DATASET EXPLORATION

A walk through of different set of analyses over NYC ROLLING SALES Data for academic submission in course “Mathematics for Data Analytics”

Submitted By:
Jagbeer Singh
St ID: 200543422

Instructor: Mr. Jonathan Gladstone

Name: Jagbeer Singh

Student #: 200543422

Group #: C

Submission Date: 4th Feb, 2023

Dataset Exploration Project– Dataset Description & FINER Questions

What does the Data Look Like?

The Dataset particularly consists of the records stored about the sales of building or apartments in the New York City, over the time period of 12 months.

What is the dimension of the Data?

The Dataset is made up of twenty-two columns and stores a data of around Eight four Thousand rows. [084,000 **X** 022]

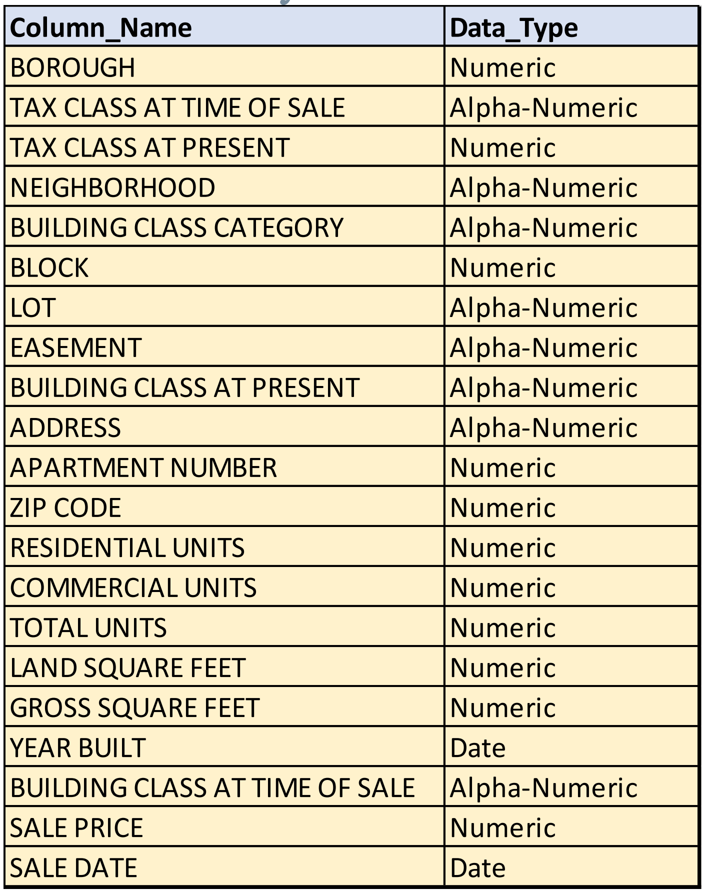
Where did you obtain the Data From?

The Dataset has been obtained from website which is primarily popular for Data Science and Analysis of it. The Data sources from the department of finance which stores the data about the New York Home sales. It can be referenced from the given link. (1)

Data set Assumptions:

The Dataset seems a complete entity on it's own to be considered for analysis. The data not only has a good share of numerical but it also has few categorical variables. There isn't necessity of finding another data in order to research and generate findings for Finer questions.

Data Dictionary:



Column_Name	Data_Type
BOROUGH	Numeric
TAX CLASS AT TIME OF SALE	Alpha-Numeric
TAX CLASS AT PRESENT	Numeric
NEIGHBORHOOD	Alpha-Numeric
BUILDING CLASS CATEGORY	Alpha-Numeric
BLOCK	Numeric
LOT	Alpha-Numeric
EASEMENT	Alpha-Numeric
BUILDING CLASS AT PRESENT	Alpha-Numeric
ADDRESS	Alpha-Numeric
APARTMENT NUMBER	Numeric
ZIP CODE	Numeric
RESIDENTIAL UNITS	Numeric
COMMERCIAL UNITS	Numeric
TOTAL UNITS	Numeric
LAND SQUARE FEET	Numeric
GROSS SQUARE FEET	Numeric
YEAR BUILT	Date
BUILDING CLASS AT TIME OF SALE	Alpha-Numeric
SALE PRICE	Numeric
SALE DATE	Date

What are different columns of the Dataset? And what do they represent? (2)

Borough:

It specifies the name of the borough in which the property is situated.

Neighborhood:

This particular attribute talks about the neighbourhood in which the property is located. It is a terminology assessed by department of finance.

Building Class Category:

This particular fields shares information in order to classify the building type into broad categories for an easy segregation, for example, One Family Homes. It also helps in storing the files for the sales with categories labeled to it.

Tax Class at Present:

There are around four tax classes which are assigned to the property on the basis of it's use. [Class 1, Class 2, Class 3 and Class 4]. It also helps in identifying the financial tax slab for the given property holder.

- Class 1: Most of the residential properties which include up to three units like one-, two, or even three-family homes and/or small stores, offices one/two attached apartments to it. This also includes and un-occupied land which is zoned for residential purposes, and condos which have around three stories.
- Class 2: This includes any property which is primarily residential, for example Cooperatives & Condominiums.
- Class 3: This includes any property which has equipment owned by any telephone, gas, and/or electricity company.
- Class 4: This includes all other properties which are not included in class one, two and three, like factories, warehouses, garages, etc.

Block:

This attribute talks about the sub-division of the column "Borough", on which any real property is situated. Finance Department uses Borough-to-Block-to-Lot divisions to label or classify all real property in the NYC.

It's Observed as a fact that Addresses describe the location of a property with respect to street where it is located. But the block and Borough and Lot differentiate any unit of real property from another.

Lot:

A property unique location is represented by a subdivision of Tax-Block, which is Tax-"LOT".

Easement:

An easement can be described as a right, that allows a party to make limited usage of someone else's property.

Building Class at Present:

The constructive use of Property is described by the Building Classification. The general class of the properties is described by these mentioned letters, "A" talks about one-family homes, "O" describes about office buildings. "R" depicts condominiums. The 2nd position is a number, which additionally tells more intricate information about the usage of property and/or construction style

Address:

The Address of the Given property and how it is listed on the Sales File.

Zip Code:

The Zip code of the Property of the New York home sales.

Commercial Units:

The no# of commercial units within the stated property. Total Units: The total number of units at the listed property.

Land Square Feet:
The calculated land area in the sq feet

Gross Square Feet:
The total area of the property including exterior parts of the property.

Year Built:
The year in which the structure of the listed property was built.

Building Class at Time of Sale:
As the name suggests, this also references the building class which is stated in above column.

Sales Price:
Paid price for the property at the time of sale.

Sale Date:
As name suggests

$0 Sales Price:
Any 0$ sale describes that the ownership was changed without any money exchange.

What are the Finer Research Questions which can help to analyze Data and infer something useful from it?

1. Each property can have a unique Date when it was built, does the building age affect the sales price in any form or not? If yes, then how?
2. What is the relationship between the selling price and gross square feet of a property, and how much of the variance in the dependent variable can be explained by the variance in the independent variable?
3. To what extent do the tax class and sale price category of properties sold in the real estate market affect each other?

Dataset Exploration Project Part 2 (DE pt2) – Performing Univariate Analysis

Ques What is Univariate Descriptive Statistic?

When dealing with a large amount of Data, it becomes essential to be aware of all the different variables and their individual importance in the whole Data Set. Univariate Analysis allows us to do the same with the help of statistics.

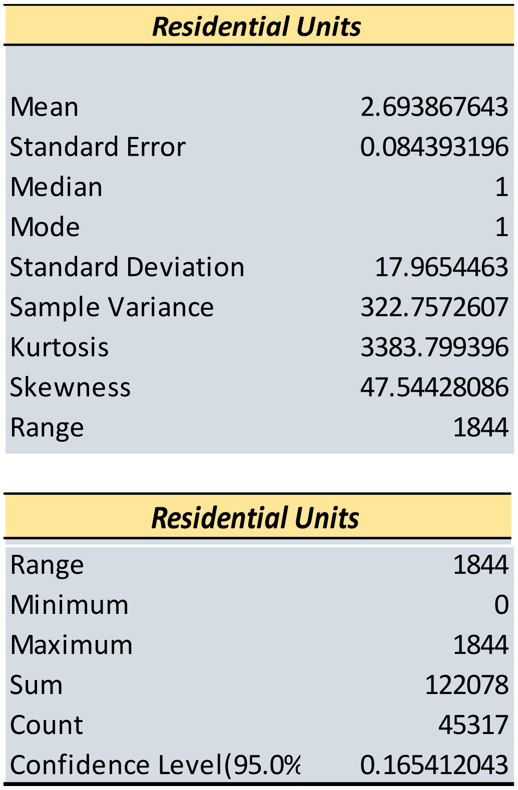
For this step, it is required to take a single Variable and analyze it through a bunch of parameters in order to take out as much information as possible and proceed with further analysis.

Data Cleaning:

The Data Cleaning is a crucial step in order to move further in analysis. This step required for me to delete the rows which had null a great deal of null or blank values. There were not may outliers but the blank values were handled by the means of deletion only, which resulted in the data to be reduced by 40%.

Here the Variable "***Residential Units***" shows information about number of units which are considered Residential at any listed property.

The Figure 1.1 shows the applied descriptive statistics over the whole data under Residential units.

