Analysis of Pierce County House Sales

Henry Shaw & Saransh Rakshak

Due: September 24rd, 2024

Part 1:

Can sales price be predicted? Choose any three variables other than sale price. For each variable, explore its association with sale price using plots, tables, and summary statistics. Include at least one plot or table for each variable. Report appropriate statistics for each variable. Summarize your findings in 2-3 sentences for each variable.

In your analysis, pay attention to outliers, missing values, and excessive zeros. In both questions, include in the text a short description of any of these features and describe the impact that they may have on your analysis.

Correlation of Variables with Sale Price

	Sale Price	House Square Feet	Square Feet (in Thousands)	Year Built	Decade Built	Bedrooms
Correlation	1	0.5263257	0.5260794	0.2288812	0.2276866	0.2054946

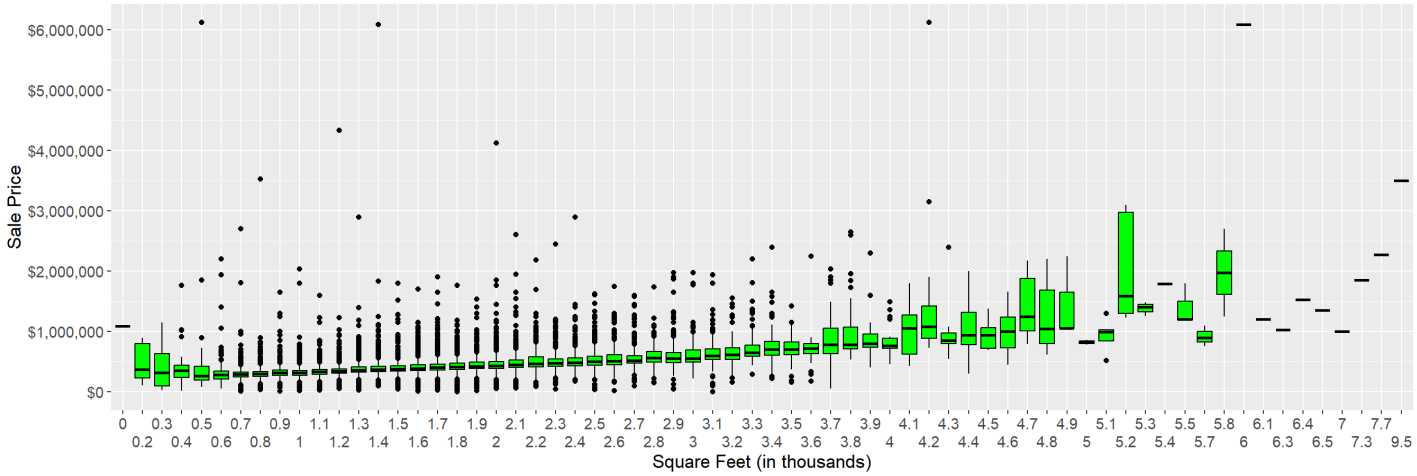
- Above is a breakdown of five variables within the Pierce County House Sales dataset provided. For Part 1 of this assignment, we decided to choose House Square Feet, Year Built, and Bedrooms for our three fields to review.
- Initially, we ran a correlation analysis to see which variables have stronger relationships with sales price. This correlation analysis showed that the three fields selected have a positive correlation to sales price. House Square Feet has a strong correlation, Year Built is moderate, while Bedrooms is somewhat weaker.

