Question 2 Craig List DataSet

Part (a).

Firstly we have to clean data and we followed following steps to clean:

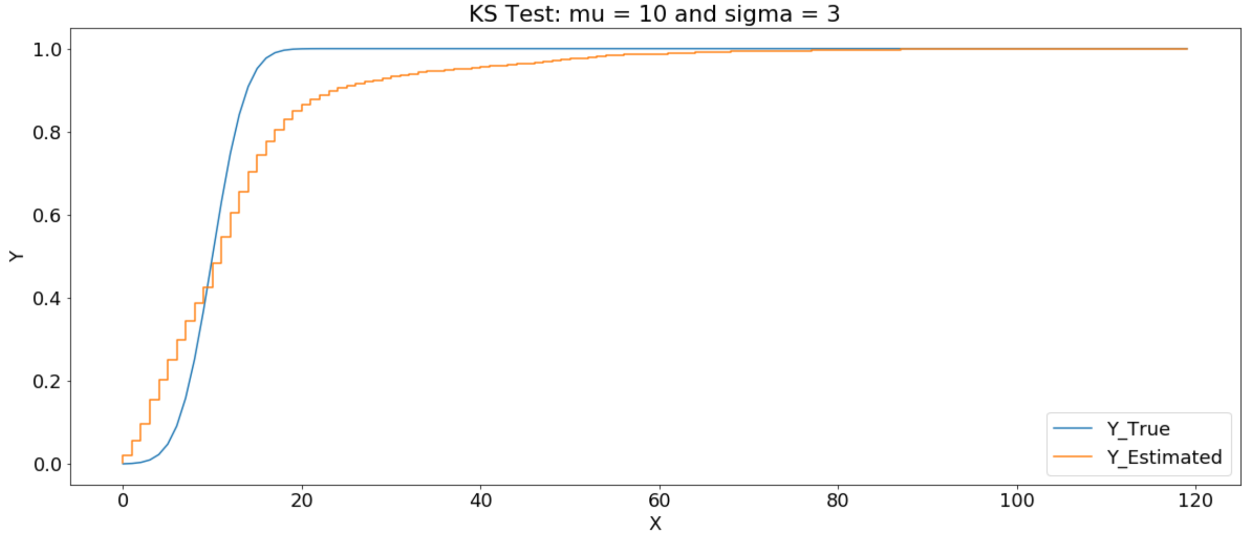
- Removed null values from year column
- Filtered year based on the range [1885, 2019]
- Calculated age by subtracting year from 2019

After this we implemented the KS-Test as taught in class. Following were the steps followed for implementing:

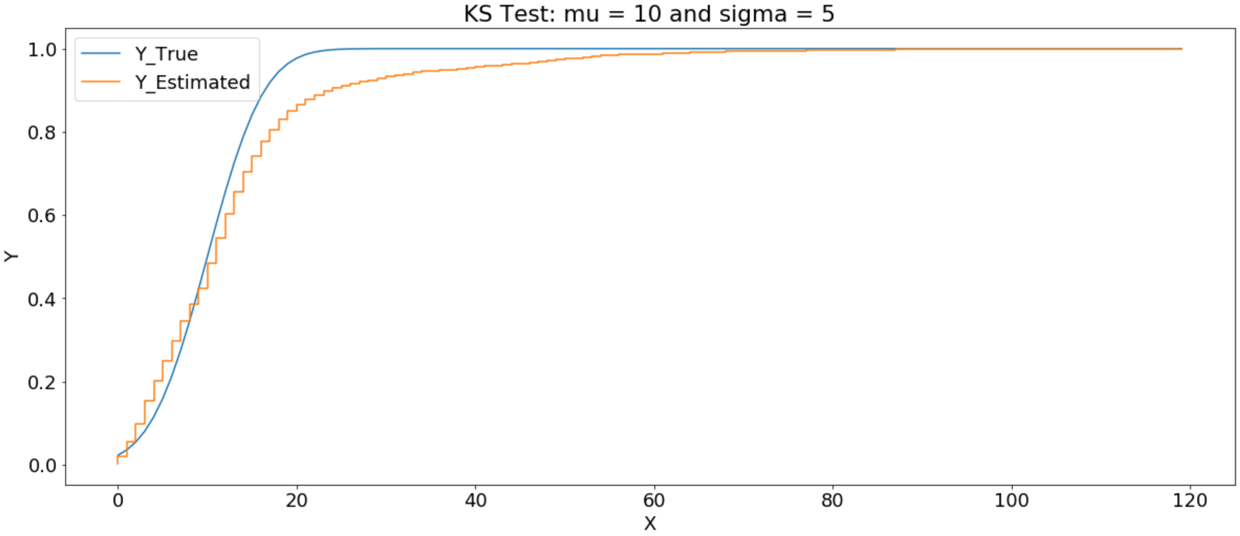
- Sorted the sample input
- Calculated value of column ‘x’ by getting unique values from sample input
- Calculated value of ‘F(x)’ by using normal distribution function for given mu and sigma values at each sample input.
- Calculated value of ‘F(y)+’ and ‘F(y)-’ from sample input
- Completed table by calculating |F(x) - F(y)+| and |F(x) - F(y)-|
- Finally calculated value of max of (|F(x) - F(y)+|, |F(x) - F(y)-|) to compare with given alpha to accept or reject the hypothesis.

