FINAL PRESENTATION

GROUP B (ACTIVE MEMBERS)

Adebowale Oluwasanmi

Miguel Acuna Angel Silva

DESCRIPTIVE ANALYSIS

(Seattle, Washington housing dataset 2014/15)

Adebowale Oluwasanmi
Miguel Acuna Angel Silva
Aman Upadhyay

OBJECTIVES

- Cleaning the data.
- Analyzing the data for parameters that a customer would find helpful when looking to buy a house.
- Descriptive analysis of relevant variables.

DATASET

- The variables inside the dataset are:
- ID: A house's ID in the dataset
- Date: Date the house was sold
- Price: Price at which it was sold
- Bedrooms: Number of bedrooms in the house
- Bathrooms: Number of bathrooms in the house
- sqft_living: Square footage of the home

- sqft_lot: Square footage of the lot
- Floors: Number of levels of the house
- Waterfront: House has a view to a waterfront (Yes/No)
- View: House has been viewed
- Condition: How good the condition is overall (1-5)
- Grade: overall grade given to the housing unit, based on King County grading system
- sqft_above: Square footage of the house excluding the basement

- sqft_basement: square footage of the basement.
- yr_built: Year in which the house was built.
- yr_renovated: Year in which the house was renovated
- Zip code
- Lat: Latitude of the house
- Long: Longitude of the house
- sqft_living15: Living room area in 2015 (implies renovations).
- sqft_lot15: Area of lot in 2015 (implies renovations).

LITERATURE REVIEW

- According to Amirhosen and Reza (2019) , for research related to houses in the US, some of the variables worthy of selection are:
  - Variables related to area
  - Number of bedrooms and bathrooms
  - Variables related to the age of the house
  - Variables related to the location of the house

DATA CLEANING

- First, we dropped irrelevant variables such as view, longitude, latitude, sq ft lot.
- Removed duplicates from the data.
- We ensured that variables were in the correct format (by choosing the right measurements, adjusting the number of decimals for easy readability).
- We checked for missing values and found that they were less than 1% for each variable, which was within the normal acceptable range.
- We corrected a few out-of-range values by checking the minimum and maximum for each variable
- Lastly, we found outliers within some variables such as price, sq ft living and we decided to keep them because in relation to the data set, it could lead us to key findings.

DESCRIPTIVE ANALYSIS

1) BEDROOMS

Table 1.
Frequency distribution for Bedroom

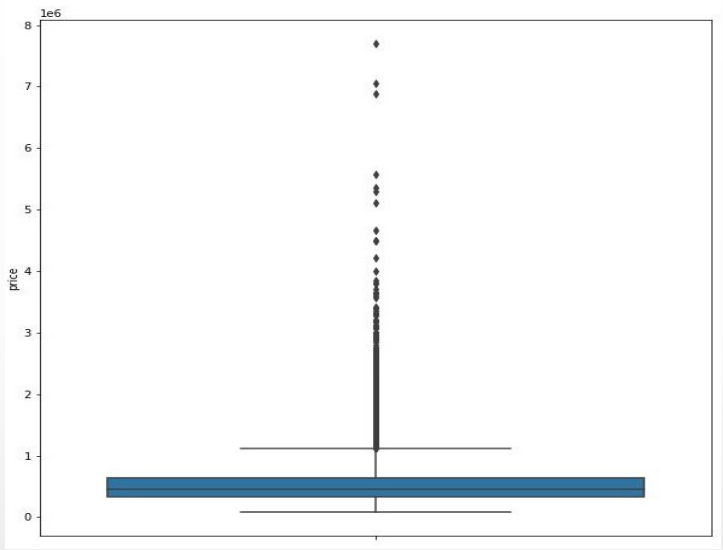
Bedroom	Frequency	Percent
1	196	.9
2	2,760	12.8
3	9,825	45.5
4	6,882	31.9
5	1,601	7.4
6	272	1.3

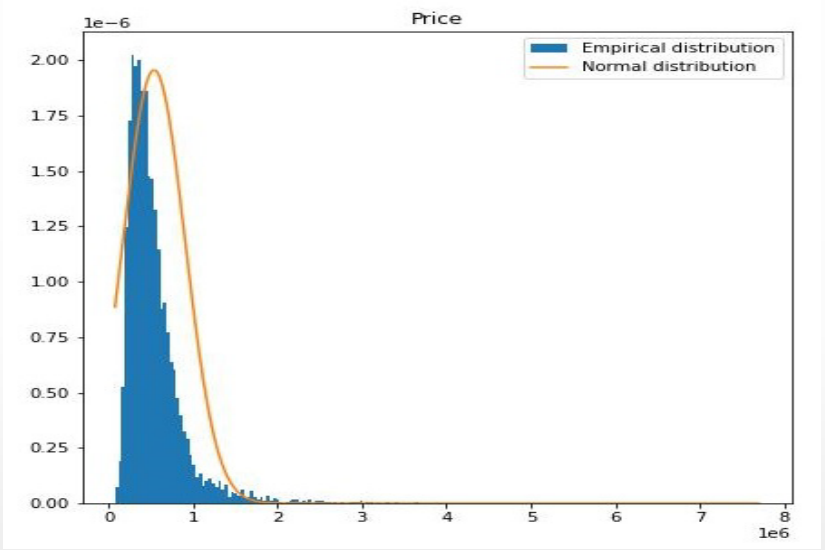
- The sample housing data consisted of 45.5% three bedroom, 31.9% four bedroom and 12.8 % two-bedroom houses (*N* = 21,597).
- This means most of the houses on the market are three- bedroom houses, followed by four-bedroom houses. This is consistent with the size of modern urban families.

DESCRIPTIVE ANALYSIS

2) PRICE

The average price of a house was 540,383.75 (*SD = 367,405.17*).





- It is easy to see that prices are very skewed to the right, with many outliers.

- For interpretation purposes, this means that houses tend to have prices on the lower end of the spectrum, but the astronomical prices on the higher end are enough to skew the whole distribution.

2) PRICE (cont'd)

- As can be seen in the previous visuals, histogram plot also confirms it, the price variable does not follow a normal distribution.

- For practical purposes, price not following a normal distribution means that missing values can be replaced with the median (the percentage of missing values was under 1%), and that further analysis must be conducted with and without outliers.

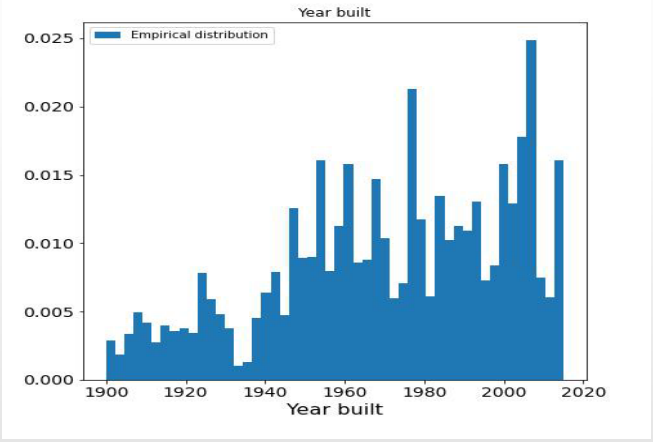
3) YEAR BUILT & YEAR RENOVATED

Table 2.

Descriptive statistics showing year built and year renovated

		Minimum	Maximum
Yr built		1900	2015
Yr renovated		0	2015

- Histogram showing year built distribution



- The sample data showed that the oldest houses were built in the year 1900, while the newest houses were built in the year 2015.
- Descriptive analysis also showed that the most recent renovations were done in the year 2015.

4) HOUSE CONDITION

Table 3.

Frequency distribution for the condition of houses.

Condition		Frequency	Percent	
	Very Poor	29	.1	
	Poor	170	.8	
	Average	14,018	64.9	
	Good	5,676	26.3	
	Excellent	1,701	7.9	

- Descriptive analysis of the housing condition showed that of 64.9% of houses were in average condition, 26.3% were in good condition and 7.9 % were in excellent condition (*N* = 21,594).

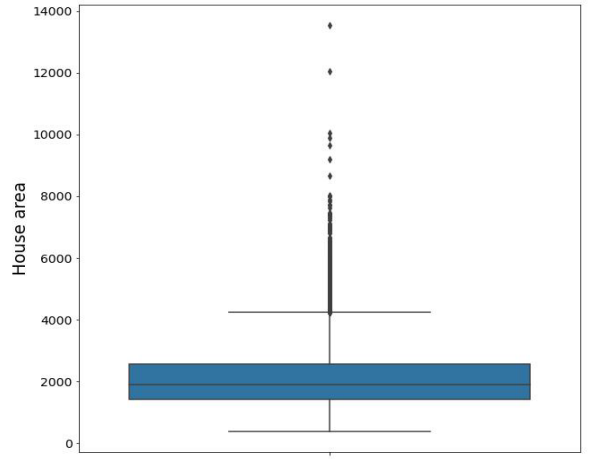
5) SQ. FOOTAGE OF THE HOUSE

Table 4.

Descriptive statistics of sq. ft living area

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
sqft_living	21440	370	10040	2061.69	876.870
Valid N (listwise)	21440				



- Average sq. ft living area is 2061.69 (*SD*=876.870)
- The median sq. ft living area is 1910, which is smaller than the median sq. ft living in Seattle as at 2020(Seattle metro lot size decreasing by over 30% over the past two decades (storagecafe.com)).
- Box plot shows the presence of many outliers which may be influenced by other variables. This will be explored during further analysis.

CROSS TABULATION

- **A)** We compared 2 categorical variables, bedrooms and condition.

bedrooms * condition Crosstabulation

			condition					
			Very Poor	Poor	Average	Good	Excellent	Total
bedrooms	1	Count	4	10	123	47	12	196
		% within bedrooms	2.0%	5.1%	62.8%	24.0%	6.1%	100.0%
		% within condition	13.8%	5.9%	0.9%	0.8%	0.7%	0.9%
	2	Count	12	51	1779	718	200	2760
		% within bedrooms	0.4%	1.8%	64.5%	26.0%	7.2%	100.0%
		% within condition	41.4%	30.0%	12.7%	12.6%	11.8%	12.8%
	3	Count	8	69	6307	2710	729	9823
		% within bedrooms	0.1%	0.7%	64.2%	27.6%	7.4%	100.0%
		% within condition	27.6%	40.6%	45.0%	47.7%	42.9%	45.5%
	4	Count	4	36	4580	1682	580	6882
		% within bedrooms	0.1%	0.5%	66.6%	24.4%	8.4%	100.0%
		% within condition	13.8%	21.2%	32.7%	29.6%	34.1%	31.9%
	5	Count	0	1	1030	418	151	1600
		% within bedrooms	0.0%	0.1%	64.4%	26.1%	9.4%	100.0%
		% within condition	0.0%	0.6%	7.3%	7.4%	8.9%	7.4%
	6	Count	1	3	158	87	23	272
		% within bedrooms	0.4%	1.1%	58.1%	32.0%	8.5%	100.0%
		% within condition	3.4%	1.8%	1.1%	1.5%	1.4%	1.3%

- Three-bedroom houses had the highest percentage of houses in Poor, Average and Excellent condition.

- **B)** We also compared variables, bedrooms and waterfront

bedrooms * waterfront Crosstabulation

			waterfront		
			No waterfront	Waterfront Present	Total
bedrooms	One Bedroom	Count	191	5	196
		% within bedrooms	97.4%	2.6%	100.0%
		% within waterfront	0.9%	3.2%	0.9%
	Two Bedrooms	Count	2727	31	2758
		% within bedrooms	98.9%	1.1%	100.0%
		% within waterfront	12.8%	20.0%	12.9%
	Three Bedrooms	Count	9755	64	9819
		% within bedrooms	99.3%	0.7%	100.0%
		% within waterfront	45.8%	41.3%	45.8%
	Four Bedrooms	Count	6817	39	6856
		% within bedrooms	99.4%	0.6%	100.0%
		% within waterfront	32.0%	25.2%	32.0%
	Five Bedrooms	Count	1551	14	1565
		% within bedrooms	99.1%	0.9%	100.0%
		% within waterfront	7.3%	9.0%	7.3%
	Six Bedrooms	Count	254	2	256
		% within bedrooms	99.2%	0.8%	100.0%
		% within waterfront	1.2%	1.3%	1.2%
Total		Count	21295	155	21450
		% within bedrooms	99.3%	0.7%	100.0%

- Descriptive analysis show that houses without waterfronts are more than houses with a waterfront on the market.
- Three-bedroom houses have the highest percentage of houses with and without a waterfront.
- This is likely because, as seen earlier, most of the houses on the market are three-bedroom houses.