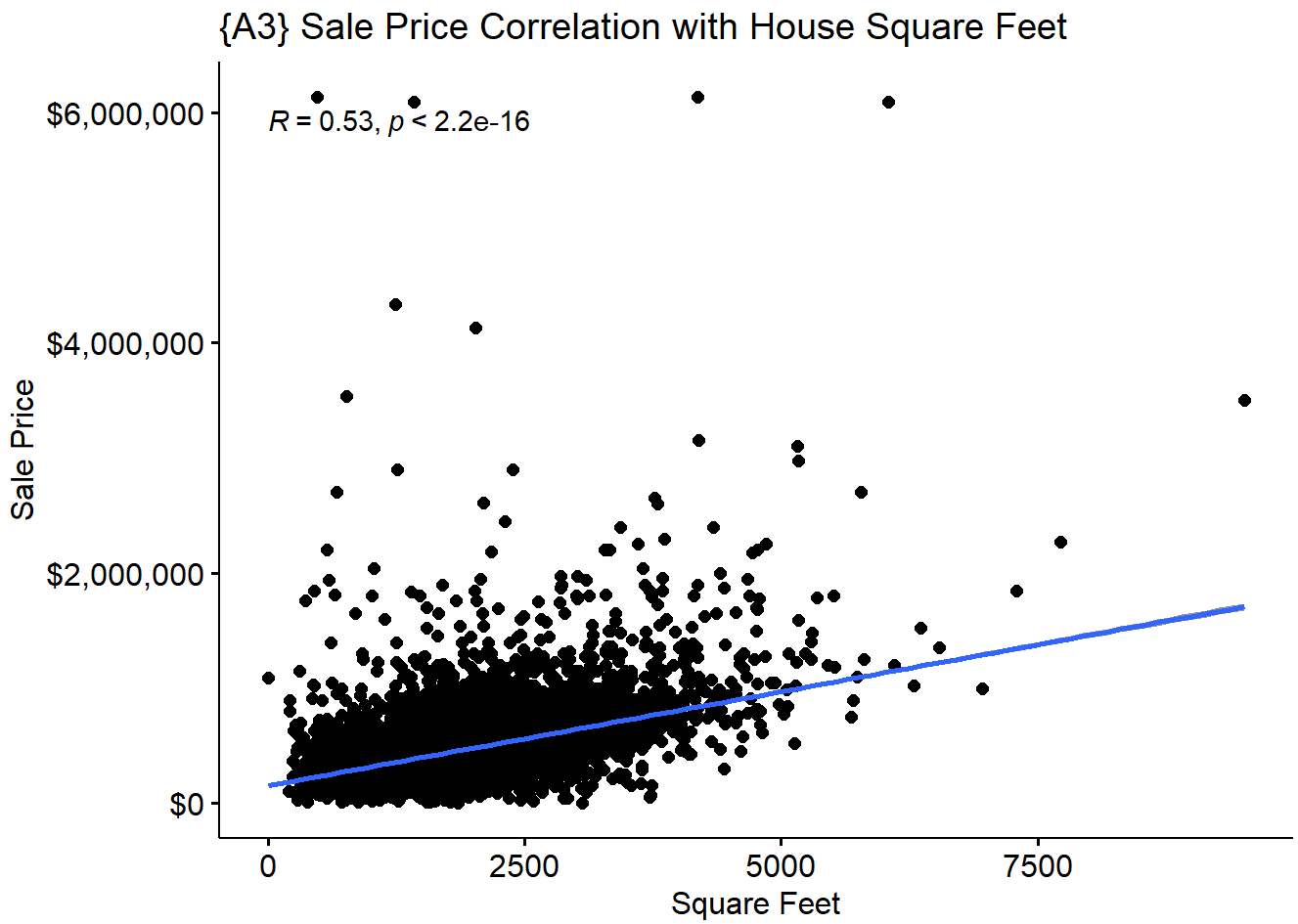
A) Sale Price and House Square Footage

- First, we will analyze House Square Footage, the variable with the highest correlation (~0.53) to Sale Price

{A1} Mean Sales Price by House Square Feet

{A2} Sale Price Distribution by Square Footage





{A4} Summary Statistics for Sale Price by House Square Feet

House Square Feet (in thousands)	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
0.0	1	1086000	1086000	NA	1086000	1086000
0.2	5	480300	372000	349676	105000	895000
0.3	17	393823	315000	311532	25000	1150000
0.4	35	399582	347000	333979	12510	1765000
0.5	69	411232	265000	742224	75350	6130000
0.6	114	347468	282500	297572	50000	2200000
0.7	226	313421	285500	224654	12000	2700000
0.8	424	305954	297000	189200	30000	3531723
0.9	554	323124	313000	130243	10000	1650000
1.0	720	322412	318500	131412	20000	2038200
1.1	710	337376	330000	119073	27750	1600000
1.2	755	360972	345000	230332	48833	4334100
1.3	836	378332	360000	149849	22000	2900000