Following were the results:

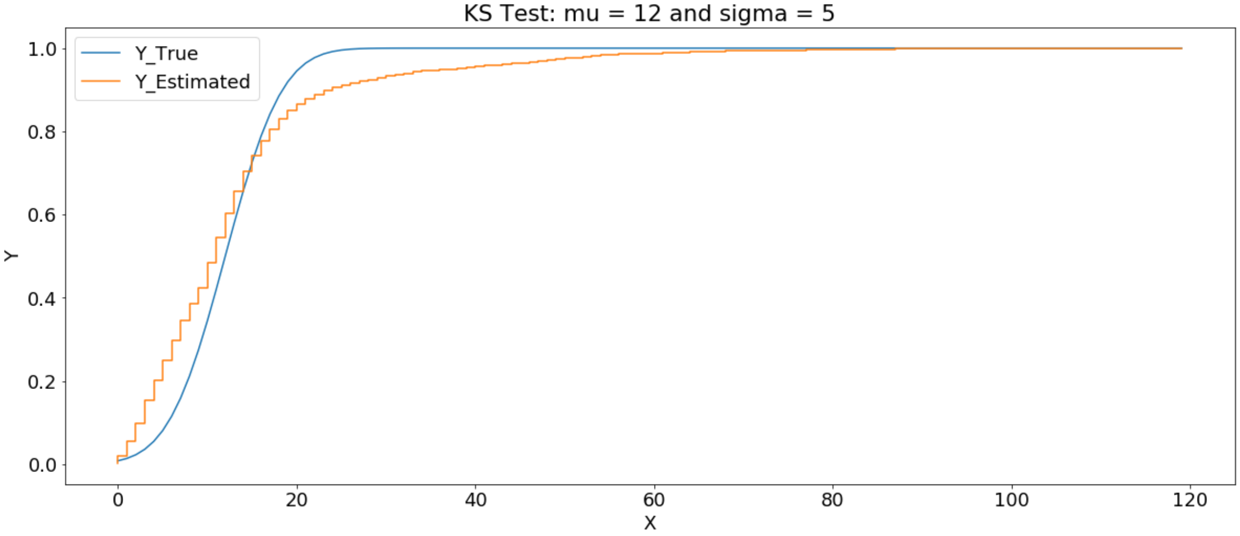
- H(o) = Age of cars available in Craigslist follows normal distribution. When N(10,3) , we got Max of |Fx - Fy-| and |Fx - Fy+| is **0.305105313955.** Which was more than given alpha(0.15). **Hence H(o) is rejected.**



- H(o) = Age of cars available in Craigslist follows normal distribution. When N(10,5) , we got Max of |Fx - Fy-| and |Fx - Fy+| is **0.185426196784.** Which was more than given alpha(0.15). **Hence H(o) is rejected.**



- H(o) = Age of cars available in Craigslist follows normal distribution. When N(12,5) , we got Max of |Fx - Fy-| and |Fx - Fy+| is **0.139219033537.** Which was less than given alpha(0.15). **Hence H(o) is accepted.**



Part (b).