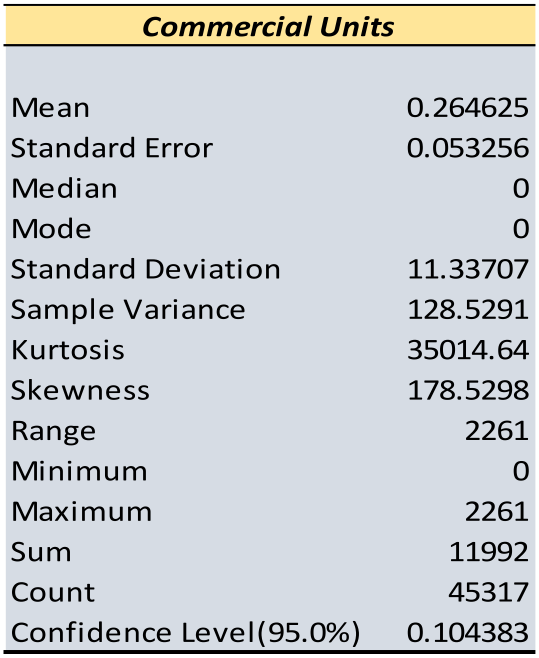


Residential Units	
Mean	2.693867643
Standard Error	0.084393196
Median	1
Mode	1
Standard Deviation	17.9654463
Sample Variance	322.7572607
Kurtosis	3383.799396
Skewness	47.54428086
Range	1844

Residential Units	
Range	1844
Minimum	0
Maximum	1844
Sum	122078
Count	45317
Confidence Level(95.0%	0.165412043

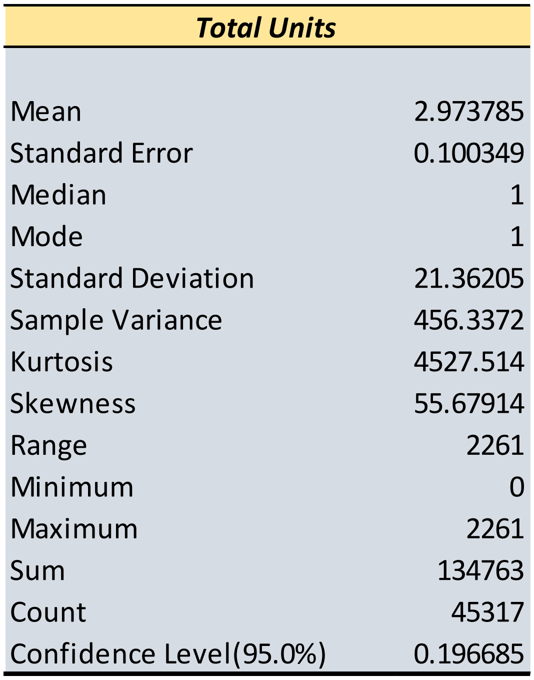
Figure (1.1)

The statistics show that there is an approximate average of 2.70 Residential units per NYC property. Whereas, as seen in figure 1.2, it becomes quite clear that average of commercial units is comparatively much less than



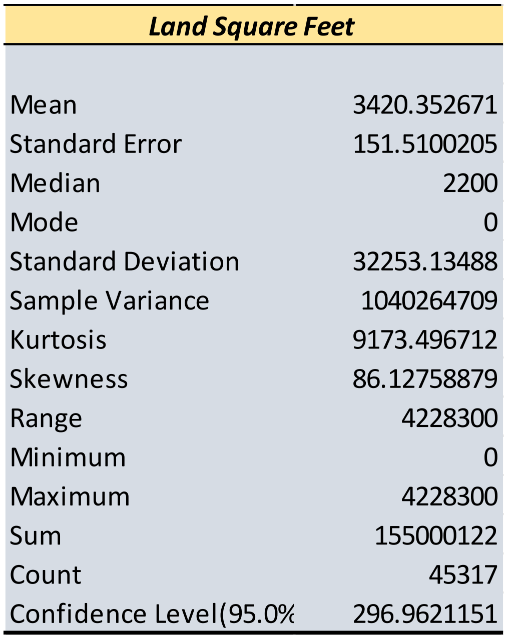
Commercial Units	
Mean	0.264625
Standard Error	0.053256
Median	0
Mode	0
Standard Deviation	11.33707
Sample Variance	128.5291
Kurtosis	35014.64
Skewness	178.5298
Range	2261
Minimum	0
Maximum	2261
Sum	11992
Count	45317
Confidence Level(95.0%)	0.104383

that of the residential ones.



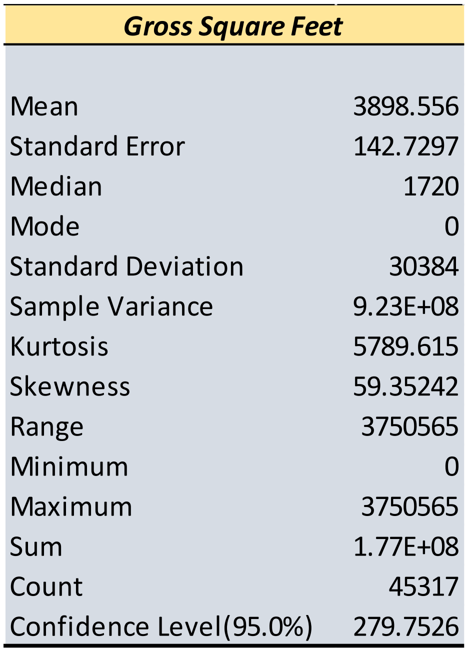
Total Units	
Mean	2.973785
Standard Error	0.100349
Median	1
Mode	1
Standard Deviation	21.36205
Sample Variance	456.3372
Kurtosis	4527.514
Skewness	55.67914
Range	2261
Minimum	0
Maximum	2261
Sum	134763
Count	45317
Confidence Level(95.0%)	0.196685

Figure (1.3)



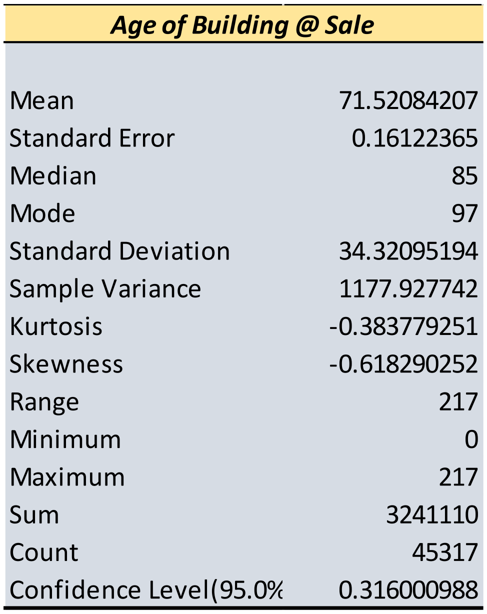
Land Square Feet	
Mean	3420.352671
Standard Error	151.5100205
Median	2200
Mode	0
Standard Deviation	32253.13488
Sample Variance	1040264709
Kurtosis	9173.496712
Skewness	86.12758879
Range	4228300
Minimum	0
Maximum	4228300
Sum	155000122
Count	45317
Confidence Level(95.0%	296.9621151

Figure (1.4)



Gross Square Feet	
Mean	3898.556
Standard Error	142.7297
Median	1720
Mode	0
Standard Deviation	30384
Sample Variance	9.23E+08
Kurtosis	5789.615
Skewness	59.35242
Range	3750565
Minimum	0
Maximum	3750565
Sum	1.77E+08
Count	45317
Confidence Level(95.0%)	279.7526

Figure (1.5)

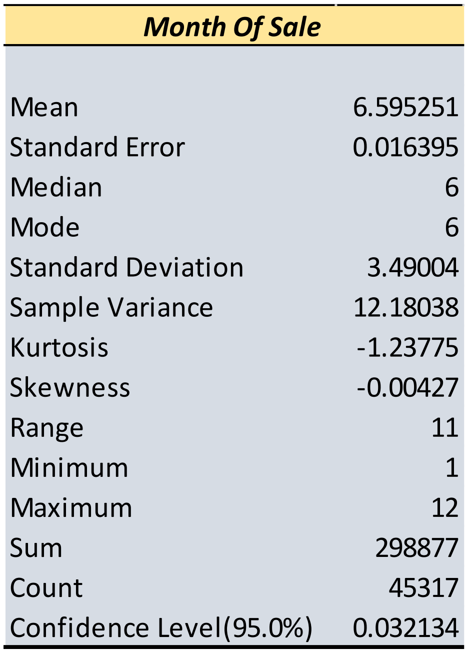


Age of Building @ Sale	
Mean	71.52084207
Standard Error	0.16122365
Median	85
Mode	97
Standard Deviation	34.32095194
Sample Variance	1177.927742
Kurtosis	-0.383779251
Skewness	-0.618290252
Range	217
Minimum	0
Maximum	217
Sum	3241110
Count	45317
Confidence Level(95.0%	0.316000988

Figure (1.6)

CNTD..

The records with missing values or zero values could have tempered with analysis in a lot of manners. The step taken was to remove such records and move forward with clean and consistent data. This resulted in removal of around 36,000



Month Of Sale	
Mean	6.595251
Standard Error	0.016395
Median	6
Mode	6
Standard Deviation	3.49004
Sample Variance	12.18038
Kurtosis	-1.23775
Skewness	-0.00427
Range	11
Minimum	1
Maximum	12
Sum	298877
Count	45317
Confidence Level(95.0%)	0.032134

Figure (1.7)

As seen in figure 1.6, the average age of the building at the time of sale is observed to be approximately 117 years, and it can also be justified with the fact that most buildings were constructed around 1910s, which makes complete sense.

Point to be Noted:

During first part of the Dataset Exploration there were more than 84,000 records in the Data, but the data had lot of inconsistencies as well. It became essential to clean it before moving towards analysis part.

Now, moving further it becomes crucial to analyze the Data with help of graphical representations as well.

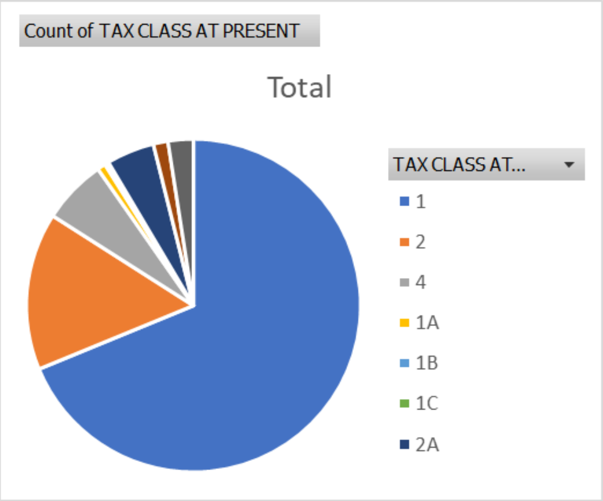


Figure (1.8)

Here, the Pie chart shown in figure 1.8, describes the distribution of data amongst the different Tax Classes under which the listed property is mentioned. The Tax Class 1, is determined to be the one with most listed properties. This particular class deals with residential units in a listed property.

