Franco Bueno Mattera
Final Term Project
Class: MET CS555

Term Project:

Real State Prices Near Boston Metropolitan Area:

Research Scenario Description:

1. Introduction:

Since the pandemic began in the United states, there have been rumors and press reports pointing at the uncertainty of the housing market these days. A report from Bloomberg from September 2020 recommends not to buy property during the pandemic due to significant uncertainty. Another press report from the New York Times points at the "Suburban Home Sales Boom" around NY Metropolitan area. In Boston, the situation appears to be similar. Reports like the one posted in Boston Magazine suggest an increased demand for single family homes (suburbs) with respect to the last year.

The main goal of this study is to confirm, based on openly available real-state data if there are any significant differences in house sale prices before and after 2020, and between high and low densely populated counties. The second goal of this project is to study if there is a change in the level of association between low and high population density counties with respect to house sale prices during the time pre and post pandemic.

2. Research Scenario Description:

This research is based on studying the variability of the housing market in the last few years. In general, highly densely populated counties are closer to Boston Metro Area than counties with lower population density (USA.com). This study focuses on the counties containing and surrounding Boston. There are six counties that meet that criteria: Suffolk, Essex, Middlesex, Bristol, and Plymouth. Some of those counties are highly populated and could be considered urban areas while others have low or very low population density and could be classified as suburban areas.

Table 1: Population Density by county in the six counties that surround Boston Metropolitan Area (USA.com).

County	Norfolk	Suffolk	Essex	Middlesex	Bristol	Plymouth
Population Density	1,536.8/sq mi	6,221.3/sq mi	914.2/sq mi	1,817.9/sq mi	797.2/sq mi	458.1/sq mi

3. Research Question:

There are two specific research question that this project intends to answer:

1. Are there any significant differences in House Sale Prices between the years 2017, 2018, 2019, 2020 and between the six different chosen counties?
2. Is there an association between house sale prices and the years 2017, 2018, 2019, 2020 for the counties Suffolk and Plymouth, separately?

4. Dataset Description:

The original dataset was obtained from RedFin.com, a website that publish information about real estate sales around the United States. The file is in a tsv format and it is freely available for download. The file contains a significant amount of hidden data, requiring accessing the file using SQL and then storing the table as a csv file.

5. Dataset Preprocessing:

The data was filtered to select the information concerning only the counties and the variables of study. The variables that were kept were properly transformed into factors, numeric and datetime objects in the case of categories, numeric values, and dates, respectively. New categorical variables such as Year and Name of County were also created (Table 1). Outliers and missing values were also eliminated.

The variable population density was obtained from another source. It was added for the only purpose of aiding in the conclusions of this study since it is strictly related with the variable Name(county).

period_begin	median_sale_price	Pop.Density	Name	Year
2017-01-02	421479.1	1536.8	Norfolk	2017
2017-01-02	553500.0	6221.3	Suffolk	2017
2017-01-02	546312.5	6221.3	Suffolk	2017
2017-01-02	505000.0	1817.9	Middlesex	2017
2017-01-02	474750.0	1817.9	Middlesex	2017
2017-01-02	363000.0	914.2	Essex	2017

Figure 1: Six first rows of the preprocessed dataset containing housing sale price information from the years 2017 to 2019 for the counties of Norfolk, Suffolk, Essex, Middlesex, Bristol, and Plymouth. Note that the variable "period_begin" refers to the date when the house was listed as available for sale.

6.Preparing data for analysis:

To prepare the data for the statistical analyses, to study the distribution and to make the code easier to follow and understand; the dataset was filtered and divided into multiple smaller datasets by the factors of the categorical Variables Year and Name(county). After that, outliers and missing values were removed.

Table: 2 Summary of the Variable Median House Sale Price per Year and County

Year		Minimum	Quantile 25	Median	Mean	Quantile 75	Maximum
2017	**Norfolk**	$399,125.00	$449,462.00	$455,934.00	$456,781.00	$470,815.00	$519,000.00
	Suffolk	$510,188.00	$571,000.00	$580,896.00	$580,082.00	$595,000.00	$681,500.00
	Essex	$339,750.00	$368,666.00	$387,142.00	$382,979.00	$395,399.00	$419,000.00
	Middlesex	$456,000.00	$498,165.00	$510,950.00	$508,878.00	$521,547.00	$575,000.00
	Bristol	$230,000.00	$258,194.00	$274,142.00	$268,834.00	$278,771.00	$295,000.00
	Plymouth	$302,750.00	$341,175.00	$349,950.00	$347,414.00	$355,028.00	$410,000.00
2018	**Norfolk**	$262,950.00	$311,000.00	$387,583.00	$407,008.00	$478,250.00	$654,497.00
	Suffolk	$254,950.00	$302,000.00	$392,167.00	$425,705.00	$509,967.00	$725,000.00
	Essex	$272,733.00	$430,000.00	$534,438.00	$518,806.00	$630,750.00	$699,000.00
	Middlesex	$270,000.00	$377,422.00	$508,256.00	$492,136.00	$614,517.00	$720,000.00
	Bristol	$259,850.00	$400,336.00	$520,750.00	$495,044.00	$599,678.00	$676,250.00
	Plymouth	$255,000.00	$306,917.00	$403,000.00	$418,202.00	$519,900.00	$660,000.00
2019	**Norfolk**	$285,000.00	$374,380.00	$482,049.00	$488,658.00	$631,563.00	$711,500.00
	Suffolk	$284,950.00	$392,884.00	$481,765.00	$483,751.00	$582,969.00	$677,496.00
	Essex	$278,725.00	$382,119.00	$494,934.00	$478,845.00	$588,250.00	$686,625.00

	Middlesex	$280,000.00	$376,250.00	$451,998.00	$467,594.00	$571,300.00	$689,500.00
	Bristol	$267,500.00	$395,375.00	$486,361.00	$478,520.00	$562,755.00	$707,750.00
	Plymouth	$281,625.00	$339,975.00	$451,823.00	$466,811.00	$557,530.00	$688,583.00
2020	**Norfolk**	$310,000.00	$404,166.00	$505,000.00	$497,909.00	$602,833.00	$690,000.00
	Suffolk	$312,500.00	$421,794.00	$529,950.00	$512,561.00	$619,000.00	$687,000.00
	Essex	$293,263.00	$433,081.00	$546,416.00	$522,350.00	$631,828.00	$722,500.00
	Middlesex	$278,500.00	$440,646.00	$530,200.00	$520,447.00	$615,063.00	$720,875.00
	Bristol	$307,500.00	$424,382.00	$499,646.00	$510,979.00	$610,550.00	$708,750.00
	Plymouth	$299,000.00	$353,680.00	$467,450.00	$478,245.00	$575,401.00	$720,000.00

Since some of the analyses required the use of statistical tests with underlying assumptions of Normality and Homoscedasticity between groups (factors), the Shapiro wilk-Normality test and the Levene's test for homogeneity of variances were the main tools used to assess those assumptions. Levene's test was chosen over Bartlett's test due to its robustness when dealing with non-normal distributions (Long and Teetor 2019). It is also important to mention that independence was assumed for all the groups. This was mainly due to the inability to control that assumption since it depends on how the data was gathered (Stephaine Glen and StatisticsHowTo.com 2015).

In most cases, the distributions of the different factors of the variable containing county names were not normally distributed, and also, the variances among the different counties were not equal (Table 3, Figures 2 and 3).

Table 3: p - values for Shapiro-Wilk's Normality test and Levene's Homogeneity of variance test. p-values that are less than 0.05 are considered significant and the null hypotheses of Normality and Homoscedasticity are rejected (Field, Miles, Field 2013a; Long and Teetor 2019).

Groups		Shapiro-Wilk	Levene's
Year	**2017**	1.75E-14	1.70E-01
	2018	< 2.2e-16	
	2019	< 2.2e-16	
	2020	1.23E-10	
County	**Norfolk**	6.14E-04	4.932E-13
	Suffolk	1.20E-03	
	Essex	8.81E-04	
	Middlesex	1.35E-02	
	Bristol	6.93E-06	
	Plymouth	1.02E-05	

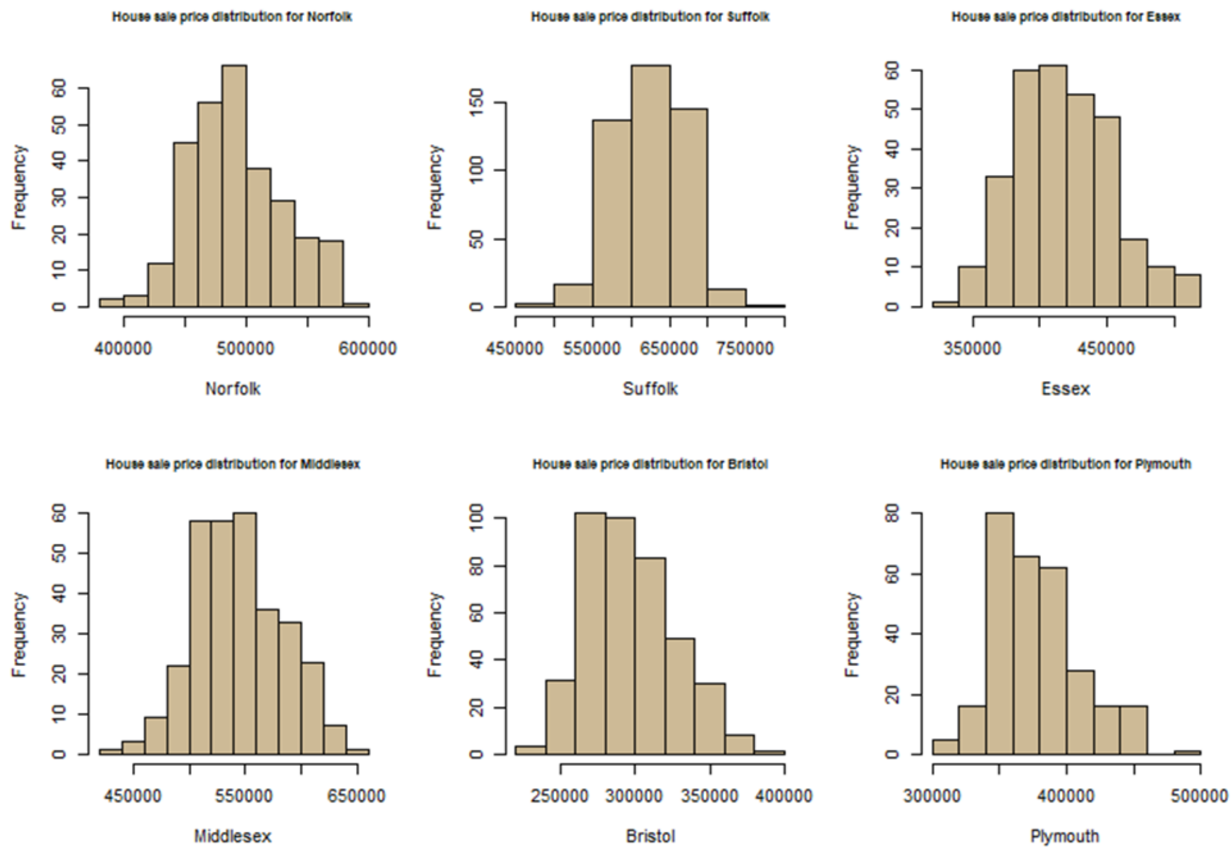

Figure 2: Histograms showing the distributions of the variable Median House Sale Price per county.