House Square Feet (in thousands)	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
1.4	925	392305	369000	229206	96065	6090000
1.5	938	390448	374997	136234	47503	1800000
1.6	933	410850	388000	133812	10000	1700000
1.7	932	430070	405000	149483	25000	1899995
1.8	910	436527	414500	127412	2000	1759990
1.9	833	448967	420000	132613	15000	1538750
2.0	713	466131	430000	206985	55000	4126500
2.1	690	489585	449950	180012	65000	2609788
2.2	653	507958	466500	180069	84613	2185000
2.3	607	504857	469900	162051	42086	2450000
2.4	545	519638	480000	176226	178000	2900000
2.5	510	533791	495990	177519	31500	1625000
2.6	461	550230	505000	176328	17000	1750000
2.7	427	560158	517379	168277	95000	1600000
2.8	341	589477	561219	166140	146761	1740000
2.9	316	602948	550500	223446	46500	1975000
3.0	293	602096	550000	200510	220000	1975000
3.1	324	636600	599250	195485	2000	1935500
3.2	219	651627	615000	185217	159247	1550000
3.3	192	719289	649000	255694	290500	2200000
3.4	119	742626	699950	283682	218750	2395000
3.5	113	724924	699950	197724	154151	1417900
3.6	69	719300	720000	232640	173338	2250000
3.7	42	887587	780000	454238	50412	2038200
3.8	52	975631	779619	463477	535000	2650000
3.9	29	884198	799540	338792	400000	2300000
4.0	29	821744	764995	235993	460000	1490000
4.1	25	1002221	1055000	387213	430000	1800000
4.2	16	1537062	1075000	1365916	725000	6130000
4.3	9	1015555	850000	540898	540000	2400000

House Square Feet (in thousands)	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
4.4	11	1061647	935000	549739	300000	2000000
4.5	5	955570	935000	279702	689900	1375000
4.6	11	1004363	998000	361371	450000	1662500
4.7	7	1424714	1250000	544931	785000	2173000
4.8	13	1206806	1042500	514426	615000	2200000
4.9	3	1450000	1050000	692820	1050000	2250000
5.0	2	819000	819000	59396	777000	861000
5.1	5	935599	988000	284805	520000	1299999
5.2	5	2038400	1592000	923355	1225000	3100000
5.3	3	1380333	1403000	117649	1253000	1485000
5.4	1	1790000	1790000	NA	1790000	1790000
5.5	3	1394029	1200000	351694	1182089	1800000
5.7	3	915000	895000	175855	750000	1100000
5.8	2	1975000	1975000	1025304	1250000	2700000
6.0	1	6090000	6090000	NA	6090000	6090000
6.1	1	1200000	1200000	NA	1200000	1200000
6.3	1	1025000	1025000	NA	1025000	1025000
6.4	1	1525000	1525000	NA	1525000	1525000
6.5	1	1350000	1350000	NA	1350000	1350000
7.0	1	1000000	1000000	NA	1000000	1000000
7.3	1	1850000	1850000	NA	1850000	1850000
7.7	1	2273000	2273000	NA	2273000	2273000
9.5	1	3500000	3500000	NA	3500000	3500000

