8th International Conference on Research Developments in Applied Science, Engineering & Management (AEM- 2020)
Osmania University Centre for International Program, Osmania University Campus, Hyderabad (India)
18th January 2020 www.conferenceworld.in ISBN : 978-81-944855-0-6

Statistics for Data Science

Arnav Oberoi
(Co-Author)
Department of Computer Science and Engineering
The NorthCap University
Gurugram, India

Aryan Sharma
(Co-Author)
Department of Computer Science and Engineering
The NorthCap University
Gurugram, India

Om Sehgal
(Co-Author)
Department of Computer Science and Engineering
The NorthCap University
Gurugram, India

ABSTRACT

The field of statistics helps us analyze our data in a more efficient manner by helping us correctly study and present the outcomes.Inferential statistics helps us learn how the population might think or behave inferring from the sample data's descriptive statistics.It helps in classifying between reasonable and doubtful claims. It is important for a person willing to work in the field of Data Science to know about the basic and high-level statistics for model evaluation.

Keywords—Statistics; Data Science;Artificial Intelligence; Data Preprocessing

I. INTRODUCTION

Statistics is a form of data analysis which helps us evaluate big data using quantified models by gathering,analyzing and drawing conclusions to the data set. Descriptive statistics helps us summarize a given data set. Inferential statistics involves the study of patterns and analyzing the behavior to comprehend the data in depth. In statistics, we use samples to draw inferences about the larger set(population). We will be working on improving the statistics of the data set. Statistical modelling is all about finding the relationship between the variables whilst making accurate predictions. Machine learning algorithms are but upon statistical formulas and feature representation. A robust machine learning model is always going to be dependent on high level statistics like Pearson correlation, Hypothesis Testing etc.

II. DATA SET OVERVIEW

The data we chose for our research is available on ChrisWong.com **[1]** published by Chris, containing 224982 records after removal of null values and 12 columns. The feature columns include bbl (unique id for Borough, Block, and Lot), owner name, address, condo type, condo number, tax class, tax rate, estimated market value, billable assessed value, tax before exemption and abatements, tax before abatements and property tax.Some of the properties are city-owned.

ISBN : 978-81-944855-0-6

III. EXPLORATORY DATA ANALYSIS

To improve the statistics of our data we will first understand how our data is distributed.

A. Understanding the Columns

Our data has 12 columns with each contributing in telling us about the property id, property price and the tax. For the column tax class, we will be working on contains 5 classes, commercial properties, residential properties (more than 10 units), Co-op or condo properties with less than 11 units, Condo with 1-3 unit building and Condo with 1-3 story building. For the feature tax rate, we will work with 3 taxes, 10.68% (for commercial properties), 12.86% (Residential and Co-op and condo) and 19.16% (Condo unit and condo story building). The estimated market value for the properties in our data set lies in between 13$ and 632350000$ giving us factors like skewness and kurtosis to work on and similarly for property tax, the range for it between 0$ and 26030173$. The data set shows us that there is $21.6 billion total tax due. According to our data, the total tax before exemptions and abatements is estimated to about $34.5 billion and that is the amount the city could earn theoretically.So, we can say that the government exempted $12.9 billion. There are 161995 different owner names in our data set giving us a lot of variety.

	bbl	ownername	address	taxclass	taxrate	emv	tbea	bav	tba	propertytax	condonumber	condo
0	1000041001	ONE NY PLAZA CO. LLC	ONE NY PLAZA CO. LLC\nRYAN, LLC\n16220 N. SCOT...	commercial	10.68	5351542	254361	2380764	254361	251840	835	unit
1	1000041002	ONE NY PLAZA CO. LLC	ONE NY PLAZA CO. LLC\nRYAN, LLC\n16220 N. SCOT...	commercial	10.68	7733995	367600	3440656	367600	342763	835	unit
2	1000041003	ONE NY PLAZA CO. LLC	ONE NY PLAZA CO. LLC\nRYAN, LLC\n16220 N. SCOT...	commercial	10.68	15960040	713000	6673528	713000	713000	835	unit
3	1000041004	ONE NY PLAZA CO. LLC	ONE NY PLAZA CO. LLC\nRYAN, LLC\n16220 N. SCOT...	commercial	10.68	1372802	65249	610721	65249	53869	835	unit
4	1000041005	ONE NY PLAZA CO. LLC	ONE NY PLAZA CO. LLC\nRYAN, LLC\n16220 N. SCOT...	commercial	10.68	3144677	149467	1398982	149467	148035	835	unit