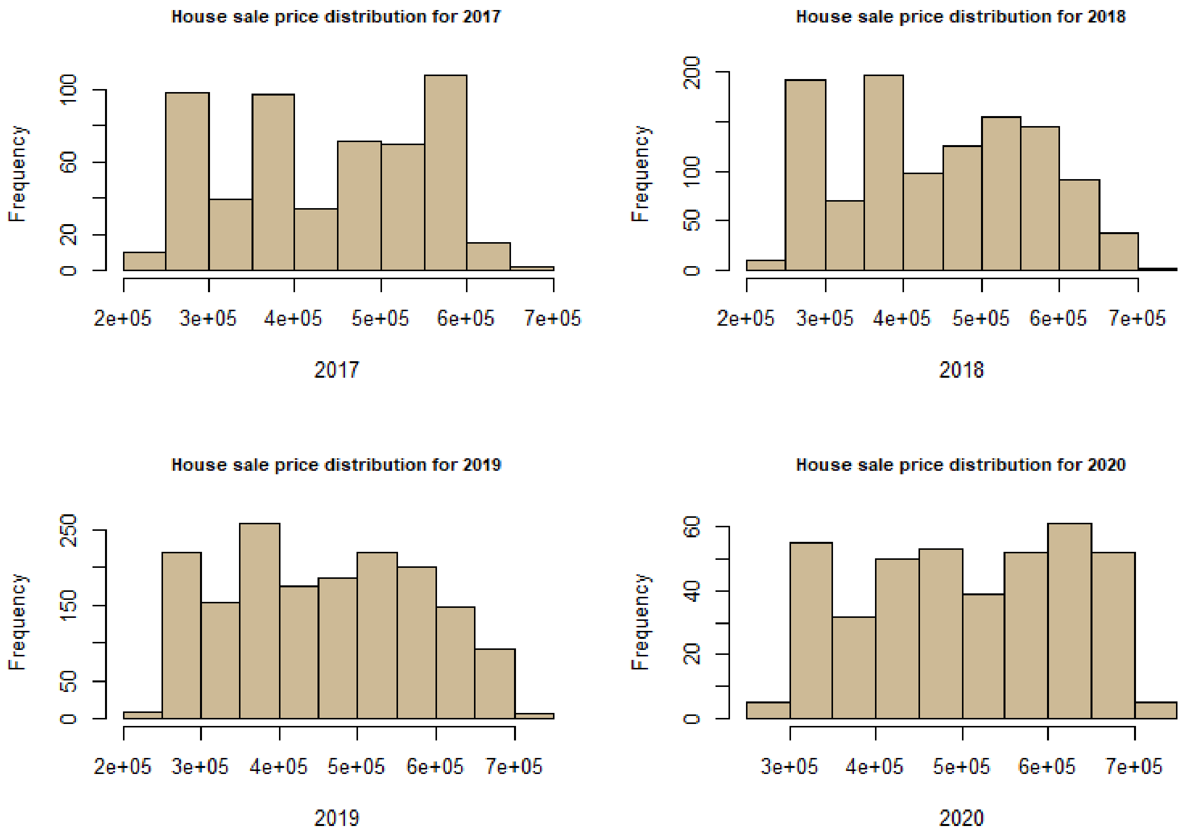

Figure 3: Histograms showing the distributions of the variable Median House Sale Price Per Year.

7.Question 1:

Are there any significant differences in House Sale Prices between the years 2017, 2018, 2019, 2020 and between the six different chosen counties?

7.1 Exploratory Data Analisis:

Looking at the boxplots from figures 4 and 5, it is possible to appreciate some variability between the groups. Figure 4 shows that the county Suffolk had the highest sales prices between the whole range of 2017 and 2020, and in Figure 5, a constant increase in median house prices through the years can be observed. It can also be seen that the increase in house prices it is slightly higher for the year 2020.

The boxplots from Figures 6 and 7 show in more detail the variability of house prices between year and county. Figure 6 shows that the variability in house prices per county is consistent over the years. Figure 7 shows that the house sale prices seem to increase over the years. Having said that, it appears that the prices for the county Suffolk seems to increase less in the year 2020 while the opposite seems to be happening in Plymouth, Middlesex and possibly Norfolk county. It is worth saying that Suffolk is the most densely populated county while Plymouth is the least densely populated of the counties surrounding Boston.

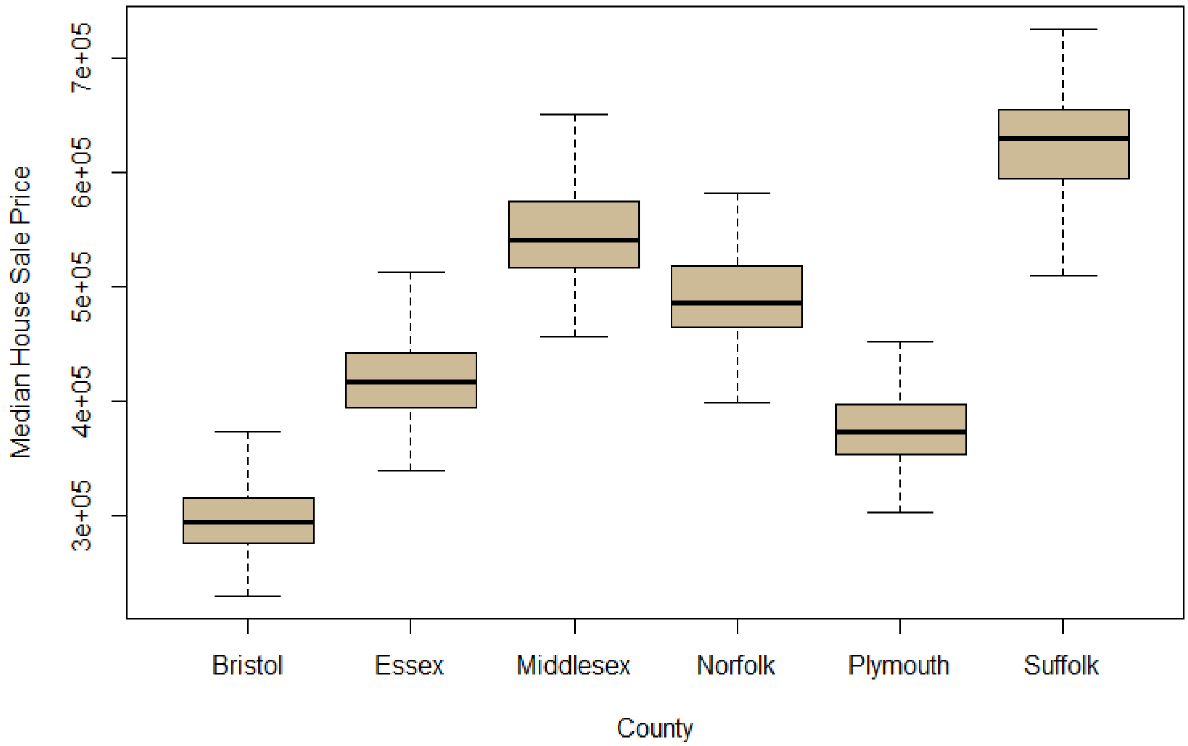

Figure 4. Boxplot Showing the variability in median house sale price between different counties surrounding Boston Metropolitan Area.

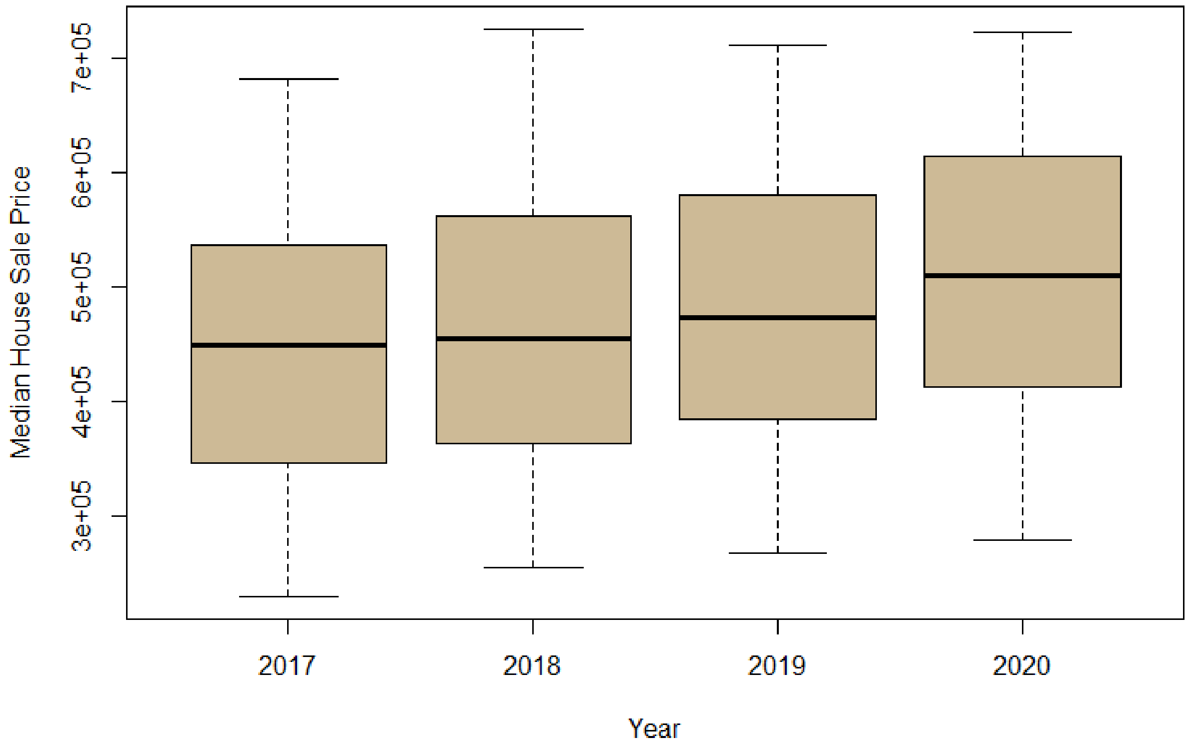

Figure 5. Boxplot Showing the variability in median house sale price between the years 2017, 2019, 2019 and 2020.

