ANOVA

A comparison of Milwaukee MLS data for select years and market segments

Melanie A. Fitzgerald

Master of Business Administration Program, Carroll University

BUS 674 – Data Analytics in Practice

John A. Michl, MSBIA, MBA, CBIP

February 9, 2025

Problem Statement

A real estate firm would like to examine sales figures for the past few years. They would like to do a test comparison using home selling prices and comparing the means across market segments at both the county and municipality levels. Market segments are defined using the Esri Tapestry Segmentation which is a "geodemographic system that identifies 68 distinctive markets in the US based on socioeconomic and demographic characteristics to provide an accurate, comprehensive profile of US consumers (*TAPESTRY SEGMENTATION the Fabric of America's Neighborhoods TM UNITED STATES of AMERICA*, n.d.)." I was tasked to examine two (2BR), three (3BR), and four (4BR) bedroom homes in the City of Milwaukee (Milwaukee County) for market segments *City Strivers*, *Family Foundations*, and *Fresh Ambitions* for years 2019-2022.

Market Segments

City Strivers (CS) are young, foreign-born residents who have embraced the American lifestyle. They live in densely populated urban neighborhoods, mostly as renters living in older, multiunit buildings, are a mix of household types, and mostly commute outside of their county to work. Half of City Strivers have some college education with labor participation slightly below the national average. (*Esri Tapestry Segmentation—Esri Demographics Regional Data* | *Documentation*, 2022)

Family Foundations (FF) is characterized by its focus on families, many of which have older children still living at home and with a slightly higher average household size. They live in major metropolitan areas in mostly single-family homes. Many are still working while a significant portion is retired. While over half have attended college, labor participation is lower as retirement rates rise. (*Esri Tapestry Segmentation—Esri Demographics Regional Data* | *Documentation*, 2022)

Fresh Ambitions (FA) are young immigrant families with multigenerational households. They are predominantly renters who mostly live in row houses located in major urban cities and subsidized by public assistance. A high school diploma is often the highest education level but they are hardworking and have a strong work ethic. (*Esri Tapestry Segmentation—Esri Demographics Regional Data* | *Documentation*, 2022)

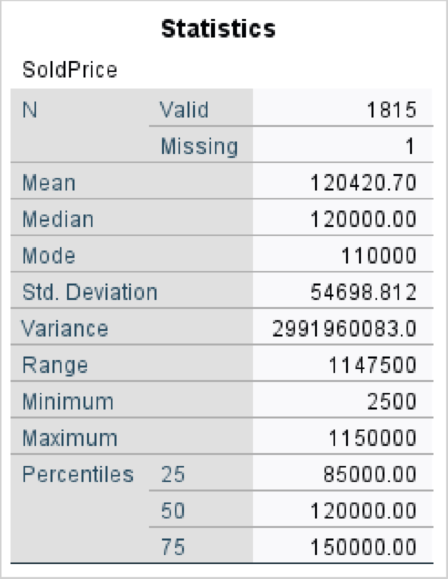
Data

The dataset is from the Milwaukee MLS and contains all home sales in the Milwaukee 7 region since 1996. SPSS is used for analysis which includes descriptive statistics and ANOVA. A new dataset was extracted to include only data from segments *City Strivers*, *Family Foundations*, and *Fresh Ambitions* and years 2019-2022. There was one outlier value of "1" in the home sales price which was coded as a missing value. Many entries had a dot (".") which means that the cell has missing data and is automatically recognized in SPSS. Data label *DominantTapestryNumberEsri*, which is the variable that identifies the market segment with a number, was simplified to *TapestryNumber* and its data type was changed from scale to nominal. Labels were attached to each number to identify each tapestry number to its market segment.

Descriptive statistics

Home selling price (*SoldPrice*)

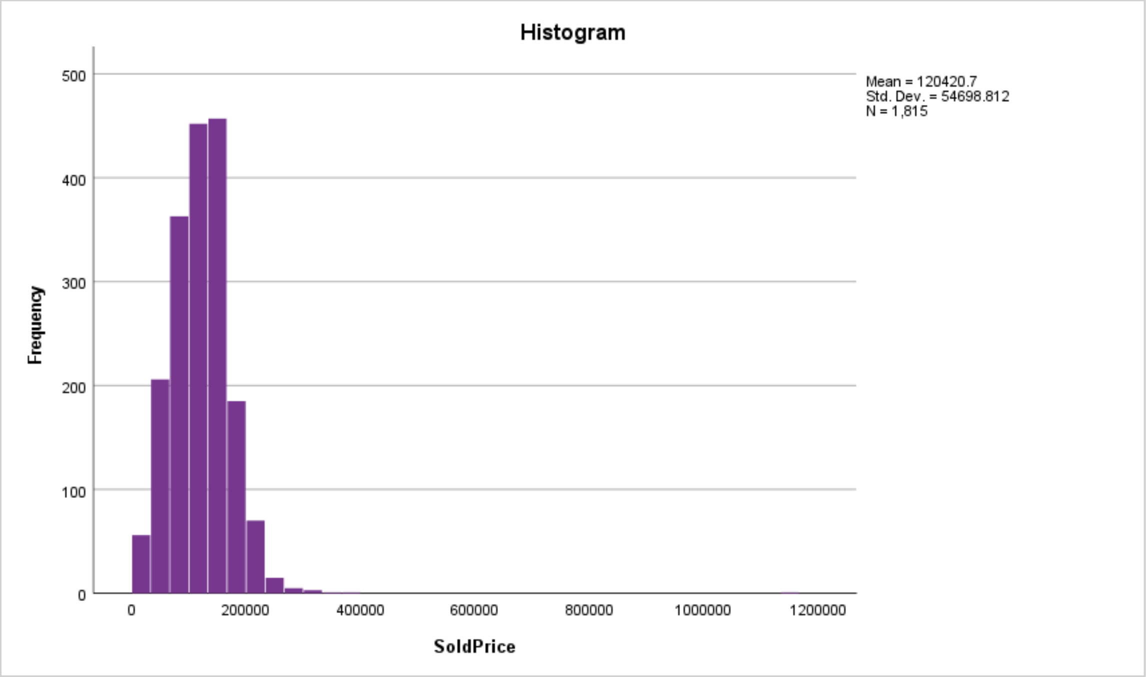
Descriptive statistics was done for the home selling price *SoldPrice* across the three market segments. SoldPrice shows a mean of $120,420 with a median of $120,000. The mean and median are very close in price suggesting a normal distribution of data as shown in the histogram below. Note that the 50th percentile is at the median value. The range of *SoldPrice* is large with a minimum *SoldPrice* of $2500 and maximum of $1,150,000. The quartile range shows that 75% of the selling price fall below $150,000.