Table 1

The above figure is a reference for how the data looks.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 224982 entries, 0 to 224981
Data columns (total 12 columns):
bbl             224982 non-null int64
ownername       224982 non-null object
address         224982 non-null object
taxclass        224982 non-null object
taxrate         224982 non-null float64
emv             224982 non-null int64
tbea            224982 non-null int64
bav             224982 non-null int64
tba             224982 non-null int64
propertytax     224982 non-null int64
condonumber     224982 non-null int64
condo           224982 non-null object
dtypes: float64(1), int64(7), object(4)
memory usage: 20.6+ MB
```

Table 2

ISBN : 978-81-944855-0-6

B. **Variable Analysis**

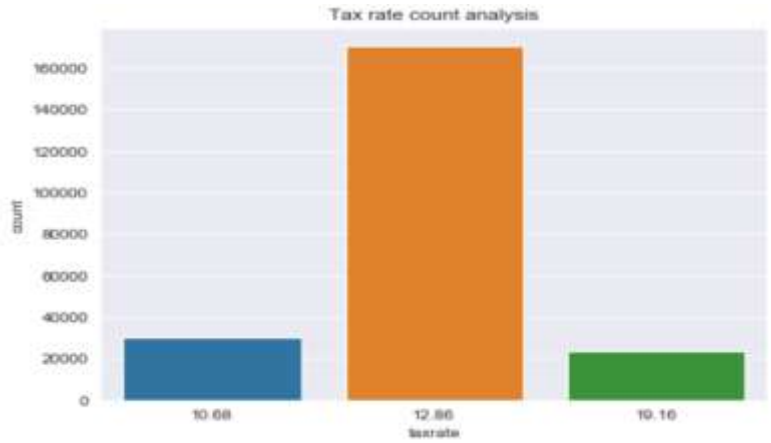

Figure 1

From the figure, we can make out that the maximum number of properties fall in the tax bracket of 12.86%.

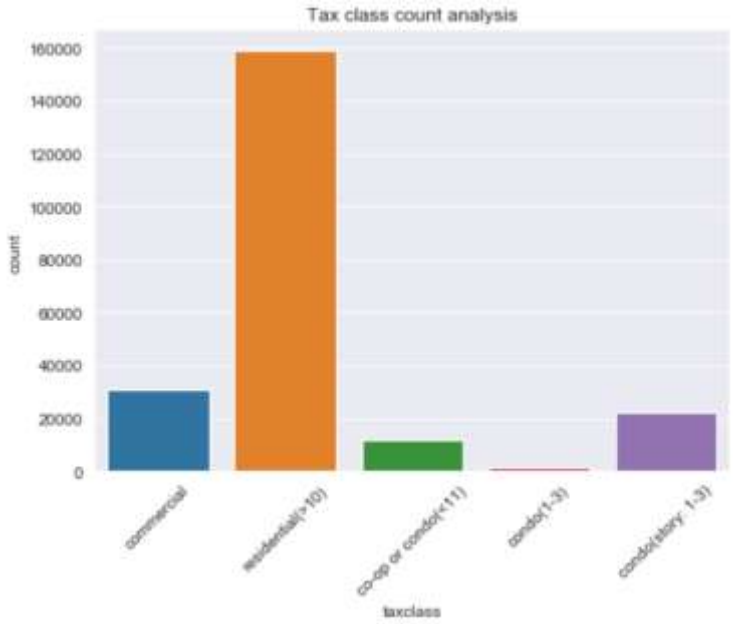

Figure 2

The maximum number of properties we are dealing with are residential with more than 10 units.

ISBN : 978-81-944855-0-6

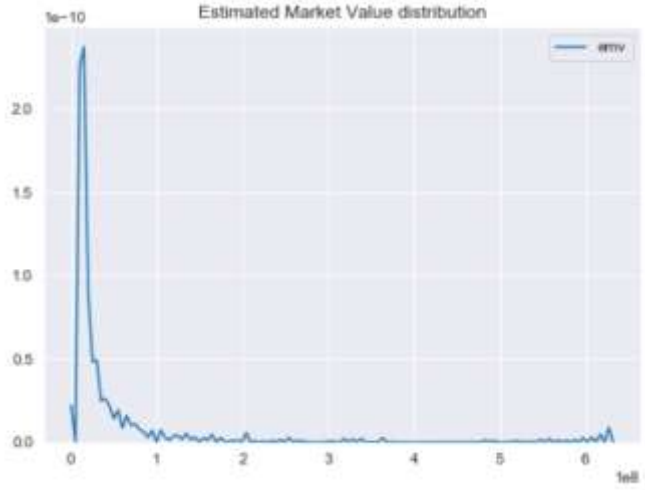

Figure 3

Distribution plot for Estimated market value shows that the maximum number of properties are evaluated to lie between 2billion to 3billion dollars.

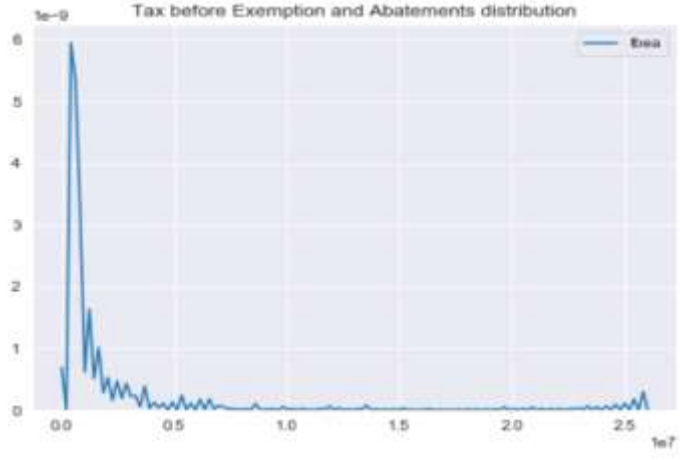

Figure 4

Distribution plot for Tax before Exemption and Abatements shows that the maximum number of properties lie between 15 million to 30 million dollars of due tax.

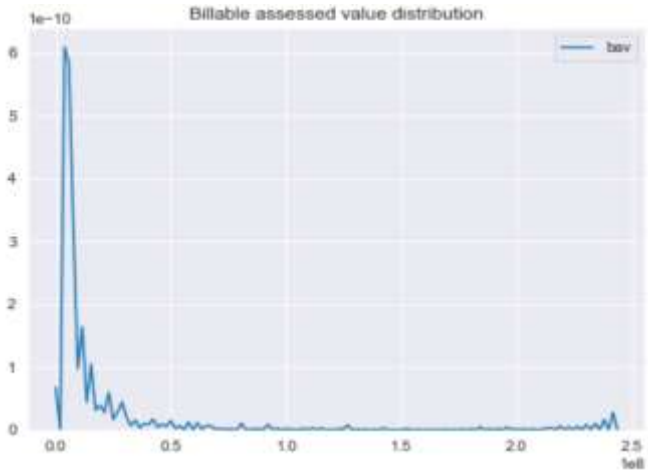

Figure 5

ISBN : 978-81-944855-0-6