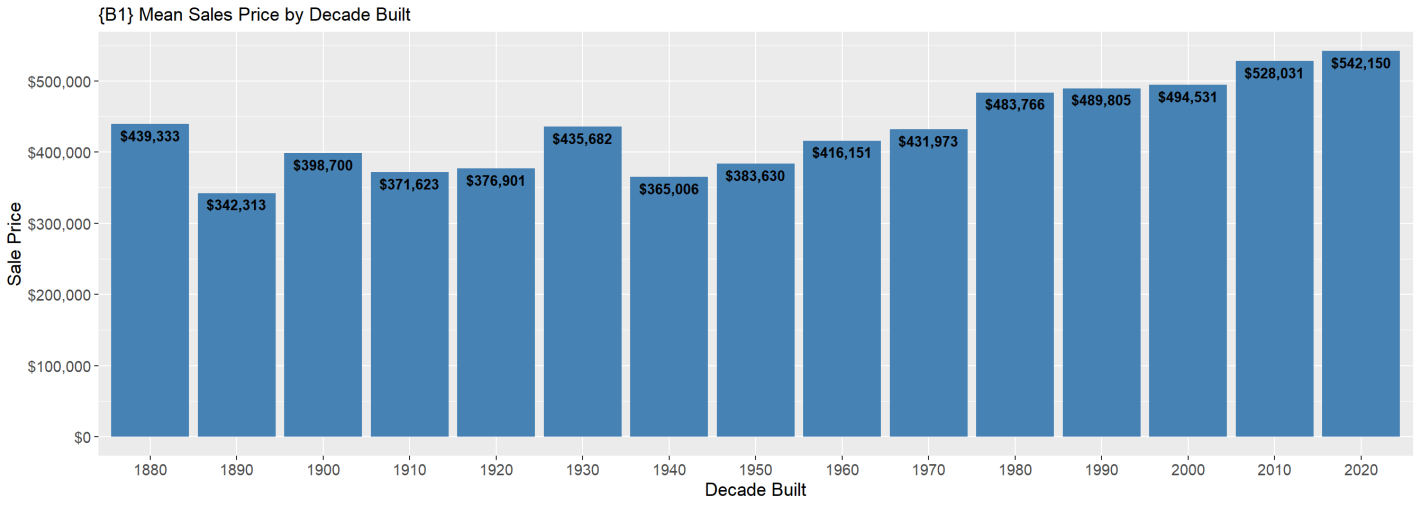
- From our graph {A3}, we can see that sale price has a moderately high correlation to house square feet (Pearson Correlation ~ 0.53). We can also observe from our graph in {A2} that sale price does not have a large range of distribution for homes with 0.6 thousand to 4.1 thousand square feet. This is because there is a greater count of instances (house sales) within that range. As a result, we can conclude that house square feet does have a general association with sale price, with sale price growing as house square feet increases, as reflected in {A3}.
- Outliers can be seen in the mean distribution in Graph {A1}, and their specific values can be found using Table {A4} at 6.0 thousand square feet. This outlier is based on a single point (i.e., a single house sale) whose mean sale price of $6,090,000 is far greater than those with slightly greater or smaller house square

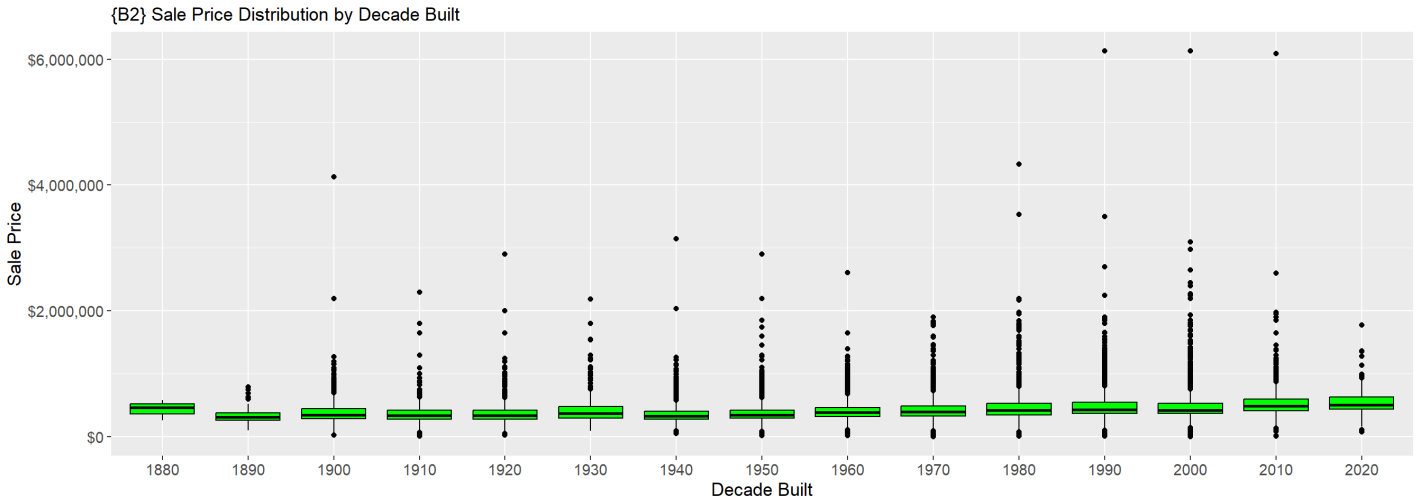
feet. The presence of this outlier reduces our Pearson Correlation Coefficient and thus weakens the association between house square feet and sale price.