Firstly we thought of cleaning data and to do this we checked price column and found that no value was null. Hence no further cleaning was required.

After this we implemented 2-sample Wald's Test as taught in class. Following was the implementation:

- Created two variables to store sample points of distribution for black cars and blue cars.
- Calculated mean and variance for both the samples(blue and black cars) and applied 2 sample Wald's test to check if H(0) = Black color cars have same value as blue color cars and H(1) = Black color cars do not have the same value as blue color cars

Following were results:

- We got value of **|W| = 0.5764777891123762**. As |W| is less than given alpha value of 1.96. Hence we **accepted** the H(0) = **Black color cars have same value as blue color cars**.
- Hence the hypothesis presented by question (**Black color cars are more valuable than blue color cars**) was **rejected.**

Part (c).

Firstly we have to clean data and we followed following steps to clean:

- Removed null values from year column
- Filtered year based on the range [1885, 2019]
- Calculated age by subtracting year from 2019

Now we created two variables X and Y. X contains values of age for transmission == manual and Y contains values of age for transmission == automatic. X and Y denotes the sample points for age of manual transmissions sold and age of automatic transmissions sold respectively.
To check that they follow same distribution we applied permutation test using value of N! = 100 and 10000.
We implemented permutation test as follows:

- Calculate value of T_obs
- Concatenated X and Y for permuting samples
- Now for value of N! = 100 and 10000 we calculated p values as taught in class

Following were results:

- H(0)= Distribution of age of automatic transmissions sold is the same as that of manual transmissions sold. For N! = 100 p-value was **0.0**. As p-value is less than given value of threshold(0.05). **Hence reject hypothesis.**
- H(0)= Distribution of age of automatic transmissions sold is the same as that of manual transmissions sold. For N! = 10000 p-value was **0.0**. As p-value is less than given value of threshold(0.05). **Hence reject hypothesis.**

1st New Hypothesis:

Resale Price is a linear combination of Resale Period, Distance Travelled, Latitude, Longitude and Number of Cylinders

Usefulness:

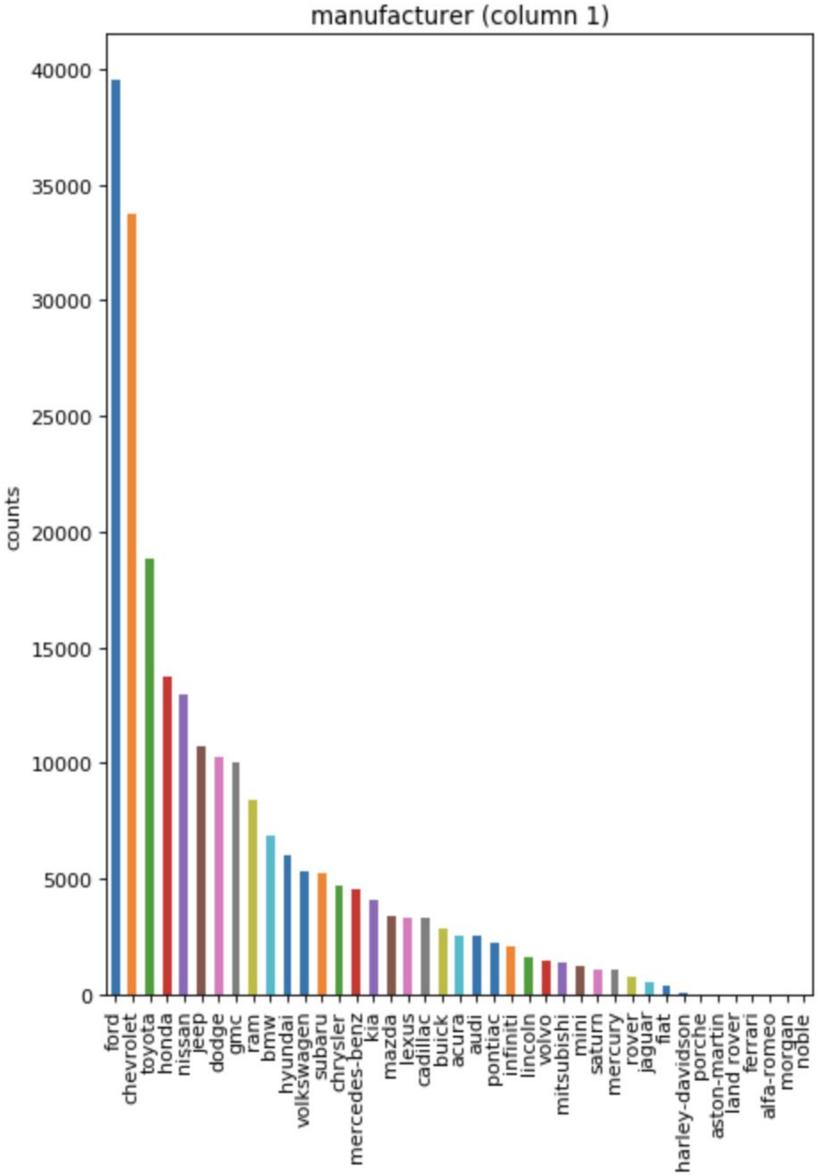
"What should be a good price which a seller can put for his car?"
If our hypothesis is accepted, we can use it to suggest prices to the car sellers on Craiglist. This can be also be useful for buyers where they can see "What is a reasonable buying price for a car with a given configuration, based on past data?"