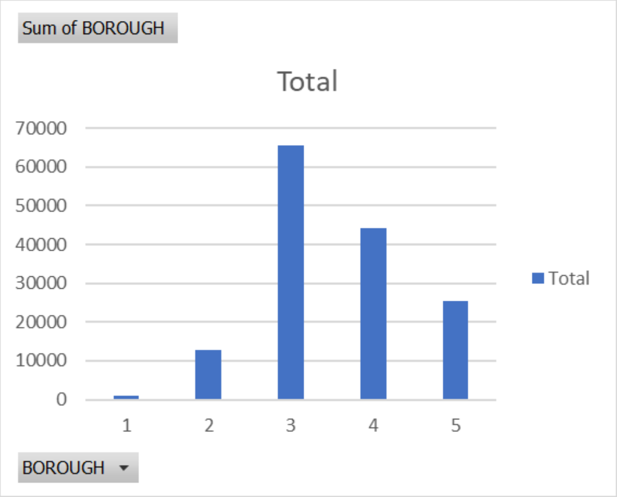
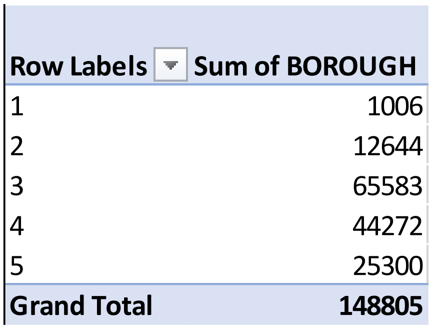


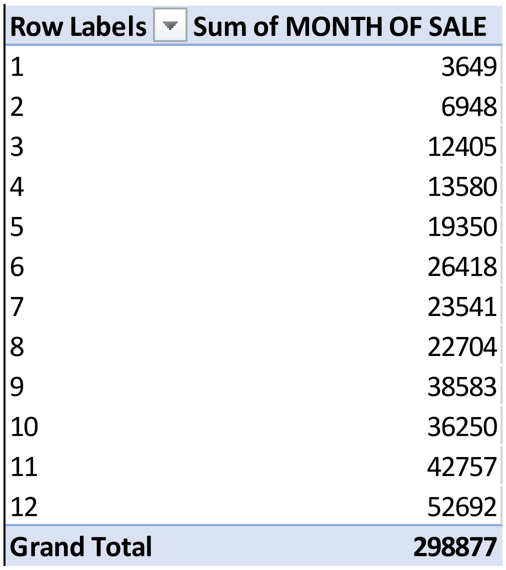
Figure (1.9)



Row Labels	Sum of BOROUGH
1	1006
2	12644
3	65583
4	44272
5	25300
Grand Total	**148805**

Figure (1.10)

As depicted in the graph (figure 1.9), it becomes quite obvious that there are huge number of properties listed in BOROUGH 3. This can help a lot to determine what kind of properties made this borough so popular in sales, when doing multivariate analysis.



Row Labels	Sum of MONTH OF SALE
1	3649
2	6948
3	12405
4	13580
5	19350
6	26418
7	23541
8	22704
9	38583
10	36250
11	42757
12	52692
Grand Total	**298877**

Figure (1.11)

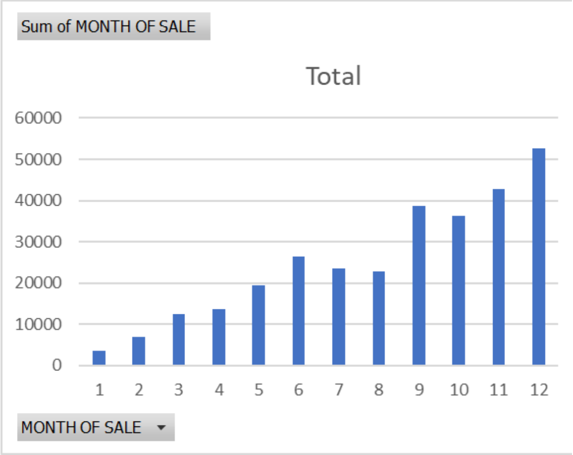


Figure (1.12)

1913
1916
1919
1922
1925
1928
1931

Figure 1.13

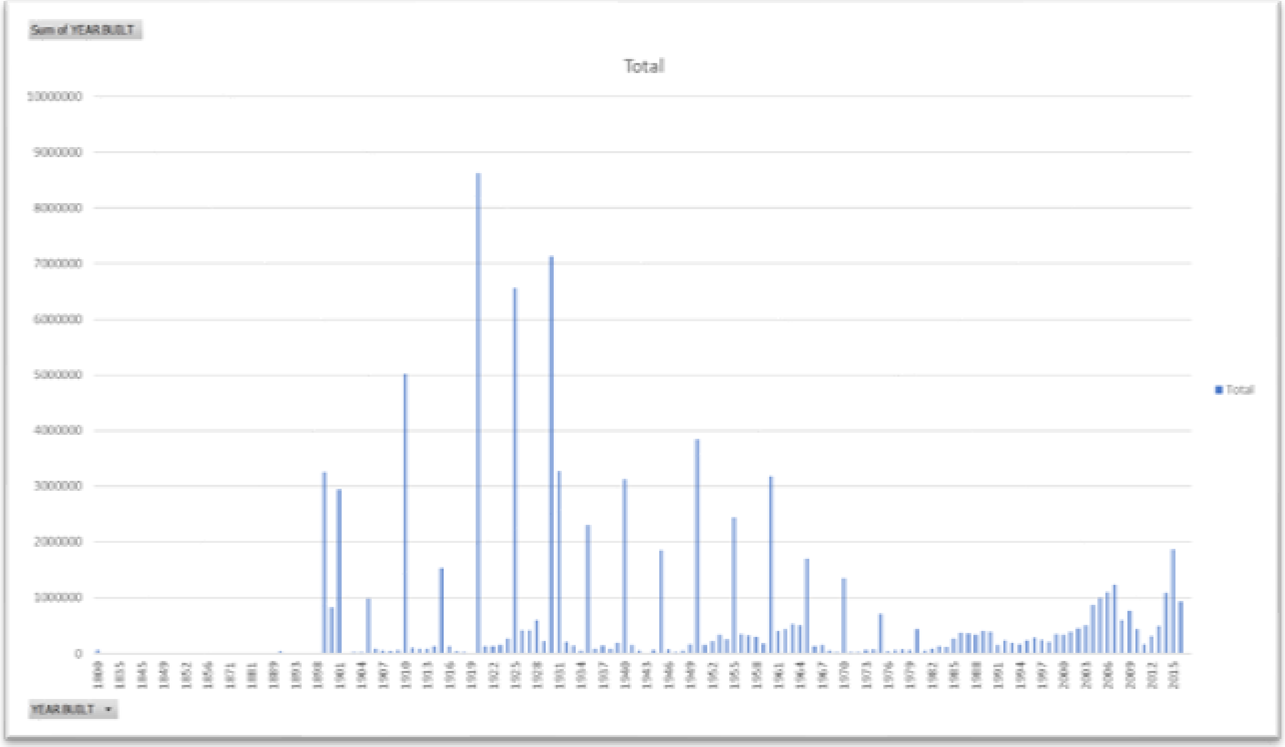

Sum of YEAR BUILT
Total
10000000
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Total
1800
1835
1845
1849
1852
1856
1871
1881
1889
1893
1898
1901
1904
1907
1910
1913
1916
1919
1922
1925
1928
1931
1934
1937
1940
1943
1946
1949
1952
1955
1958
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
YEAR BUILT

Figure (1.14)

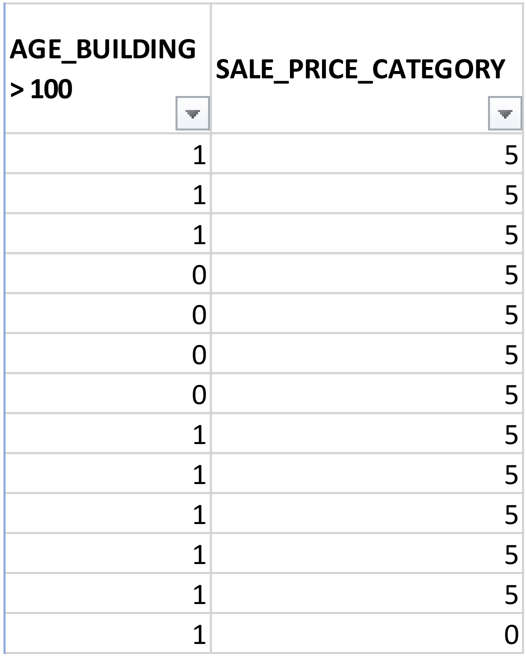
Multi Variate Analysis and Hypothesis testing

What is muti-Variate Analysis?

The Multi Variate Analysis can be described as the statistical study of multiple attributes present in the data in order to generate useful findings. It also provides a comprehensive view point on the relationship among various variables at once.

Herein, the analysis has been conducted over the dataset and the hypothesis testing has also been performed.

The analysis requires creation of two more variables as shown below:



AGE_BUILDING > 100	SALE_PRICE_CATEGORY
1	5
1	5
1	5
0	5
0	5
0	5
0	5
1	5
1	5
1	5
1	5
1	5
1	0

FORMULAE USED:

# AGE_BUILDING>100 → =IF(AGE_BUILDING>100,1,0)
This tells whether the age of building is grater than 100 or not.

#SALE_PRICE_CATEGORY →

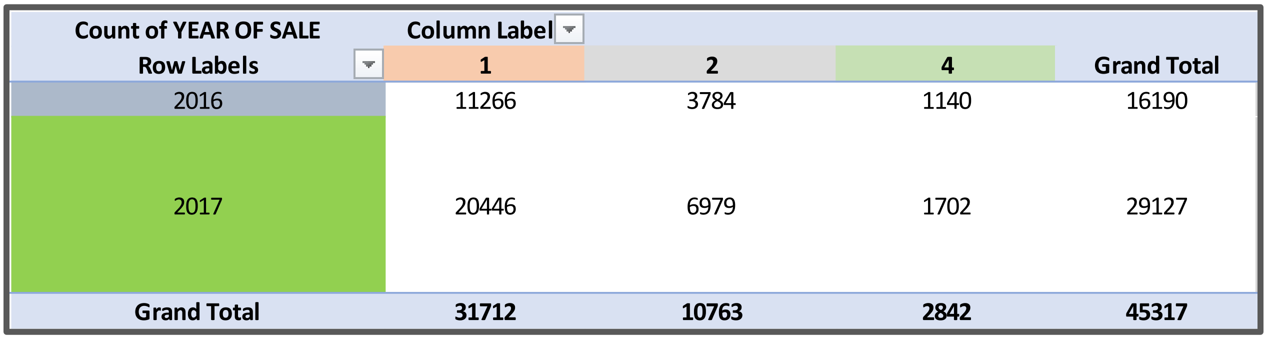
=IF(T2<=5000,0,IF(AND(T2>=5001,T2<=50000),1,IF(AND(T2>=50001,T2<=100000),2,IF(A
ND(T2>=100001,T2<=150000),3,IF(AND(T2>=150001,T2<=1000000),4,5)))))

The above formula divides the sale_price into five categories.