REFERENCES

- Amirhosein Jafari, Reza Akhavian, (2019) "Driving forces for the US residential housing price: a predictive analysis", Built Environment Project and Asset Management, https://doi.org/10.1108/BEPAM-07-2018-0100.

- Seattle's Lot Sizes Shrinking: Big-Yard Homes Increasingly Far From The Urban Core by Maria Gatea (2021) , Seattle metro lot size decreasing by over 30% over the past two decades (storagecafe.com)

THANK YOU

GROUP B

TEST FOR NORMALITY

Adebowale Oluwasanmi
Miguel Angel Acuña Silva

GROUP B

TEST FOR NORMALITY

Adebowale Oluwasanmi
Miguel Angel Acuña Silva

Objectives

To determine :

- What are the characteristics of a piece of real estate that has the most influence on price?.
- Do the relevant variables follow a normal distribution?.
- Correlation between relevant variables.

Sampling method

- We used random sampling technique in SPSS to get the sample size (approx. 10% of the entire population).

- Sample size (N = 2,105).

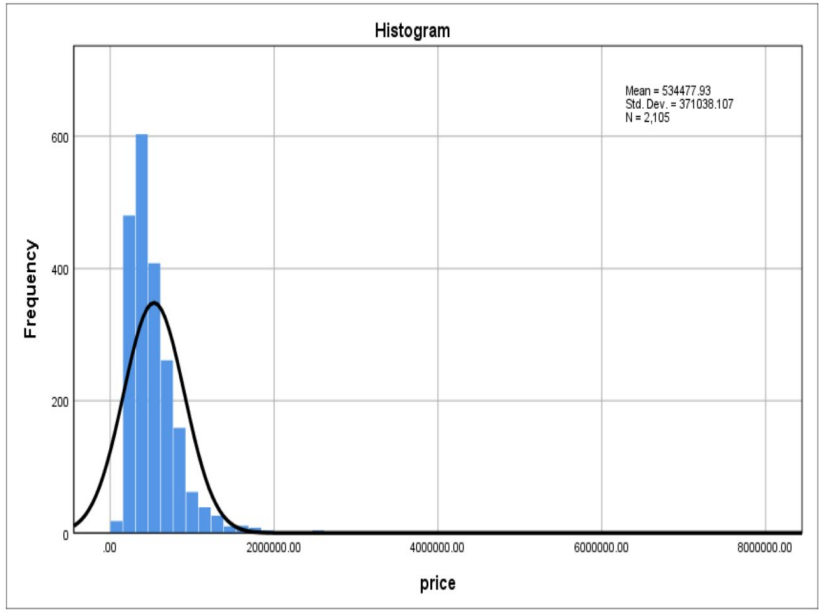
INTRODUCTION

- According to Burinskiene, Rutzkiene and Venckauskaite (2011), important factors that have influence on price are location, prestige, age of building and if it was renovated or not.
- We will be testing some of the parameters formerly mentioned, including price for normality. We'll be also including the sq. footage of the house, to be able to compare with a variable outside the suggested ones.

1. PRICE (Normality tests)

NULL HYPOTHESIS (*Ho*) – The value for prices are normally distributed within the population.

❖ Histogram showing price distribution.



❖ Table showing Statistics and Std. error

Skewness	5.131	.053
Kurtosis	58.229	.107

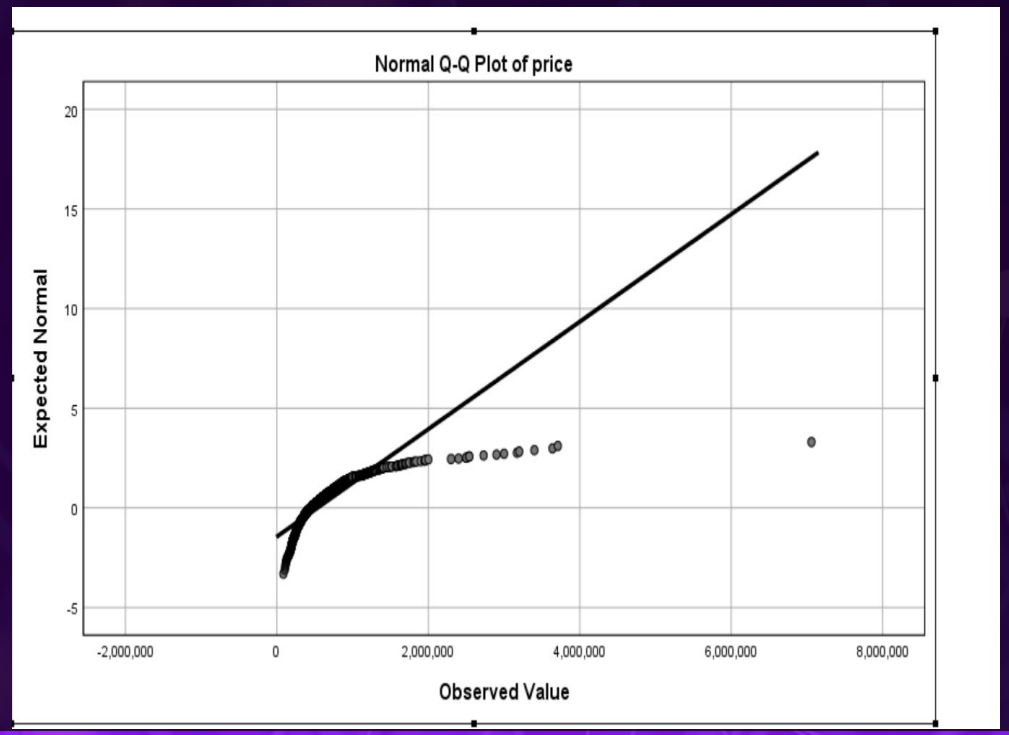
- *Z value (Skewness)*
  = 5.131/0.053
  = **96.8**
- *Z value (Kurtosis)*
  = 58.229/0.107
  = **544.19**

- The z-value for both skewness and kurtosis exceed the ±1.96 threshold.
- The z-values also confirm that the distribution has positive skewness and positive kurtosis.

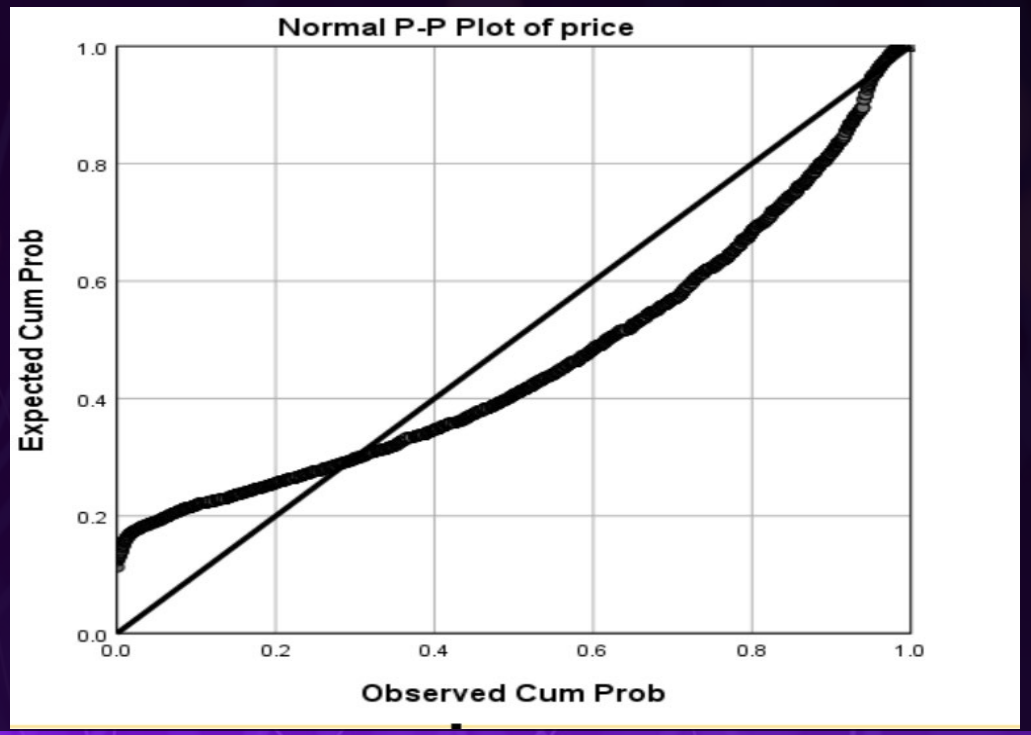
Findings

- The results and observations above show that the distribution is **rightly skewed** , this is because extreme values of prices in the tail affect the mean more than the median.
- Results also show that the distribution is highly kurtotic (positive kurtosis), which indicates the presence of a lot of outliers.

- Q-Q plot show a significant departure of price values from normality **at the tails.**

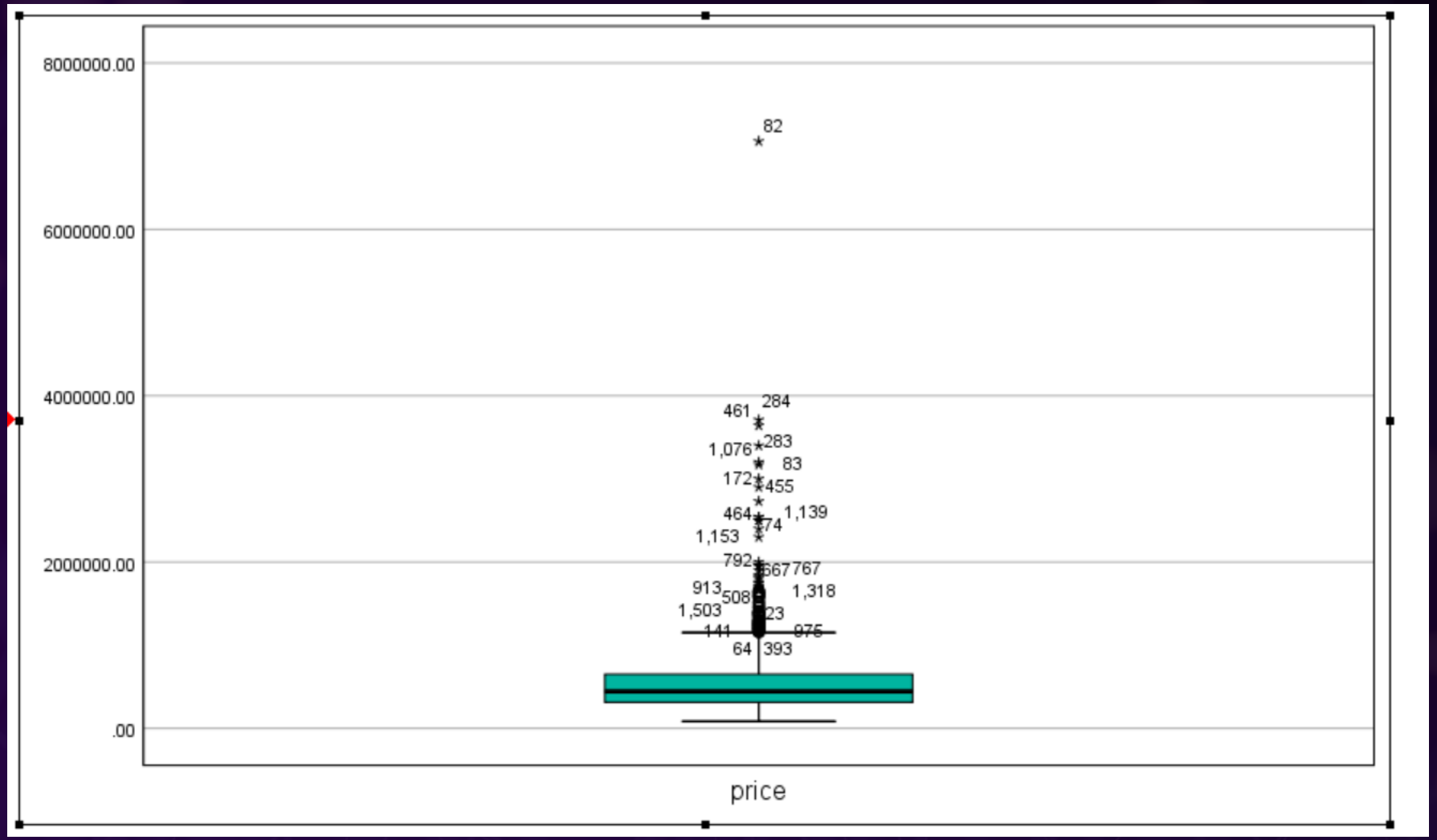


- P-P plot show a significant departure of price values from normality **at the centre**.