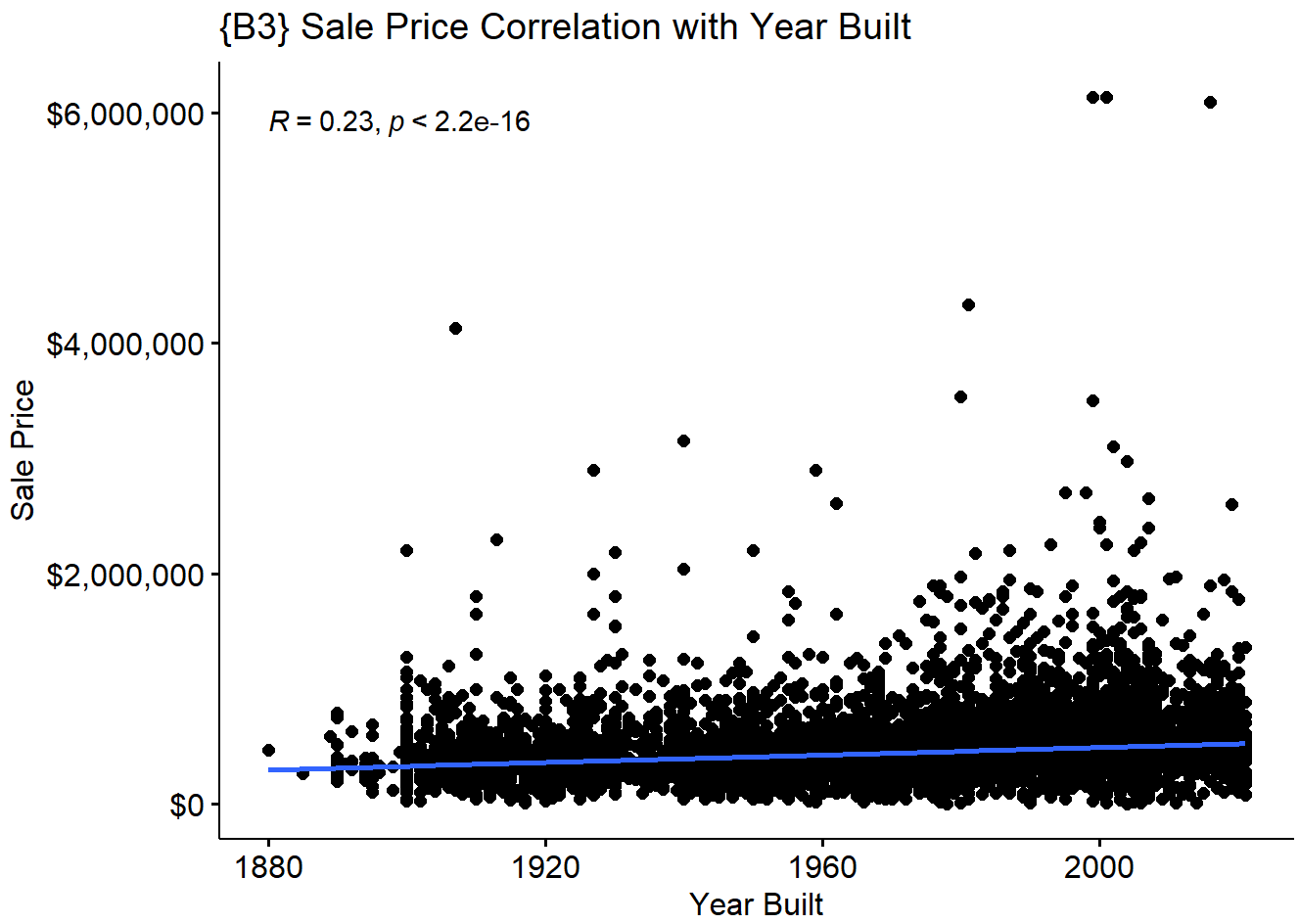
- Other high-value sales over $6 million, such as those in the 0.5, 1.4, and 4.2 thousand square foot categories, can also be considered outliers. However, unlike the outlier at 6.0 thousand square feet, these outliers have multiple other values in their bin, so they do not greatly affect our overall analysis between house square feet and sale price. The exact reasoning for these outliers cannot be deduced from the information at hand. However, we can speculate that these sales may have involved special circumstances, such as the house being located in a high-value area, being furnished with modern amenities, or some other factor that significantly increased their value.
- The exact reasoning for any of these outliers cannot be deducted from the information at hand, however, we can predict these sales were special situations where the house may be located geographically in a high value area or the house could be furnished with modern amenities, or some other reason that increases their value far greater than expected.