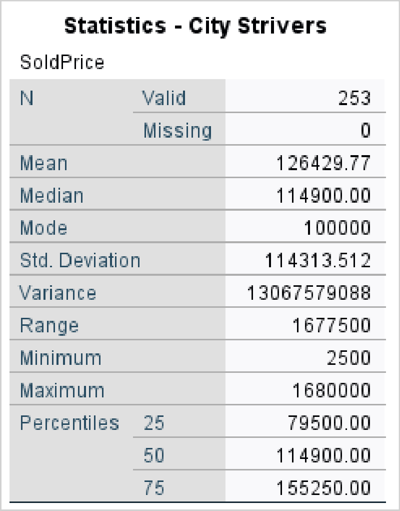
Statistics

SoldPrice

N	Valid	1815
	Missing	1
Mean		120420.70
Median		120000.00
Mode		110000
Std. Deviation		54698.812
Variance		2991960083.0
Range		1147500
Minimum		2500
Maximum		1150000
Percentiles	25	85000.00
	50	120000.00
	75	150000.00



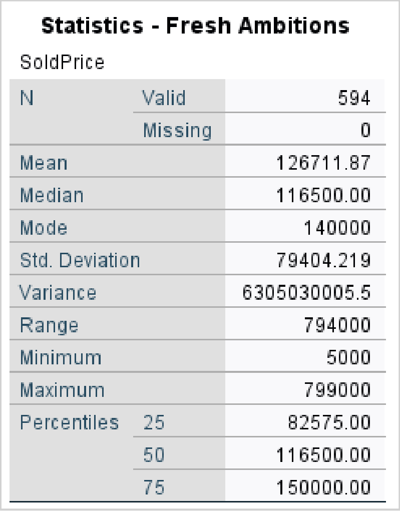
Descriptive statistics done for each market segment shows CS has a mean *SoldPrice* of $123,430 with a median of $114,900. The range is large with a minimum price of $2500 and a maximum price of $168,000 with 75% of the home prices falling below $155,250. FA show a mean *SoldPrice* of $126,712 with a median of $116,500. The range is smaller than CS with its minimum price of $5000 and maximum price of $799,000 with 75% of prices falling below $150,000. The mean *SoldPrice* in FF is $128,694 with a median of $128,500. The mean and median are very close in price suggesting a normal distribution. Note that its 50th percentile is at the median value.



Statistics - City Strivers

SoldPrice

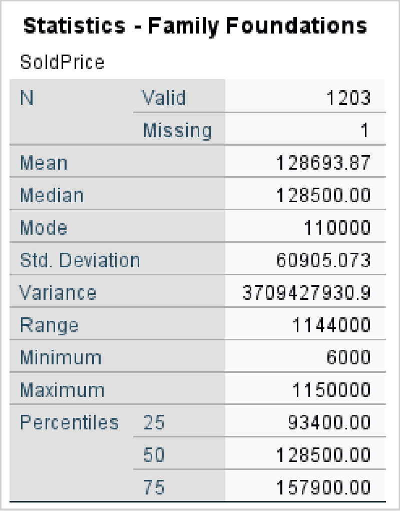
N	Valid	253
	Missing	0
Mean		126429.77
Median		114900.00
Mode		100000
Std. Deviation		114313.512
Variance		13067579088
Range		1677500
Minimum		2500
Maximum		1680000
Percentiles	25	79500.00
	50	114900.00
	75	155250.00



Statistics - Fresh Ambitions

SoldPrice

N	Valid	594
	Missing	0
Mean		126711.87
Median		116500.00
Mode		140000
Std. Deviation		79404.219
Variance		6305030005.5
Range		794000
Minimum		5000
Maximum		799000
Percentiles	25	82575.00
	50	116500.00
	75	150000.00



Statistics - Family Foundations

SoldPrice

N	Valid	1203
	Missing	1
Mean		128693.87
Median		128500.00
Mode		110000
Std. Deviation		60905.073
Variance		3709427930.9
Range		1144000
Minimum		6000
Maximum		1150000
Percentiles	25	93400.00
	50	128500.00
	75	157900.00

Hypothesis Statement

Single factor

The descriptive statistics above show the mean values of home selling prices for all market segments and for each segment considering all bedroom sizes 2-4. However, are the mean selling prices when grouped by bedroom size significantly different from one another? Does the selling price change based on the bedroom size for all segments? A hypothesis for this question can be tested to determine the difference between groups:

***H_0*:** There *is no significant difference* in the mean price between groups

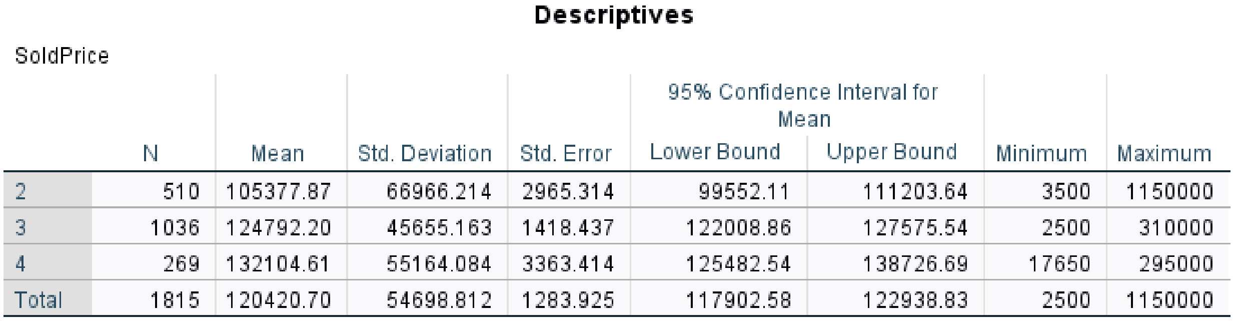
***H_a*:** There *is a difference* in the mean price between groups

where group is the number of bedrooms. The null hypothesis ***H_0*** states that the mean selling price based on a group, in this case the number of bedrooms, has no significant difference. The alternative hypothesis ***H_a*** states that the mean selling price for at least one of the groups is different.

One-way ANOVA

Number of bedrooms (*Bedrooms*)

ANOVA is used to do a test comparison of mean selling prices for 2BR, 3BR, and 4BR homes across the three market segments (CS, FF, FA) for years 2019-2022. A significance level of 0.05 is assumed. *SoldPrice* is defined as the dependent variable and *Bedrooms* as the groups. Descriptive statistics shows N=1815 total entries and the mean selling price μ for each group where μ_{2BR} = \$105,378, μ_{3BR} = \$124,792, and μ_{4BR} = \$132,104.

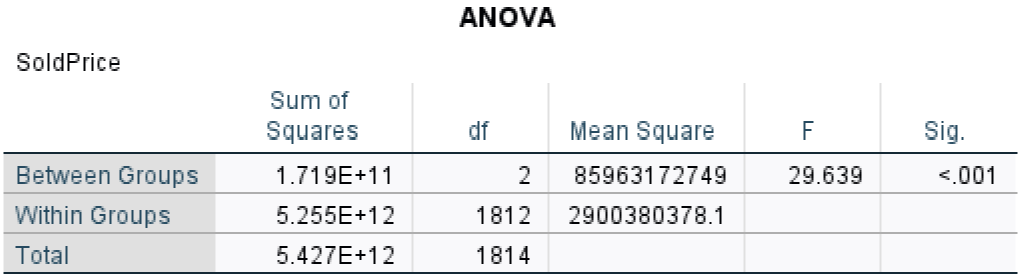


Descriptives

SoldPrice

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
2	510	105377.87	66966.214	2965.314	99552.11	111203.64	3500	1150000
3	1036	124792.20	45655.163	1418.437	122008.86	127575.54	2500	310000
4	269	132104.61	55164.084	3363.414	125482.54	138726.69	17650	295000
Total	1815	120420.70	54698.812	1283.925	117902.58	122938.83	2500	1150000

The ANOVA shows an *F*-statistic between the groups of 29.639 and a significance level less than 0.05 (p < 0.001) indicating that the differences between the groups are statistically significant and the null hypothesis can be rejected for the alternative hypothesis.

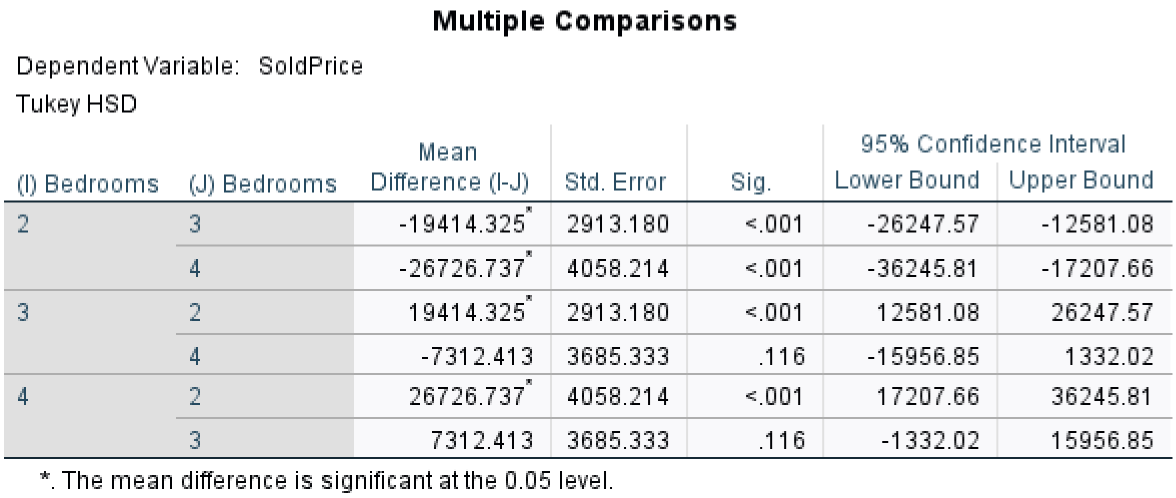


ANOVA

SoldPrice

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.719E+11	2	85963172749	29.639	<.001
Within Groups	5.255E+12	1812	2900380378.1		
Total	5.427E+12	1814			

