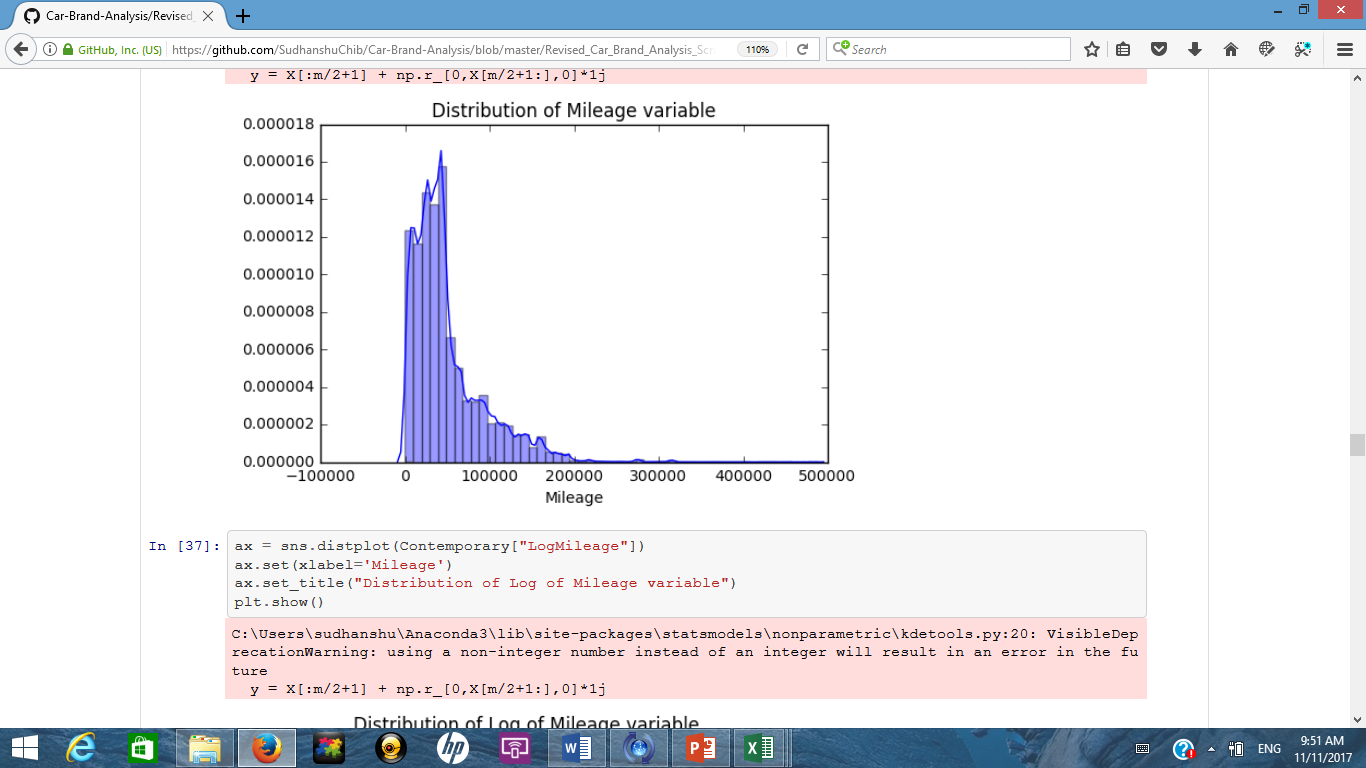
Post step 2: 233,827 car listings remained in the data set where the

Average car in the data set was priced at $38,762 and had run 46,217 miles

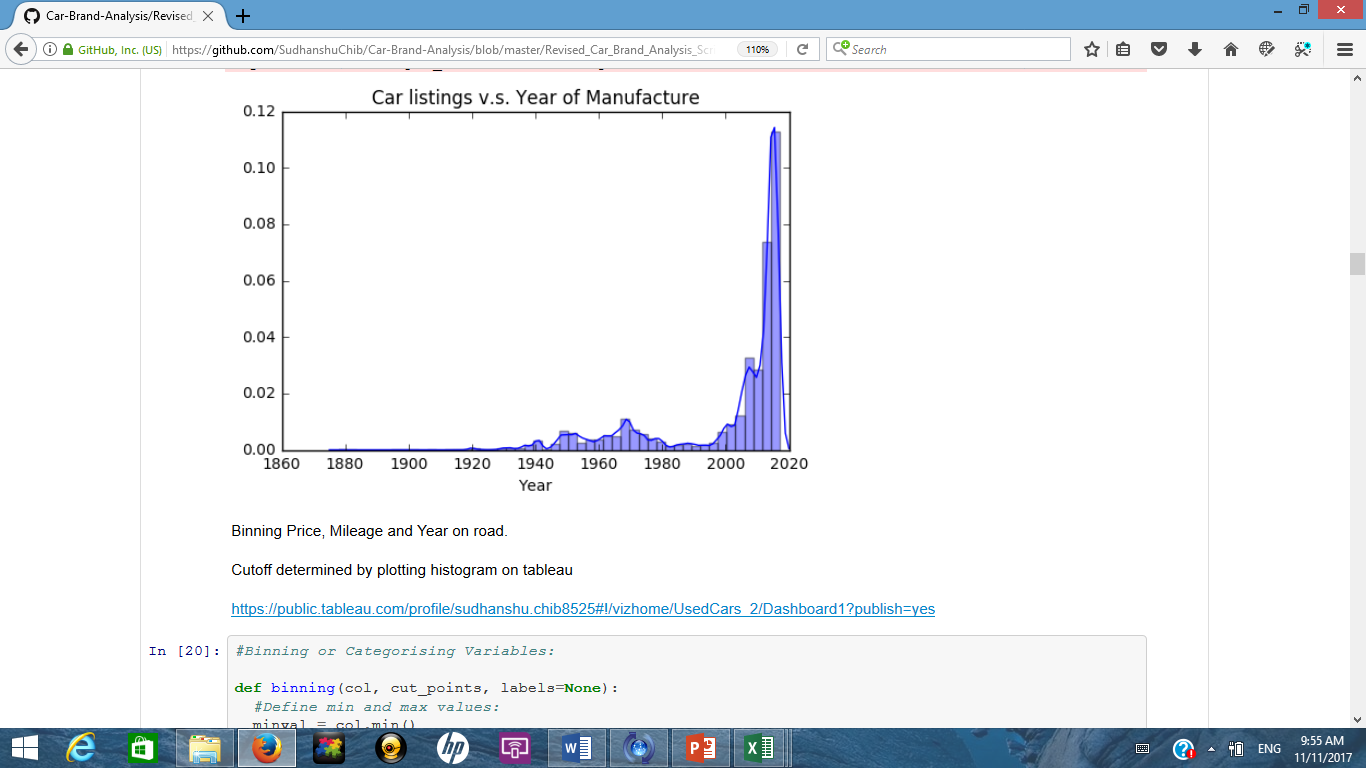
1. Distribution of car price: Major chunk of the cars are priced less than $50k on the used car market. The distribution is skewed right



1. Distribution of car Mileage: Majority of the cars on used cars market are driven less than 50L miles. The distribution is again skewed to right.



1. Distribution of car listings as per year of manufacture: Most of the listed cars were manufactured in the last 10 years. There are slight peaks near late 60’s manufactured cars. On exploring that population it was observed that those were majorly for Chevrolet Corvette cars manufactured in 1967-72 period as they are considered collectable.