B) Sale Price and Year Built

- Next, we will view the association between the Year the house was Built and Sale Price.







{B4} Summary Statistics for Sale Price by Decade Built

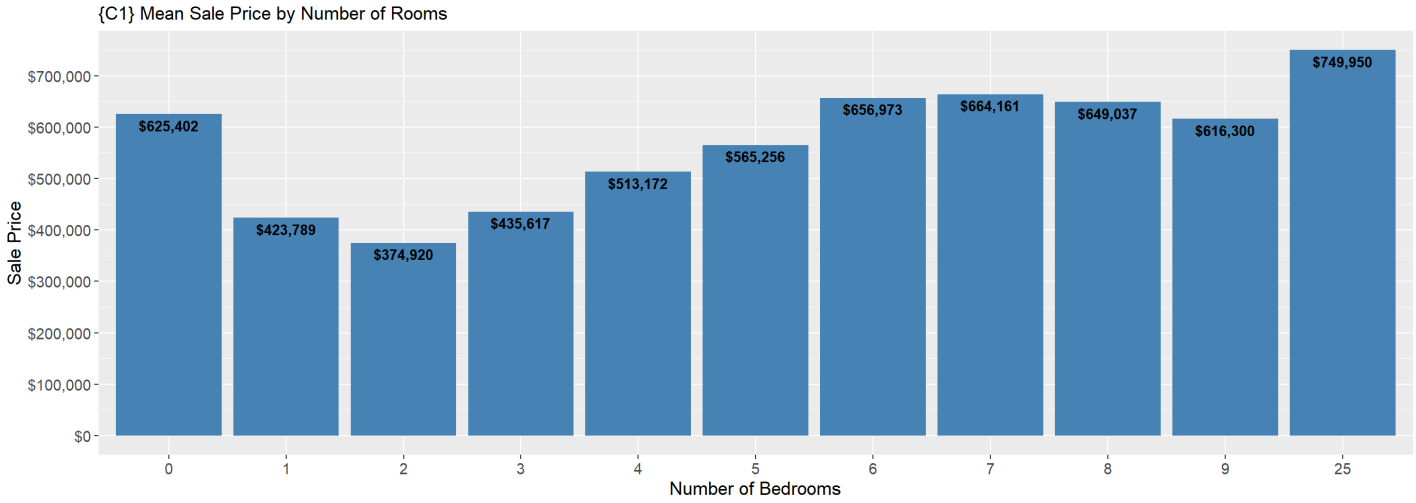
Decade Built	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
1880	3	439333	468000	161914	265000	585000
1890	56	342313	315625	140248	101951	795000
1900	552	398700	349475	251145	25000	4126500
1910	556	371623	335000	192131	12000	2300000
1920	798	376901	340000	200241	30000	2900000
1930	302	435682	369000	251970	92500	2185000
1940	965	365006	330000	194117	50000	3150000
1950	1056	383630	350000	198743	20000	2900000
1960	1199	416151	385000	177965	22000	2609788
1970	1509	431973	399800	190875	2000	1900000
1980	1264	483766	419725	295780	10000	4334100
1990	2228	489805	431150	249578	12752	6130000
2000	2776	494531	425000	262619	2000	6130000