BOX PLOT (PRICE)

- Box plot show that the tail is longer on the right, which confirms that the distribution has a positive skew.



STATISTICAL TESTS (PRICE)

- **Shapiro-Wilk & Kolmogorov-Smirnov tests**

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
price	.154	2105	.000	.678	2105	.000

a. Lilliefors Significance Correction

- As shown above, both the Shapiro-Wilk and Kolmogorov-Smirnov tests agree that price violates the assumption of normality with $P<0.05$ for both tests.

CONCLUSION

- Visual inspection of the histogram, normal Q-Q plots, P-P plots, box plots and statistical tests such as the Shapiro-Wilk & Kolmorov-Smirnov tests showed that the prices were not normally distributed, with a skewness of 5.131 (SE= 0.053) and kurtosis of 58.229 (SE= 0.107).

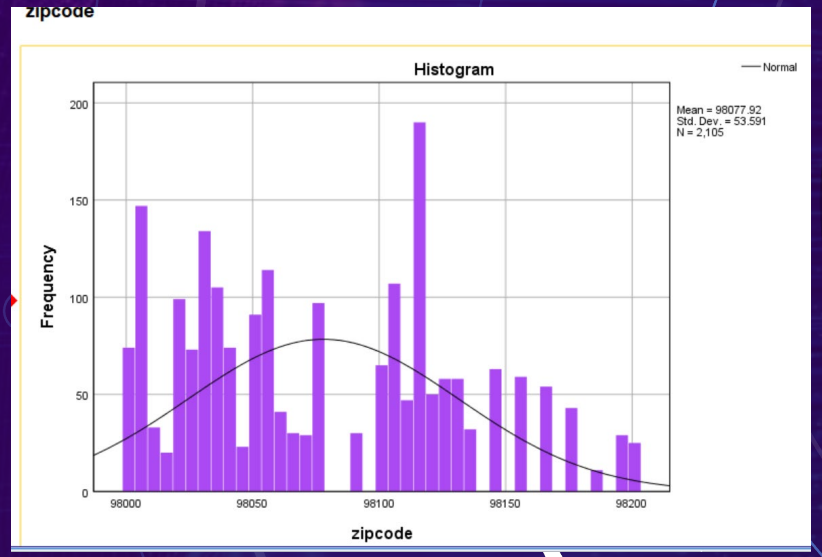
- In conclusion, we reject the null hypothesis.

2) ZIPCODE (Location)

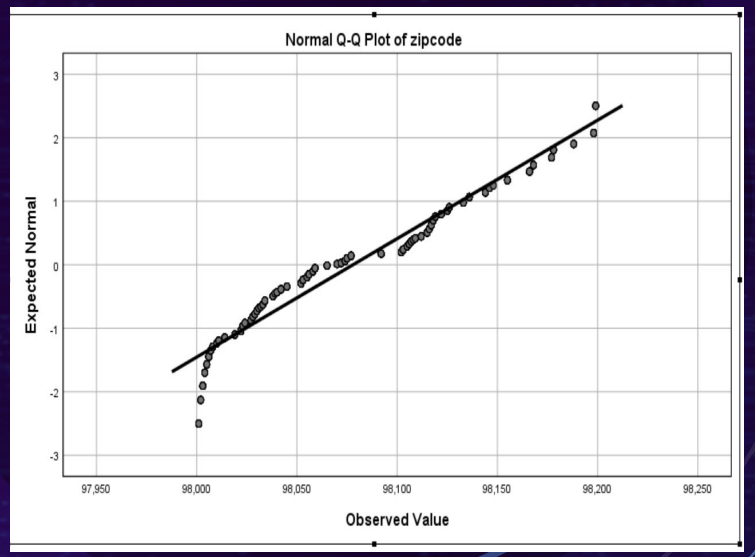
NULL HYPOTHESIS (*Ho*) – The values for zip code follow a normal distribution.

- VISUAL TESTS FOR NORMALITY

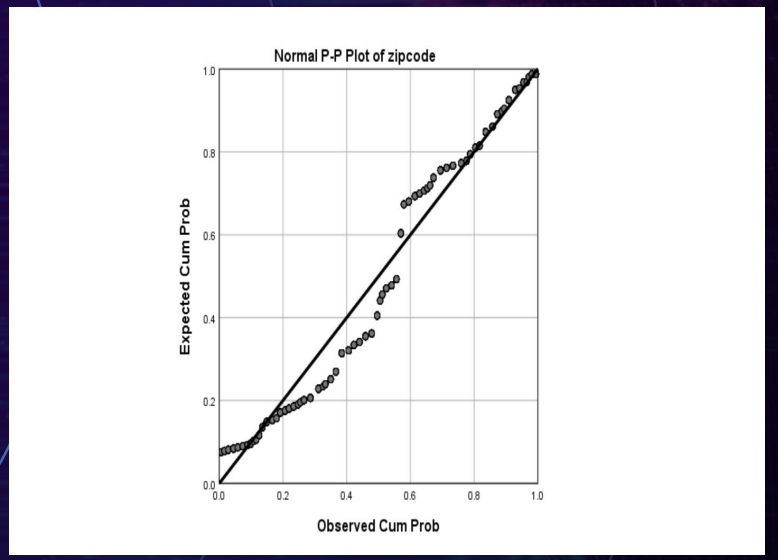
- **Histogram showing zip code distribution**



- **Q-Q plot for zip code**

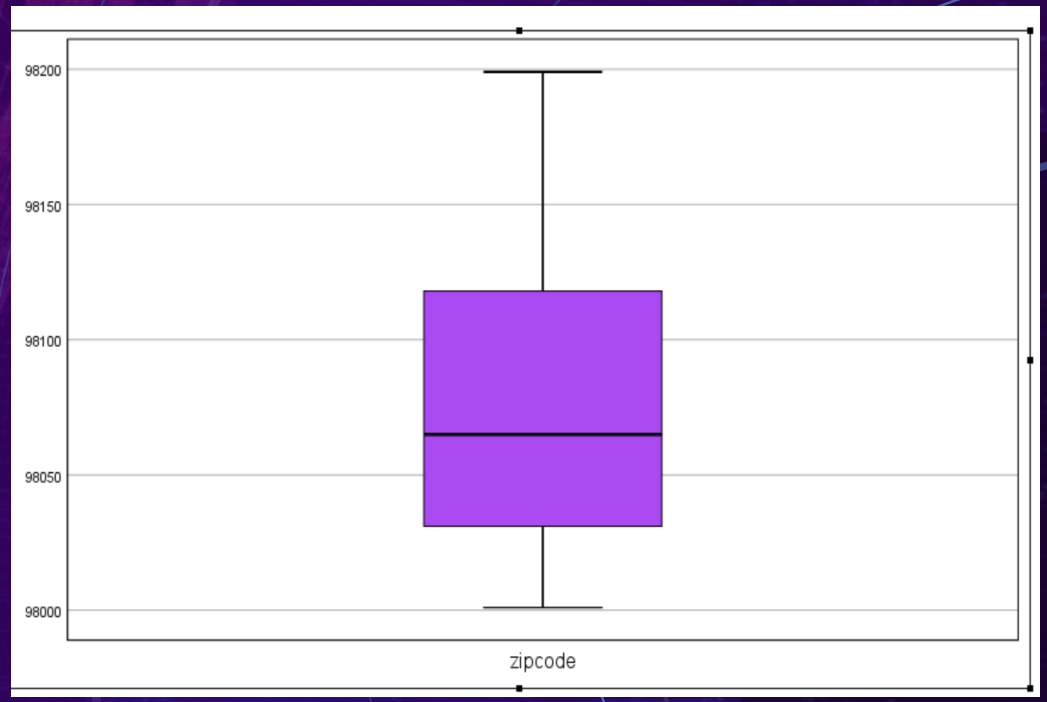


- **P-P plot for zip code**



VISUAL TESTS FOR NORMALITY (CONT'D)

❖ **Box Plot showing zip code distribution**



FINDINGS

- The histogram reveal that the zip code distribution is slightly skewed to the right.
- Q-Q plots shows a slight departure from normality at the tails of the distribution.
- P-P plot shows significant departure from normality at the centre of the distribution.
- The longer tail to the right on the box plot visual confirms that the distribution for zip code values is rightly skewed.

STATISTICAL TESTS (ZIPCODE)

Shapiro-Wilk & Kolmorov - Smirnov tests

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
zipcode	.126	2105	.000	.943	2105	.000

a. Lilliefors Significance Correction

- The Shapiro-Wilk and Kolmogorov-Smirnov tests confirm that zip code violates the assumption of normality with $P<0.05$ for both tests.

Z Value (Skewness & Kurtosis)

Skewness	.368	.053
Kurtosis	-.919	.107

- *Z value (Skewness)*
  = 0.368/0.053
  = **6.94**
- *Z value (Kurtosis)*
  = 0.919/0.107
  = **-0.81**

- The z-value for skewness confirms that the zip code distribution is positively skewed, with 6.94 exceeding ± 1.96.
- While the z-value for kurtosis falls between the threshold of ± 1.96.

CONCLUSION

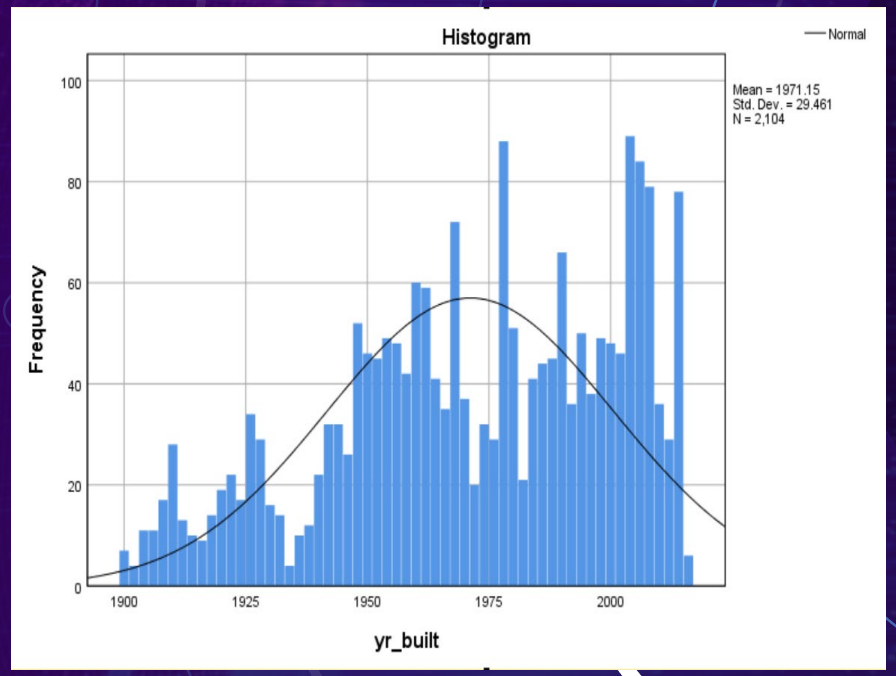
- Visual tests i.e histogram, normal Q-Q plots, P-P plots, box plots and statistical tests such as the Shapiro-Wilk & Kolmorov-Smirnov tests showed that values for zip code were not normally distributed, with a skewness of 0.368 (SE= 0.053) and kurtosis of -0.919 (SE= 0.107).
- Therefore, we reject the null hypothesis.

3) YEAR BUILT

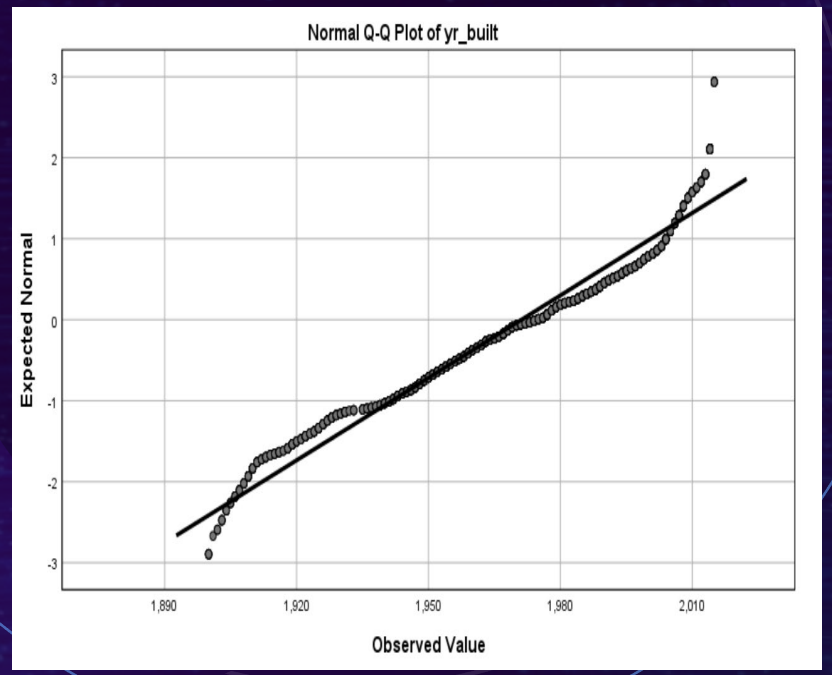
NULL HYPOTHESIS (*Ho*) – The values for year-built variable follow a normal distribution.