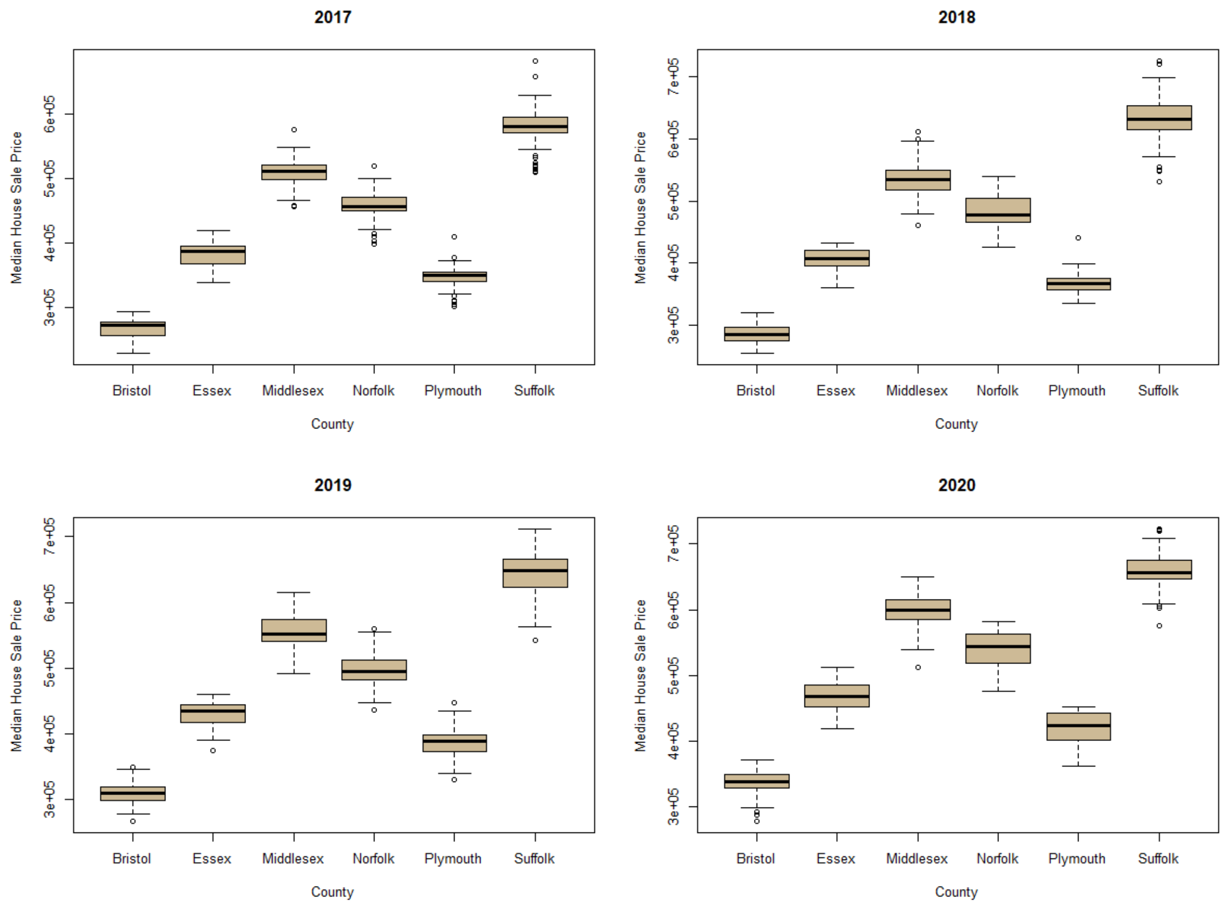

Figure 6: Boxplot showing the variability of Median House Sale Prices by year for the six counties surrounding Boston Metropolitan Area.

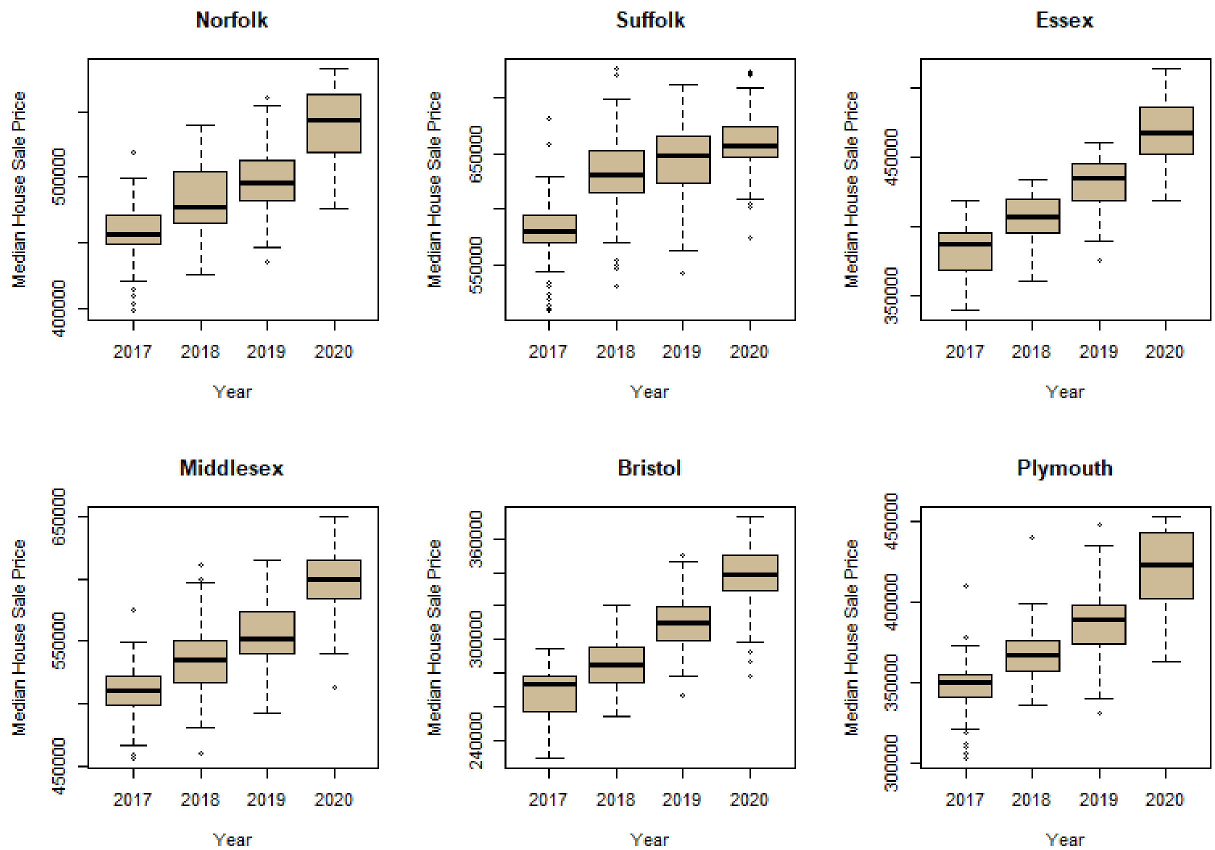

Figure 7: Boxplot showing the variability of Median House Sale Prices by county across the years 2017, 2018, 2019 and 2020.

7.1 Statistical Analyses:

Since the goal of question one is to test for significant differences in house sale prices per county and year, a two-way analysis of variance was performed, followed by post-hoc tests. Since most of the assumptions for parametric tests were not met, formal parametric and non-parametric hypotheses were performed in parallel.

7.1.1 Formal test to determine if there are there any significant differences in House Sale Prices between the years 2017, 2018, 2019, 2020 and between the six different counties chosen.

Setting up the hypotheses and the α level:

$\alpha = 0.05$

Hypothesis for parametric Test:

Factor: Name(county) after controlling for Year:

Ho: $\mu_{Norfolk} = \mu_{Suffolk} = \mu_{Essex} = \mu_{Middlesex} = \mu_{Plymouth} = \mu_{Bristol} = 0$

H1: $\mu_{county\ i} \neq \mu_{county\ j}$ (at least two of the counties present significant differences in median house prices)

Factor: Year after controlling for Name(county)

Ho: $\mu_{2017} = \mu_{2017} = \mu_{2018} = \mu_{2020} = 0$

H1: $\mu_{year\ i} \neq \mu_{Year\ j}$ (at least two of the years present significant differences in median house prices)

Hypothesis for non-parametric Test:

Factor: Name(county) after controlling for Year:

Ho: $Population_{Norfolk} = Population_{Suffolk} = Population_{Essex} = Population_{Middlesex} = Population_{Plymouth} = Population_{Bristol} = 0$

H1: $Population_{county\ i} \neq Population_{county\ j}$ (at least two of the counties present significant differences in house prices)

Factor: Year (after controlling for county)

Ho: $Population_{2017} = Population_{2017} = Population_{2018} = Population_{2020} = 0$

H1: $\mu_{year\ i} \neq \mu_{Year\ j}$ (at least two of the years present significant differences in house prices)

Selecting the appropriate test statistic:

For Parametric Test:

The initial intention was to perform a Two-way ANOVA between the dependent variable Median Sale Price and the factor variables Name(county) and Year. However, quantitative Interactions were found between the counties Middesex, Essex and the year 2020, and between county Suffolk and the years 2018, 2019, 2020 (Figure 5). A stratified one-way analysis of variance followed by pairwise t-test with Bonferroni correction was performed instead of the Two-Way ANOVA.

The F statistic was calculated to perform the one-way ANOVA. The equation is given by:

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-1}} = \frac{MSB}{MSW}$$

With k-1 and n – k degrees of freedom

where:

n = number of observations of all groups

k = number of groups

SSB = sum of squares between groups

SSW = sum of squares within groups

MSB = Mean square between groups

MSW = mean square within groups

source: (Mohammad Alaghemandi and Boston University)

The post-hoc analysis was done by performing pairwise t test:

$$t = \frac{(\bar{x}_i - \bar{x}_1)}{\sqrt{S_p^2(\frac{1}{n_i} + \frac{1}{n_1})}}$$

With n – k degrees of freedom

Where:

k = Groups (all the groups, not just the pairwise groups)

$\bar{x}_i$ = Sample mean of the ith group

$\bar{x}_j$ = Sample mean of the control group

S_p^2 = Variance calculated by ANOVA

n_i = Number of observations of the ith group

n_j = Number of observations of the jth

source: (Mohammad Alaghemandi and Boston University)

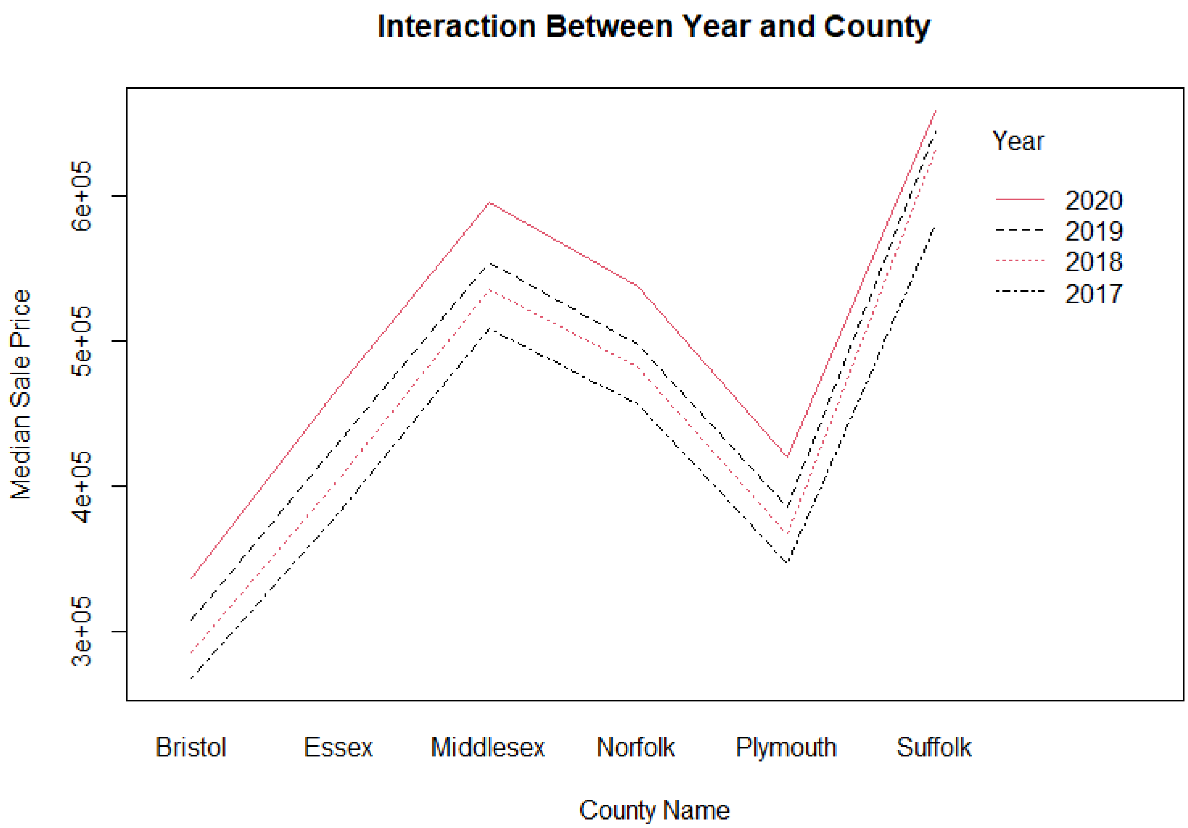

Figure 5: Interaction plot between the factor variables County and Year. Interactions were significant (p-value: < 2.2e-16).