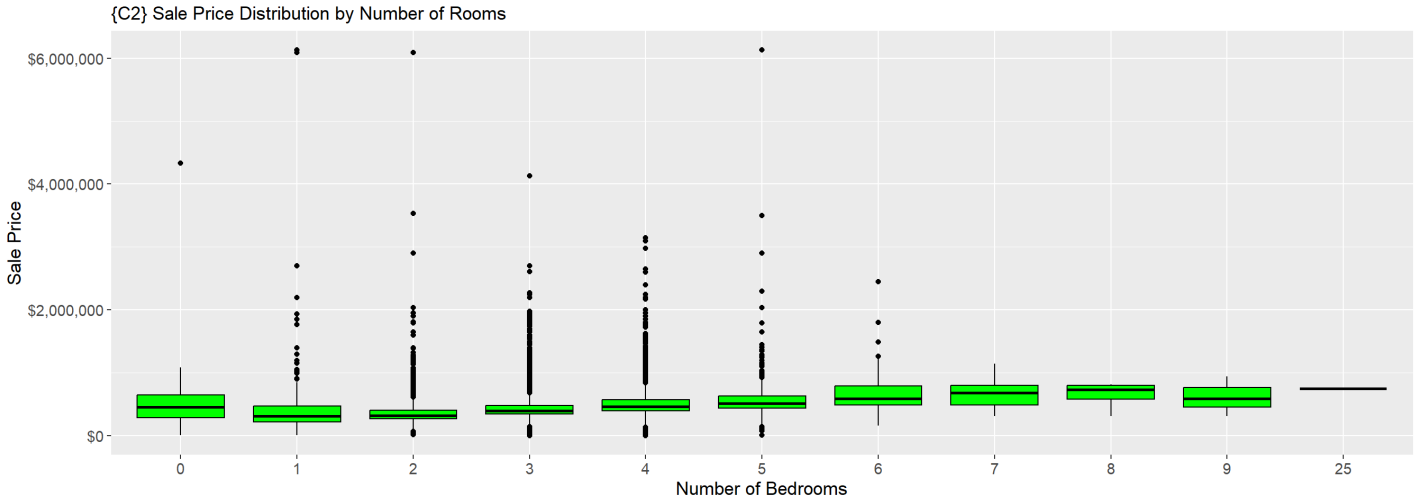
Decade Built	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
2010	2319	528031	490000	237335	10000	6090000
2020	1231	542150	508620	159584	80000	1775000