A comparison of the groups using a Tukey Post-Hoc analysis shows the relationship of means between the groups. 2BR compared to each 3BR and 4BR has a significance level less than 0.05 (*p* < 0.001) for each comparison indicating that their difference in means is statistically significant, i.e. the observed means is not due to chance and there is a real difference between the groups. However, the significance level between 3BR and 4BR is greater than 0.05 (*p* = 0.116) indicating that the difference in means is not statistically significant and the observed means between these groups are not statistically different.



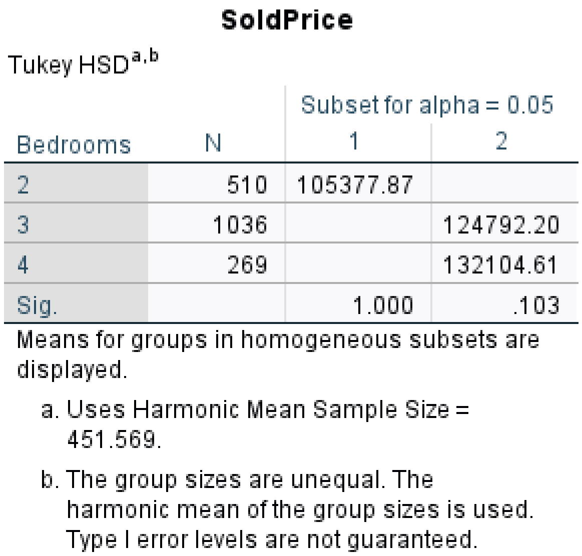
Multiple Comparisons

Dependent Variable: SoldPrice
Tukey HSD

					95% Confidence Interval	
(I) Bedrooms	(J) Bedrooms	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
2	3	-19414.325*	2913.180	<.001	-26247.57	-12581.08
	4	-26726.737*	4058.214	<.001	-36245.81	-17207.66
3	2	19414.325*	2913.180	<.001	12581.08	26247.57
	4	-7312.413	3685.333	.116	-15956.85	1332.02
4	2	26726.737*	4058.214	<.001	17207.66	36245.81
	3	7312.413	3685.333	.116	-1332.02	15956.85

*. The mean difference is significant at the 0.05 level.

The table below shows the two subsets where 2BR is in a group by itself in Subset 1 and 3BR and 4BR are grouped together in Subset 2. Within each subset there is no significant difference between the means. From the subsets, it can be concluded that 3BR has a higher mean than 2BR and 4BR has a higher mean than 2BR. However, it cannot be concluded that 3BR and 4BR have significantly different means based on their significance levels.



SoldPrice

Tukey HSD[a,b]

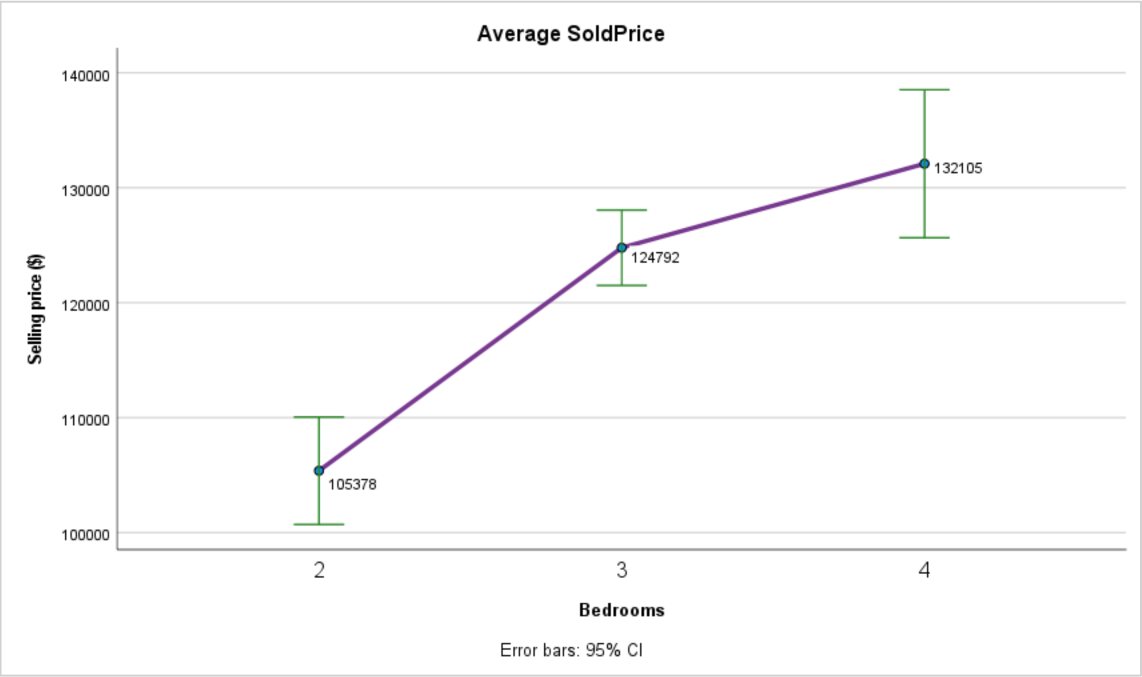
		Subset for alpha = 0.05	
Bedrooms	N	1	2
2	510	105377.87	
3	1036		124792.20
4	269		132104.61
Sig.		1.000	.103

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 451.569.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

The graph below shows how the dependent variable *SoldPrice* changes across the different bedroom sizes with the error bars showing the range which the actual mean is likely to occur for the given confidence level of 95%.