For Non-Parametric Test:

Performed a stratified non-parametric Kruskal-Wallis Analysis between the dependent variable Median Sale Price and the factor variables Name(county) and Year, followed by a Multiple Comparison Test After Kruskal-Wallis using critical differences among groups (described in pgirmess package (Field, Miles, Field 2013b)

The H statistic was calculated in order to perform Kruskal-Wallis Test. The equation is given by:

$$H = \frac{12}{N(N+1)} * \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(N+1)$$

Decision rule with χ^2 distribution with k-1 degrees of freedom

Where:

k = Groups

R_i = Sum ranks for each group

N = Total sample size

n_i = sample size of group i

Source: (Field, Miles, Field 2013b; Zar 1998)

The post-hoc analysis was done by performing a Multiple comparison test after Kruskal-Wallis:

$$|\mathrm{R}_i - \mathrm{R}_j| \geq z_{\frac{\alpha}{\mathrm{k(k-1)}}} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Where:

Ri-j: Is the mean rank of the groups

k = Groups (all the groups, not just the pairwise groups)

N = Total sample size

n_i = sample size of group i

z = z statistic

α = 0.05

Source: (Field, Miles, Field 2013c)

Setting the decision rule:

For Parametric Test:

For stratified analysis.

Factor: county, stratus: sub-dataframe = Year:

2017:

$F_{5,538,0.05}$= 2.23077

Decision rule: If Calculated F >= 2.23077, reject Ho. If calculated F < 2.23077, do not reject Ho.

If p-value < 0.05, reject Ho

***Post-Hoc*:**

$t_{538,0.5/2}$= 1.964383 Decision rule: If calculated t >= 1.964383, Reject Ho. If calculated t < 1.964383, do not reject Ho

If p-value < 0.05/2, reject Ho

2018:

$F_{5,570,0.05}$= 2.229831

Decision rule: If Calculated F >= 2.229831, reject Ho. If calculated F < 2.229831, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{570,0.5/2}$= 1.964135

Decision rule: If calculated t >= 1.964135, Reject Ho. If calculated t < 1.964135, do not reject Ho

If p-value < 0.05/2, reject Ho

2019:

$F_{5,545,0.05}$= 2.230555

Decision rule: If Calculated F >= 2.230555, reject Ho. If calculated F < 2.230555, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{545,0.5/2}$= 1.964326

Decision rule: If calculated t >= 1.964326, Reject Ho. If calculated t < 1.964326, do not reject Ho

If p-value < 0.05/2, reject Ho

2020:

$F_{5,398,0.05}$= 2.236665

Decision rule: If Calculated F >= 2.236665, reject Ho. If calculated F < 2.236665, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{398,0.5/2}$= 1.965942

Decision rule: If calculated t >= 1.965942, Reject Ho. If calculated t < 1.965942, do not reject Ho

If p-value < 0.05/2, reject Ho

source: (Mohammad Alaghemandi and Boston University)

For stratified analysis.

Factor: Year, stratus: sub-dataframe = County:

Norfolk:

$F_{3,283,0.05}$= 2.636504

Decision rule: If Calculated F >= 2.636504, reject Ho. If calculated F < 2.636504, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{283,0.5/2}$= 1.968382

Decision rule: If calculated t >= 1.968382, Reject Ho. If calculated t < 1.968382, do not reject Ho

If p-value < 0.05/2, reject Ho

Suffolk:

$F_{3,484,0.05}$= 2.623327

Decision rule: If Calculated F >= 2.623327, reject Ho. If calculated F < 2.623327, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{484,0.5/2}$= 1.964877

Decision rule: If calculated t >= 1.964877, Reject Ho. If calculated t < 1.964877, do not reject Ho

If p-value < 0.05/2, reject Ho

Essex:

$F_{3,296,0.05}$= 2.635106

Decision rule: If Calculated F >= 2.635106, reject Ho. If calculated F < 2.635106, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{296,0.5/2}$= 1.968011

Decision rule: If calculated t >= 1.968011, Reject Ho. If calculated t < 1.968011, do not reject Ho

If p-value < 0.05/2, reject Ho

Middlesex:

$F_{3,305,0.05}$= 2.634209

Decision rule: If Calculated F >= 2.634209, reject Ho. If calculated F < 2.634209, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{305,0.5/2}$= 1.967772

Decision rule: If calculated t >= 1.967772, Reject Ho. If calculated t < 1.967772, do not reject Ho

If p-value < 0.05/2, reject Ho

Bristol

$F_{3,402,0.05}$= 2.627103

Decision rule: If Calculated F >= 2.627103, reject Ho. If calculated F < 2.627103, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{402,0.5/2}$= 1.965883

Decision rule: If calculated t >= 1.965883, Reject Ho. If calculated t < 1.965883, do not reject Ho

If p-value < 0.05/2, reject Ho

Plymouth

$F_{3,281,0.05}$= 2.63673

Decision rule: If Calculated F >= 2.63673, reject Ho. If calculated F < 2.63673, do not reject Ho.

If p-value < 0.05, reject Ho

Post-Hoc:

$t_{281,0.5/2}$= 1.968442

Decision rule: If calculated t >= 1.968442, Reject Ho. If calculated t < 1.968442, do not reject Ho

If p-value < 0.05/2, reject Ho

For Non-parametric Test:

For stratified analysis.

Factor: county, stratus: sub-dataframe = Year

$\chi^2_{0.5,5\,=}$ 11.0705

Decision rule: If calculated χ^2 >= 11.0705, Reject Ho. If calculated χ^2< 11.0705, do not reject Ho

If p-value < 0.05, reject Ho

Post-Hoc:

If calculated $|R_i - R_j| \geq z_{\frac{0.05}{4(4-1)}} \sqrt{\frac{N_{county}(N_{county}+1)}{12} (\frac{1}{n_{year(i)}} + \frac{1}{n_{year(j)}})}$, Reject Ho.

For stratified analysis.

Factor: Year, stratus: sub-dataframe = County

$\chi^2_{0.5,3} =$ 7.814728

Decision rule: If calculated χ^2 >= 7.814728, Reject Ho. If calculated χ^2< 7.814728, do not reject Ho

If p-value < 0.05, reject Ho

Post-hoc:

If calculated $|R_i - R_j| \geq z_{\frac{0.05}{4(4-1)}} \sqrt{\frac{N_{year}(N_{year}+1)}{12} (\frac{1}{n_{county(i)}} + \frac{1}{n_{county(j)}})}$, Reject Ho.

Computing the test statistic:

Parametric Tests:

Table 3 shows that stratified ANOVA found significant differences in House Sale Prices for the factor variable within the sub-dataframes Year and County.

Table 3: Stratified ANOVA table for the sub-dataframes Year and County containing the groups county and year respectively. Significant p-values are show in bold.

Stratified ANOVA Table			**Degrees of Freedom**	**Sum of Squares**	**Mean Squares**	**F-value**	**p-value**
Year	**2017**	County	5	7.09E+12	1.42E+12	3522	**<2e-16**
		Residuals	538	2.17E+11	4.03E+08		

	2018	County	5	8.78E+12	1.76E+12	3240	**<2e-16**
		Residuals	570	3.09E+11	5.42E+08		
	2019	County	5	7.97E+12	1.59E+12	2824	**<2e-16**
		Residuals	545	3.07E+11	5.64E+08		
	2020	County	5	5.33E+12	1.07E+12	1773	**<2e-16**
		Residuals	398	2.40E+11	6.02E+08		
County	**Norfolk**	County	3	2.30E+11	7.65E+10	119.7	**<2e-16**
		Residuals	283	1.81E+11	6.39E+08		
	Suffolk	County	3	4.28E+11	1.43E+11	176.9	**<2e-16**
		Residuals	484	3.90E+11	8.07E+08		
	Essex	County	3	2.67E+11	8.90E+10	238.1	**<2e-16**
		Residuals	296	1.11E+11	3.74E+08		
	Middlesex	County	3	2.69E+11	8.96E+10	147	**<2e-16**
		Residuals	305	1.86E+11	6.09E+08		
	Bristol	County	3	2.37E+11	7.90E+10	324.6	**<2e-16**
		Residuals	402	9.78E+10	2.43E+08		
	Plymouth	County	3	1.80E+11	6.01E+10	158	**<2e-16**
		Residuals	281	1.07E+11	3.80E+08		

Pairwise comparisons:

Pairwise t-tests with Bonferroni correction showed significant differences in House Sale Prices for all pairs for the factor variable within the sub-dataframes Year and County (p-value <2e-16).

Non-Parametric Tests:

Kruskal Wallis H test showed significant differences in House Sale Prices for the factor variable within the sub-dataframes Year and County (p-value < 0.0001).

Pairwise comparisons:

Pairwise comparisons showed significant differences in House Sale Prices for most pairs for the factor variables county within the sub-dataframes containing each year separately (Table 4). Additionally, Table 5 showed significant differences in House Sale Prices for all pairs for the factor variables year within sub-dataframes containing each county separately (Table 5).

Table 4: Pairwise comparisons showing differences in House Sale Prices per county for the sub-dataframes years. Note that critical differences are the threshold necessary to determine that the differences are big enough to be significant.