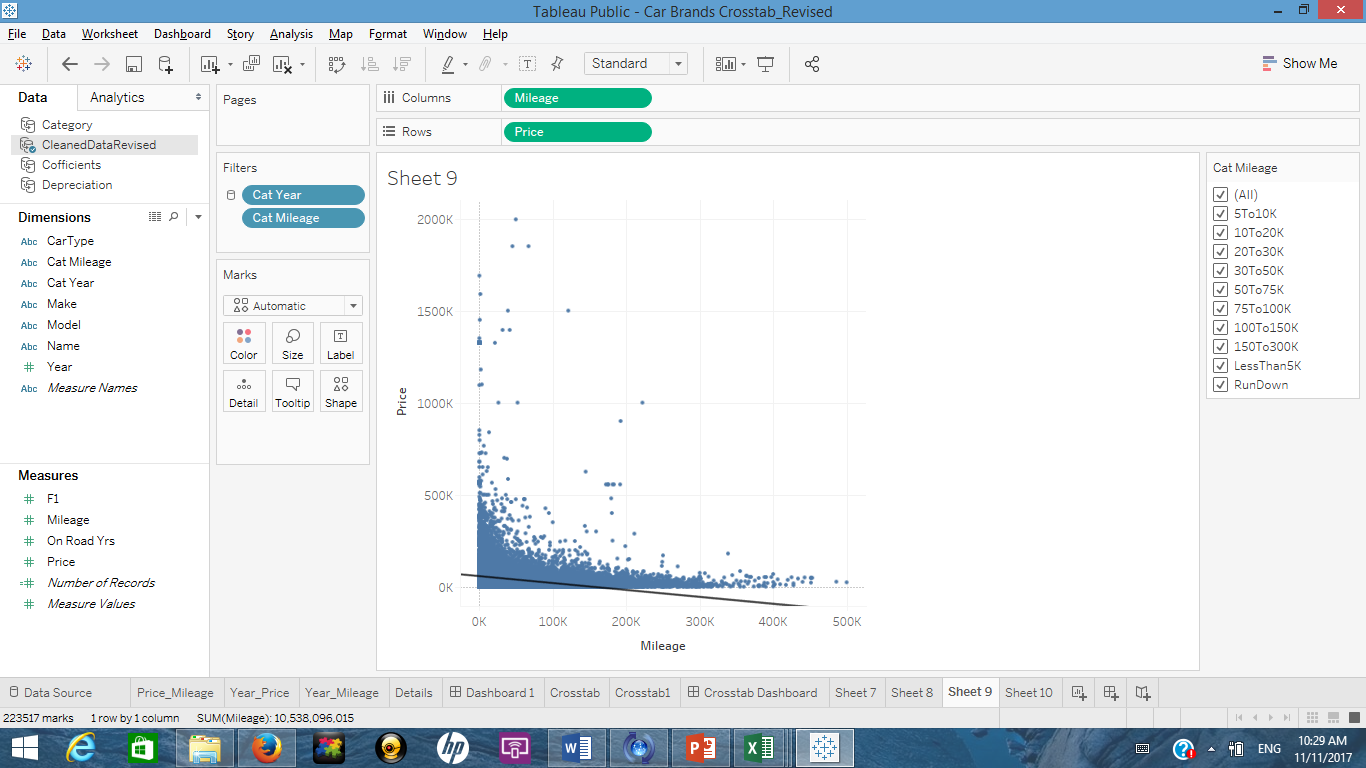
1. Years on road and mileage data was binned or categorised as per ranges shown below:

Mileage: Less than 5k miles, 5to10k miles, 10to20k miles, 20to30k miles, 30to50k miles, 50to75k miles, 75to100k miles, 100to150k miles, 150to300k miles and cars having over 300k miles were tagged as RunDown.

Year on road: Less than 2 years, 2 to 5 years, 5 to 10 years, 10 to 15 years, 15 to 20 years, 20 to 50 years, and cars having year of manufacture greater than 50 years from 2017 are tagged as heritage.

These cutoff were determined by observing histograms and car dealer websites.

1. Another categorical variable was created based on values of “Number of Years on Road” with cut-off at 15 yrs. Cars with value less than 15yrs were tagged as “Contemporary” and older cars were tagged as “Collectable”
2. Trend between price and mileage was also observed and a linear trend line was fitted to the data to observe the fit. As expected, price tends to decrease with mileage but the exact values of computed trend lines are not really a guiding light as the underlying population of data (skewed right) does not support normality assumption.



Trend line details: Only for guidance not to be used for prediction

**Trend Lines Model**

A linear trend model is computed for Price given Mileage. The model may be significant at p <= 0.05.

|  |  |
| --- | --- |
| **Model formula:** | ( Mileage + intercept ) |
| **Number of modeled observations:** | 223517 |
| **Number of filtered observations:** | 0 |
| **Model degrees of freedom:** | 2 |
| **Residual degrees of freedom (DF):** | 223515 |
| **SSE (sum squared error):** | 8.97424e+14 |
| **MSE (mean squared error):** | 4.01505e+09 |
| **R-Squared:** | 0.0626978 |
| **Standard error:** | 63364.4 |
| **p-value (significance):** | < 0.0001 |

**Individual trend lines:**

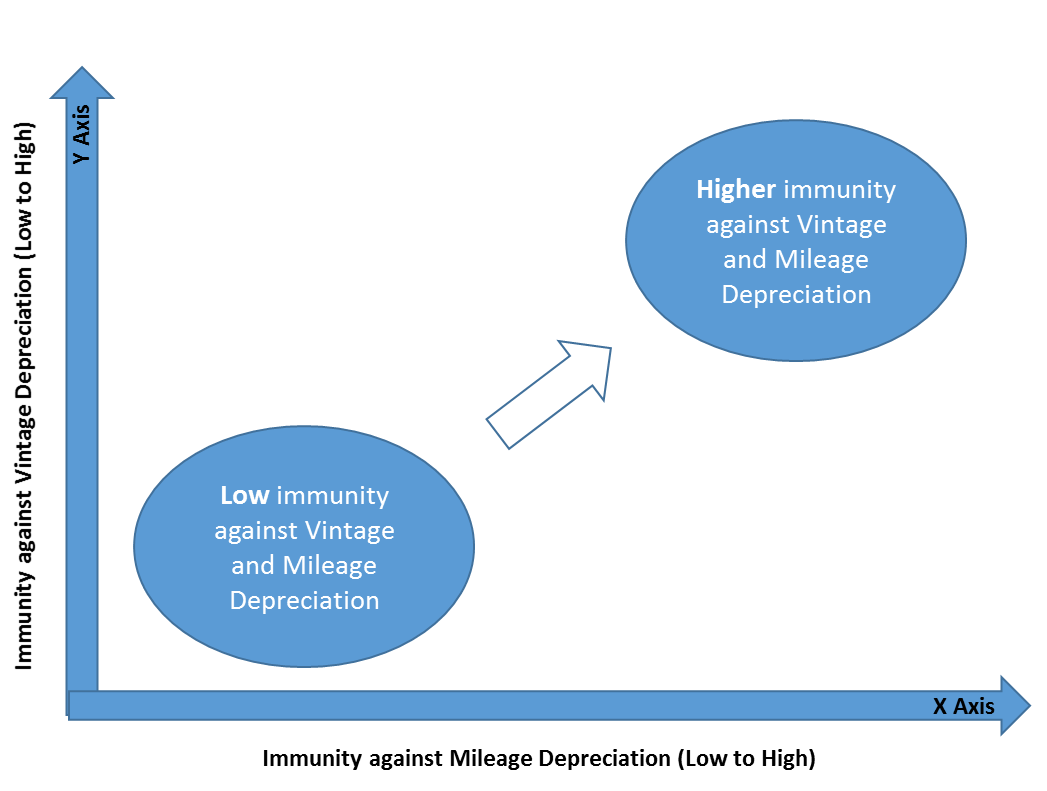
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Line** | | **Coefficients** | | | | |
| **Row** | **Column** | **p-value** | **DF** | **Term** | **Value** | **StdErr** | **t-value** | **p-value** |
| Price | Mileage | < 0.0001 | 223515 | Mileage | -0.375903 | 0.0030742 | -122.276 | < 0.0001 |
|  | | | | intercept | 57164.6 | 197.41 | 289.574 | < 0.0001 |