Distribution plot for Billable assessed value shows that for maximum number of properties have a roundabout billable assessed value between 1 billion dollars to 2 billion dollars.

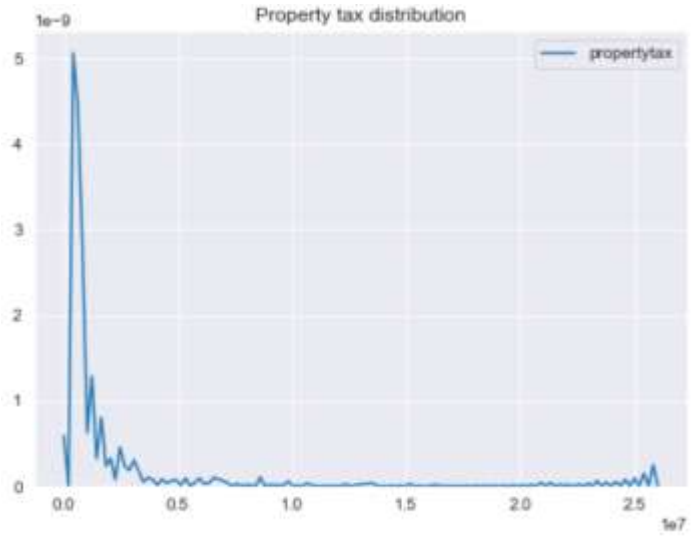

Figure 6

Distribution plot for Property tax shows that the tax for maximum number of properties is about a 150 million.

The plot below shows us that there are no null values in the data set.

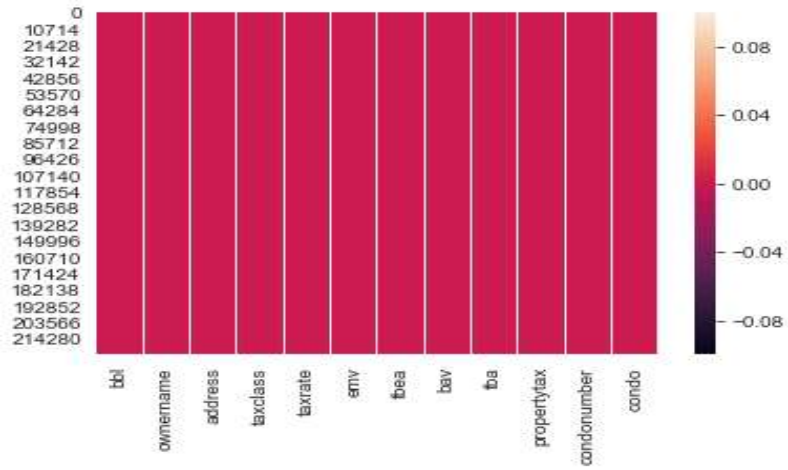

Figure 7

IV. Descriptive Statistics

A. Introduction

Machine learning is all about making predictions and so the role of descriptive statistics is crucial. Descriptive statistics help us organize our data and summarize in the best possible manner. The data we are dealing with is unimodal and is not normally distributed as per the distribution plots above.

ISBN : 978-81-944855-0-6

The descriptive statistics we will be working upon are:

1. Normal Distribution
2. Central Tendency (Median, mean and mode)
3. Measure of Variability (range, interquartile range, standard deviation and variance)
4. Kurtosis and Skewness

B. **Central Tendancy(Median, mean and mode)**

Central Tendency of a given data are the basic statistics which tell us the tendency of data points to lie around its mean, mode and median.

1. **Mean**: It is type of average used when the data is normally distributed.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Here X bar is the mean, Xi is the value at index i and n is the sample size.

- Mean Estimated Market Value (emv): 157785$
- Mean Tax before Exemption and Abatements (tbea): 6099$
