**Exploratory Data Analysis - Inferential Statistics Report**

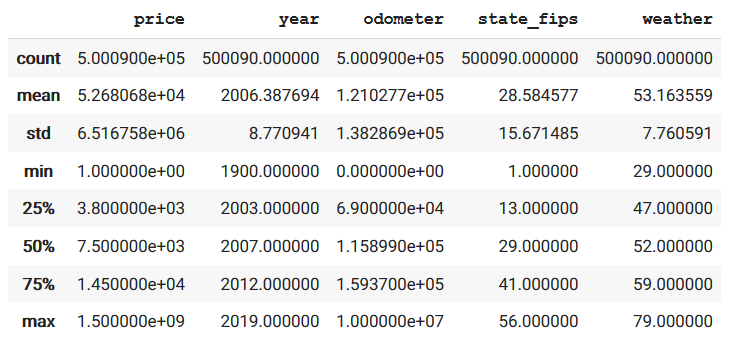
The purpose of this report is to Identify variables in the data to

* Answer to a project question.
* Identify strong correlations between pairs of independent variables or between an independent and a dependent variable.
* Practice identifying the most appropriate tests to use to analyze relationships between variables.

The main purpose of the project is to predict price of used cars. With inferential statistics, I will explore the data and figure out important variables for predicting price of used cars.

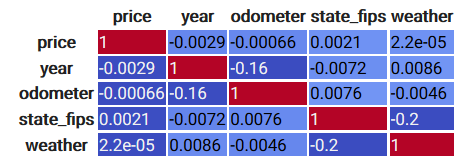
**Overall looking of the data**

For the beginning, overall looking of the data may help us gain an understanding about the dataset. Therefore, here is a brief description of the numerical variables.



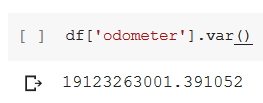
Because this is a a skewed dataset in terms of manufature year, focusing on the median value can be a good choice. The 50% row is the median value here. As we can see there, on average, used cars are 2007 model, and their price is around $7500. This chart also tells us that the variance of price and odometer is too high. We need to keep this mind while analyzing results of inferential statistics.

For the second step, looking at the correlation between numerical variables may be useful. Here is the correlation matrix:



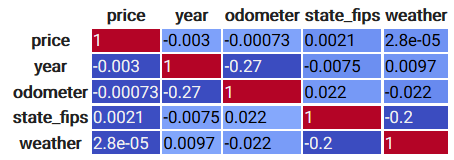
As we can see here, there is no a strong correlation between numerical variables. The highest correlation is between state\_fips and weather. However, state fips is not a numerical variable in its nature. Therefore, we can’t take this relation seriously.

As we said before, variance of odometer is too high. Let’s check actual variance value of the odometer:

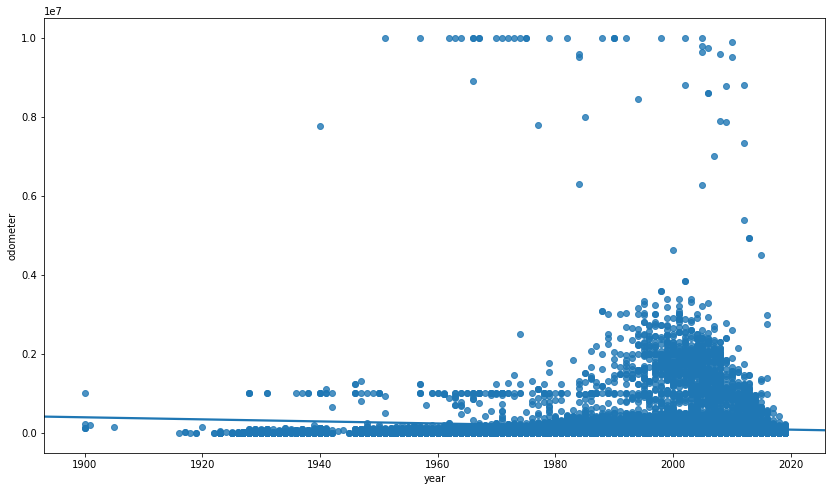


May be, decreasing the variance of the odometer makes it more informative. Therefore, let’s just take the cars that has odometer value is less than 300k miles. Then, let’s check the correlation matrix again.



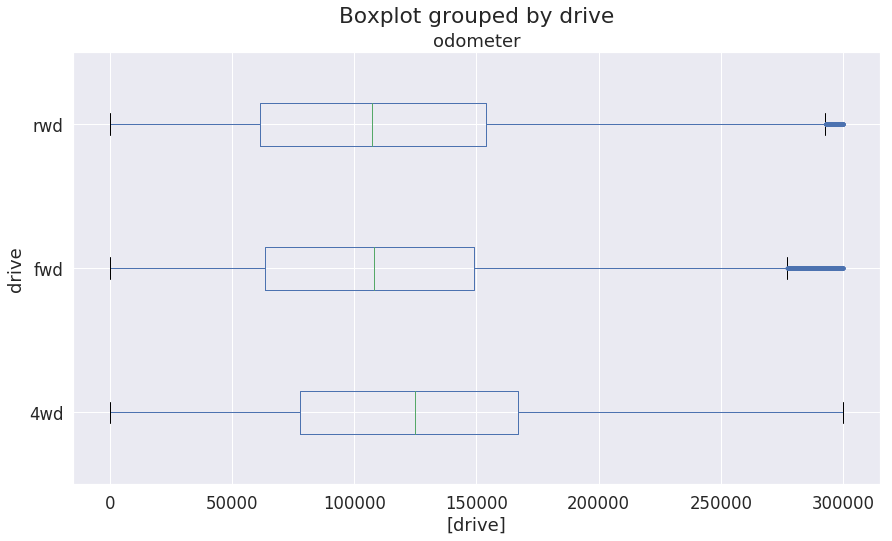


As we can see here, correlation values of odometer have increased with the modification. However, it’s still not a significant correlation.