## Data evaluation methodology:

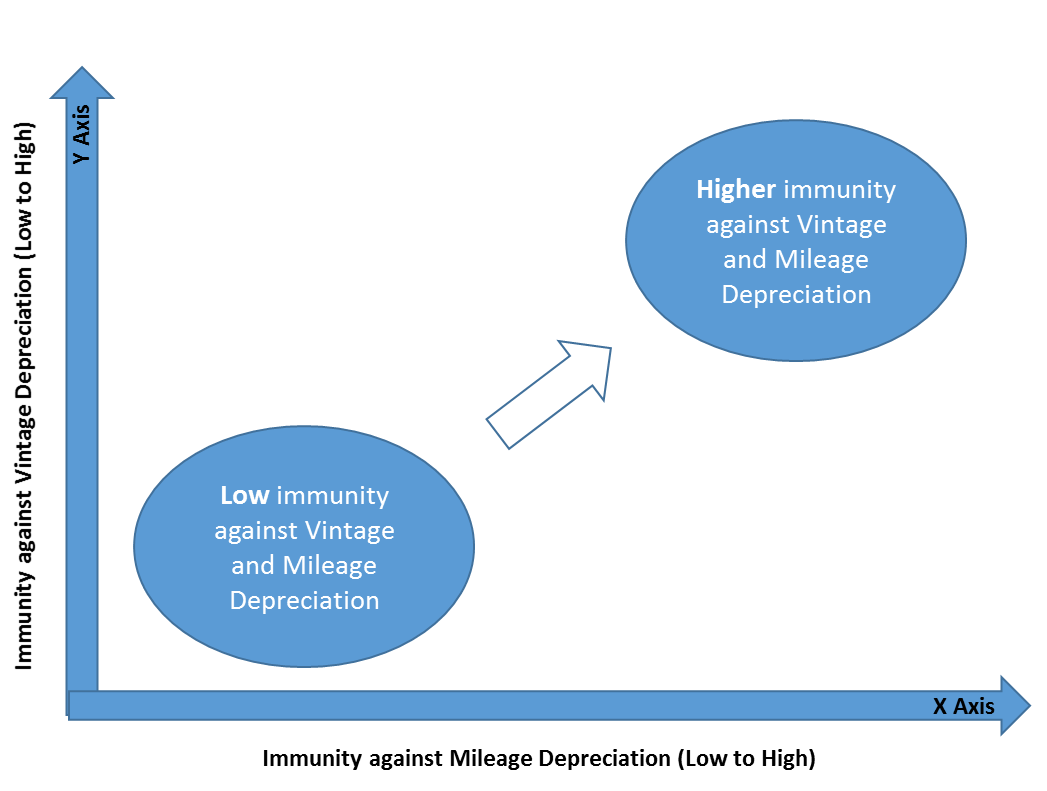
To benchmark car brands against one another a crosstab was plotted on a brands immunity with respect to mileage depreciation and vintage depreciation.

Mileage depreciation is defined as the price decline in car with increase in mileage on the car. Vintage depreciation is defined as decline in price with increasing years on road.

For this analysis the data is aggregated at car brand level

The axis on the cross tab signify how immune a car from particular brand is against depreciation w.r.t miles run and number of years on road

#### Axis calculations: X Axis



**Coefficient is 0**

**Coefficient is high in magnitude**

To calculate a brand’s position on the X Axis liner regression is used.

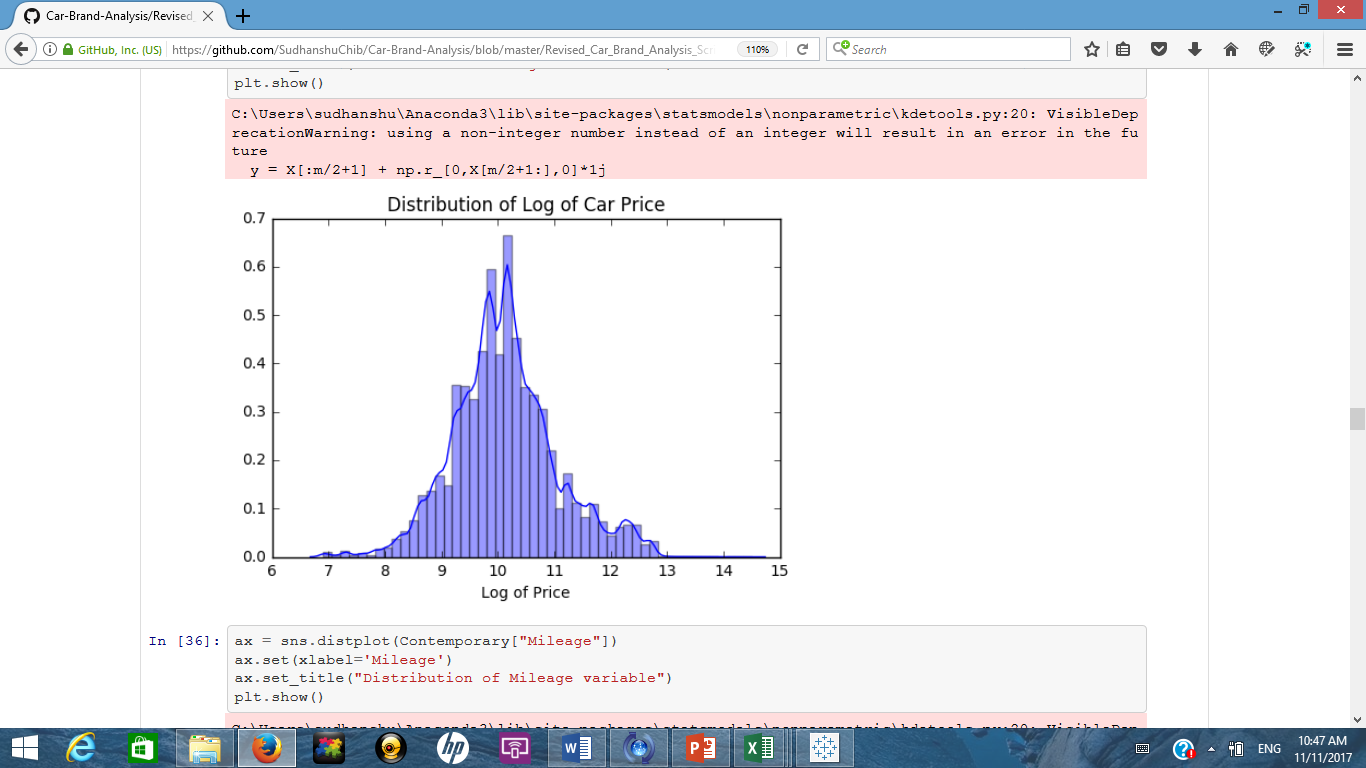
Log of car prices is regressed against log of number of miles on the car.