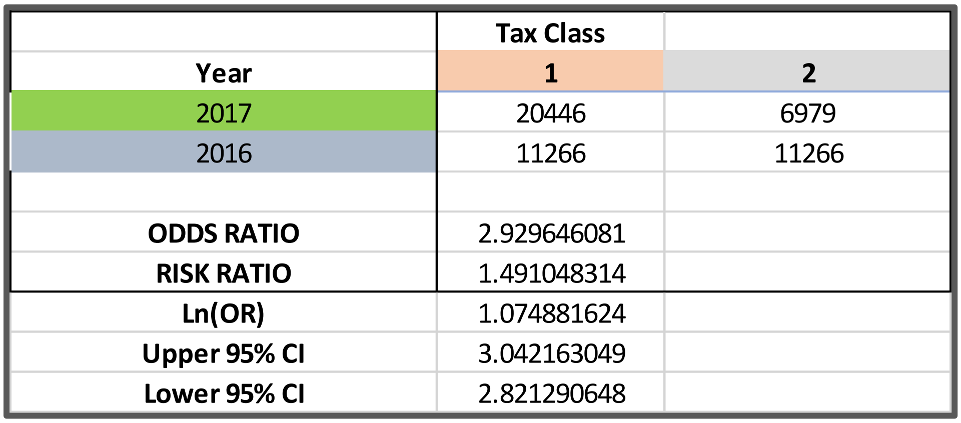
Odds and Risk Ratio:

Categorical variable: YEAR OF SALE
Numerical variable: TAX CLASS AT SALE

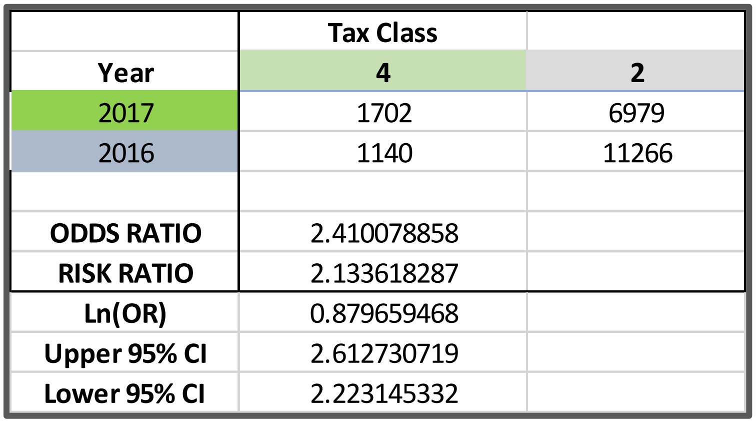
Pivot table:



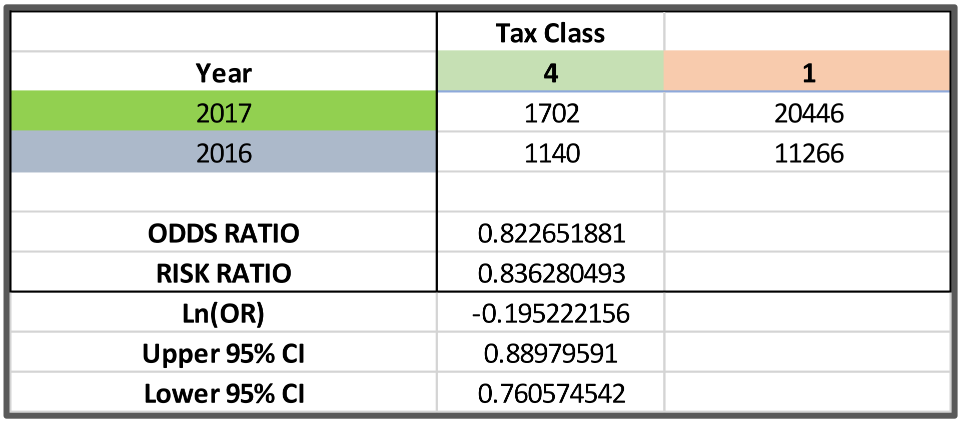
Count of YEAR OF SALE	Column Label			
Row Labels	1	2	4	Grand Total
2016	11266	3784	1140	16190
2017	20446	6979	1702	29127
Grand Total	**31712**	**10763**	**2842**	**45317**



	Tax Class	
Year	**1**	**2**
2017	20446	6979
2016	11266	11266
ODDS RATIO	2.929646081	
RISK RATIO	1.491048314	
Ln(OR)	1.074881624	
Upper 95% CI	3.042163049	
Lower 95% CI	2.821290648	



	Tax Class	
Year	**4**	**2**
2017	1702	6979
2016	1140	11266
ODDS RATIO	2.410078858	
RISK RATIO	2.133618287	
Ln(OR)	0.879659468	
Upper 95% CI	2.612730719	
Lower 95% CI	2.223145332	



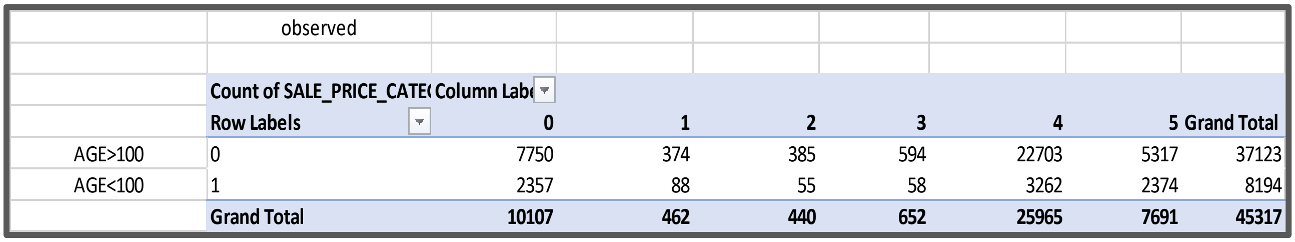
	Tax Class	
Year	4	1
2017	1702	20446
2016	1140	11266
ODDS RATIO	0.822651881	
RISK RATIO	0.836280493	
Ln(OR)	-0.195222156	
Upper 95% CI	0.88979591	
Lower 95% CI	0.760574542	

The above shown analysis focuses on few major calculations around Odds Ratio and Risk Ratio among two variables. The Data talks about 4 Tax Classes at the time of Sale, hence the analysis is divided and done in three parts.

Null Hypotheses would be that "*There is no difference between the number of sales amongst two stated years, i.e., 2016 and 2017*". But the analysis indicates otherwise, it states that there has been an incredible amount of difference between number of sales in both years for various tax classes.

Moving further, the upcoming analysis comprises of Chi-square and T tests.

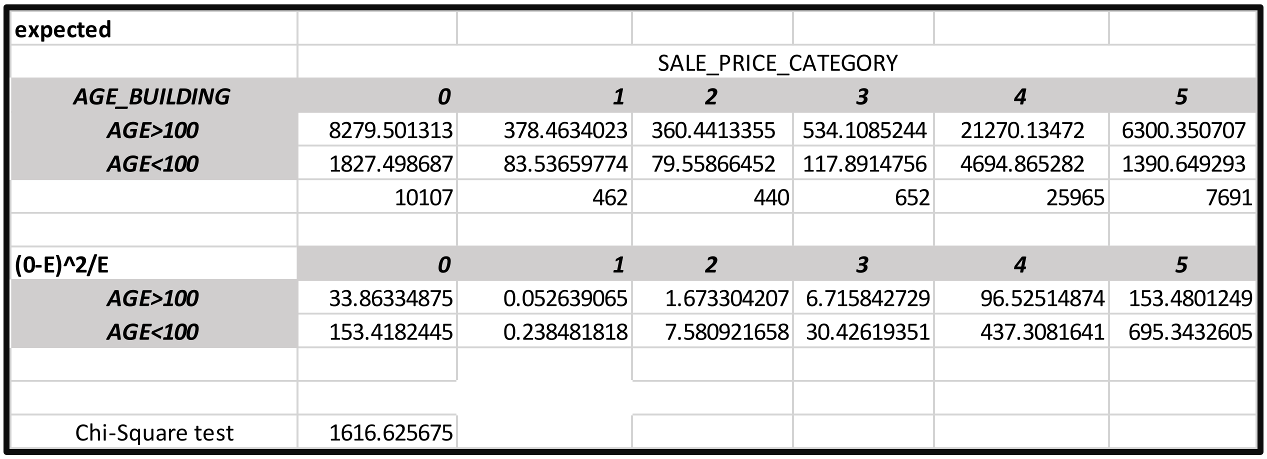
Here, the observed and expected values are used to calculate the chi square test value. The two variables "Sale_price_category" and "Age>100" are used to perform the analysis.



	observed							
	Count of SALE_PRICE_CATE(	Column Labe▾						
	Row Labels ▾	0	1	2	3	4	5	Grand Total
AGE>100	0	7750	374	385	594	22703	5317	37123
AGE<100	1	2357	88	55	58	3262	2374	8194
	Grand Total	10107	462	440	652	25965	7691	45317

The expected values are calculated with below stated formula:

Expected Value = (Row Total x Column Total) / Grand Total

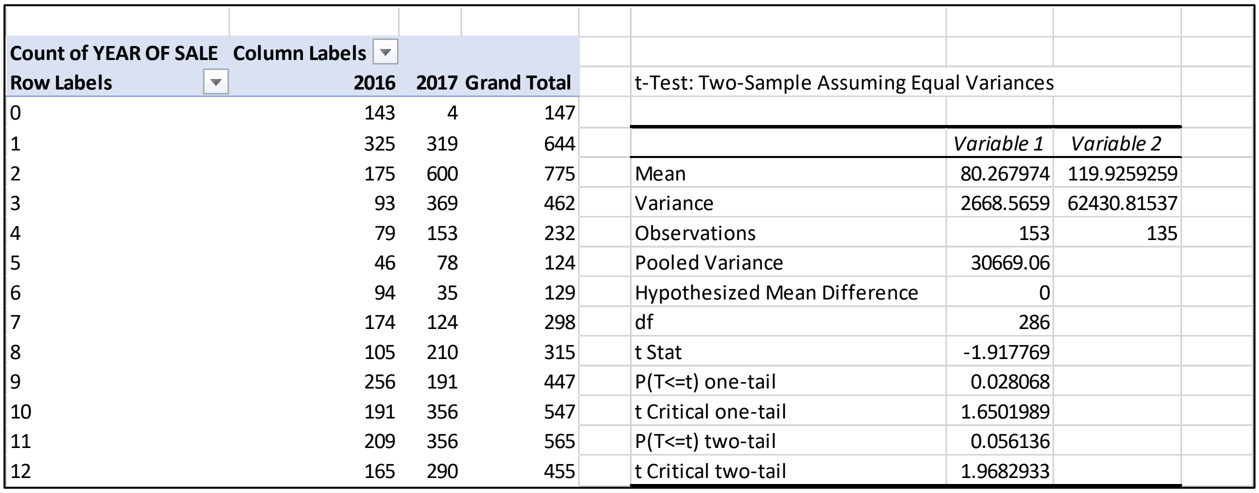


expected						
	SALE_PRICE_CATEGORY					
AGE_BUILDING	*0*	*1*	*2*	*3*	*4*	*5*
AGE>100	8279.501313	378.4634023	360.4413355	534.1085244	21270.13472	6300.350707
AGE<100	1827.498687	83.53659774	79.55866452	117.8914756	4694.865282	1390.649293
	10107	462	440	652	25965	7691
(0-E)^2/E	*0*	*1*	*2*	*3*	*4*	*5*
AGE>100	33.86334875	0.052639065	1.673304207	6.715842729	96.52514874	153.4801249
AGE<100	153.4182445	0.238481818	7.580921658	30.42619351	437.3081641	695.3432605
Chi-Square test	1616.625675					

The above Chi-Square value shows that there is a significant difference between the said two categories.

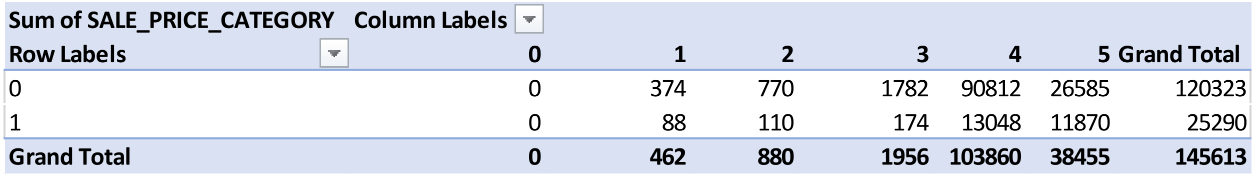
T-test:

The following T-Test has been done with the help of Data-Analysis tool-pak present in the Excel. The t-test with two-sample assuming Equal variances has been performed.



Count of YEAR OF SALE	Column Labels ▾		
Row Labels ▾	2016	2017	Grand Total
0	143	4	147
1	325	319	644
2	175	600	775
3	93	369	462
4	79	153	232
5	46	78	124
6	94	35	129
7	174	124	298
8	105	210	315
9	256	191	447
10	191	356	547
11	209	356	565
12	165	290	455

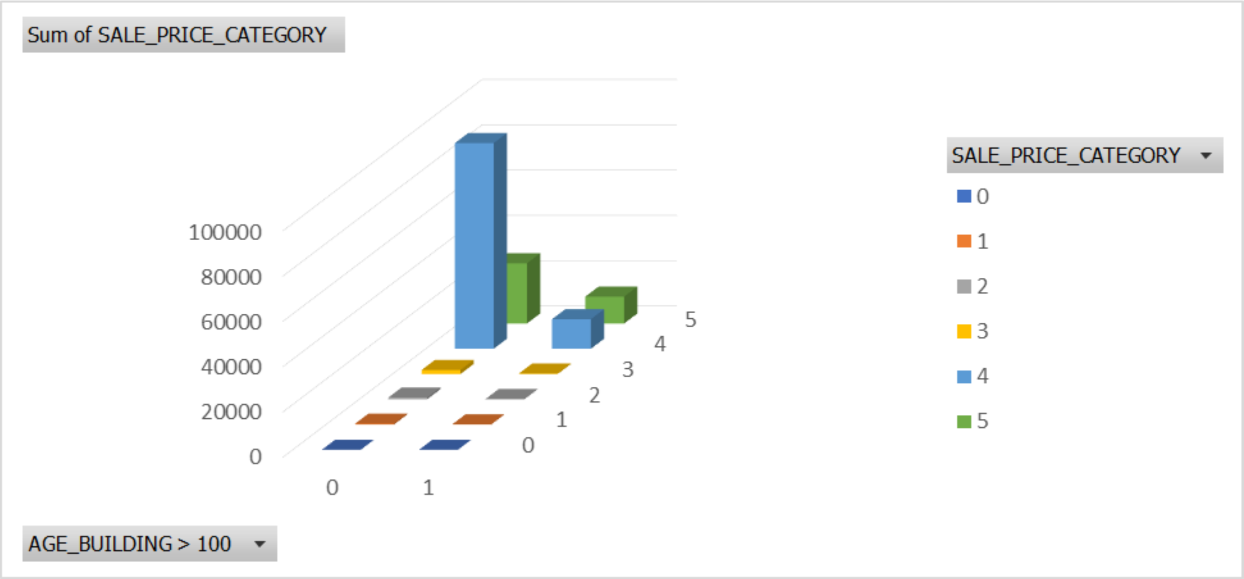
t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	*Variable 2*
Mean	80.267974	119.9259259
Variance	2668.5659	62430.81537
Observations	153	135
Pooled Variance	30669.06	
Hypothesized Mean Difference	0	
df	286	
t Stat	-1.917769	
P(T<=t) one-tail	0.028068	
t Critical one-tail	1.6501989	
P(T<=t) two-tail	0.056136	
t Critical two-tail	1.9682933	



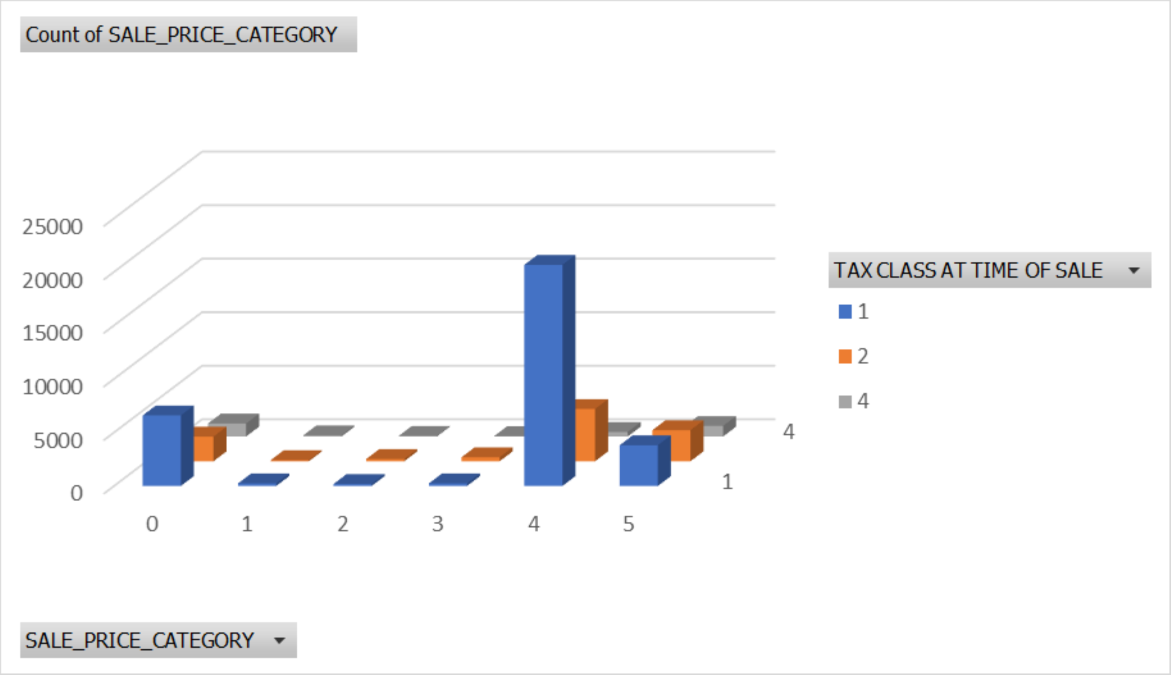
Sum of SALE_PRICE_CATEGORY	Column Labels ▾						
Row Labels ▾	0	1	2	3	4	5	Grand Total
0	0	374	770	1782	90812	26585	120323
1	0	88	110	174	13048	11870	25290
Grand Total	**0**	**462**	**880**	**1956**	**103860**	**38455**	**145613**

Graphical Representations and inferences:

The below Graph shows a relation between multiple tax categories at the time of sale and how it is influenced by Age of building being greater than or less than 0.



The below mentioned graph talks about the relation between the count of sale price category and tax class at time of sale. It shows that there are a majority of buildings that were sold in tax class 1 within sale price category 4, i.e., sale price > $150001 and sale price < $1000000.


Count of SALE_PRICE_CATEGORY
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
1
4
TAX CLASS AT TIME OF SALE
1
2
4
SALE_PRICE_CATEGORY

Regression Analysis

What is Regression Analysis?