- Mean billable assessed value (bav): 47105$
- Mean tax before abatements (tba): 3332$
- Mean property tax: 3035$

2. **Median**: It is type of average used when the data is skewed.

$$median = \frac{n+1}{2}$$

The above equation is for odd sample size (n).

$$median = \frac{n}{2}$$

The above equation is for even sample size (n).

- Median Estimated Market Value (emv): 427247.3127$
- Median Tax before Exemption and Abatements (tbea): 18206.9347$
- Median billable assessed value (bav): 152443.6029$
- Median tax before abatements (tba): 14517.7792$
- Median property tax: 13856.2448$

3. **Mode**: It is type of average used to find out the maximum occurring value.

ISBN : 978-81-944855-0-6

$$mode = maximum(X)$$

- Mode for Estimated Market Value (emv): 637$
- Mode for Tax before Exemption and Abatements (tbea): 31$
- Mode for billable assessed value (bav): 287$
- Mode for tax before abatements (tba): 31$
- Mode for property tax: 0$

Among the three central tendencies, mode is the only form of average that can be used with categorical data.

C. **Measure of Variability(range, interquartile range, standard deviation and variance)**

Range is the difference between the largest and the smallest points for the given data. The **interquartile range** is the statistical dispersion between upper(75^{th}) and lower(25^{th}) quartiles.

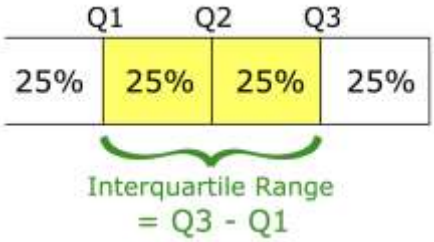

Figure 8**[2]**

While the range measures where the beginning and the end of your variable is, the interquartile range is a measure of where the major of the data points lie.