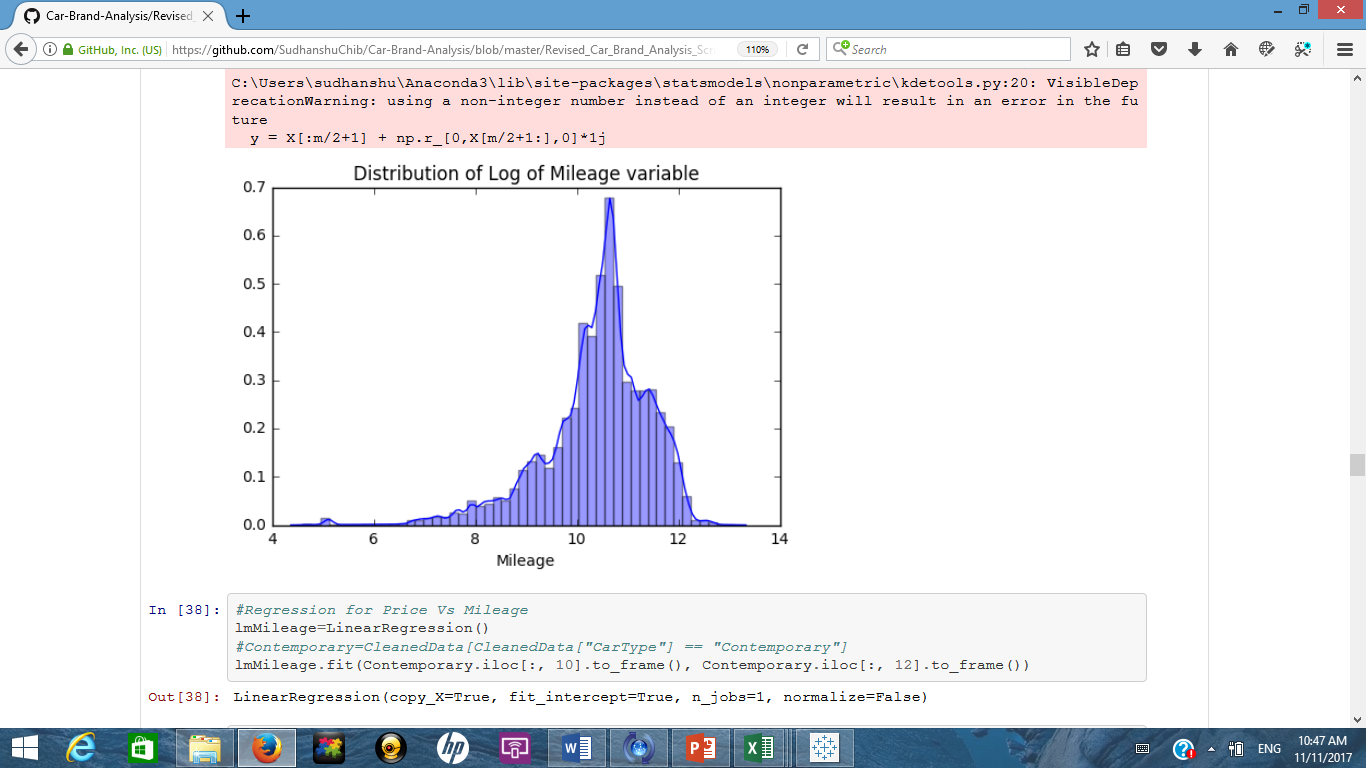
Coefficient of # of miles in the equation is plotted on the X axis for each brand

As the coefficients are negative (price of a car goes down with increase in number of miles driven) higher negative coefficients imply more depreciation for the brand.

Reason for using log of price instead of price:

One of the assumptions of linear regression is that underlying data is normally distributed. As we observed earlier that price and mileage were both not normally distributed. Taking log of price and mileage changes the distribution to resemble a normally distributed population





Overall coefficient for when price is regressed on mileage is : -0.57509752

Code is used to iterate through the data to subset it on car brands and run individual regression. To ensure that there are sufficient data points (theory says 30 or more) car brands that have atleast 100 car listings in the data are part of this analysis.

Calculated coefficients are x axis points for different car brands.

#### Axis calculations: Y Axis



To calculate a brand’s position on the Y Axis concept of single line depreciation is used

Average price of cars for a brand is calculated across defined buckets of 0-2 Yrs, 2-5 yrs, 5-10 yrs and 10 to 15 yrs.

Depreciation is calculated by dividing the average price of the subsequent bucket by the average price of the previous bucket and subtracting the result by 1 (the result is multiplied by 100 to express it in %). For eg. Avg price of Toyota cars in 0-2yrs bracket divided by Avg price of Toyota cars in 2-5yrs bracket accounts for depreciation across the first set of categories

Similarly depreciation is calculated and an average of these depreciation is taken as indicator for a particular brand

To adjust the axis (mark depreciation as low to high on Y axis) the final % value is subtracted from 50 (this is done to scale the axis). So if a car brand A depreciated on an average of 40% in value its position on y axis will be indexed to 10 whereas a brand that depreciated 25% will be at 25. Hence higher position is desired on the Y axis.

## Output and Insights

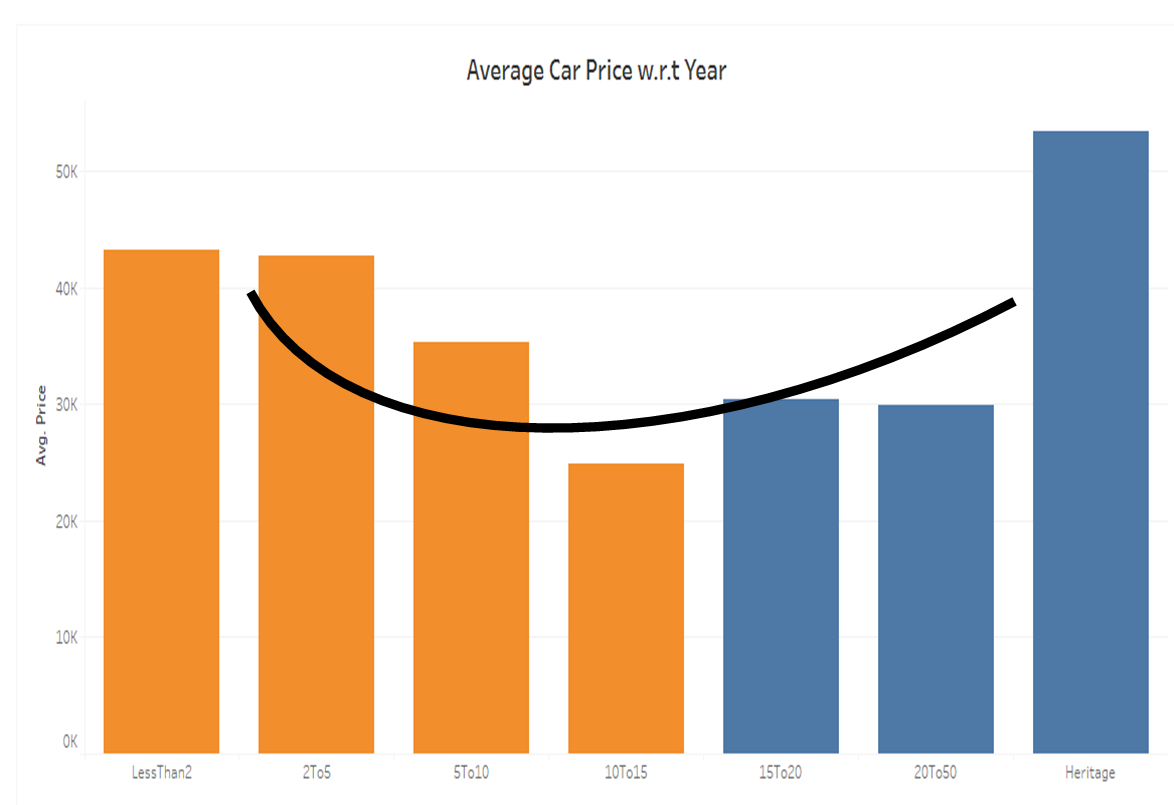
[Car Brands crosstab:](https://public.tableau.com/profile/sudhanshu.chib8525#!/vizhome/CarBrandsCrosstab_Revised/Crosstab1)