Hypothesis Statement

Two-factor

A comparison of mean *SoldPrice* against two independent variables *Bedrooms* and *TapestryNumber* is done to determine if these groups have an effect on the mean selling price. Does *Bedrooms* and *TapestryNumber* have an effect on mean SoldPrice? Do different combinations of elements from each group affect the mean *SoldPrice*? A hypothesis for these questions can be tested to determine the difference between *SoldPrice* to the independent groups *Bedrooms* and *TapestryNumber*:

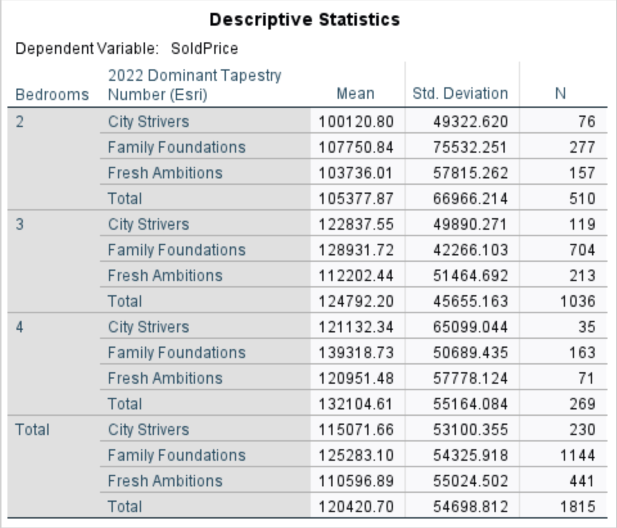
> ***H_0*:** There *is no significant difference* in the mean price across bedrooms sizes and market segments and their interactions.
>
> ***H_a*:** There *is a difference* in the mean price across bedrooms sizes and market segments and their interactions.

The null hypothesis ***H_0*** states that the mean selling price based on both the number of bedrooms and market segment, and the interaction between the two groups, has no significant difference. The alternative hypothesis ***H_a*** states that the mean selling price based on both the number of bedrooms and market segment, and the interaction between the two groups, for at least one of the interactions is different.

Factorial ANOVA

Number of bedrooms (*Bedrooms*) and Market Segments (*TapestryNumber*)

In the factorial ANOVA, *Bedrooms* and *TapestryNumber* are used as factors with *SoldPrice* as the dependent variable. A significance level of 0.05 is assumed. Descriptive statistics show the mean *SoldPrice* for the different interactions between *Bedrooms* and *TapestryNumber*. For the interaction between 2BR and CS, FF, and FA, it has N=510 entries with mean selling price *μ* for each interaction where $\mu_{2BR/CS}$ = $100,121, $\mu_{2BR/FF}$ = $107,751, and $\mu_{2BR/FA}$ = $103,736.

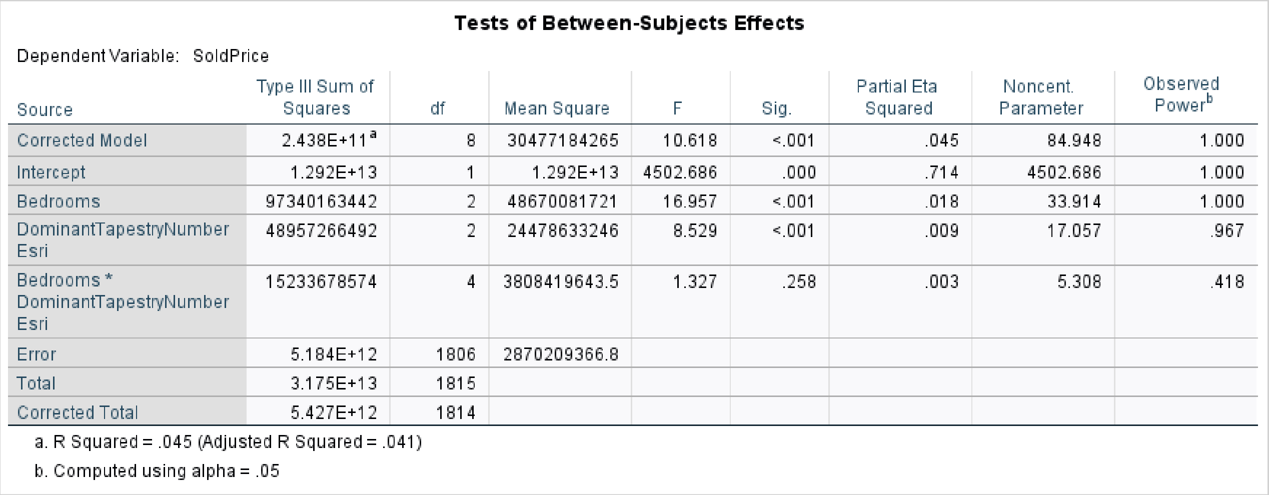


Descriptive Statistics

Dependent Variable: SoldPrice

Bedrooms	2022 Dominant Tapestry Number (Esri)	Mean	Std. Deviation	N
2	City Strivers	100120.80	49322.620	76
	Family Foundations	107750.84	75532.251	277
	Fresh Ambitions	103736.01	57815.262	157
	Total	105377.87	66966.214	510
3	City Strivers	122837.55	49890.271	119
	Family Foundations	128931.72	42266.103	704
	Fresh Ambitions	112202.44	51464.692	213
	Total	124792.20	45655.163	1036
4	City Strivers	121132.34	65099.044	35
	Family Foundations	139318.73	50689.435	163
	Fresh Ambitions	120951.48	57778.124	71
	Total	132104.61	55164.084	269
Total	City Strivers	115071.66	53100.355	230
	Family Foundations	125283.10	54325.918	1144
	Fresh Ambitions	110596.89	55024.502	441
	Total	120420.70	54698.812	1815