```
taxrate          0.0
emv         229711.0
tbea          9269.0
bav          75444.0
tba           9349.0
propertytax   8884.0
```

Standard Deviation is the value by which the data points in the given data differ from the mean value or it is the amount of dispersion around the mean. Standard Deviation is used further in z – score testing for hypothesis testing and confidence intervals.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

Here sigma is the standard deviation, n is the sample size, x is the value and mu is the mean.

- Standard Deviation for Estimated Market Value (emv): 5051987.0077$
- Standard Deviation for Tax before Exemption and Abatements (tbea): 215662.3876$
- Standard Deviation for billable assessed value (bav): 1973789.0289$
- Standard Deviation for tax before abatements (tba): 200531.3964$
- Standard Deviation for property tax: 198430.2500$

ISBN : 978-81-944855-0-6

The values of standard deviation are of same measure of the variable unlike the values of variance.

Variance is the standard deviation squared or expectation of the squared deviation a random variable from it mean. It is also used in calculation of covariance.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

Here sigma square is the variance, n is the sample size, x is the value and mu is the mean.

- Variance for Estimated Market Value (emv): 2522572726151.887$
- Variance for Tax before Exemption and Abatements (tbea): 46493014054.31797$
- Variance for billable assessed value (bav): 3895843130870.086$
- Variance for tax before abatements (tba): 40212840238.17002$
- Variance for property tax: 39374564129.99893$

D. **Kurtosis and Skewness**

Skewness is a measure of symmetry of the bell curve. It helps us identify if the data is normally distributed or not. There are 2 types of skewness, positive and negative skewness. Positive skewness shows one tail being extended to the right where as the negative skewness shows one tail being extended to the left when plotted.

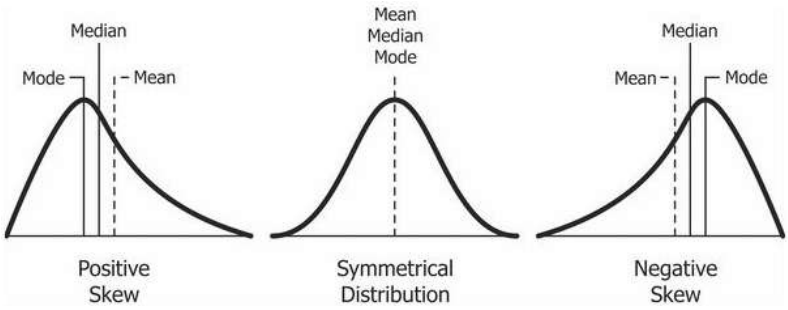

Figure 9**[3]**

The formula for skewness is:

$$skewness = \frac{3 * (mean - median)}{standard\ deviation}$$

ISBN : 978-81-944855-0-6

taxrate	1.937317
emv	68.185904
tbea	66.069007
bav	69.254048
tba	73.878469
propertytax	74.978108
dtype: float64	

Table 3: Skewness of our data

Kurtosisis a measure of the outlier presence in the data set. High kurtosis means that there are outliers or data has heavy tail whereas low kurtosis means that there is lack of outliers. There are 3 types of kurtosis:

1. Mesokurtic: The distribution has kurtosis like that of the normal distribution.
2. Leptokurtic (kurtosis > 3): Tails are thicker, and the data is distributed over a vast range.
3. Platykurtic (kurtosis < 3): Tails are thinner, and the data is distributed over a small range.

$$kurtosis = \frac{\sum_{i=1}^{n} \frac{X_i - \bar{X}}{n}}{\sigma^4}$$

taxrate	3.173043
emv	6277.320887
tbea	5991.969858
bav	6499.658241
tba	7321.982356
propertytax	7502.215474
dtype: float64	

Table 4: Kurtosis of our data

E. **Normal Distribution**

Distribution shows all the possible values of the info and the way often they occur. It is done to check whether the data is normally distributed or not. Our data is not normally distributed considering factors like skewness and distribution curve. Due to the outlier we can say there is a large amount of skewness causing the data to shift towards the left. A normally distributed data has bell shaped curve with unimodal data and is symmetric.

ISBN : 978-81-944855-0-6

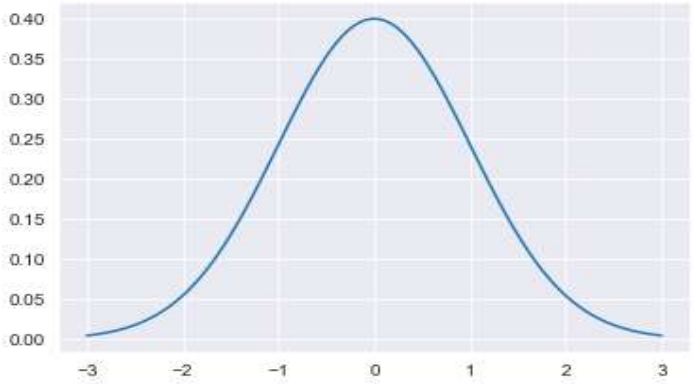

Figure 10: for normally distribution**[4]**

F. **Improving the Skewness and Kurtosis**

If we improve the skewness and kurtosis, the data set will get transformed to a normally distributed data set giving us much more comprehendible data to work on.

Steps that we will using to reduce skewness and improving kurtosis are:

1. Outlier Removal
2. Cube Root Transformation

Outlier Removal:

We will be using Inter-quartile range method to remove the outliers.We calculate the IQR then multiply the number by 1.5 and then add the result to the third quartile value and similarly for first quartile we will subtract the result from first quartile. Numbers less than or greater than the given range are considered as outliers.

$$value < (Q1 - 1.5 * IQR)$$

or

$$value > (Q3 + 1.5 * IQR)$$

The condition given above is the one we implement in python and upon implementing we get:

```
taxrate            0.0
emv           229711.0
tbea            9269.0
bav            75444.0
tba             9349.0
propertytax     8884.0
dtype: float64

(155044, 8)
```

Table 5

We will be removing outliers to see,by how does the skewness reduce. After applying the IQR method we get 69,938 outliers in our data set. Initially our skewness plot looked something like:

ISBN : 978-81-944855-0-6

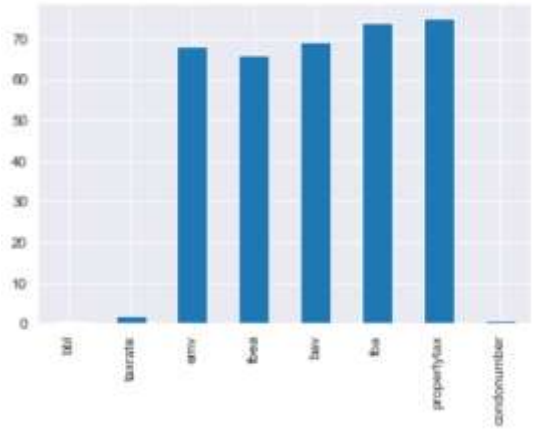

Figure 11

The skewness achieved after outlier removal:

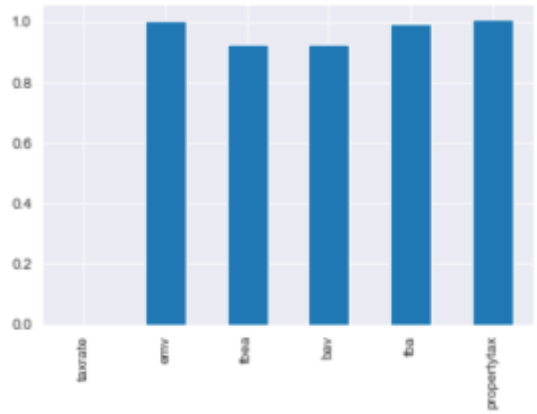

Figure 12

Cube Root Transformation:

Cube Root transformation is done to remove positive skewness. The parameters are changed from x to $x^{(1/3)}$. Cube root transformation has an upper edge over some of the transformations since it can deal with zero and negative values.

The resulting skewness post cube root transformation is:

```
taxrate        0.000000
emv            0.104933
tbea           0.081368
bav            0.081377
tba           -0.035144
propertytax   -0.101839
dtype: float64
```

Table 6

So now we can see that the skewness lies between lies between the normal distribution's range which is -0.5 to 0.5. The plot for our skewness looks something like:

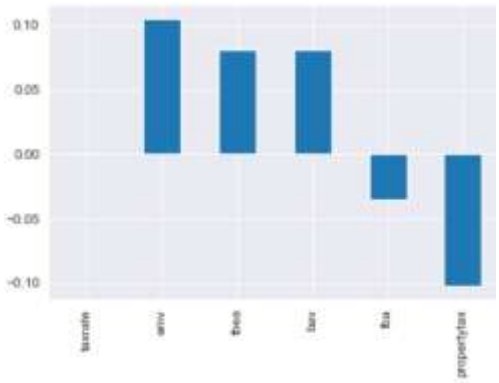

Figure 13**[5]**

V. Inferential Statistics

A. Introduction

Inferential statistics gives us an overview of how the population works based on the inferences made from sample or subset of population. Sample mean moves towards normal distribution around true mean (population mean) as the sample size increases and this is the Central Limit Theorem.

The inferential statistics we will be working upon are:

1. Confidence Interval
2. Hypothesis Testing using z-score testing
3. Pearson's Correlation

B. Confidence Interval

Confidence interval gives an approximate range of values which includes the unknown parameters for the population based upon results gathered from sample or subset of population. It inevitably hinges about the mean of the sample. This approximate range of values are obtained by adding and subtracting a margin of error from the point estimate.

$$C.I = \bar{x} + Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

Here x bar is the point estimate and sigma by square root n is the standard error.Margin of error is evaluated when the Z-score**[6]** of confidence level is multiplied by standard error of the mean. Confidence level expresses the probability of unknown population parameters to lie in between the approximate range of values.

Standard error of mean is calculated by dividing the standard deviation by the root of the total number of observations. Confidence limits are the upper and lower bound of the range of values obtained as the result of confidence interval calculation. We assume the confidence level to be 95% while the significance level (1 – confidence level) to be 5%. When the population mean is unknown, and the population standard deviation is known given that the number of observations is greater than 30, z-distribution is used to find the confidence interval. Since we are dealing with the same situation

ISBN : 978-81-944855-0-6

in our dataset where the number of observations is 155044, we tend to use z-distribution while finding out the confidence interval.

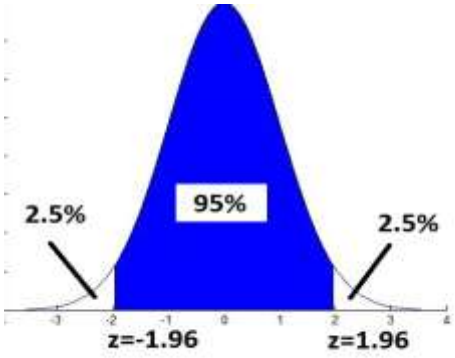

Figure 14**[7]**

While calculating the confidence interval, we calculate the Z-score for 95% confidence level to be 1.96 where the level of significance is computed to be 2.5% on each side of the curve. After calculations the resultant ranges come out to be -

C. **Hypothesis Testing using z-score testing**

Hypothesis testing is an act in statistics whereby a person tests an assumption regarding a population parameter. The methodology employed by the person depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.
Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population**[8]**. The test tells the person whether his primary hypothesis is right or not. Statistical analyst tests a hypothesis by measuring and examining a random sample of the population being analyzed.

- Here are the broad steps that are involved in hypothesis testing. Determine the critical region for your decision (We need a certain level of certainty)
- Decide on the hypothesis you're going to test (This is the claim that we're putting on trial)
- Choose your test statistic (We need to pick the statistic that best tests the claim.)
- Find the p-value of the test statistic (We need to see how rare our results are, assuming the claims are true.)
- See whether the sample result is within the critical region (We then see if it's within our bounds of certainty.)
- Make your decision

We applied hypothesis testing on the column of estimated market value (emv).

1. Decide on the hypothesis
null hypothesis: - 40.91% of the properties have an estimated market value between 10 and 20.
Alternate hypothesis: - More than 40.91% of the properties have an estimated market value between 30 and 75.
H0: p = 0.4091
H1: p > 0.4091

ISBN : 978-81-944855-0-6

2. Choose the test statistics

If we use X to represent the number of properties in the sample, this means that we can use X as our test statistic. There are 155044 properties in the sample and the probability of success is 0.4091. As X follows a binomial distribution, this means that the test statistic is actually:

X ~ B(155044, 0.4091)

So our test statistic is X ~ B(155044, 0.4091)

3. Determine the Critical Region