## 

Ford and Mercedes Benz seem to have maximum immunity against Mileage and Vintage depreciation

Among premium car brands Ferrari and Audi also seem to be doing well as compared to Lamborghini and BMW

Aston Martin, Rolls Royce, Bentley and Infiniti seem to be the least immune brands against depreciation



Generally the price of a car decreases over a period of time (10-15 yrs)

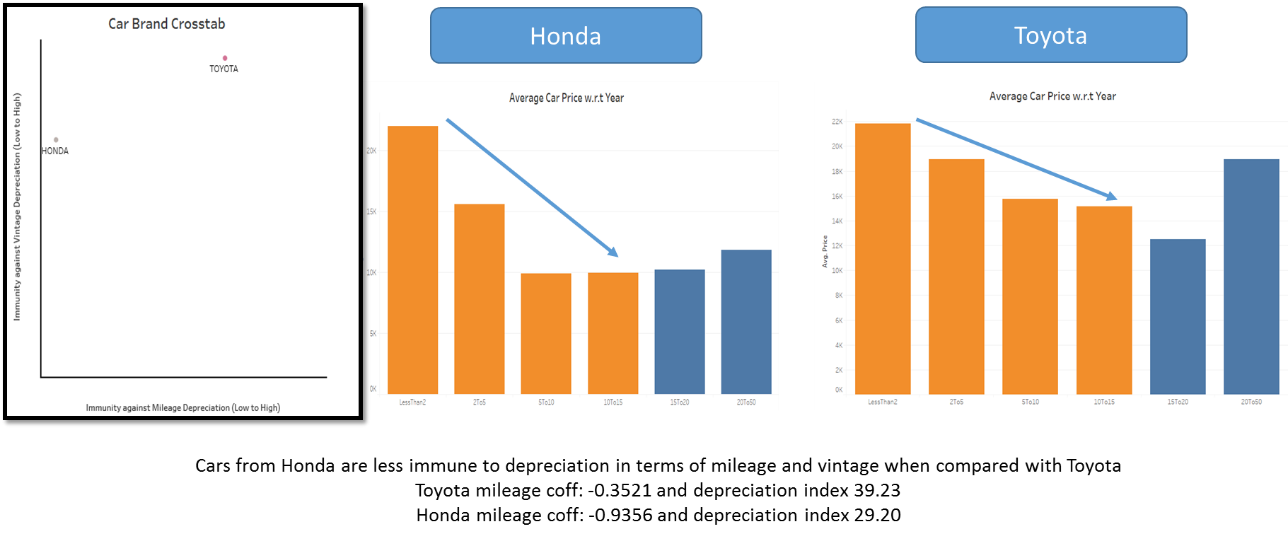
Some car brands appreciate in value post 15 to 20 yrs



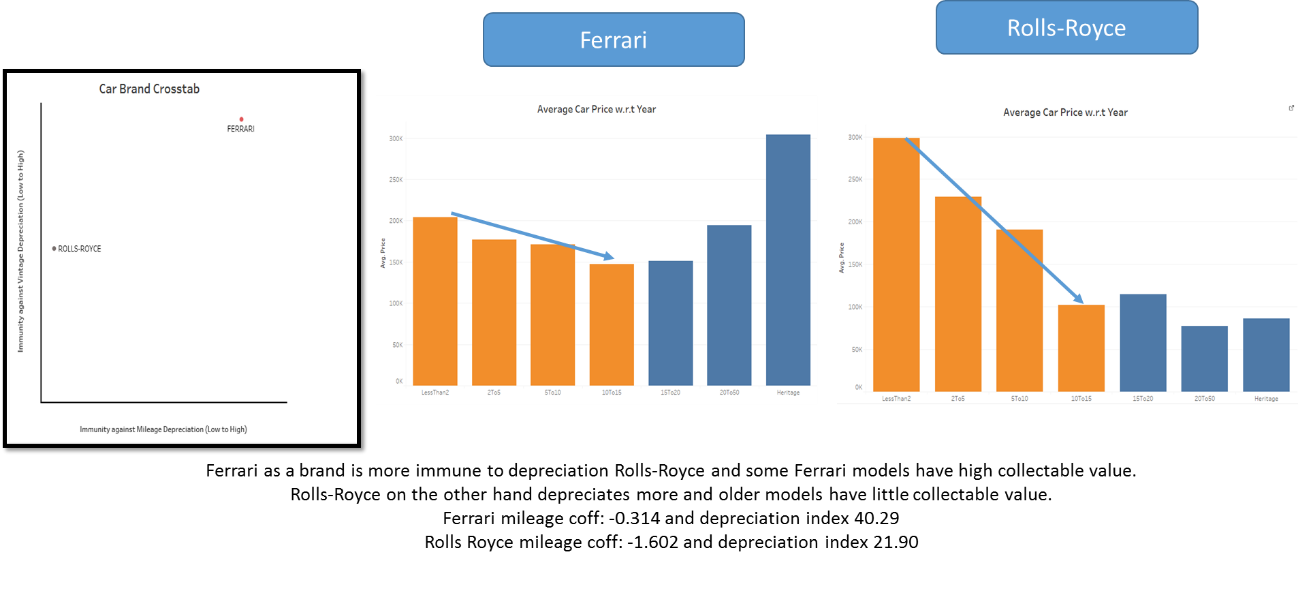
Average Price commanded by cars tends to decrease as number of on road miles increases

#### Brand Comparison:

Honda v.s. Toyota



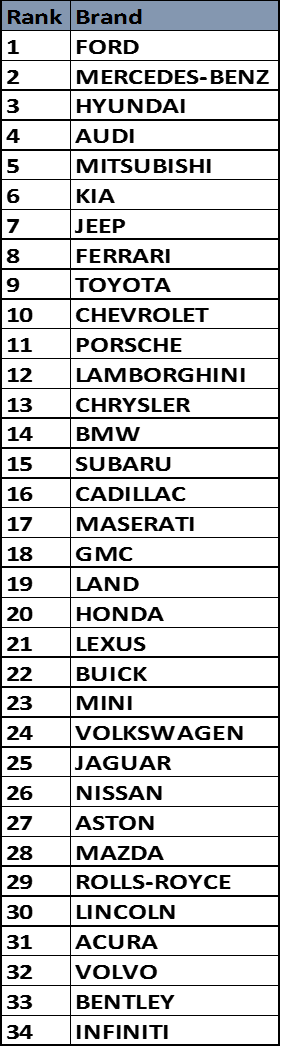
**Ferrari v.s. Rolls-Royce**



### Car Brands Ranking:

To rank car brands distance from origin was calculated for each brand on the crosstab.