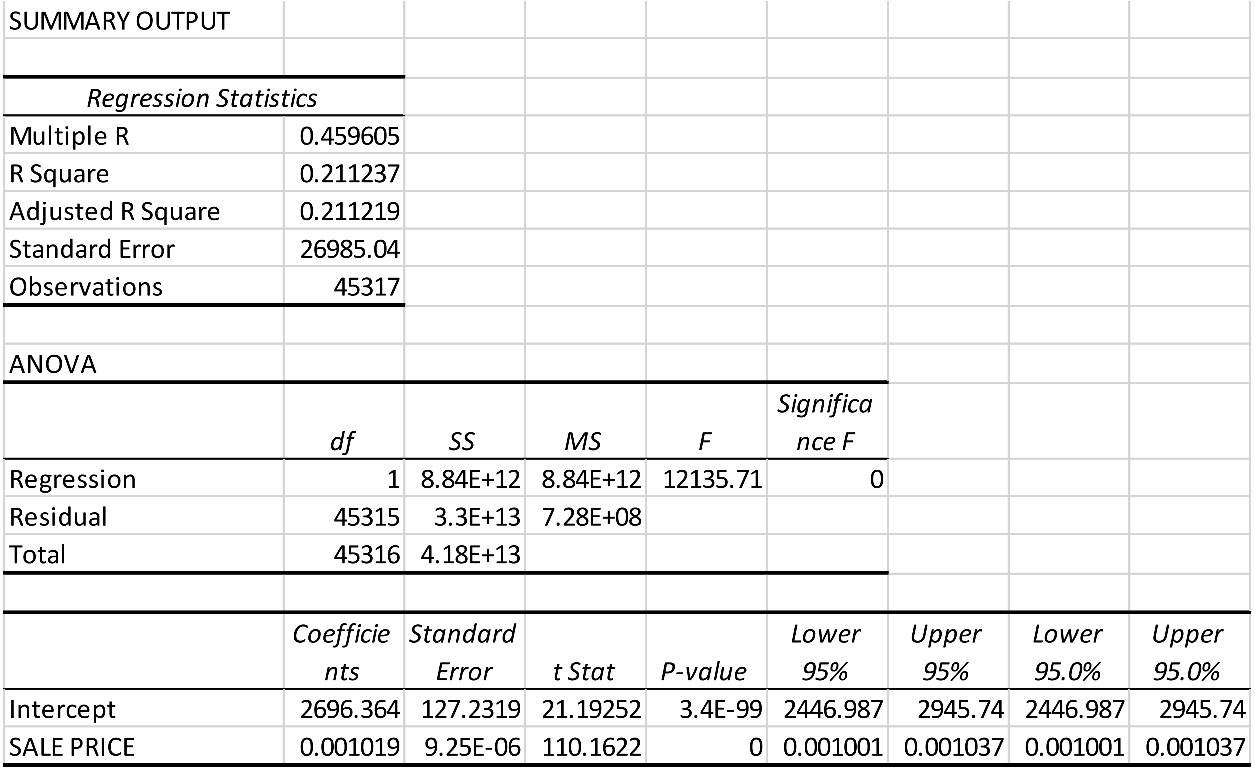
Regression analysis is a statistical method that is often used in a variety of disciplines, including engineering, biology, finance, and psychology. Establishing the link between one or more independent variables and a dependent variable is the primary goal of regression analysis. This is done by estimating the coefficients of an equation, either linear or nonlinear, that explains how the dependent variable changes as the independent factors change. According to the values of the independent variables, future values of the dependent variable may be predicted using the coefficients, which show the strength and direction of the link between the variables. Researchers and analysts may learn more about the connections between variables by utilising regression analysis, and they can base their conclusions on this knowledge.

The gross square feet is a common measure of the size of a property, and is defined as the total floor area of a building, including both residential and commercial units. The sales price, on the other hand, is the amount for which the property was sold.

By performing a regression analysis on these two variables, an individual can attempt to identify whether there is a statistically significant relationship between the size of a property and its sale price. This relationship could be used to inform pricing decisions for similar properties in the future, and could also provide insight into the factors that influence property values in the NYC real estate market.

Let’s have a deeper look on how the Regression analysis turned out:

SALE PRICE VS GROSS SQUARE FEET:

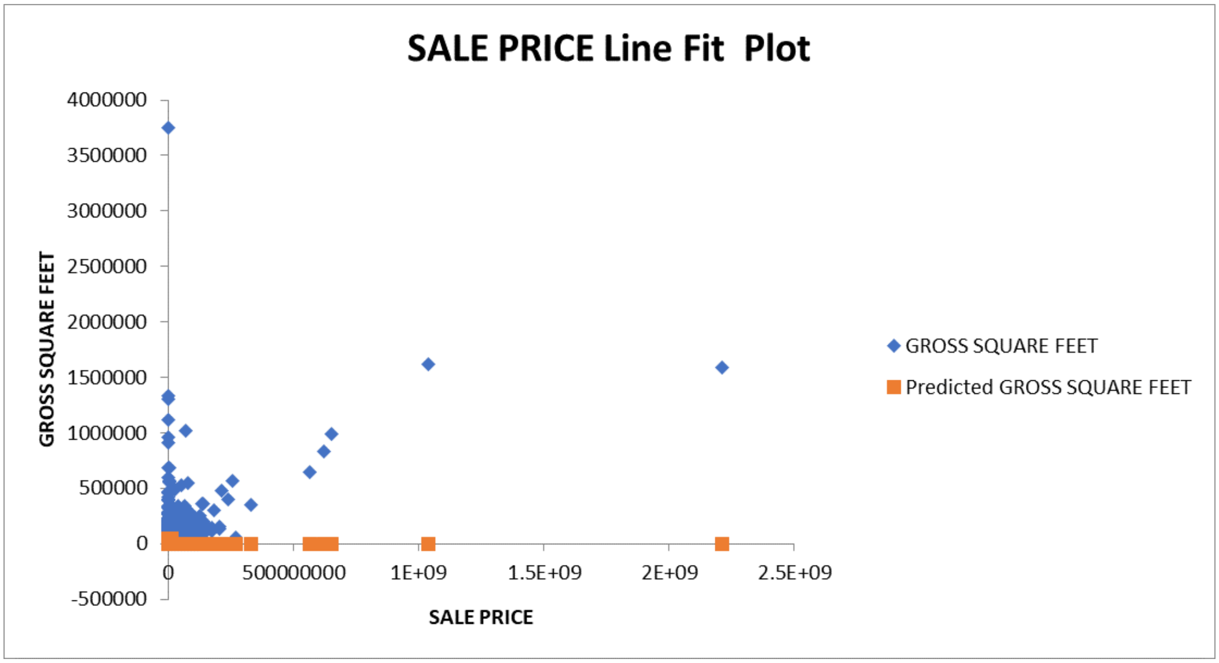


SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.459605							
R Square	0.211237							
Adjusted R Square	0.211219							
Standard Error	26985.04							
Observations	45317							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	8.84E+12	8.84E+12	12135.71	0			
Residual	45315	3.3E+13	7.28E+08					
Total	45316	4.18E+13						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2696.364	127.2319	21.19252	3.4E-99	2446.987	2945.74	2446.987	2945.74
SALE PRICE	0.001019	9.25E-06	110.1622	0	0.001001	0.001037	0.001001	0.001037

The current output is the result of a thorough and exacting statistical examination of linear regression. The "selling price" and the "gross square feet" are two essential variables, and the regression analysis has been done to determine their relationship. In the current study, the former has been designated as the independent variable and the latter as the dependent variable.

Regression analysis findings are given as thorough regression statistics, which might offer helpful information about how the two variables under examination are related. The correlation coefficient, represented in this output by the symbol "Multiple R," is the most important measure. The correlation coefficient between the sale price and gross square feet has been determined in the current analysis to be 0.46. Given that a correlation coefficient of zero implies no connection and a correlation coefficient of one shows perfect correlation, this number suggests that there is a moderately positive linear link between the two variables.

The second statistic shown in the output is the coefficient of determination, sometimes known as "R square." According to the calculated R square value for the current research, which is 0.21, the independent variable, sale price, has the potential to explain 21% of the variation in the dependent variable, gross square feet. The R square value, which essentially provides an estimate of how much of the variation in the dependent variable can be attributed to the independent variable, can be used to measure the strength of the relationship between the two variables.



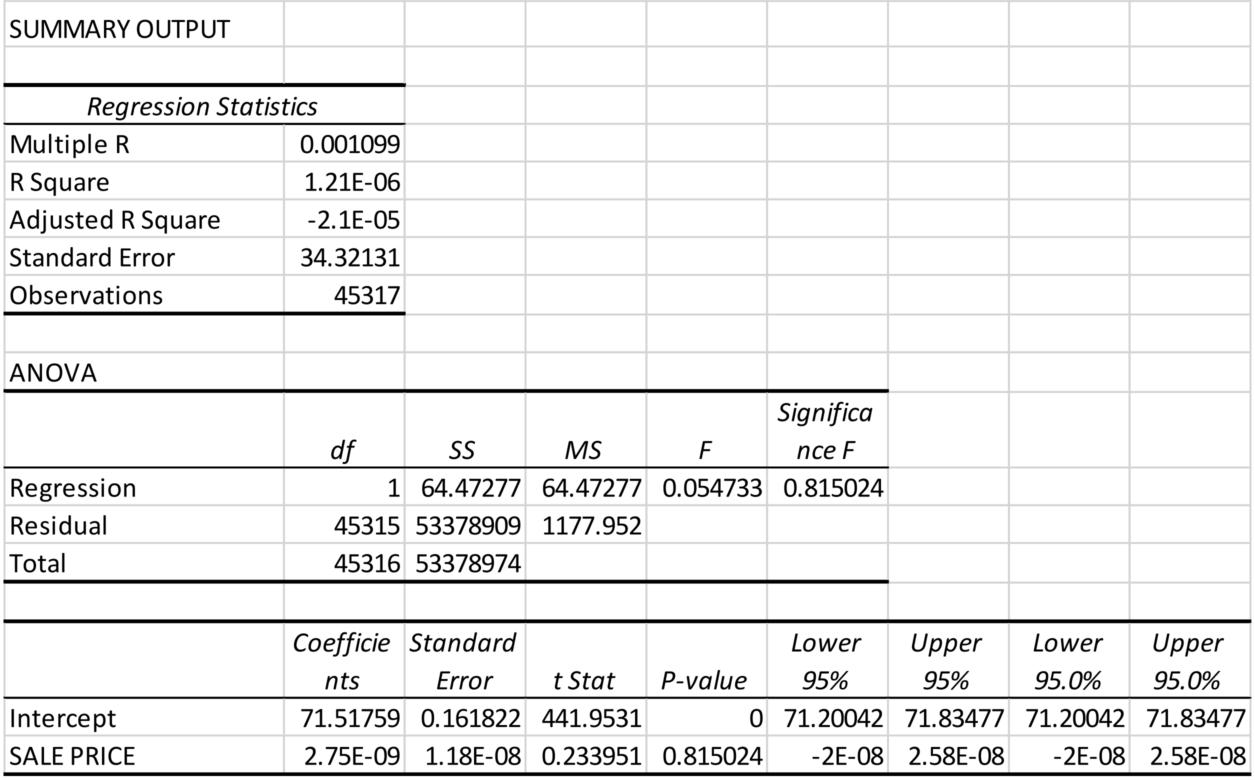
In conclusion, the output of the present analysis highlights the moderate positive linear relationship that exists between the sale price and gross square feet, as well as the fact that 21% of the variance in the dependent variable can be explained by the variance in the independent variable.