Pairs		**Observed**	**Critical Difference**	**Significant**
2017	Bristol-Essex	159.76316	69.07723	TRUE
	Bristol-Middlesex	313.23418	68.30305	TRUE
	Bristol-Norfolk	241.20139	70.19446	TRUE
	Bristol-Plymouth	94.90972	70.19446	TRUE
	Bristol-Suffolk	419.74088	59.36843	TRUE
	Essex-Middlesex	153.47102	74.12935	TRUE
	Essex-Norfolk	81.43823	75.87567	TRUE

	Essex-Plymouth	64.85344	75.87567	FALSE
	Essex-Suffolk	259.97772	65.98837	TRUE
	Middlesex-Norfolk	72.03279	75.17153	FALSE
	Middlesex-Plymouth	218.32445	75.17153	TRUE
	Middlesex-Suffolk	106.5067	65.17751	TRUE
	Norfolk-Plymouth	146.29167	76.89418	TRUE
	Norfolk-Suffolk	178.53949	67.157	TRUE
	Plymouth-Suffolk	324.83115	67.157	TRUE
2018	Bristol-Essex	175.88068	69.44862	TRUE
	Bristol-Middlesex	339.34831	69.22893	TRUE
	Bristol-Norfolk	268.02027	73.04859	TRUE
	Bristol-Plymouth	101.7375	71.37409	TRUE
	Bristol-Suffolk	452.21212	62.60413	TRUE
	Essex-Middlesex	163.46763	73.43387	TRUE
	Essex-Norfolk	92.13959	77.04535	TRUE
	Essex-Plymouth	74.14318	75.45959	FALSE
	Essex-Suffolk	276.33144	67.22475	TRUE
	Middlesex-Norfolk	71.32804	76.84738	FALSE
	Middlesex-Plymouth	237.61081	75.25745	TRUE
	Middlesex-Suffolk	112.86381	66.99776	TRUE
	Norfolk-Plymouth	166.28277	78.78538	TRUE
	Norfolk-Suffolk	184.19185	70.93767	TRUE
	Plymouth-Suffolk	350.47462	69.21211	TRUE
2019	Bristol-Essex	166.44014	69.31758	TRUE
	Bristol-Middlesex	325.45725	68.82352	TRUE
	Bristol-Norfolk	252.53253	69.83353	TRUE
	Bristol-Plymouth	97.51779	69.57272	TRUE
	Bristol-Suffolk	431.74752	61.10287	TRUE
	Essex-Middlesex	159.01711	73.8918	TRUE
	Essex-Norfolk	86.09239	74.83344	TRUE
	Essex-Plymouth	68.92235	74.59012	FALSE
	Essex-Suffolk	265.30738	66.75989	TRUE
	Middlesex-Norfolk	72.92472	74.37603	FALSE
	Middlesex-Plymouth	227.93946	74.13121	TRUE
	Middlesex-Suffolk	106.29027	66.24676	TRUE
	Norfolk-Plymouth	155.01474	75.06985	TRUE
	Norfolk-Suffolk	179.21499	67.29545	TRUE
	Plymouth-Suffolk	334.22973	67.02477	TRUE
2020	Bristol-Essex	121.79555	59.72379	TRUE
	Bristol-Middlesex	243.49423	58.85479	TRUE
	Bristol-Norfolk	184.67027	57.80589	TRUE

	Bristol-Plymouth	70.02681	60.34793	TRUE
	Bristol-Suffolk	317.81923	53.02141	TRUE
	Essex-Middlesex	121.69868	63.39355	TRUE
	Essex-Norfolk	62.87473	62.42096	TRUE
	Essex-Plymouth	51.76874	64.78216	FALSE
	Essex-Suffolk	196.02368	58.01832	TRUE
	Middlesex-Norfolk	58.82396	61.59003	FALSE
	Middlesex-Plymouth	173.46742	63.9819	TRUE
	Middlesex-Suffolk	74.325	57.12339	TRUE
	Norfolk-Plymouth	114.64347	63.0184	TRUE
	Norfolk-Suffolk	133.14896	56.04209	TRUE
	Plymouth-Suffolk	247.79242	58.66061	TRUE

Table 5: Pairwise comparisons showing differences in House Sale Prices per year for the sub-dataframes counties. Note that critical differences are the threshold necessary to determine that the differences are big enough to be significant.

Pairs		**Observed**	**Critical Difference**	**Significant**
Norfolk	2017-2018	64.4756	36.24583	TRUE
	2017-2019	103.80177	35.89593	TRUE
	2017-2020	174.32378	37.61637	TRUE
	2018-2019	39.32617	35.64437	TRUE
	2018-2020	109.84818	37.37639	TRUE
	2019-2020	70.52202	37.03718	TRUE
Suffolk	2017-2018	173.06392	45.37533	TRUE
	2017-2019	220.59458	45.64325	TRUE
	2017-2020	272.66796	50.48034	TRUE
	2018-2019	47.53066	46.06057	TRUE
	2018-2020	99.60404	50.85798	TRUE
	2019-2020	52.07339	51.09717	TRUE
Essex	2017-2018	60.96322	35.83803	TRUE
	2017-2019	134.33153	36.77189	TRUE
	2017-2020	205.125	40.1007	TRUE
	2018-2019	73.36831	35.47103	TRUE
	2018-2020	144.16178	38.91129	TRUE
	2019-2020	70.79347	39.77305	TRUE
Middlesex	2017-2018	67.49026	36.43617	TRUE
	2017-2019	118.08181	37.27267	TRUE
	2017-2020	196.57996	40.36499	TRUE
	2018-2019	50.59155	36.19708	TRUE
	2018-2020	129.0897	39.37396	TRUE

	2019-2020	78.49815	40.1493	TRUE
Bristol	2017-2018	81.61115	41.6612	TRUE
	2017-2019	185.82792	42.22816	TRUE
	2017-2020	269.22614	46.00278	TRUE
	2018-2019	104.21677	41.76062	TRUE
	2018-2020	187.61499	45.57397	TRUE
	2019-2020	83.39821	46.09283	TRUE
Plymouth	2017-2018	61.8	35.32178	TRUE
	2017-2019	115.75	35.53563	TRUE
	2017-2020	186.40909	38.93913	TRUE
	2018-2019	53.95	34.59937	TRUE
	2018-2020	124.60909	38.08663	TRUE
	2019-2020	70.65909	38.28504	TRUE

Conclusions:

Factor: Name(county) after controlling for Year

For both, parametric and non-parametric tests, the null hypothesis that stated that there were no differences between House Sale Prices between the years 2017, 2018, 2019 and 2020 is rejected in favor of the null hypothesis which states that the House Sale Prices varied at least between two years.

Factor: Year (after controlling for County)

For both, parametric and non-parametric tests, the null hypothesis that stated that there were no differences between House Sale Prices between the counties Norfolk, Suffolk , Essex, Middlesex, Bristol and Plymouth is rejected in favor of the null hypothesis which states that the House Sale Prices varied at least between two years.

In the case of the pairwise comparisons, there were some small inconsistencies between parametric and non-parametric approaches. Parametric approaches found differences between all pairs of groups for both factor variables. This was not the case for the non-parametric pairwise tests (Table 4). County pairs such as Essex vs Plymouth and Middlesex vs Norfolk showed no differences in house sale prices whatsoever. The reason for this could be that those pairs have very similar population densities (low and high population density respectively).

Since the assumptions for non-parametric tests were not met. The conclusions yielded by the non-parametric approaches should be more significant than the parametric approaches.

It has been concluded that there are indeed significant differences in house sale prices among the years between different counties. However, the pairwise comparisons from tables 4 and 5 show that those differences are present for almost all counties and years and are highly significant. This makes it hard to assess which differences are more significant than others.

To help with this issue, Question 2 intends to determine the direction and magnitude of association between two extremes: counties Plymouth and Suffolk, which are, due to its differences in population density, the ultimate representatives of urban and suburban regions around Boston Metropolitan area.

8.Question 2:

Is there an association between house sale prices and the years 2017, 2018, 2019, 2020 for the counties Suffolk and Plymouth, separately?

Figure 6 shows that most of the relationships between County and Year are not linear, because of this, a Spearman Rank Correlation was used to answer the question stated above.

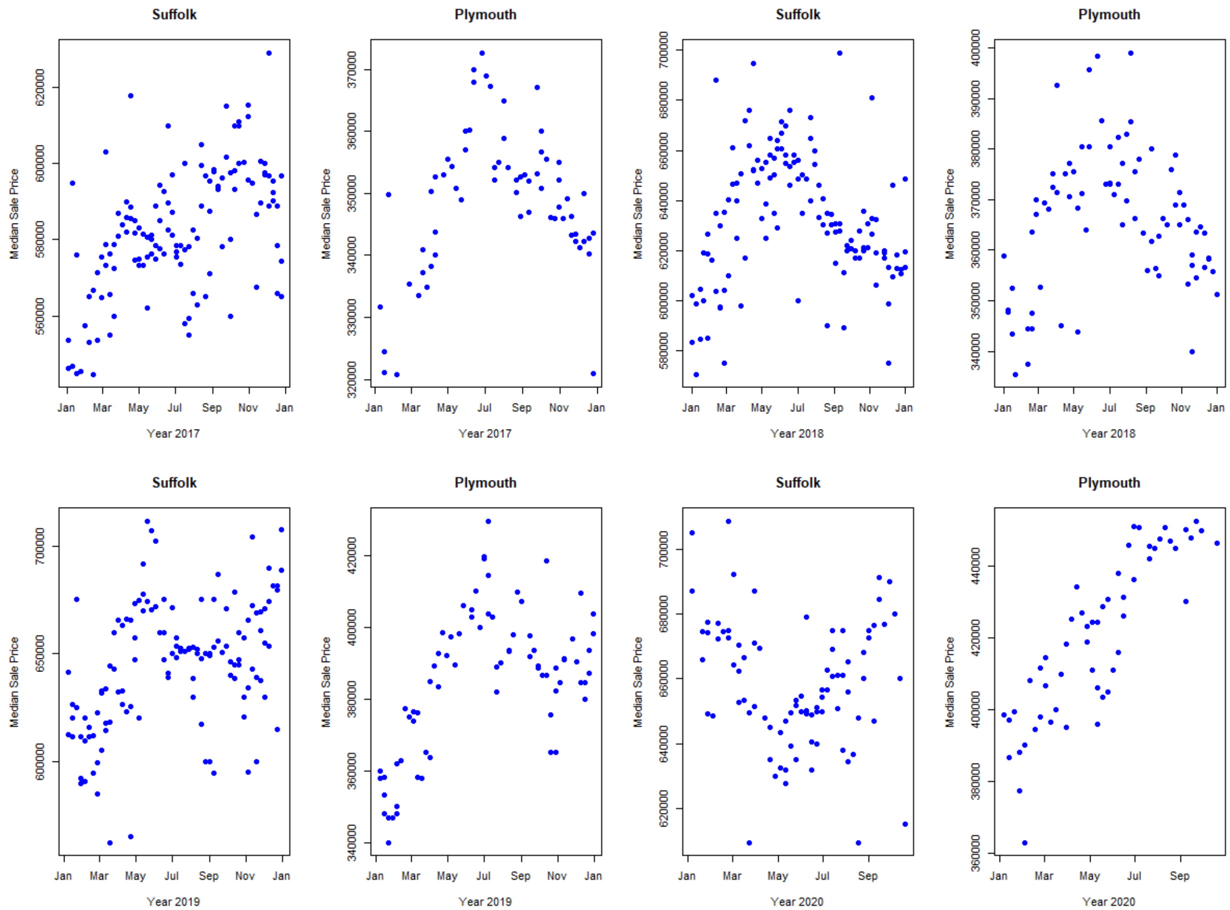

Figure 6: Plots showing the change in House Sale Prices by year and for the counties of Norfolk and Suffolk. Outliers were removed before plotting the variables.

Multiple correlation tests were done to determine if, by county, an association between time and house sale price was found. Depending on the pattern shown in the figure, if the plot seemed to show a negative association, a left sided test was used, if the association seemed to be mostly positive, then a right sided test was used.

It was assumed that the sample was taken randomly from the population.

Setting up the hypotheses and the α level:

$\alpha = 0.1$

Hypotheses:

Ho: $\rho = 0$ (There is no correlation between time and House Sale Price for the years 2017, 2018, 2019, 2020)

$H1_a$: $\rho < 0$ (There is a negative correlation between time and House Sale Price for the years 2017, 2018, 2019, 2020) or

$H1_b$: ρ > 0 (There is a positive correlation between time and House Sale Price for the years 2017, 2018, 2019, 2020)

Selecting the appropriate test statistic:

Since the relationships are in most cases non-linear a Spearman Rank correlation was used

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where:

d_i = differences in ranks for x_i and y_i observations

n = sample size

Source: (James McClave and Terry Sincich 2013)

Setting the decision rule:

If calculated $r_s < -r_{s,\,0.1}$ if $H1_a$ alternative hypothesis was chosen or $r_s > r_{s,\,0.1}$ if $H1_b$ alternative hypothesis was chosen, reject Ho

Also, if p-value < 0.1, reject ho (which was the method used to evaluate significance in this case)

Note that $r_{s,\,\alpha}$ are values for the Spearman's Rho Table

Computing the test statistics:

Table 6: Calculated Spearman's correlation coefficients for the relationships between time and House Sale Prices with its respective p-values. Significant values are shown in bold.