Higher the distance better the brands ranking on the ranking system:



### Validation of Ranking System

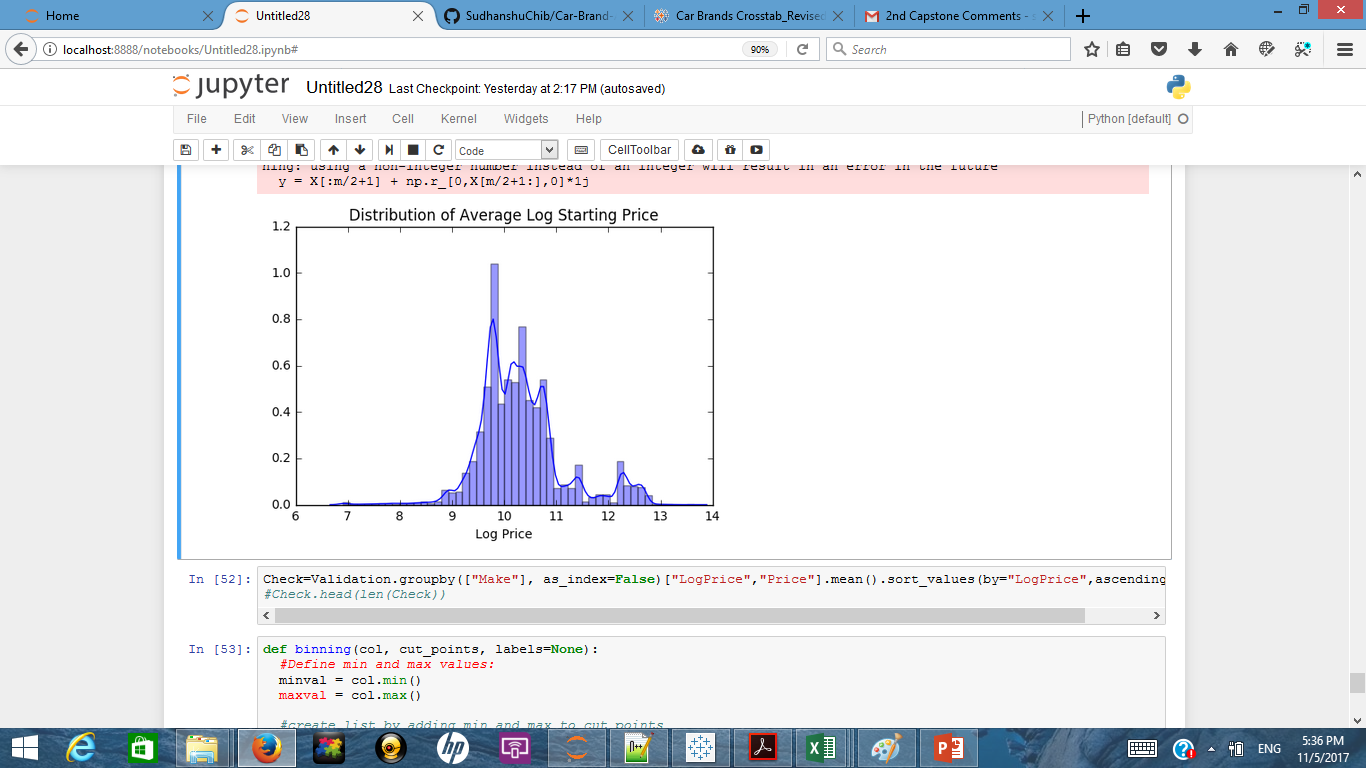
For all car brands the average price for the brand was calculated by pooling together cars that have been on road for less than or equal to one year

This dataset (car brands + avg starting price) had the following distribution:



As the distribution looked skewed to the right (more brands manufacture affordable cars) log of average price was considered

Below is the resulting distribution



With the population distribution now resembling more of normal distribution a quartile segmentation scheme was adopted

Car brands were divided into four buckets based on their log average car price

The four buckets are 1st Quartile: 75-100 percentile, $55,000 and beyond

2nd Quartile: 50-75 percentile, $35,000 to $55,000

3rd Quartile: 25-50 percentile, $24,000 to $35,000

4th Quartile: 0-25 percentile, Less than $24,000

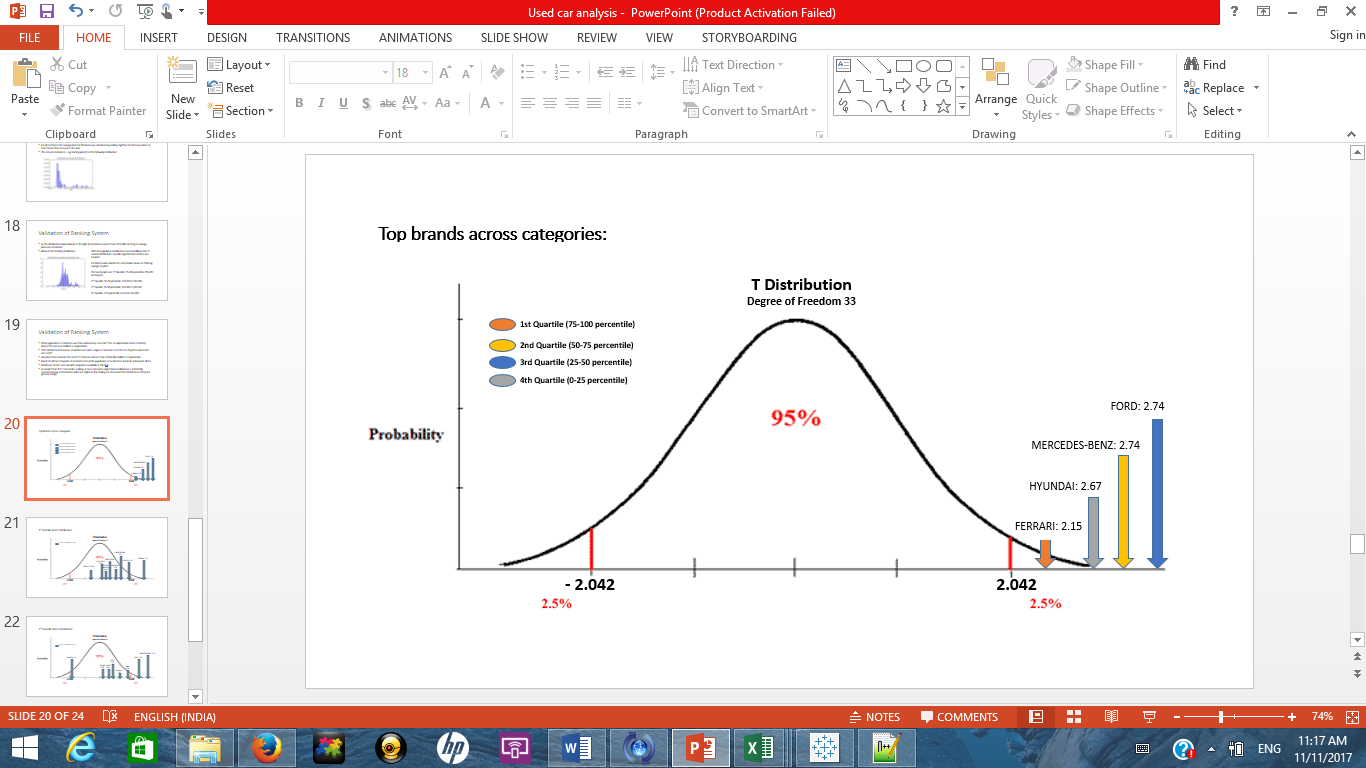
Entire population of car brands was then evaluated by a two tail T test on depreciation index to identify brands that are most resilient to depreciation. 95% confidence interval was considered and with a degree of freedom as 33 the cut off points were 2.042 and -2.042