The factorial ANOVA shows an R^2=0.045 which means that 4.5% of the variance in the dependent variable *SoldPrice* can be explained by the independent variables *Bedrooms* and *TapestryNumber* and has a nearly zero effect size where the effect size “is a quantitative measure of the magnitude of the experimental effect. The larger the effect size the stronger the relationship between two variables (Mcleod, 2019).” Both *Bedrooms* (F = 16.957, p < 0.001) and *TapestryNumber* (F = 8.529, *p* < 0.001) show to have a statistically significant effect on *SoldPrice* suggesting that the difference in means between the two groups are significant. However, the interaction between Bedrooms and TapestryNumber (F = 1.327m p = 0.258) is not statistically significant with a *p*-value greater than 0.05 suggesting that their interactions do not have an effect on *SoldPrice* and the null hypothesis can be rejected in favor of the alternative hypothesis.



Tests of Between-Subjects Effects

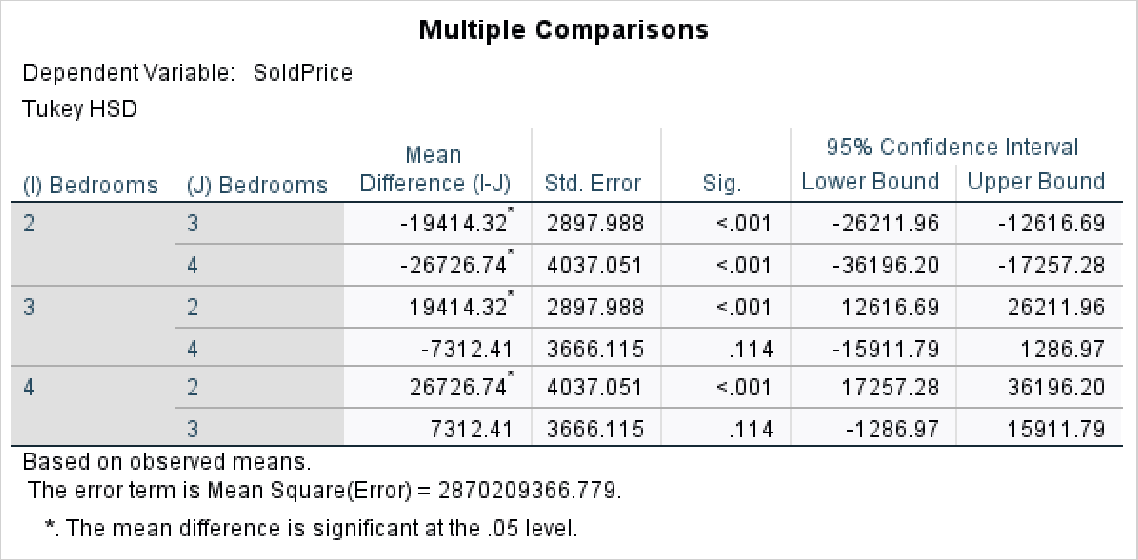
Dependent Variable: SoldPrice

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	2.438E+11[a]	8	30477184265	10.618	<.001	.045	84.948	1.000
Intercept	1.292E+13	1	1.292E+13	4502.686	.000	.714	4502.686	1.000
Bedrooms	97340163442	2	48670081721	16.957	<.001	.018	33.914	1.000
DominantTapestryNumber Esri	48957266492	2	24478633246	8.529	<.001	.009	17.057	.967
Bedrooms * DominantTapestryNumber Esri	15233678574	4	3808419643.5	1.327	.258	.003	5.308	.418
Error	5.184E+12	1806	2870209366.8					
Total	3.175E+13	1815						
Corrected Total	5.427E+12	1814						

a. R Squared = .045 (Adjusted R Squared = .041)

b. Computed using alpha = .05

A comparison of the groups using a Tukey Post-Hoc analysis shows the relationship of means between the groups. For *Bedrooms*, the comparison is similar to the single-factor ANOVA with 2BR having a statistically significant difference in means with both 3BR and 4BR with *p* < 0.001 for each, and a difference in means that is not statistically significant between 3BR and 4BR with *p* = 0.114 which is greater than the assumed level of significance 0.05.