To further examine the relationship between odometer and year, Let's graph the odometer and year on a scatter plot and add a regression line on it: 

As we can see from the above graph, we cannot say that there is a significant relation between odometer and year.

We can investigate the nature of the odometer variable further. For this purpose, a boxplot can be useful. We can odometers in category of drive type.

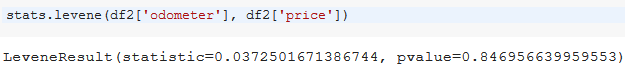


From this graph, we can say that on average 4WD cars can last longer in terms of milage. This is not a surprising result.

In conclusion, for now, we couldn’t identify a significant relation between numerical variables. In the next part, the relation between categorical variables will be investigated.

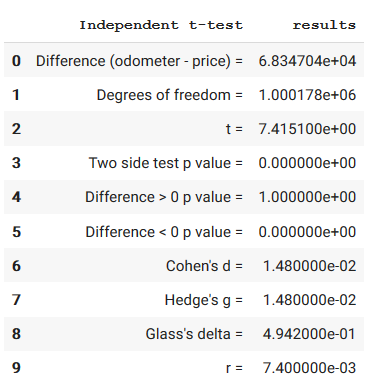
**Applying T-test**

At this point, looking at the relation between price and odometer can be useful. Applying independent t-test can be applied. However, before t-test, we need to make assumption check.



The Levene test is not significant meaning there is homogeneity of variances and we can proceed. If the test were to be significant, a viable alternative would be to conduct a Welch’s t-test.

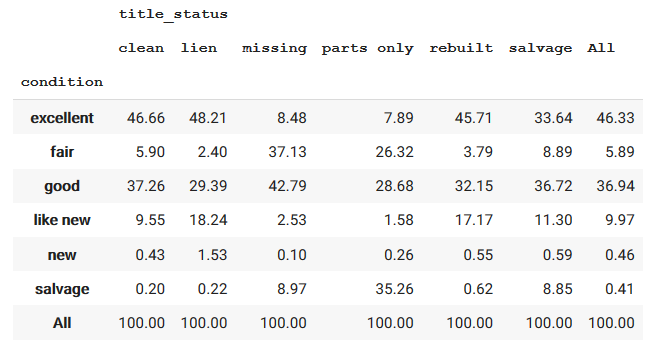
Here is the result of the t-test:

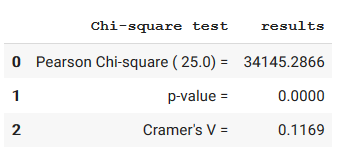


By looking at the table above, we see that p-value is 0. This indicates that odometer can be a strong predictor of the price of used cars.

**Chi-square Test of Independence using Researchpy**

Chi-square is an appropriate method for investigating the relation between categorical variables. As beginning, the relation between ‘title\_status and ‘condition variables will be investigated. As we can see from the

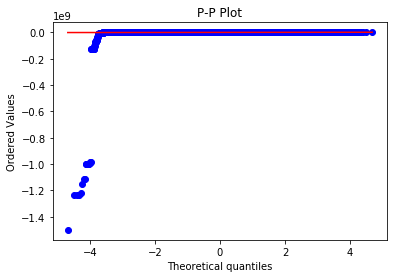




The chi-square test tells us there is a significant relation between title status and condition, which makes sense.

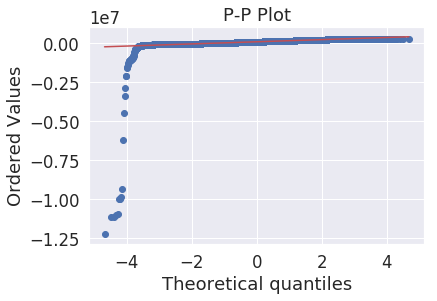
### Normal distribution of residuals

We also need to check normality assumption for odometer and price.



The p-p plot shows that the data doesn't satisfy normality. We have problem with initial values. May be, scaling the data can solve this issue.

Here is the scaled plot:



This is slightly better than the previous situation but we still have violation of the normality assumption here. We can also perform the Shapiro-Wilk test for checking normality.

stats.shapiro(diff) : (0.0014088153839111328, 0.0)

The output is not labeled, but the first value is the Wilk test statistic and the second value is the p-value. Since the test statistic produce a significant p-value, the data is indicated to be not normally distributed.

In conclusion, in order to predict used cars’ price, odometer and title status can be good predictors. However, this relation needs further investigation. The analysis also showed that title status and condition of the car are significantly related. This relation may be useful for predicting cars’ prices. We need to keep in mind that our numerical variables do not have a notable correlation among them. We need different methods that can utilize non-correlated variables while making prediction.