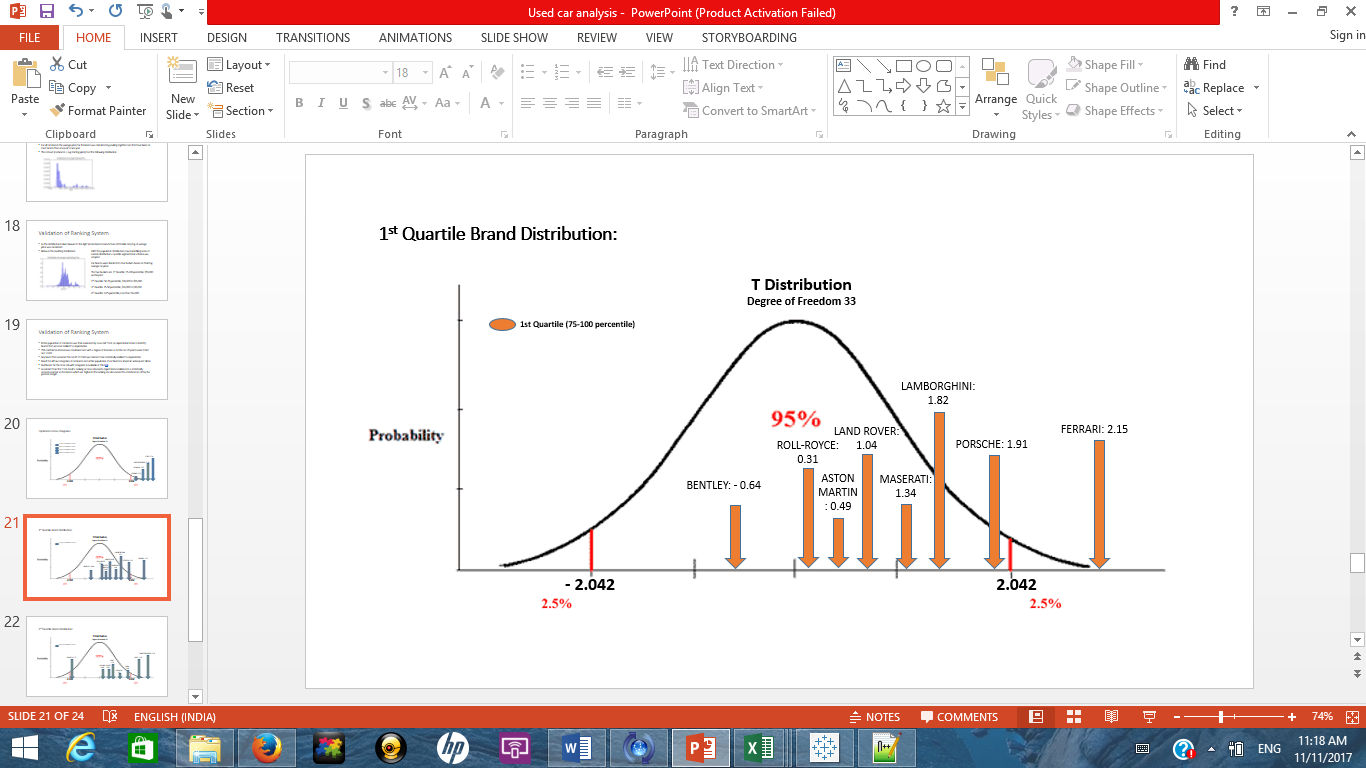
Any brand that exceeded the cutoff of 2.042 was termed to be statistically resilient to depreciation

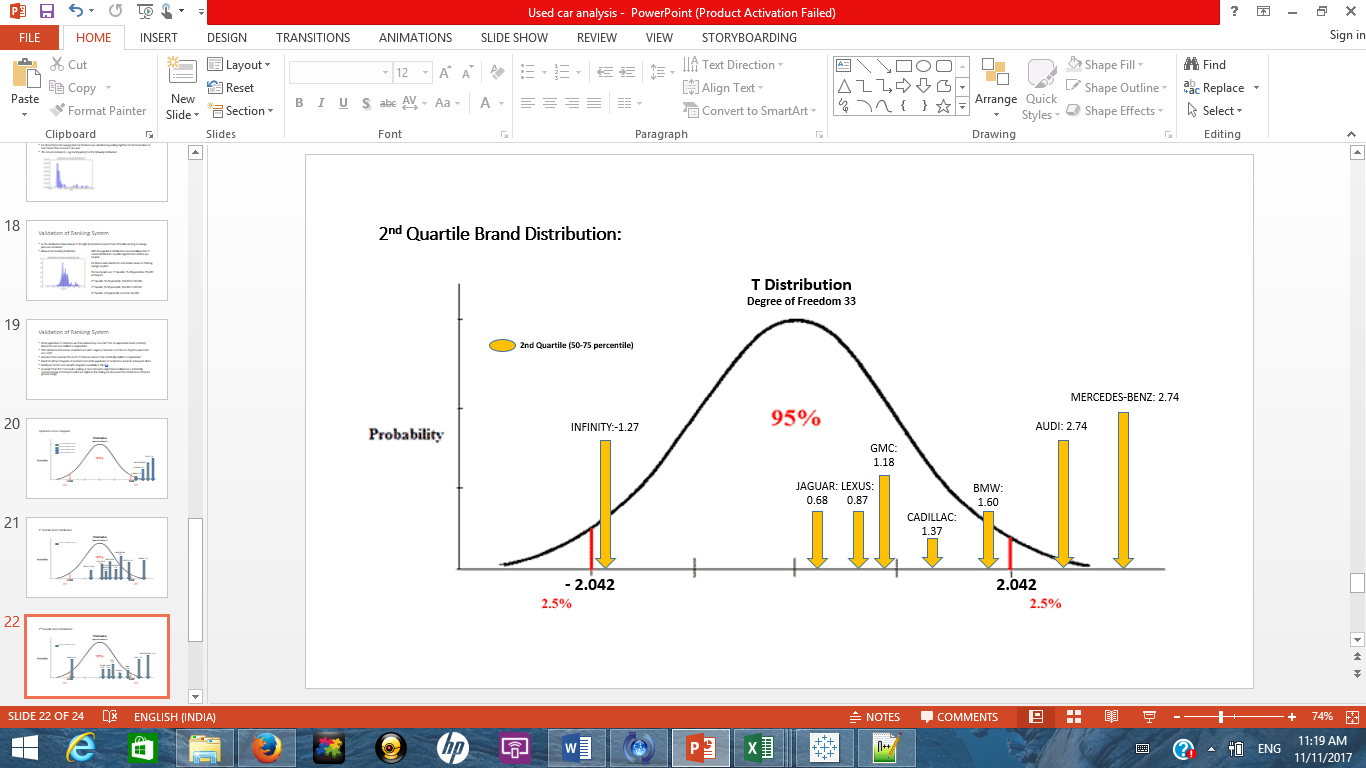
Result for all four categories of car brands and entire population of car brands is shared in subsequent slides