❖ VISUAL TESTS

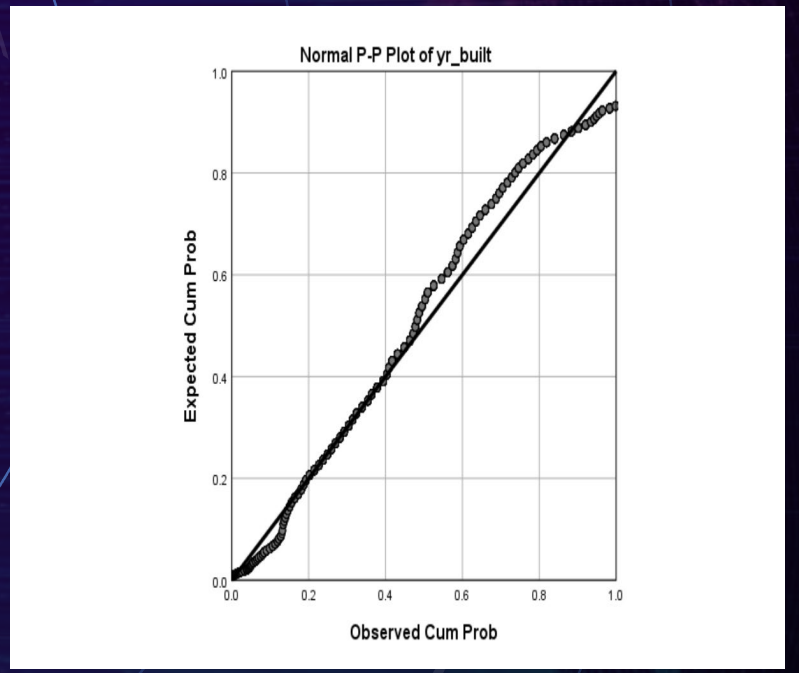
❖ **Histogram showing year-built distribution**



❖ **Q-Q plot for year-built**

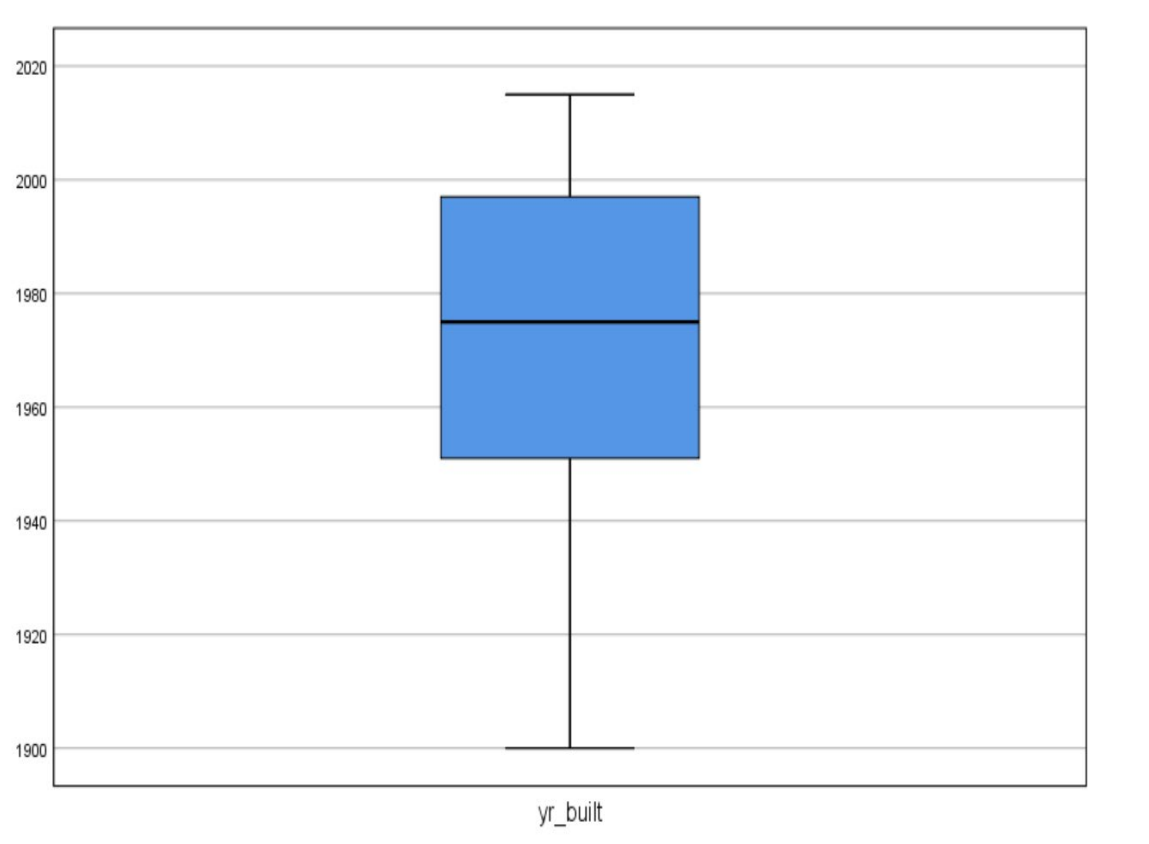


❖ **P-P plot for year-built**



VISUAL TESTS (CONT'D)

❖ **Box Plot showing year-built distribution**



FINDINGS

- The histogram reveal that the year-built distribution is significantly skewed to the left,
- Q-Q plots shows a slight departure from normality at the tails of the distribution.
- P-P plot shows significant departure from normality at the centre of the distribution.
- The longer tail to the left on the box plot visual confirms that the distribution for year-built values is left skewed.

STATISTICAL TESTS (YEAR BUILT)

Shapiro-Wilk & Kolmorov - Smirnov tests

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
yr_built	.077	2104	.000	.954	2104	.000

a. Lilliefors Significance Correction

- The Shapiro-Wilk and Kolmogorov-Smirnov tests confirm that year-built distribution violates the assumption of normality with P<0.05 for both tests.

Z Value (Skewness & Kurtosis)

Skewness	-.450	.053
Kurtosis	-.699	.107

- *Z value (Skewness)*
  = -0.450/0.053
  = **-8.4**
- *Z value (Kurtosis)*
  = -0.699/0.107
  = **-6.53**

- The z-value for skewness confirms that the year-built distribution is negatively skewed, with -8.4 exceeding1.96.
- The z-value for kurtosis -6.53,also exceeds the threshold of 1.96.

Conclusion

- Visual tests i.e., histogram, normal Q-Q plots, P-P plots, box plots and statistical tests such as the Shapiro-Wilk & Kolmorov-Smirnov tests showed that values for year built were not normally distributed, with a skewness of -0.450 (SE= 0.053) and kurtosis of -0.699 (SE= 0.107).

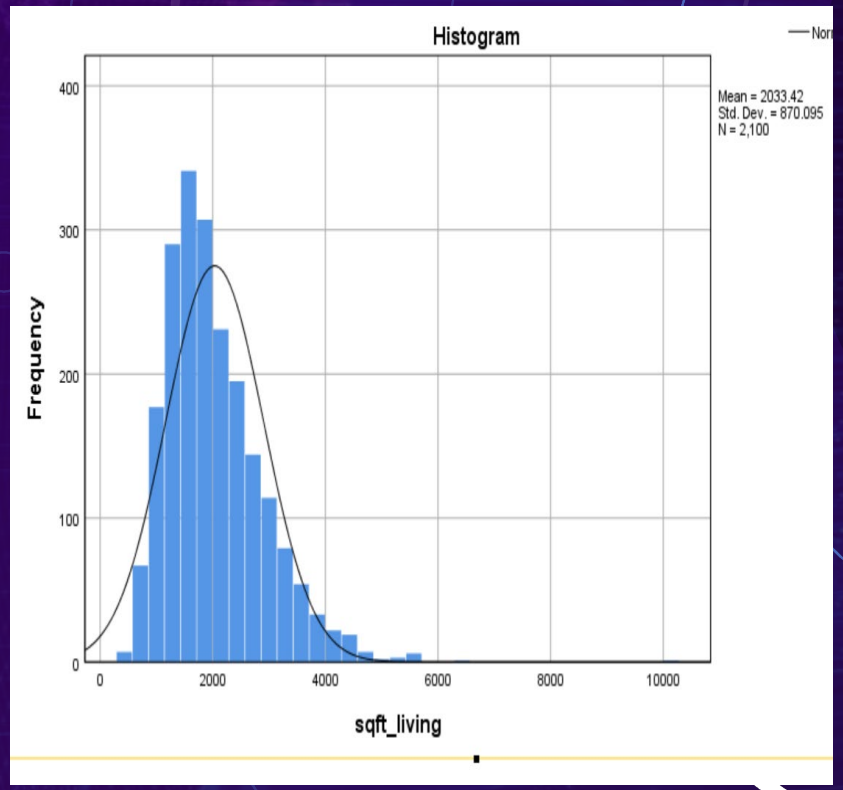
- We reject the null hypothesis.

4) SQ. FT LIVING (SQ. FOOTAGE OF THE HOUSE)

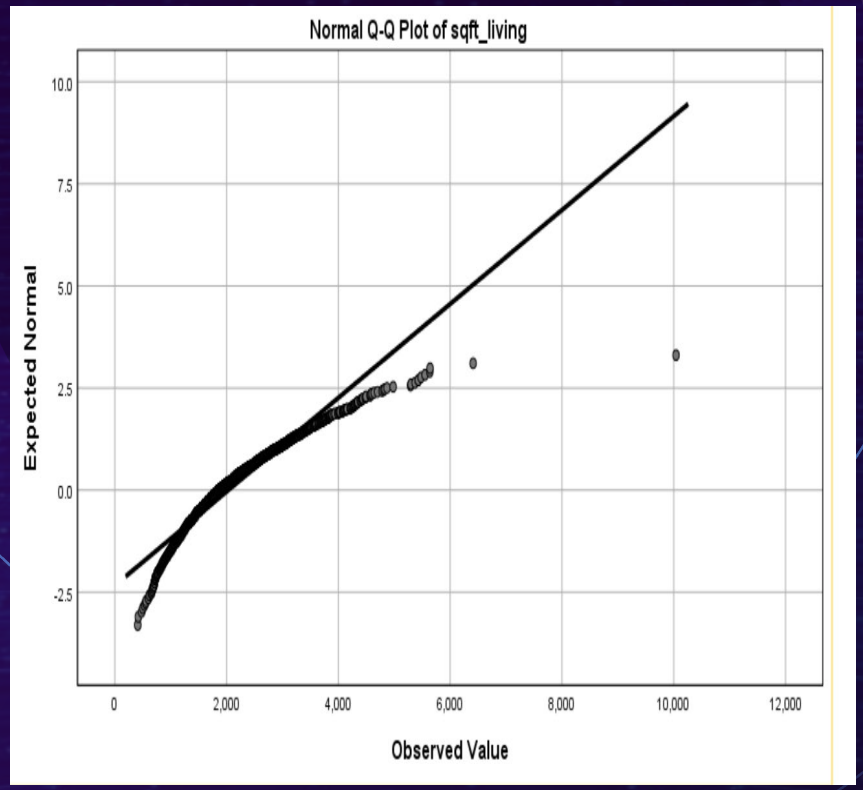
NULL HYPOTHESIS (*Ho*) – The values for square footage of the home variable follow a normal distribution.