Spearman Rank Correlation		County	
		Suffolk	**Plymouth**
2017	ρ	0.521	0.056
	p-value	**1.66E-10**	0.330
2018	ρ	-0.071	0.021
	p-value	0.214	0.427
2019	ρ	0.431	0.492
	p-value	**1.90E-07**	**3.76E-06**
2020	ρ	-0.100	0.867
	p-value	0.182	**< 2.2E-16**

Conclusions:

Table 6 shows that for the county Suffolk, for the years 2017 and 2019 there is enough evidence to reject the null hypothesis. There is a positive relationship between house sale prices and the years 2017 and 2019. There is no sufficient evidence to reject the null hypothesis and suggest a negative correlation for the years 2018 and 2020.

For the county Plymouth, there is not enough evidence to reject the null hypothesis and suggest a positive correlation for the years 2017 and 2018. There is however enough evidence to reject the null hypothesis and suggest a positive correlation for the years 2019 and 2020.

Summarizing, the general results for the county Suffolk showed some evidence of a significant change in the relationship between those two variables during the time pre and post pandemic. There was some positive correlation for the year 2019 and a negative correlation for the year 2020 (although that correlation was not significant). This may indicate an important change in the time pre, post pandemic but a similar change can be seen for the years 2017 and 2020.

The correlations for the years 2019 and 2020 for the county Plymouth were both significant, but for the year 2020, the correlation coefficient was substantially high (almost twice as high than for the previous year).

General Conclusions:

It was possible to determine that there were indeed significant differences in House Sale Prices for the 6 counties studied and for all the years.

By performing the correlation test there was indeed a change in the level of association for the counties of Plymouth and Suffolk. The House prices for the county of Suffolk went from a positive correlation in 2019 to no correlation in the year 2020. For the county of Plymouth, the change in the association was more evident. It went from no evident correlation to a significant correlation in 2019 followed by a significant and very high correlation for the year 2020.

Since time is linear, it is possible to appreciate some change in counties. Suffolk is the most densely populated county in this study and Plymouth is the least densely populated. It is not possible to say that this change is caused by the events from 2020 but there is enough evident to suggest taking a deeper look on the situation, especially when considering to by property in the suburbs or sell property in the city.

References

- Homogeneity of variance [Internet]. Available from: http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_variance/.
- Field A, Miles J, Field Z. 2013a. Discovering statistics using R. Los Angeles; London; New Delhi: SAGE. 681 p.
- Field A, Miles J, Field Z. 2013b. Discovering statistics using R. Los Angeles; London; New Delhi: SAGE.
- Field A, Miles J, Field Z. 2013c. Discovering statistics using R. 1. publ., reprint ed. Los Angeles, Calif. [u.a.]: Sage Publ. 188 p.
- James McClave, Terry Sincich. 2013a. Nonparametric statistics. In: Statistics. 14 p.
- James McClave, Terry Sincich. 2013b. Statistics. 223 p.
- Long JD, Teetor P. 2019. R cookbook. Second edition ed. Beijing: O'Reilly.
- Mihir Zaveri, New York Times. Suburban home sales boom as people move out of N.Y.C.
- Mohammad Alaghemandi, Boston University. Module-5: One-way analysis of variance. lecture - 9.
- Real State Data Set [Internet]. Available from: https://www.redfin.com/news/data-center/.
- SOFIA RIVERA, Boston Magazine. 2020. Three major effects coronavirus is having on the Boston housing market.
- Assumption of Independence [Internet]; c2015 [cited 2020 Nov 27,]. Available from: https://www.statisticshowto.com/assumption-of-independence/.
- Teresa Ghilarducci, Bloomberg.com. 2020. A pandemic is a terrible time to buy real estate. Bloomberg.Com.
- Massachusetts Population Density County Rank [Internet]. Available from: http://www.usa.com/rank/massachusetts-state--population-density--county-rank.htm.
- Zar JH. 1998. Biostatistical analysis. 4. ed. ed. Englewood Cliffs, New Jersey: Regents/Prentice Hall. 223 p.