Let's use a significance level of 95% in our hypothesis test. This means that 40.91% of properties have emv in between 10 to 20 in the sample is in the highest 95% of the probability distribution, then we will reject the null hypothesis

α= 95% or 0.95

In this we will use upper one - tailed test

4. Finding the p- value Using Discrete Sampling **[9]**

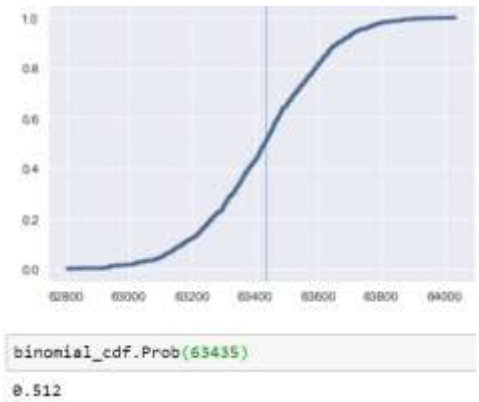

Figure 15

Step 5: Is the sample result in the critical region?

Now that we've found the p-value, we can use it to see whether the result from our sample falls within the critical region. If it does, then we'll have enough evidence to reject the claims of the doctors.

Our critical region is the upper tail of the probability distribution, and we're using a significance level of 95%. This means that we can reject the null hypothesis if our p-value is > 0.95. As our p-value is 0.512, this means that the number of properties that have an estimated market value between 10 and 20, doesn't fall within the critical region.

Step 6: Make your decision

We've now reached the final step of the hypothesis test. We can decide whether to accept the null hypothesis or reject it in favor of the alternative.

The p-value of the hypothesis test falls outside the critical region of the test. This means that there isn't enough evidence to reject the null hypothesis. In other words:

We accept the claim.

ISBN : 978-81-944855-0-6

D. **Pearson's Correlation**

Correlation is the dependence of one variable on the other. For example, in supervised machine learning the independent variables are used to predict the dependent variable and so in this case the correlation between the two makes it easier for us to optimize our model.

Correlation is of 3 types:

1. Neutral correlation: No relationship between the two features
2. Negative correlation: Change in one cause change in other in opposite direction
3. Positive correlation: Change in one cause change in other in same direction

$$\rho_{X,\,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Correlation formula **[10]**

The formula above is for calculating the correlation by Pearson's method where the numerator is the covariance of the two variables and the denominator is the product of the standard deviation of the two variables. To calculate the correlation, we need to calculate the covariance of the two variables.

$$Cov(x,y) = \frac{\Sigma\,(x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

Covariance formula **[11]**

The above formula is for calculation of covariance for any two features x and y. In our case we are choosing to find the correlation between billable assessed value and tax before abatement. The covariance of the two is 41.6131 and the Pearson correlation is 0.6815.

VI. CONCLUSION

Statistics plays a vital role in the field of Data Science when it comes to feature engineering, data cleaning and optimization. In our research, we have shown the importance of statistics and how it can be applied to a data set in order to get information about it. For best analysis of a given data set, one must understand the central tendencies, skewness, correlation etc. Linear or Non-Linear regression both assume that the residuals follownormal distribution **[12],** so it important to normalize data. Hypothesis testing is done to verify whether there is enough statistical evidence to claim a belief to be true or false.

VII. ACKNOWLEDGEMENT

We would like to acknowledge our families for their continuous support and our teachers for their continuous guidance and help to complete this research. We would like to thank Mr. Wong for making this data set for our research. We are thankful to the NorthCap University, Gurugram for giving us this opportunity.We would like to thank the almighty for showering us with the greed to learn.

ISBN : 978-81-944855-0-6

REFERENCES

[1] Chris Wong "Liberating Data From NYC Property Tax bills", https://chriswhong.com/open-data/liberating-data-from-nyc-property-tax-bills/
[2] https://statsmethods.wordpress.com/2013/05/09/iqr/
[3] https://en.wikipedia.org/wiki/File:Relationship_between_mean_and_median_under_different_skewness.png
[4] https://pythonforundergradengineers.com/plotting-normal-curve-with-python.html
[5] http://www.ijsrp.org/research-paper-0319.php?rp=P878342
[6] http://heather.cs.ucdavis.edu/probstatbook
[7] https://www.mathandstatistics.com/learn-stats/finding-z-critical-values
[8] https://www.investopedia.com/terms/h/hypothesistesting.asp
[9] https://www.investopedia.com/terms/p/p-value.asp
[10] https://miro.medium.com/max/288/1*uf1MiJerVbP2uAdlfAJLpw.png
[11] https://cdn.educba.com/academy/wp-content/uploads/2019/05/Covariance-Formula.jpg
[12] https://statisticsbyjim.com/basics/normal-distribution/
[13] https://www.elsevier.com/books/elementary-statistics/hope/978-0-08-012131-4