- VISUAL TESTS

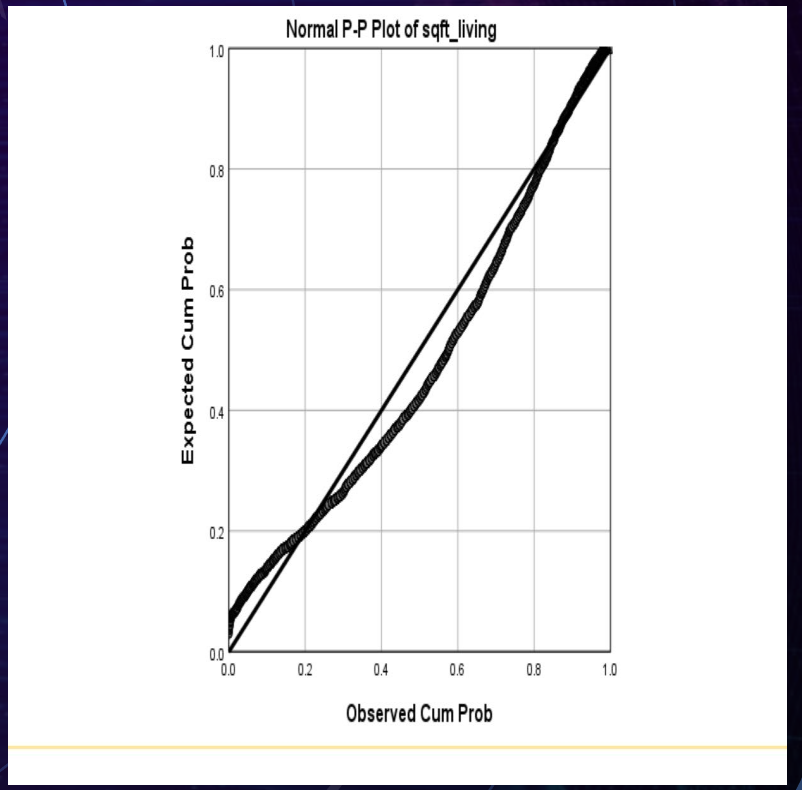
- **Histogram showing sq. ft distribution**



- **Q-Q plot for sq. ft living**

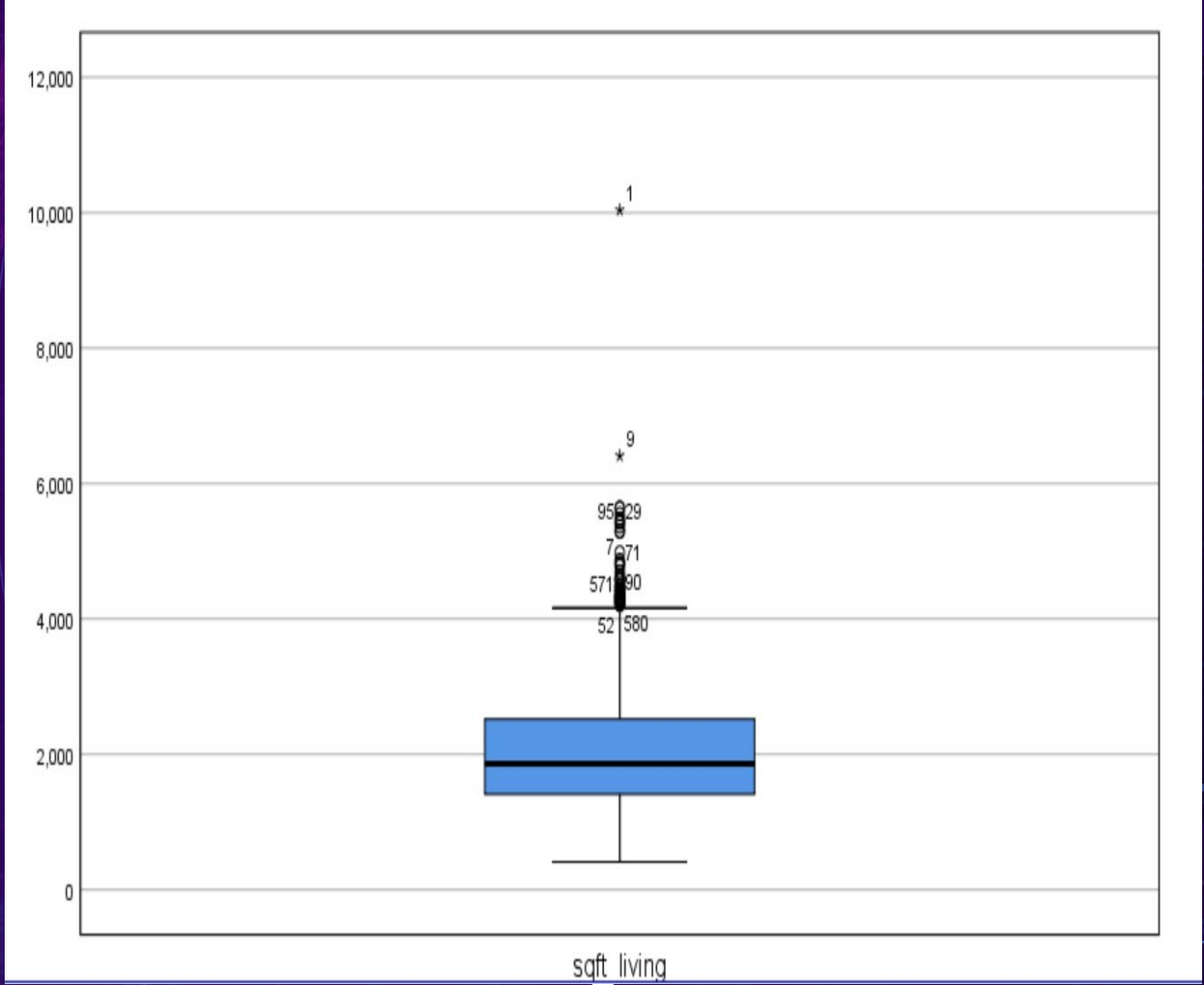


- **P-P plot for sq. ft living**



VISUAL TESTS (CONT'D)

❖ **Box Plot showing sq. ft living distribution**



FINDINGS

- The histogram reveal that the sq. ft living distribution is significantly skewed to the right,
- Q-Q plots shows a very significant departure from normality.
- P-P plot shows very significant departure from normality.
- The longer tail to the right of the boxplot confirms the sq. ft living variable is rightly skewed, with many outliers.

STATISTICAL TESTS (Sq. ft living)

Shapiro-Wilk & Kolmorov - Smirnov tests

- The Shapiro-Wilk and Kolmogorov-Smirnov tests both indicate a p-value of less than 0.05, small enough to reject the null hypothesis that the sq. ft living variable is normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
sqft_living	.083	2100	.000	.929	2100	.000
sqft_lot	.372	2100	.000	.222	2100	.000

a. Lilliefors Significance Correction

Z Value (Skewness & Kurtosis)

Skewness	1.290	.053
Kurtosis	4.317	.107

- *Z value (Skewness)*
  = 1.290/0.053
  = **24.3**
- *Z value (Kurtosis)*
  = 4.317/0.107
  = **40.3**

Conclusion

- Visual tests such as histogram, normal Q-Q plots, P-P plots, box plots and statistical tests such as the Shapiro-Wilk & Kolmogrov-Smirnov tests showed that values for sq. ft living variable were not normally distributed, with a skewness of 1.290 (SE= 0.053) and kurtosis of 4.317 (SE= 0.107).

- We reject the null hypothesis.

- All relevant variables to our objective violated the assumption of normality test. Therefore, non-parametric tests would be best suited for further analysis.

Probability (Calculating Z-Scores)

QUESTIONS

1) What is the probability of 4-bedroom houses, that cost more than 1,000,000?.
2) What is the probability of houses without a waterfront, measuring less than 4,980 sq. ft living?.
3) What is the probability of houses in excellent condition, that was built before 1992?.

ANSWERS

price	bedrooms	Zprice
909500.00	4	1.01074
1190000.00	4	1.76672
1050000.00	4	1.38940
792500.00	4	.69541
997950.00	4	1.24912
3400000.00	4	7.72299
1000000.00	4	1.25465

- P (x=1,000,000)
- P (z=1.25) = 0.894 * 100
  = 89.4 %
- P (x>1,000,000) = P (z>1.25)
  = 100-89.4
  = **10.6%**

➢ There is a 10.6% chance that 4-bedroom houses will cost more than 1,000,000.

Probability (Calculating Z-Scores)

2)

sqft_living	waterfront	Zsqft_living
5300	0	3.75670
5290	0	3.74520
4980	0	3.38874

- P (x=4980)
- P (z=3.39) = 0.999 * 100
- P (x<4980) = **99.9%**
- There is a 99.9% chance that houses without a waterfront will measure less than 4980 sq. ft

3)

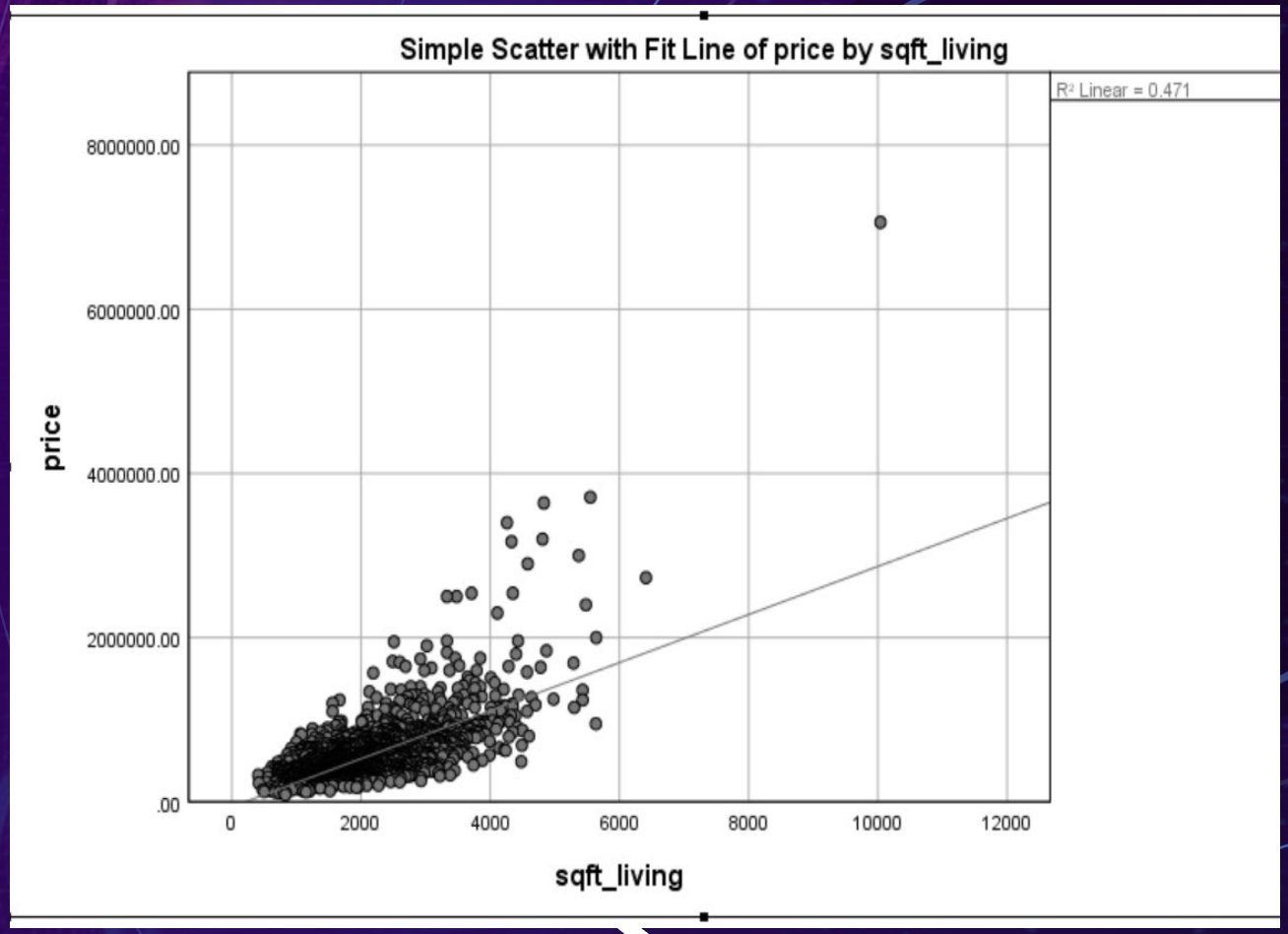
condition	yr_built	yr_renovated	zipcode	Zyr_built
5	1969	0	98040	-.07292
5	1966	0	98040	-.17475
5	1966	0	98008	-.17475
5	1952	0	98004	-.64996
5	1908	0	98112	-2.14347
5	1925	0	98109	-1.56643
5	1992	0	98027	.70778

- P (x=1992)
- P (z=0.70) = 0.758 * 100
- P (x<1992) = **75.8%**
- There is 75.8% chance that houses built before 1992 are in excellent condition.

