IFN465: Assignment 1 – Cornelis de Jager

# Task 1

### Questions 1

The proportion of cars that are kick is 12.949...%

### Question 2

Yes.

* We replaced ‘?’, ‘#VALUE’ and 0 as NaN
* PRIMEUNIT and AUCGUART were dropped due to only having a small amount of data
* WheelTypeID and PurchaseDate were duplicates of other features and were dropped
* ForSale was dropped because of the extreme data skew
* In TopThreeAmericanName Hyundai and Jeep were renamed into HYUNDAI and JEEP
* All other columns strings were converted to uppercase
* In nationality USA was replaced to AMERICA
* Transmission was renamed to Auto and Auto was set to 1, Manual set to 0
* In Color, NOT AVIL was changed to NaN
* Dropped VNST due to statisical insignificance
* Size was split inot Size and Body and Van was placed into the Body feature
* Any non-zero values in IsOnlineSale was set to 1
* Changed PurchaseTimestamp into days of week
* Replace any missing values not specifically taken care of previously to a value chosen randomly from the distribution of the column
* Columns with high correlation (98%) were averaged out into a new column. See correlation matrices below:

Acquisition Auction Price: 98% correlation

[[1. 0.98001885]

[0.98001885 1. ]]

Acquisition Retail Price: 98% correlation

[[1. 0.98551284]

[0.98551284 1. ]]

Current Auction Price: 97.5% correlation

[[1. 0.97646792]

[0.97646792 1. ]]

Current Retail Price: 98% correlation

[[1. 0.98311569]

[0.98311569 1. ]]

### Question 3

Yes.

In the histograms and the Box plots we can clearly see that there are many outliers as well as skewed data values. The outliers are most likely due to the skewness.

Both Fig. 1 and Fig. 2 shows in order lef-to-right, then top-to-bottom for the columns: ‘1. *Acquisition Auction Price*’, ‘2. *Acquisition Retail Price*’, ‘3. *Current Auction Price’* and ‘.4 *Current Retail Price*’.

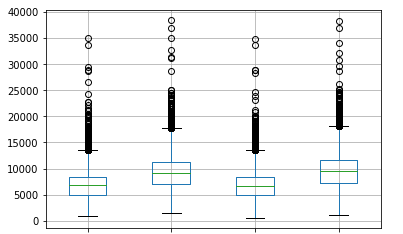


Fig. 1

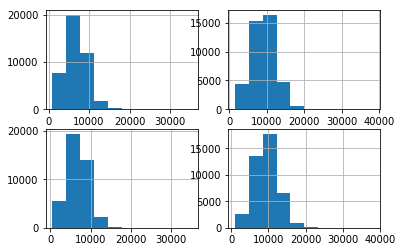
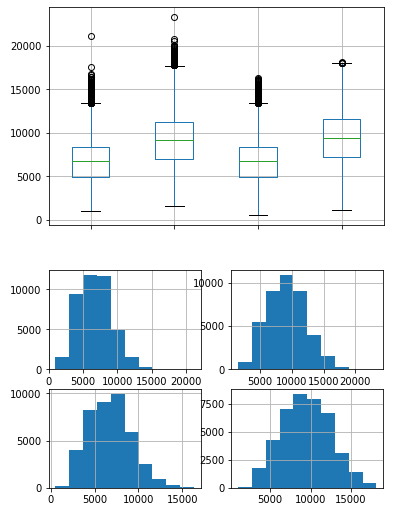


Fig. 2

It was decided that outliers will be removed first then a re-evaluation will be done. These are the results:



As seen above the distribution is allot better for : ‘3. *Current Auction Price’* and ‘.4 *Current Retail Price*’. However, some skewness still exist in : ‘1. *Acquisition Auction Price*’, and ‘3. *Current Auction Price’* . A transformation was done on these.

4.

5.

6.