Multiple Comparisons

Dependent Variable: SoldPrice

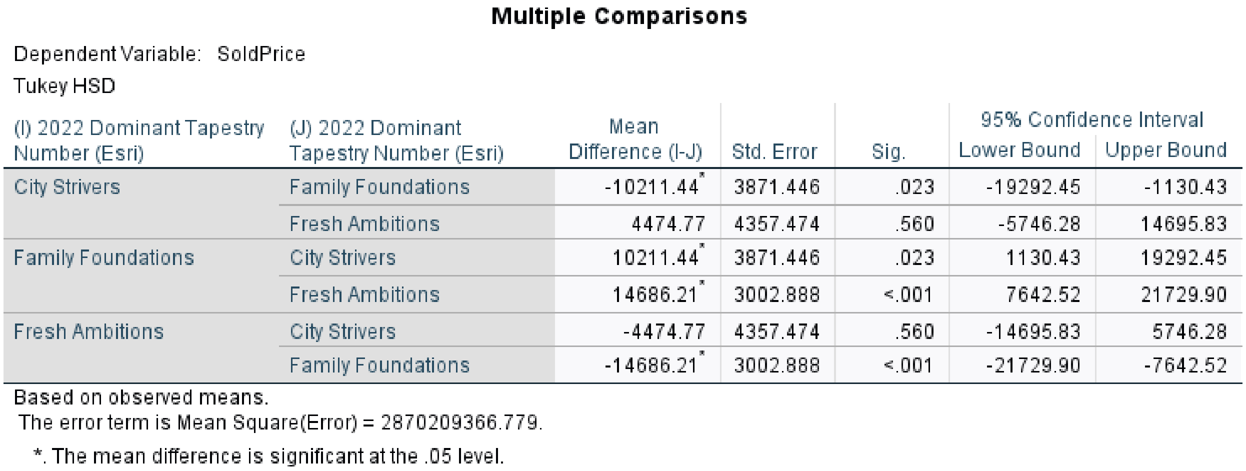
Tukey HSD

(I) Bedrooms	(J) Bedrooms	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
2	3	-19414.32*	2897.988	<.001	-26211.96	-12616.69
	4	-26726.74*	4037.051	<.001	-36196.20	-17257.28
3	2	19414.32*	2897.988	<.001	12616.69	26211.96
	4	-7312.41	3666.115	.114	-15911.79	1286.97
4	2	26726.74*	4037.051	<.001	17257.28	36196.20
	3	7312.41	3666.115	.114	-1286.97	15911.79

Based on observed means.
The error term is Mean Square(Error) = 2870209366.779.

*. The mean difference is significant at the .05 level.

For TapestryNumber, the comparison shows that the difference in means between CS and FF (*p* = 0.023) and FF and FA (*p* < 0.001) are statistically significant with *p*-values less than 0.05. However, the difference in means between CS and FA is not statistically significant with p = 0.560 which is greater than the significance level 0.05.



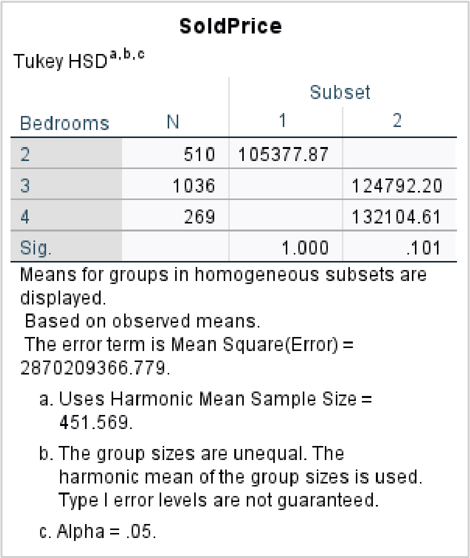
Multiple Comparisons

Dependent Variable: SoldPrice
Tukey HSD

(I) 2022 Dominant Tapestry Number (Esri)	(J) 2022 Dominant Tapestry Number (Esri)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
City Strivers	Family Foundations	-10211.44*	3871.446	.023	-19292.45	-1130.43
	Fresh Ambitions	4474.77	4357.474	.560	-5746.28	14695.83
Family Foundations	City Strivers	10211.44*	3871.446	.023	1130.43	19292.45
	Fresh Ambitions	14686.21*	3002.888	<.001	7642.52	21729.90
Fresh Ambitions	City Strivers	-4474.77	4357.474	.560	-14695.83	5746.28
	Family Foundations	-14686.21*	3002.888	<.001	-21729.90	-7642.52

Based on observed means.
The error term is Mean Square(Error) = 2870209366.779.
*. The mean difference is significant at the .05 level.

The table below shows the subsets for each group. The subsets for *SoldPrice* is the same as in the single-factor ANOVA. The subsets for *TapestryNumber* show CS and FA grouped together in Subset 1 and FF in a group by itself in Subset 2. Within each subset there is no significant difference between the means. From the *TapestryNumber* subsets, it can be concluded that FA has a lower *SoldPrice* mean than FF and CS has a lower *SoldPrice* mean than FF. However, it cannot be concluded that FA and CS have significantly different means based on their significance levels.



SoldPrice

Tukey HSD[a,b,c]

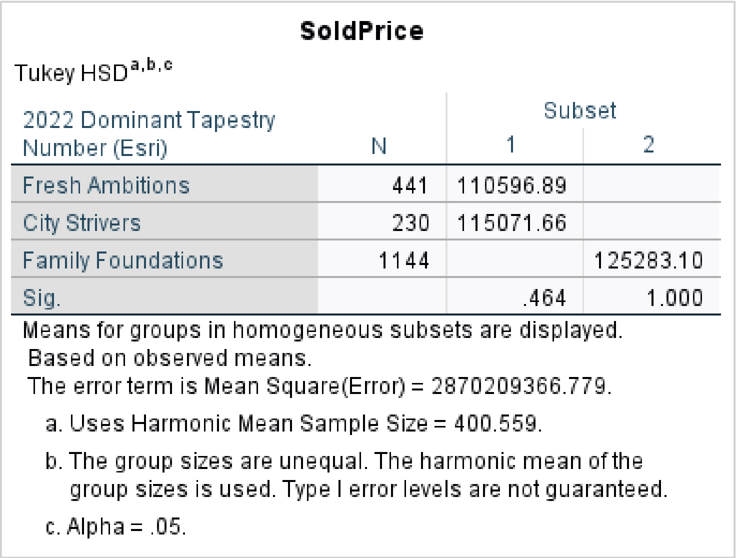
Bedrooms	N	Subset	
		1	2
2	510	105377.87	
3	1036		124792.20
4	269		132104.61
Sig.		1.000	.101

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 2870209366.779.

a. Uses Harmonic Mean Sample Size = 451.569.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.