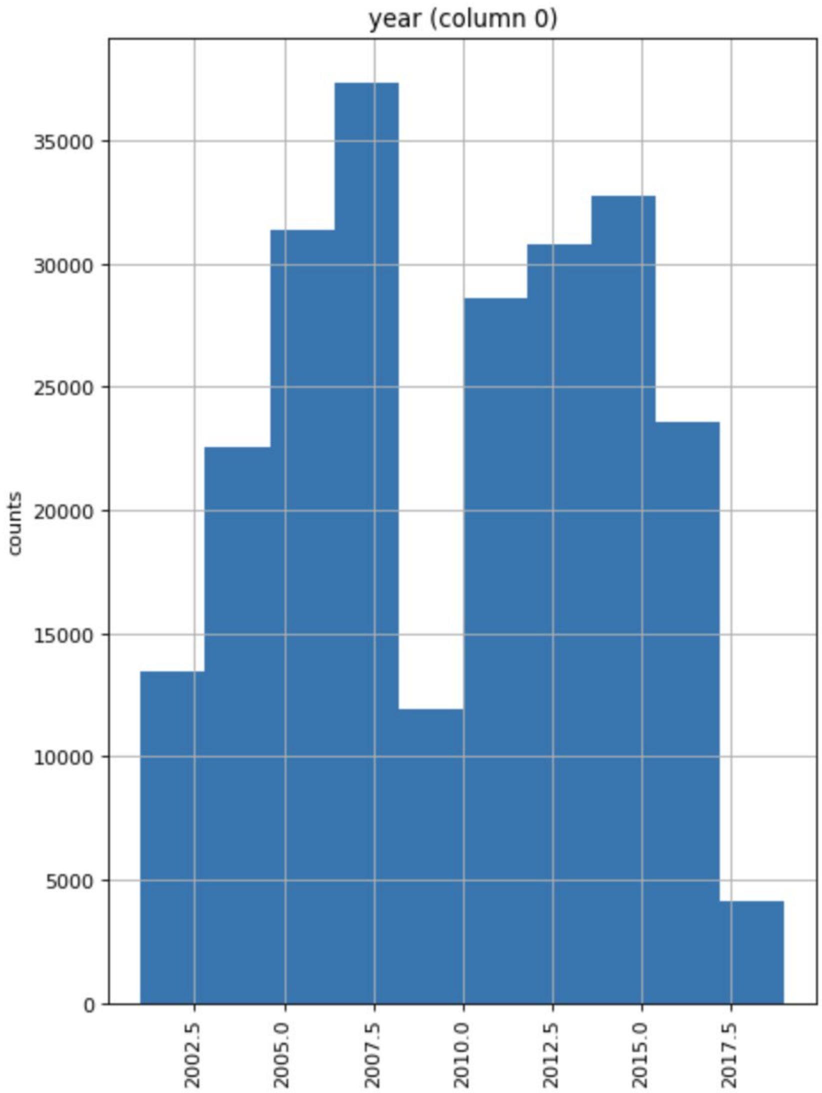
Work Summary:

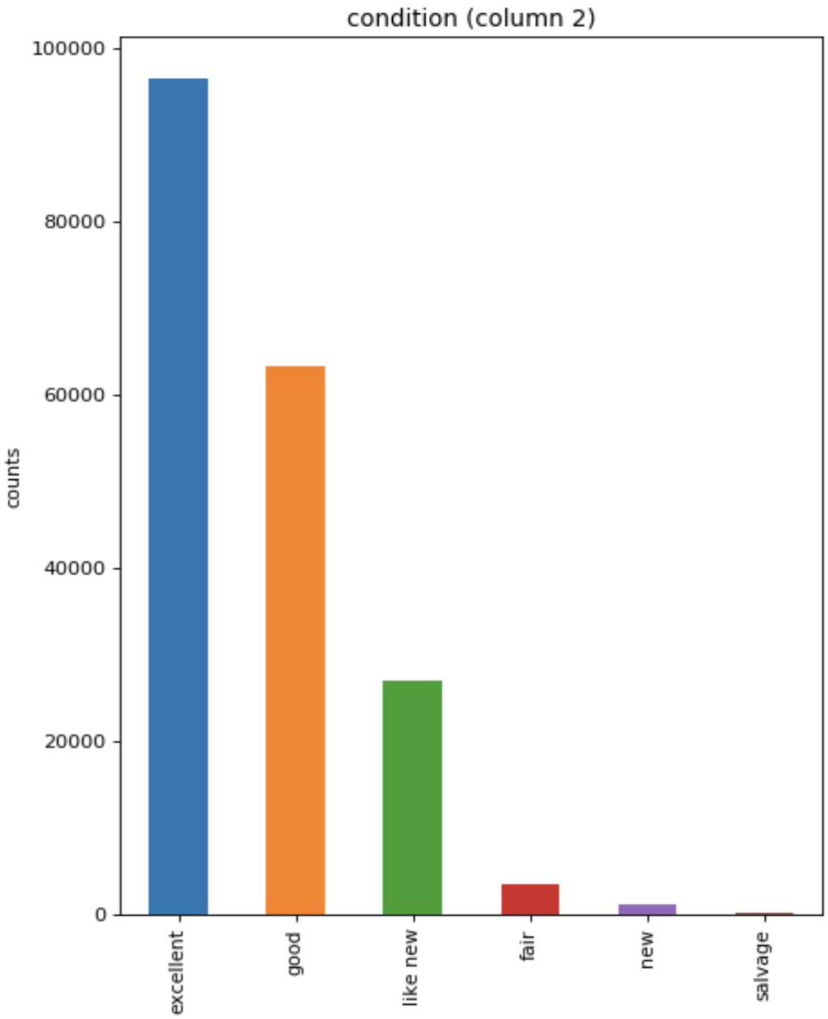
We have cleaned the full dataset and played around with different categorical and numerical features. We have found these 5 features to be highly correlated to the price which sellers put up on Craiglist.

We have come with a linear regression model which uses 80% train data and 20% test data. The model has a mean absolute error of 3839 dollars using our own implementation.