Correlation (Scatter Plot)

1) ❑ Since the price and the sq. ft living variables seem to behave similarly, is there any correlation?

❑ To investigate if the sq. footage of the house and price variables are related, we chose a scatterplot.



Correlations

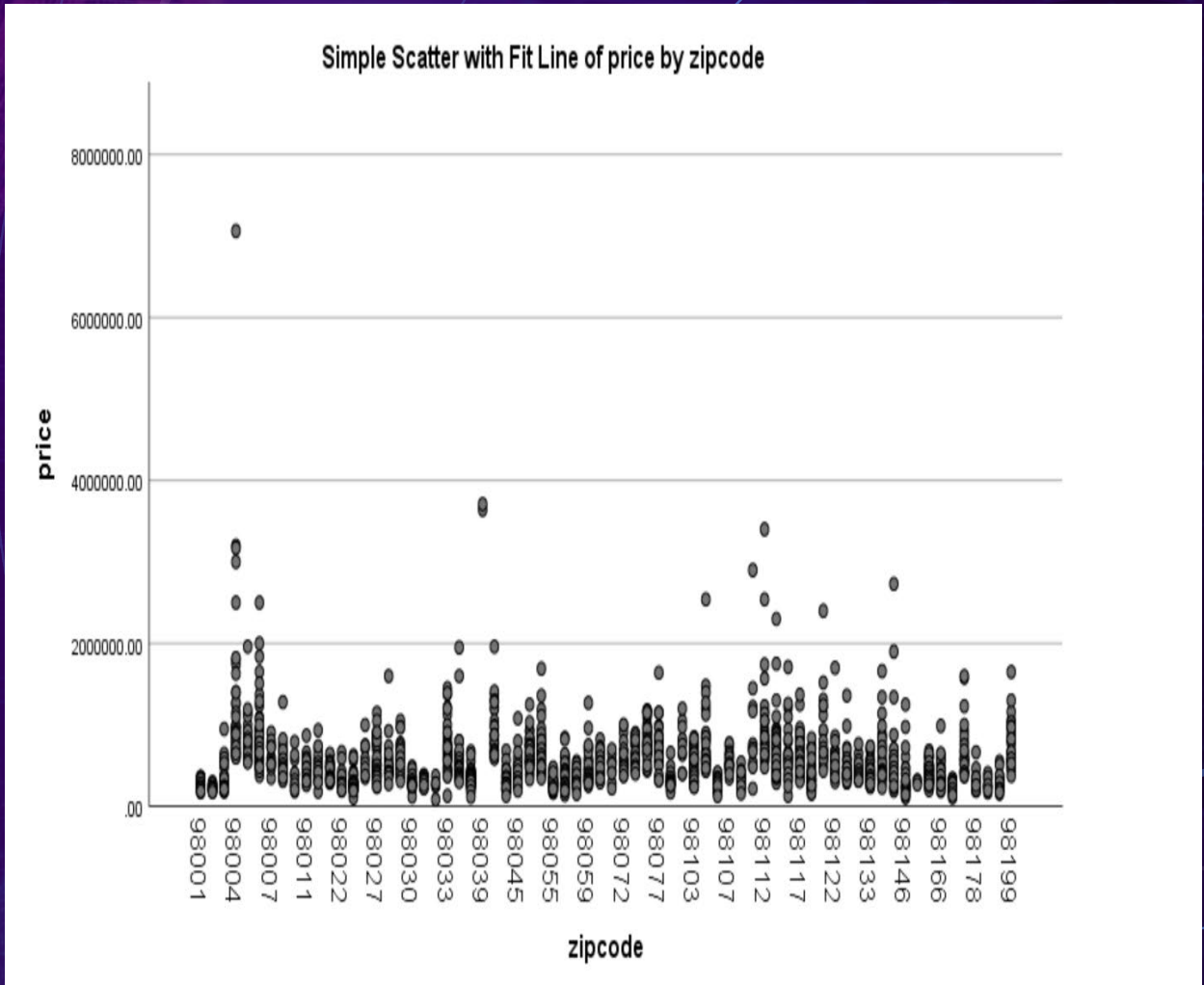
			price	sqft_living
Spearman's rho	price	Correlation Coefficient	1.000	.646**
		Sig. (2-tailed)	.	.000
		N	2105	2104
	sqft_living	Correlation Coefficient	.646**	1.000
		Sig. (2-tailed)	.000	.
		N	2104	2104

**. Correlation is significant at the 0.01 level (2-tailed).

- Spearman's r = 0.646, which is a fairly strong correlation i.e., 64.6% of the variance in price can be explained by sq. footage of the house.
- We can conclude that increase in sq. footage of houses may result in corresponding increase in prices.

Correlation (Scatter Plot)

2) ❑ We also investigated correlations between price and zip code.



Correlations

			price	zipcode
Spearman's rho	price	Correlation Coefficient	1.000	-.023
		Sig. (2-tailed)	.	.294
		N	2105	2105
	zipcode	Correlation Coefficient	-.023	1.000
		Sig. (2-tailed)	.294	.
		N	2105	2105

- Spearman's r = - 0.23
- Both visual and correlation analysis confirm that there is no correlation between price and zip code.

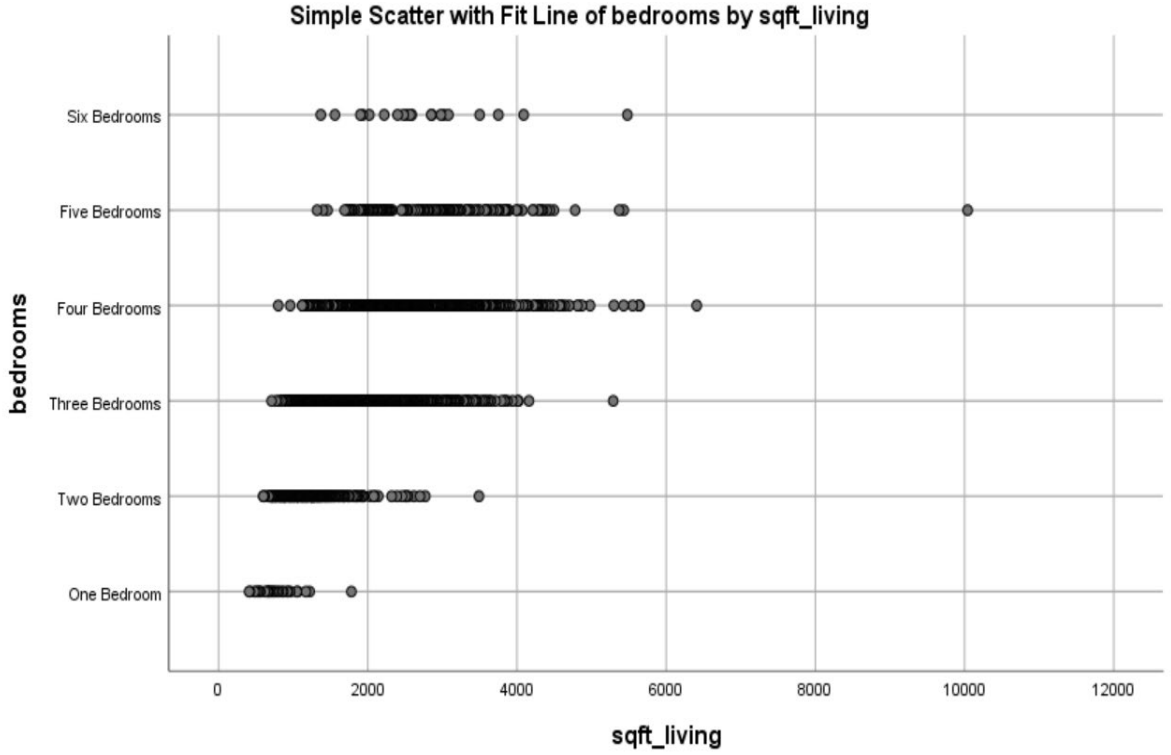
Correlation (Scatter Plot)

3) ❑ Correlation between bedrooms and sq. ft living area

Correlations

			bedrooms	sqft_living
Spearman's rho	bedrooms	Correlation Coefficient	1.000	.652**
		Sig. (2-tailed)	.	.000
		N	2105	2104
	sqft_living	Correlation Coefficient	.652**	1.000
		Sig. (2-tailed)	.000	.
		N	2104	2104

**. Correlation is significant at the 0.01 level (2-tailed).



- Spearman's r = 0.652, which is a fairly strong correlation i.e., 65.2% of the variance in sq. ft living area can be explained by the bedrooms.
- We can conclude that increase in the number of bedrooms may result in corresponding increase in sq. ft living area of houses.

References

Marija Buriskiene, Vitalija Rutzkiene and Jurate Venckauskaite (2011). Models of factors influencing real state price. *Environmental engineering.* p. 873-878.

Thank you

Parametric tests

Adebowale Oluwasanmi

Miguel Angel Acuna Silva

GROUP B

OBJECTIVES

To carry out :

- Observed power & Effect size analysis.
- One-Sample T-test
- Paired sample T-test
- Unpaired sample T-test
- One-factorial ANOVA
- Two-factorial ANOVA

(All tests were carried out on the assumption that relevant variables followed normal distribution)

Observed power & Effect size

- Observed Power & Effect size between price and bedrooms

Tests of Between-Subjects Effects

Dependent Variable: price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2.908E+13[a]	5	5.816E+12	46.849	.000	.100	234.243	1.000
Intercept	1.102E+14	1	1.102E+14	887.315	.000	.297	887.315	1.000
bedrooms	2.908E+13	5	5.816E+12	46.849	.000	.100	234.243	1.000
Error	2.606E+14	2099	1.241E+11					
Total	8.910E+14	2105						
Corrected Total	2.897E+14	2104						

a. R Squared = .100 (Adjusted R Squared = .098)

b. Computed using alpha = .05

- Observed power shows a result of 1. This means that we have a 100% chance of detecting significant effects and a 0% chance of committing a type II error which is good.
- Effect size result is 0.100
- Even though P<0.05, the effect size is relatively small, therefore the findings has limited to no practical significance.

- Observed Power & Effect size between price and zip code

Tests of Between-Subjects Effects

Dependent Variable: price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	1.233E+14[a]	69	1.787E+12	21.855	.000	.426	1507.974	1.000
Intercept	4.384E+14	1	4.384E+14	5362.949	.000	.725	5362.949	1.000
zipcode	1.233E+14	69	1.787E+12	21.855	.000	.426	1507.974	1.000
Error	1.664E+14	2035	8.176E+10					
Total	8.910E+14	2105						
Corrected Total	2.897E+14	2104						

a. R Squared = .426 (Adjusted R Squared = .406)

b. Computed using alpha = .05

- Observed power also shows a result of 1.
- Effect size is 0.426
- Here also, P<0.05 and effect size is relatively low , meaning that the finding may have little to no practical significance in the real world.

T-tests

➢ One Sample T-test

According to the office of financial management (https://ofm.wa.gov/), the median price of a house in 2021 in Washington state was $560,400 us dollars.

Q: We want to find out if the average price of homes has changed due to economic conditions?.

H*(o)* : The average price of homes is equal to or greater than 560,400.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
price	2105	534477.9349	371038.1071	8087.09352

One-Sample Test

Test Value = 560400

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
price	-3.205	2104	.001	-25922.06508	-41781.6005	-10062.5296

- As shown in the statistics result, the sample mean is lower than the specified mean, with P<0.05
- But critical t value(-1.65) > t statistic, however the findings were still statistically significant.
- Therefore, we reject the null hypothesis.
- This means that the economic conditions has caused the average home price to decrease in Washington.

T-tests

➢ Paired Sample T-test

Q 1: We wanted to find out if the average sq. ft for homes was different after renovation?.

H*(o)* : The average sq. ft before and after renovation is equal.

H*(A)* : The average sq. ft before and after renovation is different.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	sqft_living	2032.99	2104	869.650	18.959
	sqft_living15	1526.66	2104	321.930	7.018

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	sqft_living - sqft_living15	506.325	1040.518	22.684	461.838	550.811	22.320	2103	.000

- As shown, the average sq. ft before renovation is higher than after renovation.
- T statistic > critical t value(1.96), P<0.05
- The mean sq. ft before and after renovation is different.
- Therefore, we accept the alternative hypothesis.

T-tests

➢ Unpaired Sample T-test

Q 2.1: We wanted to find out if the average sq. ft vary for homes with and without a waterfront?.

H*(o)* : The average sq. ft is the same for homes with and without a waterfront.