Annex

```
### Importing necessary libraries

library(ggplot2)
library(dbplyr)
library(tidyr)
library(tidyverse)
library(asbio)
library(data.table)
library(RSQLite)
library(zoo)
library(xts)
library(car)

setwd("C:/Users/franc/Desktop/BU_MS_ANALYTICS/1Semester/CS555/Term_project")

# Data from: https://www.redfin.com/news/data-center/, https://www.redfin.com/news/data-center-metrics-definitions/

### WARNING: Big file, if Computer is not powerful start from line 105

# Opening the table:

df <- read.table(file = 'weekly_housing_market_data_most_recent.tsv', sep = '\t', header = TRUE)
df <- df[df$region_name %like% "MA", ]

df2 <- df[df$region_name %like% "MA", ]
df2 <- df[,c(3,4,5,6,7,8,10,18,22,38)]
```

```
drv <- dbDriver("SQLite")
con <- dbConnect(drv, dbname = "housing.db")


dbWriteTable(con, "housing", df2)


# Access from database to make things faster: tsv files contain hidden data which
# makes the things run slower


query <- function(region.id, density, county.name){
  q <- dbGetQuery(con, "SELECT region_type, period_begin,
            period_end, duration, total_homes_sold,
            average_homes_sold, median_sale_price, median_days_to_close,
            average_new_listings FROM housing WHERE region_id == ?", region.id)

  q <- cbind(q, Pop.Density = rep(density, nrow(q)), Name = rep(county.name, nrow(q)))
  return(q)
}



# norfolk = region id == 1344

norfolk <- query(1344, 1536.8, "Norfolk" )


# suffolk = region id == 1346

suffolk <- query(1346, 6221.3, "Suffolk")
```

```
# middlesex = region id == 1342

middlesex <- query(1342, 1817.9, "Middlesex")


# essex = region id == 1338

essex <- query(1338, 914.2, "Essex")

# bristol = region id == 1336

bristol <- query(1336, 797.2, "Bristol")


# plymouth = region id == 1345

plymouth <- query(1345, 458.1, "Plymouth")

# Combining all rows:

dataset <- rbind(norfolk, suffolk, middlesex, essex, bristol, plymouth)


# Consulted:  https://stackoverflow.com/questions/33322248/how-to-import-a-tsv-file
#           http://www.usa.com/rank/massachusetts-state--population-density--county-rank.htm
#           r-Documentation


# Saving final dataset:
```

```
data.table::fwrite(dataset, file = 'final.csv', sep = ',')

# Write.table(df, file = 'Ma_House_prices.tsv', sep = '\t')

# Consulted: https://stackoverflow.com/questions/18587334/subset-data-to-contain-only-columns-whose-names-match-a-condition

# https://stackoverflow.com/questions/13043928/selecting-data-frame-rows-based-on-partial-string-match-in-a-column

# https://stackoverflow.com/questions/17108191/how-to-export-proper-tsv/17108345

# https://www.quora.com/How-do-you-load-a-TSV-file-into-R

# https://stackoverflow.com/questions/50021302/writing-small-dataframe-to-csv-creates-a-huge-file

# Clearing the memory:

############################# Start from here:######################################

################# 1. Preliminary steps:############################

df <- read.csv(file = 'final.csv', sep = ',', header = TRUE)

# a) Selecting columns of interest:

df <- df[,c(2,7,10,11)]

# b) Selecting years of interest (2017 to 2020):

# Note that the date column 'Period END' was removed and the date column

# 'Period Begin' was kept. The later was chosen over the former since,
```

```
# logically, it should be more likely that a deal would be closed before the
# listing period end than after that listing period is over.


# Transforming date column to datetime object:

df$period_begin <- as.Date(df$period_begin)

df <- subset(df,format(df$period_begin,'%Y')=='2017'|
             format(df$period_begin,'%Y')=='2018' |
             format(df$period_begin,'%Y')=='2019' |
             format(df$period_begin,'%Y')=='2020')

# ordering by date:

df <- df[order(df$period_begin),]


# Creating a new columns date factor:

year <- function(date.column){

  year.factor <- c()

  for (i in 1:length(date.column)){

    if(date.column[i]  >= '2020-01-01'){

      year.factor[i] = '2020'

    }else if (date.column[i]  >= '2019-01-01') {
```

```
    year.factor[i] = '2019'
   }else if (date.column[i]  >= '2018-01-01') {
    year.factor[i] = '2018'
   }else{
    year.factor[i] = '2017'
   }

 }

 return(year.factor)
}

year.factor <- year(df$period_begin)

df <- cbind(df, Year = year.factor)

# consulted :https://www.stat.berkeley.edu/~s133/dates.html
#          https://stat.ethz.ch/pipermail/r-help/2011-September/289364.html
#          https://stackoverflow.com/questions/6246159/how-to-sort-a-data-frame-by-date/6246186
#          https://www.ling.upenn.edu/~joseff/rstudy/week2.html
#          https://stackoverflow.com/questions/22235809/append-value-to-empty-vector-in-r/22235924
#          https://stackoverflow.com/questions/22235809/append-value-to-empty-vector-in-r/22235924


# Transforming 'Pop.Density', 'Name' and 'year' columns into a factor:

df$Year <- as.factor(df$Year)
df$Name <- as.factor(df$Name)
df$Pop.Density <- as.factor(df$Pop.Density)
```

```
########################### 2. Data Exploration :###############################

# Note that this is a preliminary exploration of data not shown in the report.
# other potential issues with outliers and distribution will be dealt with
# by factor before performing a particular statistical analysis.


attach(df)

summary(df)

# Check for NANs:

any(is.na(df))

# a) Distribution of numeric columns before dealing with general outliers:

h <- function(col, name){
  hist(col, xlab = sprintf("%s", name),
       main = sprintf('General distribution of %s', name),
       cex.main = 0.8, col = 'wheat3')
}


h(median_sale_price, 'Median Sale Price')


# Shapiro-Wilk Test for the distribution

shapiro.test(median_sale_price)
```

```
# Checking for general outliers

b <- function(col, name, y = '', main = 'General Boxplot of %s', graph = boxplot){
  graph(col, xlab = sprintf("%s", name), ylab = y,
        main = sprintf(main, name),
        cex.main = 0.8, col = 'wheat3')
}

b(median_sale_price, 'Median Sale Price')

# Dealing with general outliers:

df$median_sale_price[df$median_sale_price < quantile(df$median_sale_price, probs = 0.25) - 1.5*IQR(df$median_sale_price) |
                       df$median_sale_price > quantile(df$median_sale_price, probs = 0.75) + 1.5*IQR(df$median_sale_price)] <- NA

# consulted: (https://stackoverflow.com/questions/4787332/how-to-remove-outliers-from-a-dataset/31683403#31683403 )

df <- na.omit(df)

detach(df)
```

```
attach(df)



############################# 3. Analytics :####################################

###############Question a:###################

# Are there any significant differences in Median House Sale Price
# between 'Average Homes Sold' and the years 2017, 2018 , 2019, 2020?.

# The variable population density will not be considered here since it derives
# from the variable county. This is done in order to avoid using two measurement
# for the same group unit (dependency)

# i : First, testing for normality and homogeneity of variances for the groups
# of the two factor variables(Year and County) after dealing with potential outliers:



##### Outliers: ######


# Name (county):

Norfolk <- median_sale_price[Name == 'Norfolk']
Suffolk <-median_sale_price[Name == 'Suffolk']
Essex <- median_sale_price[Name == 'Essex']
Middlesex <- median_sale_price[Name == 'Middlesex']
```

```
Bristol <- median_sale_price[Name == 'Bristol']
Plymouth <- median_sale_price[Name == 'Plymouth']

par(mfrow = c(2,3))

b(Norfolk, 'Norfolk', y = 'Median Sale Price', main = 'House sale price for %s')
b(Suffolk, 'Suffolk', y = 'Median Sale Price', main = 'House sale price for %s')
b(Essex, 'Essex', y = 'Median Sale Price', main = 'House sale price for %s')
b(Middlesex, 'Middlesex', y = 'Median Sale Price', main = 'House sale price for %s')
b(Bristol, 'Bristol', y = 'Median Sale Price', main = 'House sale price for %s')
b(Plymouth, 'Plymouth', y = 'Median Sale Price', main = 'House sale price for %s')

par(mfrow = c(1,1))

county.outliers <- c(boxplot(Norfolk)$out,boxplot(Suffolk)$out, boxplot(Essex)$out,
            boxplot(Middlesex)$out, boxplot(Bristol)$out,boxplot(Plymouth)$out )

for (i in county.outliers){
  df$median_sale_price[df$median_sale_price == i] <- NA
}

df <- na.omit(df)

detach(df)
attach(df)
```

```
# Year:

year2017 <- median_sale_price[period_begin < '2018-01-01']

year2018 <- median_sale_price[period_begin < '2019-01-01']

year2019 <- median_sale_price[period_begin < '2020-01-01']

year2020 <- median_sale_price[period_begin >= '2020-01-01']



par(mfrow = c(2,2))

b(year2019 , '2017', y = 'Median Sale Price', main = 'House sale price for %s')
b(year2019 , '2018', y = 'Median Sale Price', main = 'House sale price for %s')
b(year2019 , '2019', y = 'Median Sale Price', main = 'House sale price for %s')
b(year2020, '2020', y = 'Median Sale Price', main = 'House sale price for %s')

par(mfrow = c(1,1))



### Summary for years 2017 to 2020 ####


df2017 <- subset(df, Year == '2017')
df2018 <- subset(df, Year == '2018')
df2019 <- subset(df, Year == '2019')
```

```
df2020  <- subset(df, Year == '2020')



# Overall:

summary(df2017$median_sale_price)
summary(df2018$median_sale_price)
summary(df2019$median_sale_price)
summary(df2020$median_sale_price)



# Per county and year:

summary(subset(df2017$median_sale_price, Name == 'Norfolk'))
summary(subset(df2017$median_sale_price, Name == 'Suffolk'))
summary(subset(df2017$median_sale_price, Name == 'Essex'))
summary(subset(df2017$median_sale_price, Name == 'Middlesex'))
summary(subset(df2017$median_sale_price, Name == 'Bristol'))
summary(subset(df2017$median_sale_price, Name == 'Plymouth'))

summary(subset(df2018$median_sale_price, Name == 'Norfolk'))
summary(subset(df2018$median_sale_price, Name == 'Suffolk'))
summary(subset(df2018$median_sale_price, Name == 'Essex'))
summary(subset(df2018$median_sale_price, Name == 'Middlesex'))
summary(subset(df2018$median_sale_price, Name == 'Bristol'))
summary(subset(df2018$median_sale_price, Name == 'Plymouth'))



summary(subset(df2019$median_sale_price, Name == 'Norfolk'))
```

```
summary(subset(df2019$median_sale_price, Name == 'Suffolk'))
summary(subset(df2019$median_sale_price, Name == 'Essex'))
summary(subset(df2019$median_sale_price, Name == 'Middlesex'))
summary(subset(df2019$median_sale_price, Name == 'Bristol'))
summary(subset(df2019$median_sale_price, Name == 'Plymouth'))

summary(subset(df2020$median_sale_price, Name == 'Norfolk'))
summary(subset(df2020$median_sale_price, Name == 'Suffolk'))
summary(subset(df2020$median_sale_price, Name == 'Essex'))
summary(subset(df2020$median_sale_price, Name == 'Middlesex'))
summary(subset(df2020$median_sale_price, Name == 'Bristol'))
summary(subset(df2020$median_sale_price, Name == 'Plymouth'))

##### Normality #####:

# Name (county):

par(mfrow = c(2,3))

b(Norfolk, 'Norfolk', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(Suffolk, 'Suffolk', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(Essex, 'Essex', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(Middlesex, 'Middlesex', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(Bristol, 'Bristol', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(Plymouth, 'Plymouth', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)

par(mfrow = c(1,1))

shapiro.test(Norfolk)
shapiro.test(Suffolk)
```

```
shapiro.test(Essex)
shapiro.test(Middlesex)
shapiro.test(Bristol)
shapiro.test(Plymouth)


# Year:

par(mfrow = c(2,2))

b(year2017, '2017', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(year2018, '2018', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(year2019, '2019', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)
b(year2020, '2020', y = 'Frequency', main = 'House sale price distribution for %s', graph = hist)

par(mfrow = c(1,1))


shapiro.test(year2017)
shapiro.test(year2018)
shapiro.test(year2019)
shapiro.test(year2020)

#### Homogeneity of variances #####

library(lawstat)
library(car)

# County:

leveneTest(df$median_sale_price, Name)
```

```
# Year:

# years <- cbind(year2017, year2018, year2019, year2020)

leveneTest(df$median_sale_price, Year)

# consulted: https://www.rdocumentation.org/packages/lawstat/versions/3.2/topics/levene.test

##### Graphical comparisons# of name and year #####:

boxplot(df$median_sale_price ~ df$Name, col = 'wheat3',
      ylab = 'Median House Sale Price', xlab = 'County')
boxplot(df$median_sale_price ~ df$Year, col = 'wheat3',
      ylab = 'Median House Sale Price', xlab = 'Year')


df2017 <- subset(df, Year == '2017')
df2018  <- subset(df, Year == '2018')
df2019 <- subset(df, Year == '2019')
df2020  <- subset(df, Year == '2020')

par(mfrow = c(2,2))

boxplot(df2017$median_sale_price ~ df2017$Name, col = 'wheat3',
      ylab = 'Median House Sale Price', xlab = 'County', main = '2017')

boxplot(df2018$median_sale_price ~ df2018$Name, col = 'wheat3',
      ylab = 'Median House Sale Price', xlab = 'County', main = '2018')
```

```
boxplot(df2019$median_sale_price ~ df2019$Name, col = 'wheat3',
        ylab = 'Median House Sale Price', xlab = 'County', main = '2019')

boxplot(df2020$median_sale_price ~ df2020$Name, col = 'wheat3',
        ylab = 'Median House Sale Price', xlab = 'County', main = '2020')

par(mfrow = c(1,1))

nor <- subset(df, Name == 'Norfolk')
su <- subset(df, Name == 'Suffolk')
es <- subset(df, Name == 'Essex')
mid <- subset(df, Name == 'Middlesex')
bris <- subset(df, Name == 'Bristol')
ply <- subset(df, Name == 'Plymouth')

par(mfrow = c(2,3))

boxplot(nor$median_sale_price ~ nor$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Norfolk')
boxplot(su$median_sale_price ~ su$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Suffolk')
boxplot(es$median_sale_price ~ es$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Essex')
boxplot(mid$median_sale_price ~ mid$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Middlesex')
boxplot(bris$median_sale_price ~ bris$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Bristol')
boxplot(ply$median_sale_price ~ ply$Year,col = 'wheat3', xlab = 'Year', ylab = 'Median House Sale Price', main = 'Plymouth')

par(mfrow = c(1,1))
```

```
#### Statistical Analyses: ####

# Since the assumptions for parametric tests are not met. The parametric
# test will be accompanied by a non-parametric counterpart. The results obtained
# from the non parametric version of the test will be given more consideration
# that the results obtained for the parametric tests:

# Formal hypothesis test for question a. Two way ANOVa:

# 1. Setting hypothesis and alpha level

# Alpha = 0.05



# Parametric Test:---------



# Factor: Name(county) after controlling for Year

#Ho: mu-Norfolk = mu-Suffolk = mu-Essex = mu-Middlesex = mu-Plymouth = mu-Bristol = 0

#H1: mu -i != mu-j (at least two of the counties present significant differences

# in Median Sale House Price)

# Factor : Year (after controlling for county)
```

#Ho: mu-2017 = mu2018 = mu-2019 = mu2020

#H1: mu-some.year != mu-other.year

Non-parametric Test:------

Factor: Name(county) after controlling for Year

#Ho: population-Norfolk = population-Suffolk = population-Essex

= population-Middlesex = population-Plymouth = population-Bristol = 0

#H1: population -i != population-j

(at least two of the counties present significant differences

in "Median Sale House Price")

Factor : Year (after controlling for county)

#Ho: population-2017 = population-2018 = population-2019 = population-2020

#H1: population-some.year != population-some.other.year

#Consulted: Biostatistical Analysis, Jerold H. Zar 196-197p

2. Selecting the appropriate test statistic

Two way ANOVA between dependent variable median_sale_price and the factor

variables Name(county) and Year:

If interactions are found, a stratified one-way analysis of variance will be

performed between the dependent variable median_sale_price and the factor

variables Name(county) and Year followed by a post-hoc pairwise t-test

Stratified non parametric Kruskal-Wallis analysis of variance between the

dependent variable median_sale_price and the factor variables Name(county)

and Year followed by a Multiple comparison test after Kruskal-Wallis

using critical differences among groups (described in kruskalmc {pgirmess}

package and Discovering Statistics Using R. Andy Field, Jeremy Miles &

Zoe Fields)