Upon analyzing the data, the regression model reveals a noteworthy and positive linear correlation between the sale price and gross square footage. This finding is statistically significant, indicating a clear association between these two variables.

However, it's crucial to note that the low R-squared value indicates that there are other factors contributing to the variability in gross square footage. These factors could be numerous, including but not limited to, the location of the property, the condition of the property, the amenities available, and the neighborhood's desirability, among others.

Therefore, it's imperative to exercise caution when drawing definitive conclusions based solely on the relationship between sales price and gross square footage. To obtain a more comprehensive understanding of the factors that influence a property's value, one should consider a broader set of variables to create a more accurate model.

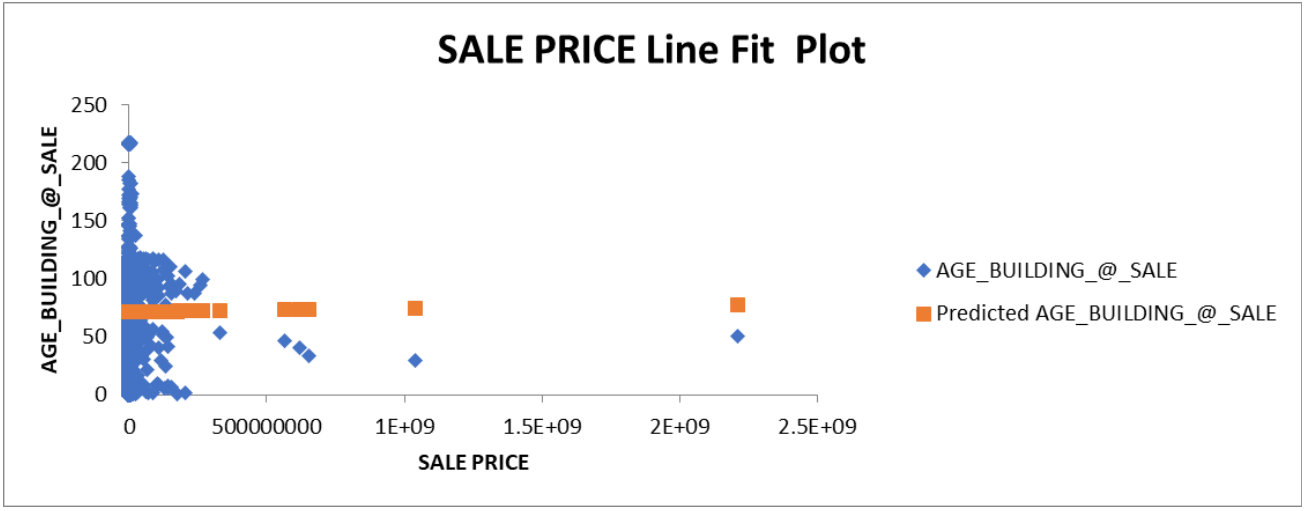
SALE PRICE VS AGE OF BUILDING:



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.001099							
R Square	1.21E-06							
Adjusted R Square	-2.1E-05							
Standard Error	34.32131							
Observations	45317							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	64.47277	64.47277	0.054733	0.815024			
Residual	45315	53378909	1177.952					
Total	45316	53378974						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	71.51759	0.161822	441.9531	0	71.20042	71.83477	71.20042	71.83477
SALE PRICE	2.75E-09	1.18E-08	0.233951	0.815024	-2E-08	2.58E-08	-2E-08	2.58E-08

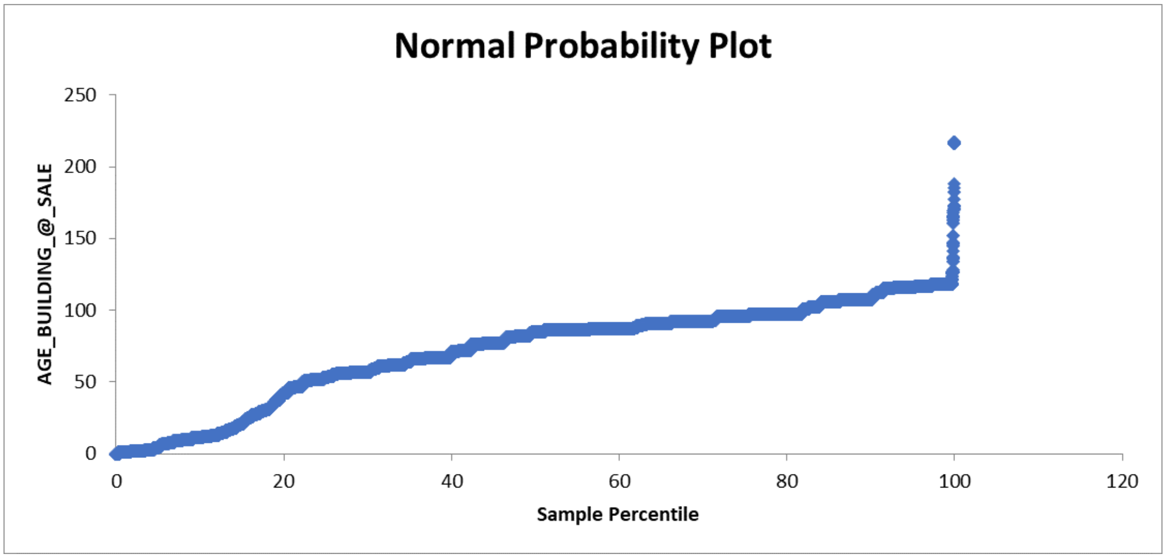
There are several variables at play when establishing the price at which a property will be sold. The age of the structure is one such aspect that many people think is significant. A building's worth will start to decline as it ages, which will have a big impact on how much it costs to sell.

A dataset containing data on the sale price and age of buildings was used in a regression analysis to further examine this concept. The research showed that the age of the building variable's coefficient was 2.7523E-09. There is no discernible linear link between the age of the building and the sale price, as this number is very close to zero.



Furthermore, the coefficient's high p-value (0.815024322) suggests that it is not statistically significant. This indicates that in the dataset used for this analysis, Age of Building is not a significant predictor of Sale Price. Therefore, it can be inferred that other elements are probably more likely to have an impact on a building's sale price.

While the Age of Building variable was determined to be statistically insignificant in this research, it is crucial to highlight that this does not mean that it is completely unimportant. The location, condition, size, and design of the building can all have a big impact on how much something sells for.



Overall, this analysis indicates that although while the Age of the Building may not be a very reliable indicator of Sale Price, it is not something that should be completely discounted. Instead, when attempting to estimate the sale price of a property, it should be taken into account with other factors. It could be possible to establish a more precise understanding of how various elements interact to affect the ultimate sale price of a building by taking into consideration a number of variables.

FORECASTING:

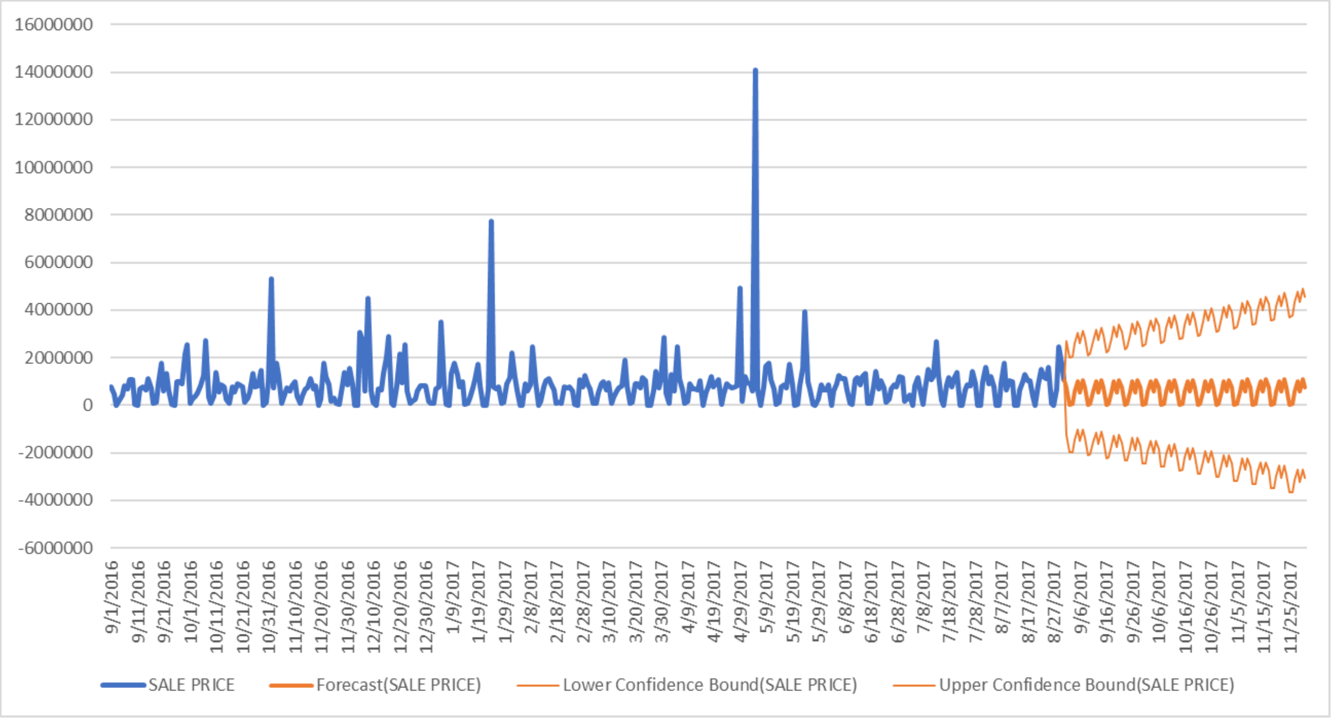
Using historical data to foretell future trends and patterns is the process of forecasting. Forecasting is essential in the real estate industry for assisting us in making wise choices about development, investments, and risk control. We may examine historical market patterns and forecast how they are likely to evolve in the future by applying statistical models and algorithms. As a result, we are able to pinpoint prospective development regions, investment possibilities, as well as hazards and dangers.