➢ We ensured equal number of sample size in both groups for robust results.

Group Statistics

	waterfront	N	Mean	Std. Deviation	Std. Error Mean
sqft_living	No waterfront	19	5161.32	492.324	112.947
	Waterfront Present	19	3092.58	1901.129	436.149

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
sqft_living	Equal variances assumed	4.046	.052	4.592	36	.000	2068.737	450.536	1155.007	2982.467
	Equal variances not assumed			4.592	20.403	.000	2068.737	450.536	1130.125	3007.349

- As shown in the statistics result, the sample mean for houses with no waterfront is higher than those with waterfronts.
- T statistic > critical t value(1.72), P<0.05
- Therefore, we reject the null hypothesis.
- This means there is a significant difference in the average sq. ft of homes with and without a waterfront.

T-tests

➢ Unpaired Sample T-test

Q 2.2: We wanted to find out if the average price for homes with and without a waterfront is different?.

H*(o)* : The average price is the same for homes with and without a waterfront.

N.B : We ensured equal number of sample size in both groups for robust results.

Group Statistics

	waterfront	N	Mean	Std. Deviation	Std. Error Mean
price	1	163	1662524.184	1120388.1749	87755.5746
	0	163	521380.546	266693.3023	20889.0316

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
price	Equal variances assumed	125.283	<.001	12.650	324	<.001	<.001	1141143.6380	90207.4970	963677.2771	1318609.9990
	Equal variances not assumed			12.650	180.300	<.001	<.001	1141143.6380	90207.4970	963145.4259	1319141.8502

- As shown in the statistics result, the mean price of the houses with waterfront is almost 3 times of that without.
- T statistic > critical t value(1.72), P<0.05
- Therefore, we reject the null hypothesis.
- This means there is a significant difference in the means of the price of houses with and without a waterfront.

ANALYSIS OF VARIANCE (ANOVA)

One-Factor ANOVA

Results of a research done by (Shishir Mathur 2019, p. 1),showed that a medium-quality house sells for approximately 25% more than a low-quality house, and **a well-maintained house sells for approximately 5% more than a house that is not well-maintained.**

Q : We wanted to find check if there is a difference in price between different housing conditions.

Categories: (Very Poor, Poor, Average, Good, Excellent).

H*(o)* : There are no significant differences between the mean price of each housing condition.

ANOVA

price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.800E+12	4	7.000E+11	5.122	.000
Within Groups	2.869E+14	2099	1.367E+11		
Total	2.897E+14	2103			

- F statistic > critical f value(2.38), P<0.05
- Therefore, we reject the null hypothesis.
- This means the avg. price of at least two groups are significantly different from each other.

Post-Hoc Test

- While ANOVA tests for significant differences between the means of individual groups, Post Hoc tests are used to dive in and uncover where the differences lie within the groups by testing each possible pair of groups.

- We used the R-E-G-W-Q post hoc test.

price

			Subset for alpha = 0.05	
	condition	N	1	2
Ryan-Einot-Gabriel-Welsch Range	Very Poor	7	301778.5714	
	Poor	26	318318.7692	
	Average	1383	530093.0289	530093.0289
	Good	533	534203.6604	534203.6604
	Excellent	155		621713.5806
	Sig.		.152	.121

Means for groups in homogeneous subsets are displayed.

- Groups that share the same column are not significantly different while groups that DO NOT share the same column are significantly different.
- There is a significant difference between the mean price of the groups {Very Poor, Poor} and {Excellent}.
- This confirms the ANOVA test that the mean price of at least two groups were significantly different.

ANALYSIS OF VARIANCE (ANOVA)

Two-Factor ANOVA

Q : Asides the number of bedroom, does housing condition also have an influence on price?.

H*(o)* : There are no significant differences in the mean price between the groups of each independent variable.

❖ **Sample size (N=2104)**

Tests of Between-Subjects Effects

Dependent Variable: price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	3.320E+13[a]	26	1.277E+12	10.341	.000	.115	268.860	1.000
Intercept	2.395E+13	1	2.395E+13	194.007	.000	.085	194.007	1.000
bedrooms	5.096E+12	5	1.019E+12	8.255	.000	.019	41.276	1.000
condition	9.959E+11	4	2.490E+11	2.016	.090	.004	8.066	.608
bedrooms * condition	2.079E+12	17	1.223E+11	.991	.466	.008	16.839	.711
Error	2.565E+14	2077	1.235E+11					
Total	8.908E+14	2104						
Corrected Total	2.897E+14	2103						

a. R Squared = .115 (Adjusted R Squared = .104)

b. Computed using alpha = .05

- We noticed that Observed power (0.608) for the condition variable was low, at sample size (N= 2104)
- This means there is a high risk of Type II error, so we increased the sample size to over 6k.

Two-Factor ANOVA (Cont'd)

- At a sample size of over 6k, the Observed power for the condition variable increased to about 0.67

- We further increased the sample size to (N=10,914). Statistics result is as shown below :

❖ **Sample size (N=10,914)**

Tests of Between-Subjects Effects

Dependent Variable: price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	1.394E+14[a]	26	5.360E+12	53.355	.000	.113	1387.226	1.000
Intercept	3.593E+13	1	3.593E+13	357.637	.000	.032	357.637	1.000
bedrooms	2.163E+13	5	4.326E+12	43.059	.000	.019	215.296	1.000
condition	1.387E+12	4	3.467E+11	3.451	.008	.001	13.803	.861
bedrooms * condition	2.532E+12	17	1.489E+11	1.483	.090	.002	25.205	.908
Error	1.094E+15	10887	1.005E+11					
Total	4.305E+15	10914						
Corrected Total	1.233E+15	10913						

a. R Squared = .113 (Adjusted R Squared = .111)

b. Computed using alpha = .05

- At a sample size of (N=10,914), Observed power for condition variable increased to 0.861
- It is clear there are no interactions between the bedrooms and condition variables, as P>0.05
- Both variables have extremely small effect sizes, which means the effects on price may have little to no practical significance.
- Also, with P<0.05 for both variables, we can deduce that number of bedrooms and housing conditions have a statistically significant influence on price.
- Therefore, we reject the null hypothesis.

❑Post-Hoc Test

➢ We also used the R-E-G-W-Q post hoc test, because the number of groups within the variables (bedrooms and housing condition) were more than three.

❖ **Between groups for bedrooms variable**

Homogeneous Subsets

price

Ryan-Einot-Gabriel-Welsch Range[a,b]

		Subset			
bedrooms	N	1	2	3	4
One Bedroom	103	294979.6990			
Two Bedrooms	1424	394318.7015			
Three Bedrooms	4993		466372.8604		
Four Bedrooms	3484			626440.9139	
Six Bedrooms	122			717302.3197	717302.3197
Five Bedrooms	788				761651.3731
Sig.		.072	1.000	.074	.618

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 100458065734.866.

a. Critical values are not monotonic for these data. Substitutions have been made to ensure monotonicity. Type I error is therefore smaller.

b. Alpha = .05.

N.B: Groups that share the same column are not significantly different while groups that DO NOT share the same column are significantly different.

- Post hoc result show a significant difference between the mean price of groups {One, Two}, {Three}, {Four, Six},{Four, Five} bedrooms.
- This confirms our conclusion that there is a significant difference in the mean prices between the groups of bedrooms variable (first factor).

❑Post-Hoc Test (cont'd)

❖ **Between groups for condition variable**

Homogeneous Subsets

price

Ryan-Einot-Gabriel-Welsch Range[a,b]

condition	N	Subset 1	2	3	4
Very Poor	10	223500.0000			
Poor	77	317723.3506			
Good	2887		509657.7967		
Average	7056			533860.1324	
Excellent	884				594689.9208
Sig.		.829	1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 100458065734.866.

a. Critical values are not monotonic for these data. Substitutions have been made to ensure monotonicity. Type I error is therefore smaller.

b. Alpha = .05.

- Here, the post hoc result also reveal a significant difference between the mean price of groups {Very Poor, Poor},{Good}, {Average}and {Excellent} housing conditions.
- This confirms the conclusion that there is a significant difference in the mean prices between the groups of condition variable (second factor).
- Therefore, we reject the null hypothesis.

ANALYSIS OF VARIANCE (ANOVA)

Two-Factor ANOVA

Q 2 : Does the zip code and year built influence the price of homes?

H*(o)* : There are no significant differences in the mean, between the groups of each independent variable.

Tests of Between-Subjects Effects

Dependent Variable: price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2.638E+14[a]	1469	1.796E+11	4.404	.000	.911	6469.221	1.000
Intercept	3.261E+14	1	3.261E+14	7998.036	.000	.927	7998.036	1.000
zipcode	1.111E+14	69	1.610E+12	39.485	.000	.811	2724.454	1.000
yr_built	1.889E+13	114	1.657E+11	4.065	.000	.422	463.356	1.000
zipcode * yr_built	1.210E+14	1286	9.412E+10	2.308	.000	.824	2968.293	1.000
Error	2.585E+13	634	4.078E+10					
Total	8.906E+14	2104						
Corrected Total	2.896E+14	2103						

a. R Squared = .911 (Adjusted R Squared = .704)

b. Computed using alpha = .05

- We have an excellent result for observed power i.e., there are no risks of type II error.
- The effect size for zip code is large (0.811), and that of year built is relatively small (0.422).
- For both variables, P<0.05 . This means that the effects on prices are statistically significant.
- Results also show a high level of interaction between both variables, accompanied with a large effect size.
- In conclusion, we reject the null hypothesis as there is a significant difference in the mean between the groups of each independent variable.

References

Office of financial management 2022, *Median home price in Washington.* Washington Data and Research, accessed 24 January 2023, https://ofm.wa.gov/washington-data-research/statewide-data/washington-trends/economic-trends/median-home-price.

Shishir, M 2019,' House price impacts of construction quality and level of maintenance on a regional housing market', *Evidence from King County, Washington,* accessed 24 January 2023, https://www.researchgate.net/publication/332487322_House_price_impacts_of_construction_quality_and_level_of_maintenance_on_a_regional_housing_market_Evidence_from_King_County_Washington.

ThankYou

Adebowale Oluwasanmi
Miguel Acuna Angel Silva

NON-PARAMETRIC TESTS

Miguel Angel Acuña Silva
Adebowale Oluwasanmi

Objectives

- To find out which characteristics of a house have the greater impact on its price.

- To perform non-parametric tests and correlation analysis in the variables of our data set.

Kruskal-Wallis test

Is there a difference in price, based on the housing conditions?

Ranks

	condition	N	Mean Rank
price	1	29	5716.00
	2	170	5246.76
	3	14011	10893.04
	4	5674	10357.89
	5	1701	12061.19
	Total	21585	

Test Statistics[a,b]

	price
Kruskal-Wallis H	255.670
df	4
Asymp. Sig.	<.001

a. Kruskal Wallis Test

b. Grouping Variable: condition

H_0: There is no difference in the rank sums of the houses in different conditions.

- It is clear by looking at the mean ranks, we can see the differences in the rank sum.
- **Since p < 0.05, we reject the null hypothesis.**
- This means there are differences in price for each varying housing condition.

Kruskal-Wallis test (cont'd)

Pairwise Comparisons of condition

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Poor-Very Poor	467.619	1243.383	.376	.707	1.000
Poor-Good	-5061.648	481.746	-10.507	.000	.000
Poor-Average	-5574.634	477.548	-11.673	.000	.000
Poor-Excellent	-6748.390	497.928	-13.553	.000	.000
Very Poor-Good	-4594.029	1152.166	-3.987	.000	.001
Very Poor-Average	-5107.015	1150.417	-4.439	.000	.000
Very Poor-Excellent	-6280.771	1159.026	-5.419	.000	.000
Good-Average	512.986	97.662	5.253	.000	.000
Good-Excellent	-1686.742	171.519	-9.834	.000	.000
Average-Excellent	-1173.756	159.347	-7.366	.000	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

- This table confirms our conclusion to reject the null hypothesis.
- It dives in to show differences in prices within the groups by comparing one group to another.
- After comparing all groups, we can see that there are significant differences in price , as $P<0.05$ for almost every condition except {Poor, Very Poor}.