#

Testing for interactions:

```
model <- lm(median_sale_price ~ Name + Year + Name*Year, data = df)

interaction.plot(df$Name, df$Year, df$median_sale_price, col = 1:2,
        ylab = 'Median Sale Price', xlab = 'County Name', legend = T,
        trace.label = 'Year', main = 'Interaction Between Year and County')

summary(model)
```

```
# There appears to be an interaction between the

# county Suffolk and the years 2018, 2019 and 2020. Also, an interaction between

# the counties Norfolk and Essex with the year 2020. Suffolk is the most populous

# county in this analysis, which makes it very important. Due to the interactions,

# A stratified ANOVA will be performed followed by a Pairwise t test with

# Bonferroni correction


# 3. Setting the decision rule:


# if Calculated statistic of choice

# is Greater or equal than the Critical value, reject Ho. Also, if p-value < alpha,

# Reject Ho


# Critical values:



# Parametric:--------




# For stratified analysis Factor: county, stratus(sub-dataframe = year)


#2017


qf(0.95, df1 = 5, df2 = 538)


# Post-Hoc:


qt(0.975, df = 538)
```

```
#2018

qf(0.95, df1 = 5, df2 = 570)

# Post-Hoc:

qt(0.975, df = 570)

#2019

qf(0.95, df1 = 5, df2 = 545)

# Post-Hoc:

qt(0.975, df = 545)

#2020

qf(0.95, df1 = 5, df2 = 398)

# Post-Hoc:

qt(0.975, df = 398)

# For stratified analysis Factor: Year,  stratus(sub-dataframe = county)

#Norfolk

qf(0.95, df1 = 3, df2 = 283)
```

```
# Post-Hoc:

qt(0.975, df = 283)

#Suffolk

qf(0.95, df1 = 3, df2 = 484)

# Post-Hoc:

qt(0.975, df = 484)

#Essex

qf(0.95, df1 = 3, df2 = 296)

# Post-Hoc:

qt(0.975, df = 296)

#Middlesex

qf(0.95, df1 = 3, df2 = 305)

# Post-Hoc:

qt(0.975, df = 305)

#Bristol
```

```
qf(0.95, df1 = 3, df2 = 402)

# Post-Hoc:

qt(0.975, df = 402)

#Plymouth

qf(0.95, df1 = 3, df2 = 281)

# Post-Hoc:

qt(0.975, df = 281)

# Non-parametric:--------



# For stratified analysis Factor: county, stratus(sub-dataframe = year)

qchisq(0.95, df = 5)

# Post-Hoc:

critical.difference <- function(N, ni, nj, k){
  c = qnorm(1-(0.05/(k*(k-1))))*(sqrt((N*(N+1))/12*(1/ni+1/nj)))
  return(c)
}


# Example:
```

```
critical.difference(N = nrow(df2017), ni = nrow(subset(df2017, Name == 'Suffolk')),
                    nj = nrow(subset(df2017, Name == 'Norfolk')), k = 6)


# For stratified analysis Factor: Year,  stratus(sub-dataframe = county)

qchisq(0.95, df = 3)

# Post-Hoc:

# Example:

critical.difference(N = nrow(nor), ni = nrow(subset(nor, Year == '2017')),
                    nj = nrow(subset(nor, Year == '2018')), k = 4)


# Consulted: Biostatistical Analysis, Jerold H. Zar 223-226p

# 4. Computing test statistic:


# Parametric:--------


# Stratification by # Year:

#2017
```

```
a <- aov(median_sale_price ~ Name, data = df2017)
summary(a)

#2018

b <- aov(median_sale_price ~ Name, data = df2018)
summary(b)

#2019

c <- aov(median_sale_price ~ Name, data = df2019)
summary(c)

#2020
d <- aov(median_sale_price ~ Name, data = df2020)
summary(d)


#Pairwise:

pairwise.t.test(median_sale_price, Name, data = df2017, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Name, data = df2018, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Name, data = df2019, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Name, data = df2020, p.dj = 'bonferroni')


# consulted: Lecture 9 and 10 CS555. Module 5 Video tutorial code review
#	R-Documentation Two-way Interaction Plot
```

```
# Stratification by #County#

#Norfolk

summary(aov(median_sale_price ~ Year, data = nor))

#Suffolk

summary(aov(median_sale_price ~ Year, data = su))

#Essex

summary(aov(median_sale_price ~ Year, data = es))

#Middlesex

summary(aov(median_sale_price ~ Year, data = mid))

#Bristol

summary(aov(median_sale_price ~ Year, data = bris))

#Plymouth
```

```
summary(aov(median_sale_price ~ Year, data = ply))

#Pairwise:

pairwise.t.test(median_sale_price, Year, data = nor, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Year, data = su, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Year, data = es, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Year, data = mid, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Year, data = bris, p.dj = 'bonferroni')
pairwise.t.test(median_sale_price, Year, data = ply, p.dj = 'bonferroni')

# Non-Parametric:--------

# Stratification by Year:

library(pgirmess)

# 2017:

kruskal.test(median_sale_price ~ Name, data = df2017)

# Pairwise:

kruskalmc(median_sale_price ~ Name, data = df2017)
```

```
# 2018:

kruskal.test(median_sale_price ~ Name, data = df2018)

# Pairwise:

kruskalmc(median_sale_price ~ Name, data = df2018)

# 2019:

kruskal.test(median_sale_price ~ Name, data = df2019)

# Pairwise:

kruskalmc(median_sale_price ~ Name, data = df2019)

# 2020:

kruskal.test(median_sale_price ~ Name, data = df2020)

# Pairwise:

kruskalmc(median_sale_price ~ Name, data = df2020)
```

```
# Consulted: Discovering Statistics Using R. Andy Field, Jeremy Miles, Zoe Field,
# 679-681p R-Documentation, pgirmess library:
# Multiple comparison test after Kruskal-Wallis

# Stratification by #County#




#Norfolk

kruskal.test(median_sale_price ~ Year, data = nor)
kruskalmc(median_sale_price ~ Year, data = nor)

#Suffolk

kruskal.test(median_sale_price ~ Year, data = su)
kruskalmc(median_sale_price ~ Year, data = su)

#Essex

kruskal.test(median_sale_price ~ Year, data = es)
kruskalmc(median_sale_price ~ Year, data = es)

#Middlesex

kruskal.test(median_sale_price ~ Year, data = mid)
kruskalmc(median_sale_price ~ Year, data = mid)
#Bristol
```

```
kruskal.test(median_sale_price ~ Year, data = bris)
kruskalmc(median_sale_price ~ Year, data = bris)

#Plymouth

kruskal.test(median_sale_price ~ Year, data = ply)
kruskalmc(median_sale_price ~ Year, data = ply)

# 5. Conclusions:

# Significant differences were found for all global pairwise comparison,
# reject Ho



library(car)

###################################b#########################################

# Is there a pattern or association between Median Sale Price for the
# counties of Suffolk and Plymouth and time (Years
# 2019 and 2020)?


su2017 <- subset(df2017,  Name == 'Suffolk')
ply2017 <- subset(df2017,  Name == 'Plymouth')


su2018 <- subset(df2018,  Name == 'Suffolk')
ply2018 <- subset(df2018,  Name == 'Plymouth')
```

```
su2019 <- subset(df2019,  Name == 'Suffolk')

ply2019 <- subset(df2019,  Name == 'Plymouth')


su2020 <- subset(df2020,  Name == 'Suffolk')
ply2020 <- subset(df2020,  Name == 'Plymouth')



# Is there a relationship? between time and median sale price:


lp <- function(period, sale.price, y, main){
  plt = plot(period, sale.price, xlab = sprintf('Year %s', y),
           ylab = 'Median Sale Price', col = 'blue',
           main = main, pch = 16)
  return(plt)
}






### Correlation Tests:

# Eliminating outliers:

# Suffolk
```

```
su2017$median_sale_price[su2017$median_sale_price < quantile(su2017$median_sale_price, probs = 0.25) - 1.5*IQR(su2017$median_sale_price) |

                    su2017$median_sale_price > quantile(su2017$median_sale_price, probs = 0.75) + 1.5*IQR(su2017$median_sale_price)] <- NA

su2018$median_sale_price[su2018$median_sale_price < quantile(su2018$median_sale_price, probs = 0.25) - 1.5*IQR(su2018$median_sale_price) |

                    su2018$median_sale_price > quantile(su2018$median_sale_price, probs = 0.75) + 1.5*IQR(su2018$median_sale_price)] <- NA

su2019$median_sale_price[su2019$median_sale_price < quantile(su2019$median_sale_price, probs = 0.25) - 1.5*IQR(su2019$median_sale_price) |

                    su2019$median_sale_price > quantile(su2019$median_sale_price, probs = 0.75) + 1.5*IQR(su2019$median_sale_price)] <- NA

su2020$median_sale_price[su2020$median_sale_price < quantile(su2020$median_sale_price, probs = 0.25) - 1.5*IQR(su2020$median_sale_price) |

                    su2020$median_sale_price > quantile(su2020$median_sale_price, probs = 0.75) + 1.5*IQR(su2020$median_sale_price)] <- NA



su2017 <- na.omit(su2017)
su2018 <- na.omit(su2018)
su2019 <- na.omit(su2019)
su2020 <- na.omit(su2020)

# Plymouth
```

```
ply2017$median_sale_price[ply2017$median_sale_price < quantile(ply2017$median_sale_price, probs = 0.25) - 1.5*IQR(ply2017$median_sale_price) |

                    ply2017$median_sale_price > quantile(ply2017$median_sale_price, probs = 0.75) + 1.5*IQR(ply2017$median_sale_price)] <- NA

ply2018$median_sale_price[ply2018$median_sale_price < quantile(ply2018$median_sale_price, probs = 0.25) - 1.5*IQR(ply2018$median_sale_price) |

                    ply2018$median_sale_price > quantile(ply2018$median_sale_price, probs = 0.75) + 1.5*IQR(ply2018$median_sale_price)] <- NA

ply2019$median_sale_price[ply2019$median_sale_price < quantile(ply2019$median_sale_price, probs = 0.25) - 1.5*IQR(ply2019$median_sale_price) |

                    ply2019$median_sale_price > quantile(ply2019$median_sale_price, probs = 0.75) + 1.5*IQR(ply2019$median_sale_price)] <- NA

ply2020$median_sale_price[ply2020$median_sale_price < quantile(ply2020$median_sale_price, probs = 0.25) - 1.5*IQR(ply2020$median_sale_price) |

                    ply2020$median_sale_price > quantile(ply2020$median_sale_price, probs = 0.75) + 1.5*IQR(ply2020$median_sale_price)] <- NA

ply2017 <- na.omit(ply2017)
ply2018 <- na.omit(ply2018)
ply2019 <- na.omit(ply2019)
ply2020 <- na.omit(ply2020)



# plots:
```

```
par(mfrow = c(2,4))

lp(su2017$period_begin, su2017$median_sale_price, '2017', main = 'Suffolk')
lp(ply2017$period_begin, ply2017$median_sale_price, '2017', main = 'Plymouth')
lp(su2018$period_begin, su2018$median_sale_price, '2018', main = 'Suffolk')
lp(ply2018$period_begin, ply2018$median_sale_price, '2018', main = 'Plymouth')
lp(su2019$period_begin, su2019$median_sale_price, '2019', main = 'Suffolk')
lp(ply2019$period_begin, ply2019$median_sale_price, '2019', main = 'Plymouth')
lp(su2020$period_begin, su2020$median_sale_price, '2020', main = 'Suffolk')
lp(ply2020$period_begin, ply2020$median_sale_price, '2020', main = 'Plymouth')

par(mfrow = c(1,1))

# 2017

cor.test(as.numeric(su2017$period_begin), su2017$median_sale_price, method = 'spearman', conf.level = 0.9,
alternative = 'greater') # Significant

cor.test(as.numeric(ply2017$period_begin), ply2017$median_sale_price, method = 'spearman', alternative =
'greater') # Non Significant

# 2018

cor.test(as.numeric(su2018$period_begin), su2018$median_sale_price, method = 'spearman', conf.level = 0.9,
alternative = 'less') # Significant

cor.test(as.numeric(ply2018$period_begin), ply2018$median_sale_price, method = 'spearman', conf.level = 0.9,
alternative = 'greater') # Non Significant
```

2019

cor.test(as.numeric(su2019$period_begin), su2019$median_sale_price, method = 'spearman', conf.level = 0.9, alternative = 'greater') # Significant

cor.test(as.numeric(ply2019$period_begin), ply2019$median_sale_price, method = 'spearman', conf.level = 0.9, alternative = 'greater') # Significant

2020

cor.test(as.numeric(su2020$period_begin), su2020$median_sale_price, method = 'spearman', conf.level = 0.9, alternative = 'less')# Non-Significant

cor.test(as.numeric(ply2020$period_begin), ply2020$median_sale_price, method = 'spearman', conf.level = 0.9, alternative = 'greater')# Significant

consulted : https://www.stat.berkeley.edu/~s133/dates.html

https://stat.ethz.ch/pipermail/r-help/2011-September/289364.html

https://stackoverflow.com/questions/6246159/how-to-sort-a-data-frame-by-date/6246186

https://www.ling.upenn.edu/~joseff/rstudy/week2.html