SoldPrice

Tukey HSD[a,b,c]

2022 Dominant Tapestry Number (Esri)	N	Subset	
		1	2
Fresh Ambitions	441	110596.89	
City Strivers	230	115071.66	
Family Foundations	1144		125283.10
Sig.		.464	1.000

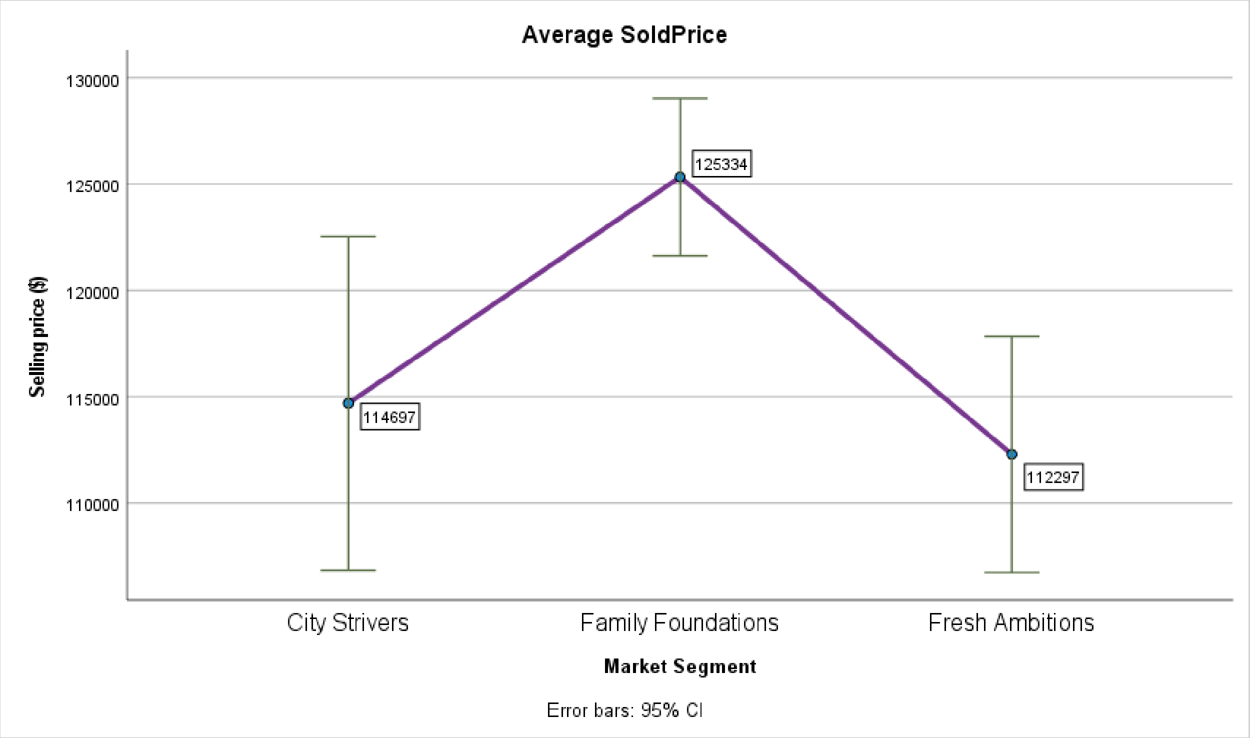
Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 2870209366.779.

a. Uses Harmonic Mean Sample Size = 400.559.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.

The graph below shows how the dependent variable *SoldPrice* changes across the different market segments with the error bars showing the range which the actual mean is likely to occur for the given confidence level of 95%. (See previous chart for *Bedrooms*.)



Conclusions

Based on the ANOVA comparisons, it can be concluded that bedroom size and market segment as individual factors have a significant impact on home selling prices, but their interaction does not. 3BR and 4BR tend to have a higher selling price compared to 2BR homes. However, the difference in mean prices between 3BR and 4BR homes is not statistically significant suggesting that increasing from 3BR to 4BR does not significantly impact the price of the home. Market segment also impacts home selling prices with both FA and CS selling homes with lower prices than FF. The interaction between the number of bedrooms and market segment were shown to not be statistically significant, which means the effect of bedroom size on selling price remains consistent across market segments.

References

Esri Tapestry Segmentation—Esri Demographics Regional Data | Documentation. (2022). Doc.arcgis.com. https://doc.arcgis.com/en/esri-demographics/latest/regional-data/tapestry-segmentation.htm

Mcleod, S. (2019, July 10). *Sampling Distribution*. Simplypsychology.org; Simply Psychology. https://www.simplypsychology.org/effect-size.html

TAPESTRY SEGMENTATION The Fabric of America's Neighborhoods TM UNITED STATES OF AMERICA. (n.d.). https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/fliers/pdfs/tapestry_segmentation.pdf