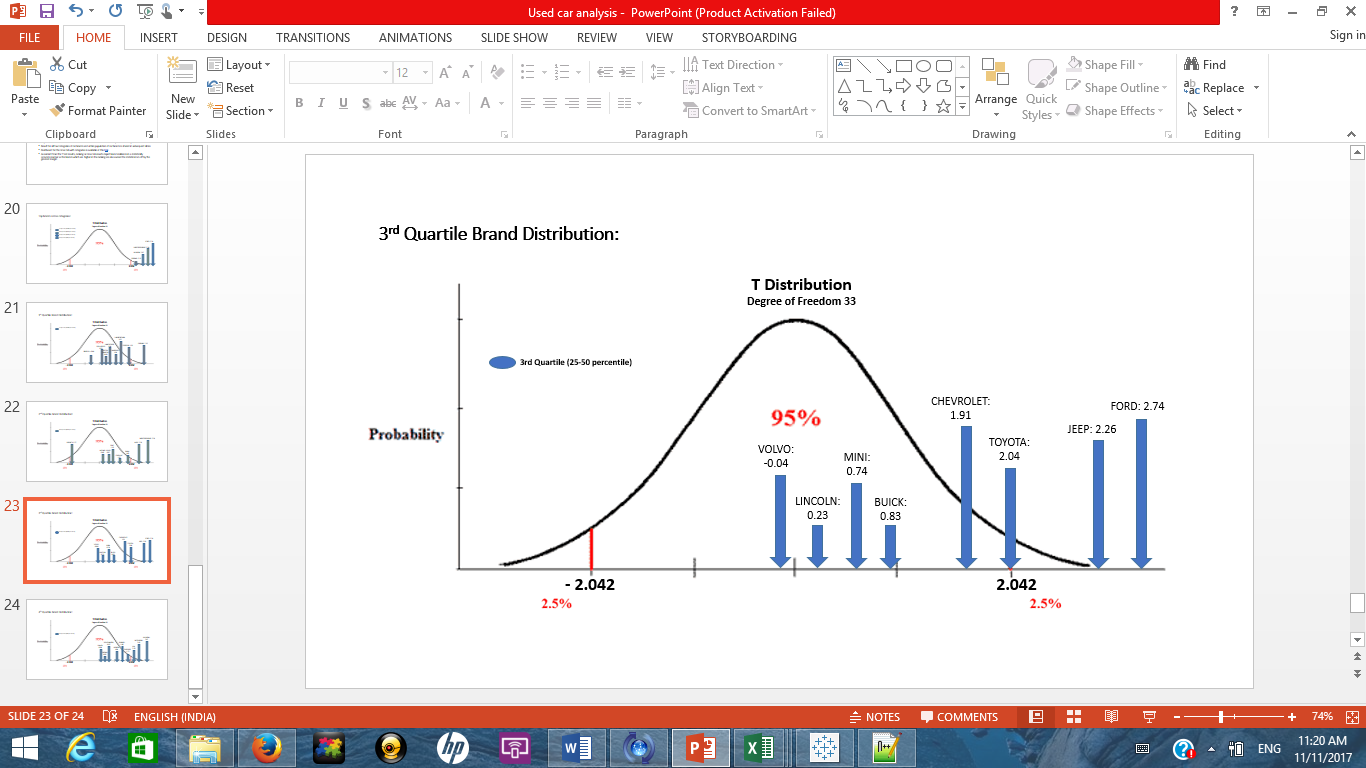
Dashboard for the cross tab with categories is available at this [**link**](https://public.tableau.com/profile/sudhanshu.chib8525)

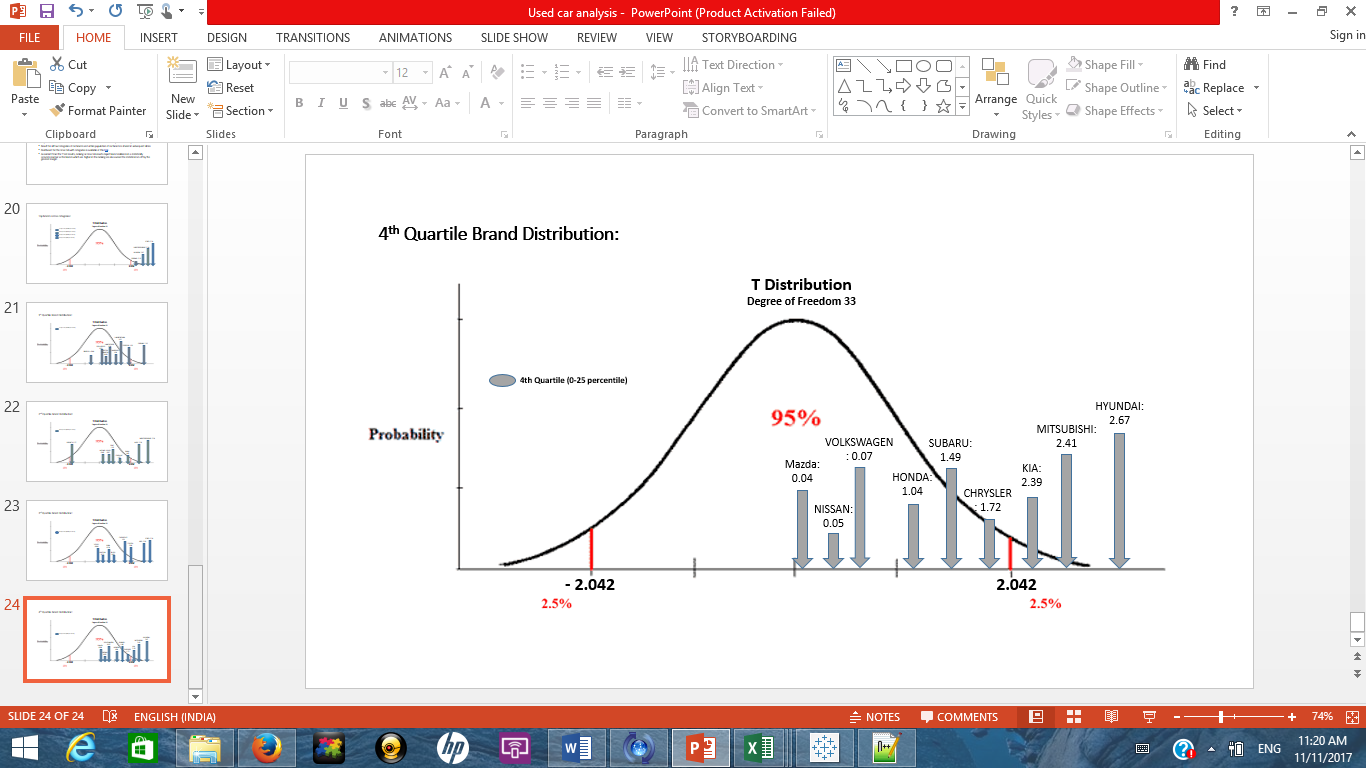
As evident from the T test results, ranking or cross tab results depict brand resilience in a statistically accurate manner as the brands which are higher in the ranking are also exceed the statistical cut-off by the greatest margin











## Assumptions:

1. eBay quoted prices are taken as representative price for various cars in this study. Generally the quoted prices are higher than the final selling price. Ideally the reserve price or actual selling price should have been considered.
2. Only mileage and year on road are considered as factors impacting car price. There are other known factors like type (sedan, suv, truck etc), color, number of previous owners etc. that are not accounted for in this study
3. Data points considered in this study are not collected as part of study hence there is no underlying mechanism for them to be homogenous. For e.g. there will be cars in this data set that are driven in rocky patches and hence would have greater wear and tear whereas similar model car driven on highways will have lesser damage. Ideally there should have been a way to homogenise the data population.

## Enhancements in future:

Capture additional information about the cars like car type, color, ownership information etc. and rerun the analysis with new available information.

Capture region specific information and provide city or area specific results. That account for terrain and region.