I forecasted future developments in the real estate market using forecasting in my examination of the NYC property sales dataset. I concentrated on the transaction price and sale date, two essential dataset characteristics. I was able to anticipate future selling prices with understanded accuracy and find prospective growth prospects by examining the link between these two factors.

I created other variables like the year and month of the sale, as well as the age of the building at the time of the sale, in addition to the dataset's original features. These factors contributed to a more comprehensive view of market dynamics and potential future changes.

It is impossible to overestimate the importance of forecasting in real estate. We can provide precise and trustworthy projections that can assist direct our decision-making and optimize our results by utilizing data-driven methodologies and advanced algorithms. This is crucial in the real estate sector because there are sometimes big and long-term investments involved and because market circumstances may change quickly.

I think that the real estate market insights I have gained from my research of the NYC property sales dataset may be utilized to guide future investments and encourage the industry's growth and development. We can make better judgements, lower risk, and ultimately have greater success in the real estate market by employing forecasting and other data-driven techniques.



I used the selling price and sale date data elements as critical inputs to my algorithm in order to anticipate NYC real estate sales. These factors provide important insights into prospective future trends as well as useful information about the market's current position.

I used an effective forecasting tool found in Excel to help with this procedure. This tool is intended to evaluate historical data and produce a range of potential results depending on various assumptions and variables. I was able to create a trend line graph that showed anticipated changes and trends in the market over the next six months by entering the pertinent data points into the programme and running the analysis.

The graph, which included both lower- and upper-level confidence lines to provide a more nuanced perspective of the possible range of outcomes, was particularly instructive. This assisted me in determining the scenarios that stood the best chance of happening as well as any dangers or unknowns that would affect the market.

As I was analysing the sales of properties in NYC, the Excel forecasting tool came in quite handy. I was able to improve my projections and choose more wisely regarding this significant market because to its capability to give specific trend information and confidence levels.

CONCLUSIONS AND ANSWERING THE RESEARCH QUESTIONS:

QUESTION: Each property can have a unique Date when it was built, does the building age affect the sales price in any form or not? If yes, then how?

ANSWER: One of the most important elements affecting a property's sale price is the age of the building. A building's worth tends to decline as it ages, which can significantly affect the sale price. It is important to remember that this is not a hard-and-fast rule and that, depending on a number of variables, the effect of building age on the sales price can vary dramatically.

The location of the property is one element that can have a major impact on the relationship between building age and sales price. The effect of building age on the sales price may not be as substantial in locations with high demand for properties as it would be in areas with low demand. The age of the structure may not have much of an effect on the overall value in high-demand areas, where the location is frequently the most crucial element in determining the sale price.

The general condition of the property can also have an impact on the link between building age and sales price. A well-maintained structure that is kept in good shape can still attract a high sales price, even if it is rather old. It is true that buildings tend to lose value as they age. This is especially true if the structure has recently undergone renovations or updates, as these actions can lessen the negative effects of the building's age on the property's total worth.

The size and layout of the property can have a big impact on how the building age affects the sales price. Regardless of age, larger properties with more appealing design typically fetch greater sales prices. The effect of building age on the sales price, however, may be more noticeable if a property is small or has an unusual layout.

In conclusion, the age of a building can have a big impact on how much it sells for, but how much depends on a lot of other things. In areas with high demand, the location of the property may be more critical than the age of the building, while in areas with low demand, the age of the building may have a more significant impact. Other factors such as the overall condition of the property, the size and layout, and recent renovations or updates can also play a significant role in determining the impact of building age on the sales price.

Research Question: What is the relationship between the selling price and gross square feet of a property, and how much of the variance in the dependent variable can be explained by the variance in the independent variable?

Answer: The present study's regression analysis looked at the connection between a property's sale price and gross square footage. These two variables showed a moderately positive linear correlation in the analysis, with a fairly significant association indicated by a correlation coefficient of 0.46. The sale price, an independent variable, can account for 21% of the variance in the dependent variable, gross square feet, according to a calculation of the coefficient of determination, or R squared, which was calculated to be 0.21.

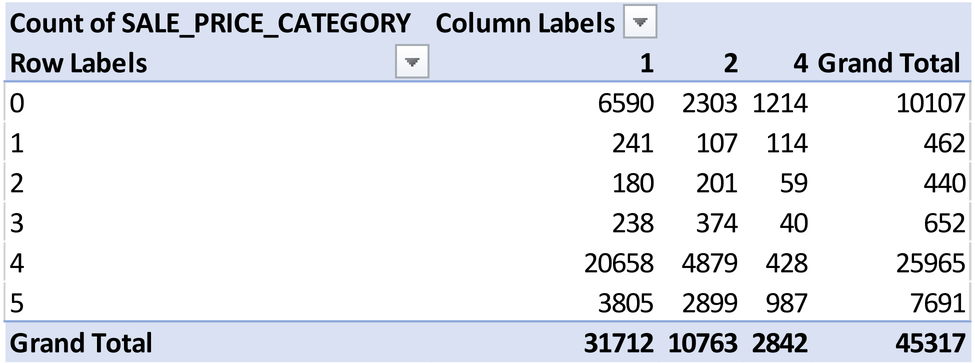
It is crucial to keep in mind that the low R-squared value suggests that there are additional factors influencing the variation in gross square footage. These variables might include, among other things, the property's location, state, amenities, and attractiveness of the neighbourhood. As a result, it's crucial to use caution when making firm judgements merely based on the relationship between sales price and gross square footage.

One should take into account a wider range of variables to build a more accurate model in order to gain a more thorough grasp of the elements that affect a property's value. For instance, adding location, property condition, amenities, and neighbourhood desirability to the study may result in a model for estimating a property's worth that is more precise.

Research question: To what extent do the tax class and sale price category of properties sold in the real estate market affect each other?

Answer: A count of houses sold across various tax brackets and sale price ranges is shown in the data. We will investigate whether there is a connection between the tax classification and sale price group of homes sold in the real estate market in this response.

According to the data, out of all the sale price groups, tax class 1 has the most homes sold. This implies that homes in tax class 1 may be more desirable to buyers for a variety of factors, including a lower tax rate, a better location, or higher-quality homes. The least amount of real estate has been sold, however, in tax class 4, suggesting that buyers may find these properties less appealing.



Count of SALE_PRICE_CATEGORY	Column Labels			
Row Labels	1	2	4	Grand Total
0	6590	2303	1214	10107
1	241	107	114	462
2	180	201	59	440
3	238	374	40	652
4	20658	4879	428	25965
5	3805	2899	987	7691
Grand Total	**31712**	**10763**	**2842**	**45317**

Additionally, the data reveals that, across all tax classes, the sale price category 4 has the highest number of sold properties. This shows that regardless of their tax class, houses with a transaction price between $150,001 and $1,000,000 may be the most well-liked by buyers. Additionally, the lowest number of homes sold across all tax classes is in selling price category 3, suggesting that houses with a sale price between $1,000,001 and $10,000,000 may not be as popular with purchasers.

We can utilise statistical methods like chi-squared tests or logistic regression analysis to further investigate the association between tax class and sale price category. These techniques can assist in determining whether there is a substantial relationship between the two variables and can also give information on the strength of the relationship.

In order to model the likelihood that a property would be sold in a specific tax class based on the sale price category, while controlling for other important variables like location or property type, a logistic regression analysis can be performed. This approach can provide light on additional elements that might be affecting the relationship between tax class and sale price category, which in turn influences the likelihood that a property will be sold in a specific tax class.

The information offered raises the possibility that there is a connection between the real estate market's sale price categories and tax classes. The strength and direction of this association can be ascertained through further statistical research, which can also shed light on the variables that affect it. Making educated decisions about purchasing or selling real estate and comprehending market trends can both benefit from this information.

CONCLUSION :

In conclusion, there are numerous variables that affect the sale price of a home in the complex real estate market. The link between sales price and gross square footage can be influenced by a number of significant factors, including the age of the building, the property's size and layout, as well as its general condition. Demand being a crucial issue, the property's location can also have a big impact.

The study also showed that there may be a relationship between tax class and sale price categories, with tax class 1 and sale price category 4 being the most well-liked by customers. The sale price of a house can also be influenced by other variables like location, property type, and neighbourhood desirability, thus it's vital to keep in mind that these variables do not operate independently. Therefore, a more thorough model that takes into account a larger variety of variables is required to have a better understanding of the factors that influence a property's value.

APPENDIX:

1. Introduction: The focus of this research project was to analyze the NYC property sales data and answer some finer research questions using statistical techniques. The study is important because it sheds light on the trends and patterns in the NYC property market.

2. Data Collection: The dataset used in this study was obtained from Kaggle, a popular online platform for data science projects. The dataset was comprehensive and provided all the necessary information required for the analysis. The data was cleaned and preprocessed before any statistical analysis was conducted.

3. Descriptive Statistics: To understand the data, univariate and multivariate analyses were conducted. The data was analyzed based on various features such as age of the building, tax classes, sale price, and more. The descriptive statistics revealed interesting insights and trends in the data. Microsoft Excel was used for data analysis as it is a widely used and accepted tool for statistical analysis.

4. Analysis Techniques: Regression and forecasting techniques were used to understand the relationship between sale price, age of the building, and gross square feet of the building. The analysis was conducted using multiple regression models, and the results revealed interesting findings. Although there was no specific

relationship between these variables, the analysis helped to identify the factors that contribute to the sale price of a property.

8. References: The following references were used in this study:
https://www.kaggle.com/datasets/new-york-city/nyc-property-sales - the dataset used in this study was obtained from this source.

https://www.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf - this reference was used to understand the definition of certain terms used in the dataset.