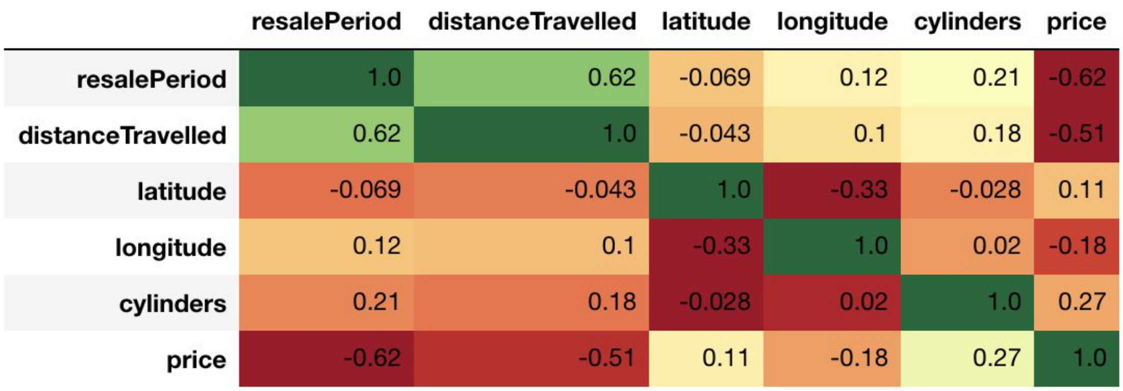
Analysis Graphs:




year (column 0)
counts
35000
30000
25000
20000
15000
10000
5000
0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5



Features Correlation:



	resalePeriod	distanceTravelled	latitude	longitude	cylinders	price
resalePeriod	1.0	0.62	-0.069	0.12	0.21	-0.62
distanceTravelled	0.62	1.0	-0.043	0.1	0.18	-0.51
latitude	-0.069	-0.043	1.0	-0.33	-0.028	0.11
longitude	0.12	0.1	-0.33	1.0	0.02	-0.18
cylinders	0.21	0.18	-0.028	0.02	1.0	0.27
price	-0.62	-0.51	0.11	-0.18	0.27	1.0

Expectations from Correlation
1. As “resalePeriod” increases, price decreases. Hence the correlation should be negative.
2. As distance traveled increases, the price should decrease. Hence the correlation should be negative
3. As the number of cylinders increases, the price should increase. Hence the correlation should be positive.

Observations from Correlation
We observe that our correlation values follows our expectation in terms of sign and magnitude.

We tried finding correlation with numerous categorical attributes as well by converting categorical attributes to integers. However, there was no such correlation.

Linear Regression Model:

Taking log of price for prediction
Our linear regression predicts negative values as well. We should ideally take log of price and then apply linear regression on it. This would keep the price always positive.

Features used = ['resalePeriod', 'distanceTravelled', 'latitude', 'longitude', 'cylinders']

Training Data Size (189136, 6)
Training Error 3825.2091055

Testing Data Size (47284, 6)
Testing Error 3839.85745067

2nd New Hypothesis:

Craigslist website traffic per hour can be estimated by the time series prediction of the number of ads posted per hour in the past.

Background & Observation

Traffic of a website is usually estimated by the number of visits and page views on the website per unit of time.

On plotting the number of ads posted every hour, we observe that the frequency of ads posted has increased exponentially with time over the last 2 years. This, in turn, causes the traffic on the website to increase drastically over time.

Assumption - More the number of ad posts, more is the traffic on the website.

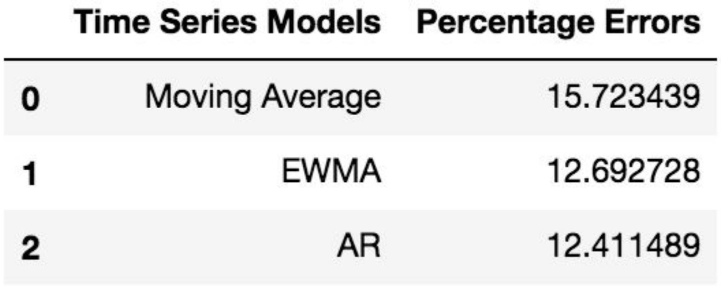
Usefulness

This prevents the downtime of Craiglist servers due to unexpected ad posts(writes). An effective estimation of Craigslist website traffic can help them scale their infrastructure to be reliable and fault tolerant to the upcoming large traffic.
Craigslist can plan it's deployment and maintenance cycles well in advance
Owing to the knowledge of increase in traffic, business actions (Ex. Ad targeting) can be scaled accordingly.