Wilcoxon test

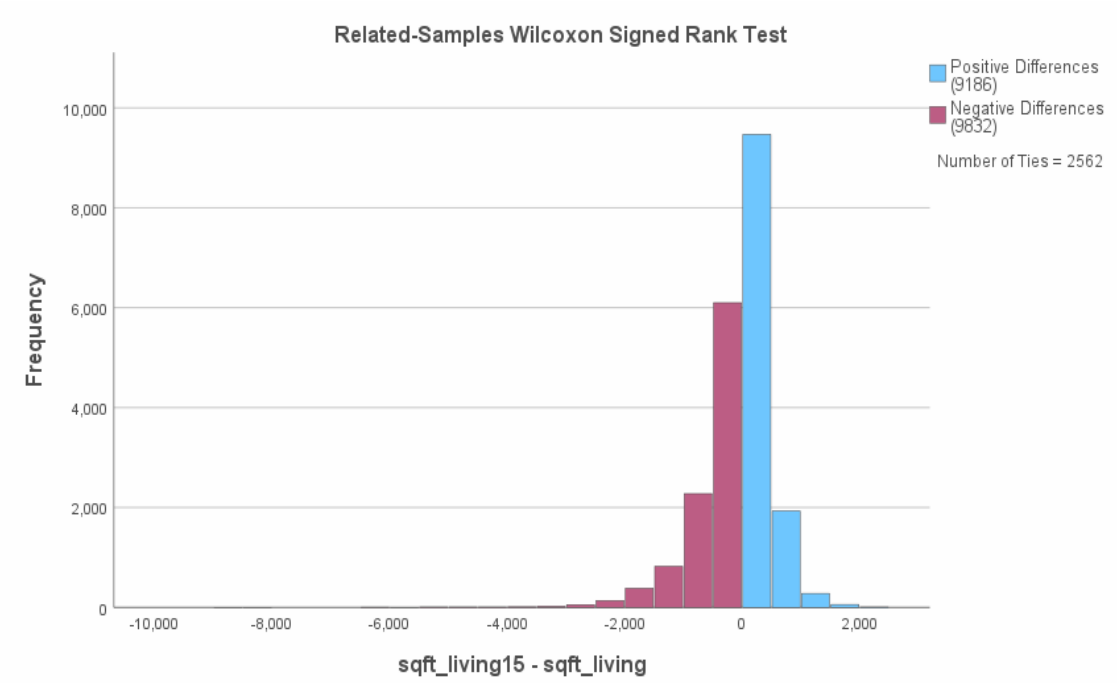
Are there differences between the sq. ft living area before and after renovation?

- H_0: The central tendencies of the sq. ft area of houses before and after renovation are the same.

Related-Samples Wilcoxon Signed Rank Test Summary

Total N	21580
Test Statistic	79984498.000
Standard Error	757121.265
Standardized Test Statistic	-13.791
Asymptotic Sig.(2-sided test)	<.001

- **Since p < 0.05, we reject the null hypothesis.**



Mann Whitney test

Is there a difference in price , if houses have a waterfront or not?

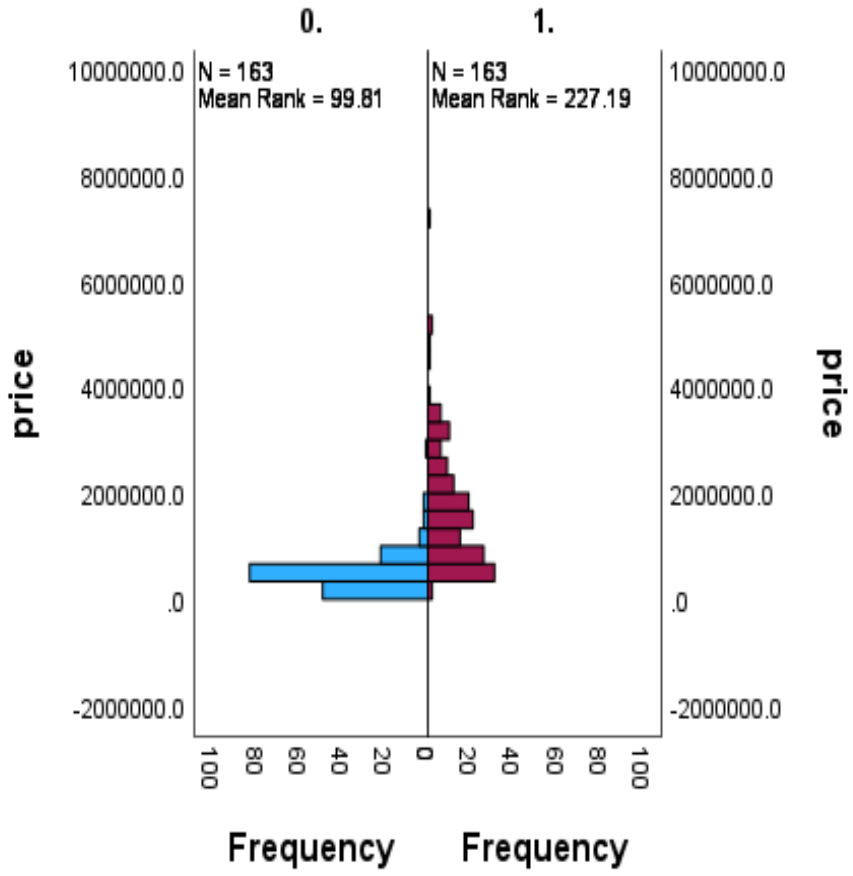
- H_0: There is no difference in the price sum of the ranks for houses with and without a waterfront.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of price is the same across categories of waterfront.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.



Mann Whitney test (cont'd)

- Based on the analysis above, we reject the null hypothesis, as $P<0.05$
- This means that there is a difference in the price sum of ranks for houses with and without a waterfront.

N.B : We couldn't perform the Friedman Test, because we need 3 or more dependent variables which we do not have in our current data.

CORRELATION

- Are there strong correlations between price and other variables?

Correlations

			price	sqft_living	sqft_lot	yr_built	zipcode	condition
Spearman's rho	price	Correlation Coefficient	1.000	.644**	.075**	.102**	-.009	.018**
		Sig. (2-tailed)	.	<.001	<.001	<.001	.191	.008
		N	21588	21579	21572	21582	21582	21585
	sqft_living	Correlation Coefficient	.644**	1.000	.305**	.353**	-.207**	-.063**
		Sig. (2-tailed)	<.001	.	<.001	<.001	<.001	<.001
		N	21579	21587	21571	21582	21580	21584
	sqft_lot	Correlation Coefficient	.075**	.305**	1.000	-.037**	-.319**	.115**
		Sig. (2-tailed)	<.001	<.001	.	<.001	<.001	<.001
		N	21572	21571	21581	21576	21574	21579
	yr_built	Correlation Coefficient	.102**	.353**	-.037**	1.000	-.317**	-.394**
		Sig. (2-tailed)	<.001	<.001	<.001	.	<.001	<.001
		N	21582	21582	21576	21591	21584	21588
	zipcode	Correlation Coefficient	-.009	-.207**	-.319**	-.317**	1.000	-.023**
		Sig. (2-tailed)	.191	<.001	<.001	<.001	.	<.001
		N	21582	21580	21574	21584	21590	21587
	condition	Correlation Coefficient	.018**	-.063**	.115**	-.394**	-.023**	1.000
		Sig. (2-tailed)	.008	<.001	<.001	<.001	<.001	.
		N	21585	21584	21579	21588	21587	21594

**. Correlation is significant at the 0.01 level (2-tailed).

- As shown on the table, there are statistically significant correlations between price and sq. ft living, sq. ft lot, condition, year-built variables.
- Price→sq. ft living = 64.4%. This was the only medium positive correlation.
- All other correlations between price and other variables were weak, even though they were statistically significant.

CORRELATION (Cont'd)

Correlations

			price	bedrooms	floors	waterfront
Spearman's rho	price	Correlation Coefficient	1.000	.337**	.319**	.112**
		Sig. (2-tailed)	.	.000	.000	.000
		N	21441	21441	21441	21441
	bedrooms	Correlation Coefficient	.337**	1.000	.225**	-.014*
		Sig. (2-tailed)	.000	.	.000	.037
		N	21441	21450	21450	21450
	floors	Correlation Coefficient	.319**	.225**	1.000	.023**
		Sig. (2-tailed)	.000	.000	.	.001
		N	21441	21450	21450	21450
	waterfront	Correlation Coefficient	.112**	-.014*	.023**	1.000
		Sig. (2-tailed)	.000	.037	.001	.
		N	21441	21450	21450	21450

**. Correlation is significant at the 0.01 level (2-tailed).

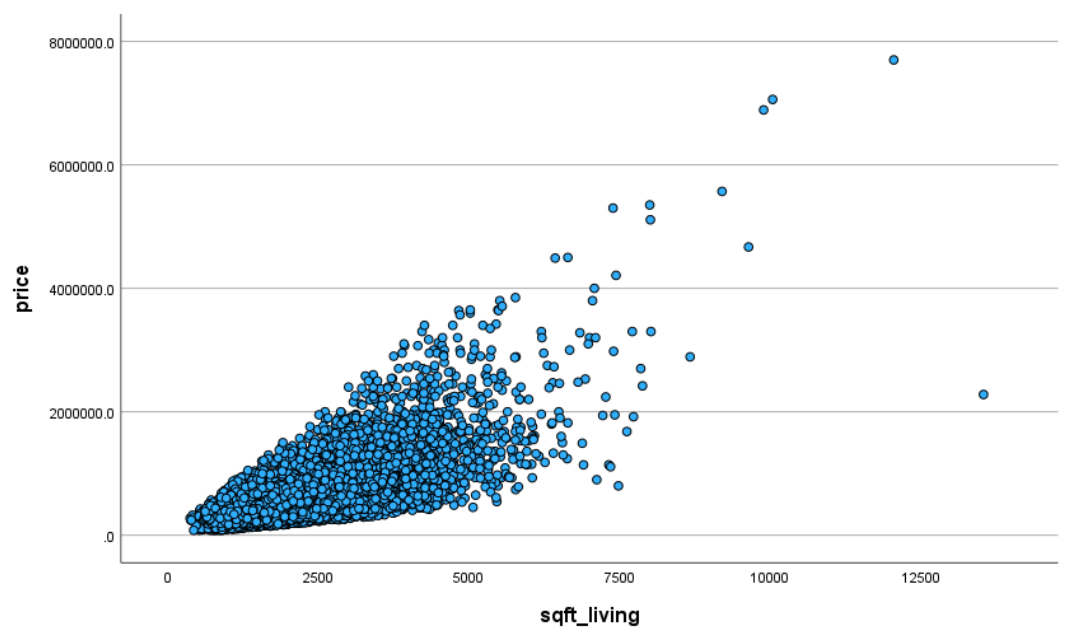
*. Correlation is significant at the 0.05 level (2-tailed).

- Here, we also have weak correlations with price and other variables.
- This means that we have just one strong correlation that is good enough to build a price prediction model.
- Price→sq. ft living = 64.4%

REGRESSION

SIMPLE LINEAR REGRESSION

- A.) Does the sq. ft living area influence the price of homes?



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.702[a]	.493	.493	261613.2431

a. Predictors: (Constant), sqft_living

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.434E+15	1	1.434E+15	20958.010	<.001[b]
	Residual	1.477E+15	21577	68441488969		
	Total	2.911E+15	21578			

a. Dependent Variable: price

b. Predictors: (Constant), sqft_living

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-43941.755	4411.159		-9.961	<.001
	sqft_living	280.840	1.940	.702	144.769	<.001

a. Dependent Variable: price

Regression (cont'd)

MULTIPLE LINEAR REGRESSION

- Do the variables bedrooms, floors and waterfront influence the price of a home?.
- **N.B :** We do not have enough multiple variables with strong correlations to build a good prediction model, so this analysis was done for educational purposes.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.026[a]	.001	.001	338299.9959	.001	4.723	3	21437	.003

a. Predictors: (Constant), Waterfront_D, Floors_D, Bedrooms_D

b. Dependent Variable: price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.622E+12	3	5.405E+11	4.723	.003[b]
	Residual	2.453E+15	21437	1.144E+11		
	Total	2.455E+15	21440			

a. Dependent Variable: price

b. Predictors: (Constant), Waterfront_D, Floors_D, Bedrooms_D

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	527666.123	2857.552		184.657	.000
	Bedrooms_D	12817.060	4971.035	.018	2.578	.010
	Floors_D	19385.953	13954.303	.010	1.389	.165
	Waterfront_D	67881.942	27275.411	.017	2.489	.013

a. Dependent Variable: price

- The ANOVA table shows that bedrooms, floors and waterfront have statistically significant effects on the price of homes, as $P<0.05$
- Adjusted R square= 0.001 i.e., This is no surprise, as we confirmed earlier, that all the above variables had weak correlations with price.
- Therefore, its obvious that the independent variables are not good predictors of price.
- This is likely due to the large amount of unexplained variance in the dependent variable

Thank you.

Miguel Angel Acuña Silva
Adebowale Oluwasanmi