Work

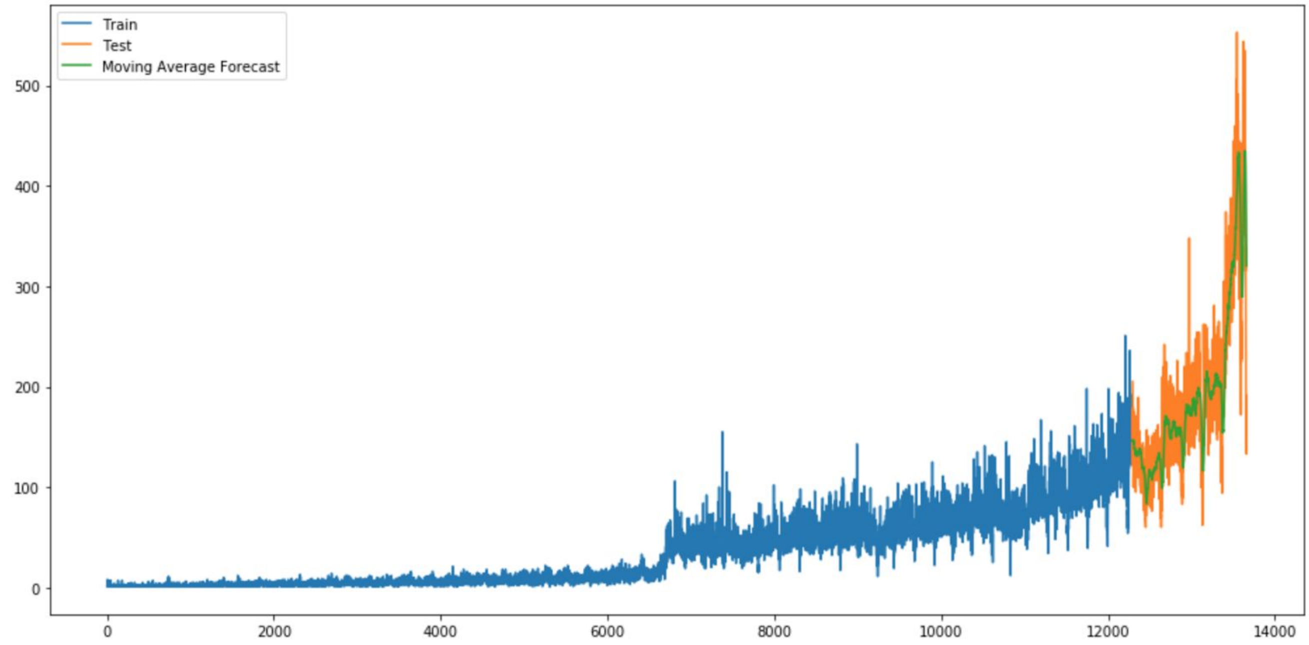
Our estimation particularly uses 3 Time Series Techniques - Moving Average, EWMA and Auto Regression to answer "What is the number of ads posted on Craigslist every hour?"

Results Found:

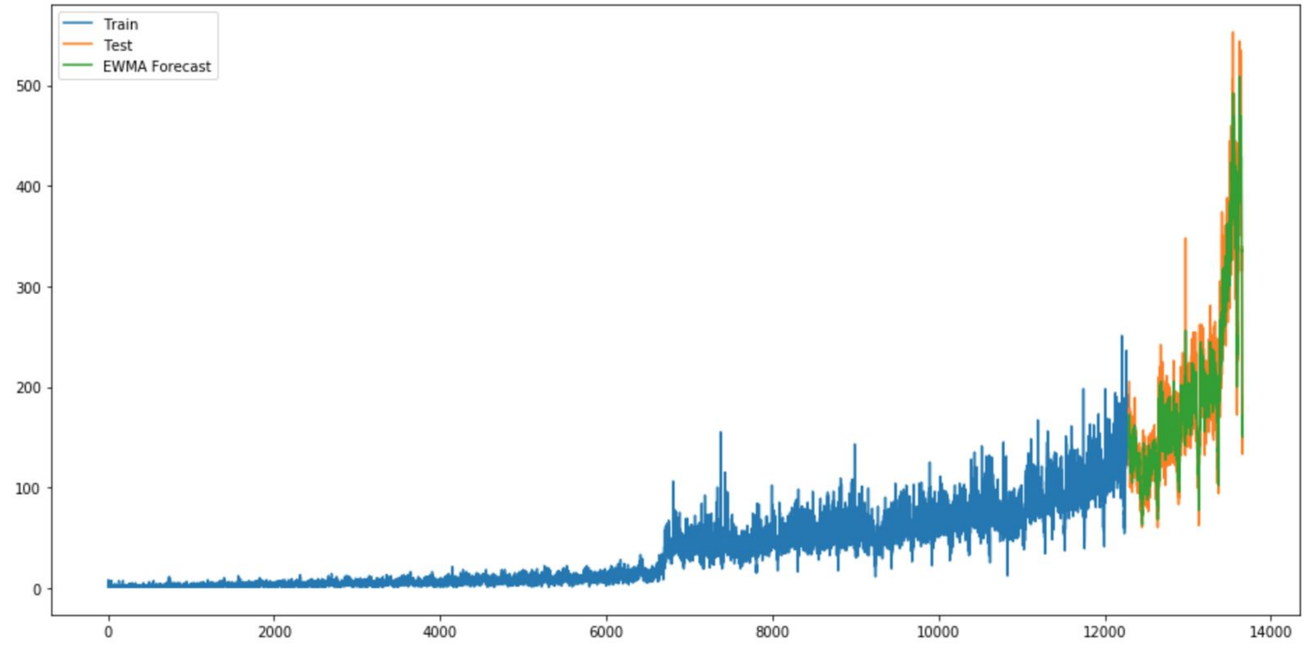


	Time Series Models	Percentage Errors
0	Moving Average	15.723439
1	EWMA	12.692728
2	AR	12.411489

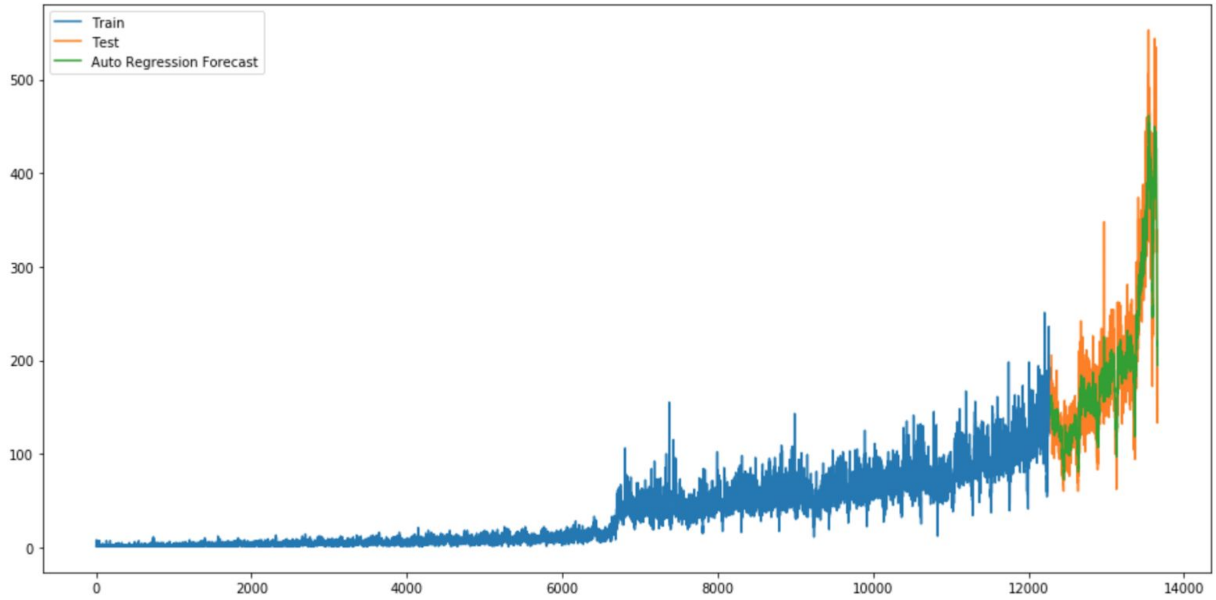
Moving average Forecast



EWMA Forecast



Auto regression forecast


Train
Test
Auto Regression Forecast