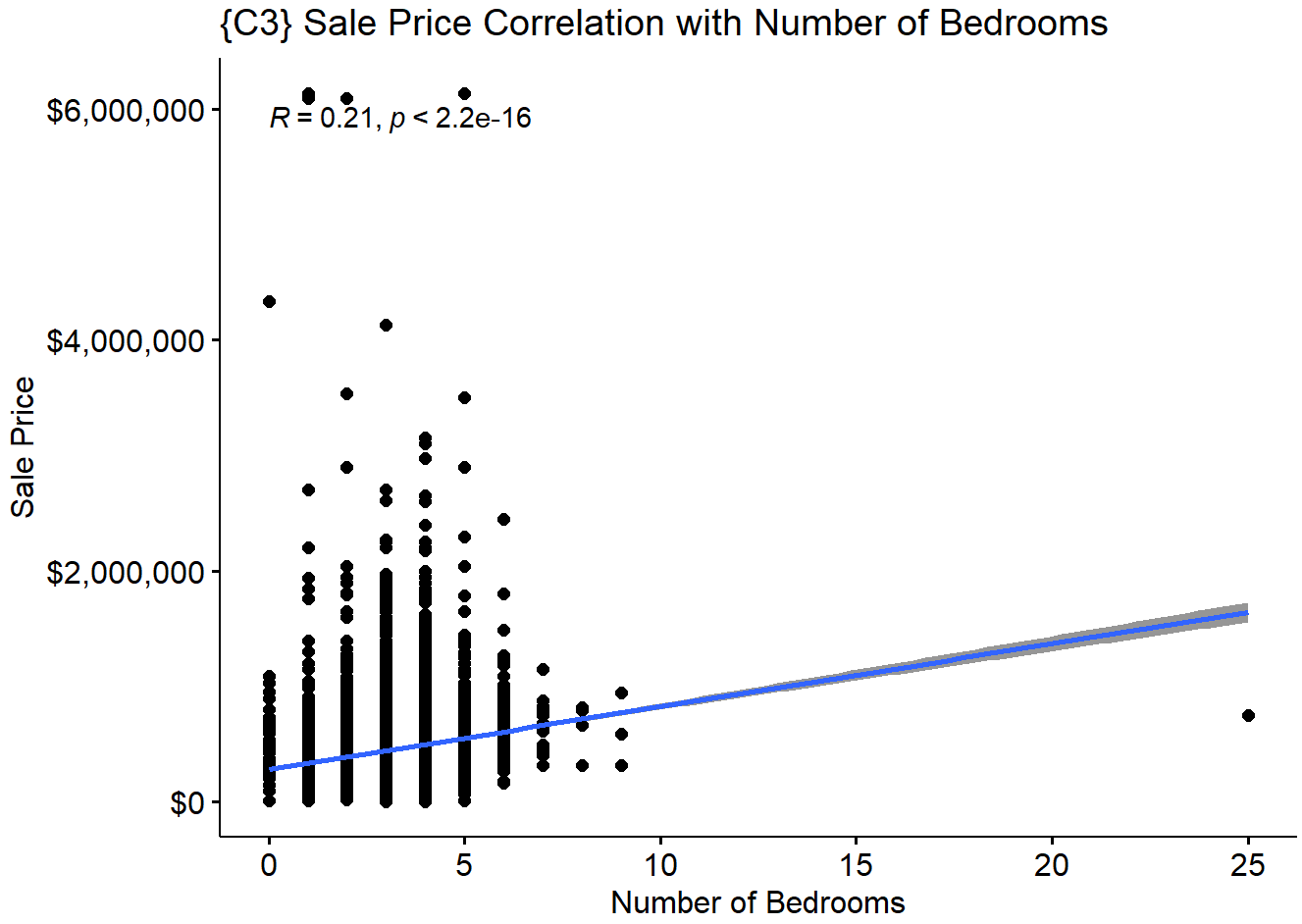
- From our graph in {B3}, we can conclude that the year built has a slight association with sale price based on its Pearson Correlation (R = 0.23). We can also observe that houses built in the 1880s and 1930s have a comparably higher mean sale price than other houses built between 1880 and 1980. Thus, we can infer that there is something unique about houses built in the 1880s and 1930s that has allowed them to retain a higher value in today's market.
- Outliers can also be identified from our graphs. From the distribution graph {B2}, we see that the decades 1990, 2000, and 2010 each have a single house sale price of over 6 million-much higher than the mean sale price for those decades. These individual house sales may be special cases, such as custom-built homes or houses located in high-value areas. Since the bins for the 1990s, 2000s, and 2010s contain a significant number of data points, these outliers do not significantly impact our overall analysis.
- Additionally, outliers can be seen in graph {B1}, where the sale price for houses built in the 1880s is unusually high. However, table {B4} shows that this decade has only one data point, likely a special case influenced by factors such as location or historical value. As a result, this outlier can be disregarded.

C) Sale Price and Number of Bedrooms

- Lastly, we will view the association between Number of Bedrooms and Sale Price







{C4} Summary Statistics for Sale Price by Number of Bedrooms

Bedrooms	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
0	47	625402	459000	826819	12510	4334100
1	337	423789	312500	537333	10000	6130000
2	1900	374920	324995	255351	22000	6090000
3	8504	435617	398000	189366	2000	4126500

Bedrooms	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
4	4858	513172	465000	217712	2000	3150000
5	1028	565256	512922	291257	12752	6130000
6	113	656973	587450	303596	162000	2450000
7	19	664161	679000	204336	316000	1147500
8	4	649037	730000	231370	317150	819000
9	3	616300	590000	315273	315000	943900
25	1	749950	749950	NA	749950	749950

- The summary statistics {C4} indicate that as the number of bedrooms increases, the mean and median sale prices generally rise. The correlation between the number of bedrooms and sale price, while statistically significant (Pearson Correlation Coefficient = 0.21), is weak. This suggests that while the number of bedrooms has some effect on sale price, other factors (such as location or property type) likely play a more influential role. Properties with 0 bedrooms have a relatively high mean sale price of $652,402, likely representing special types of properties that distort the expected trend.
- The housing data point with 25 bedrooms is an outlier due to its unusually high number of rooms and its singular occurrence in the dataset. Additionally, outliers in sale price, as shown in graph {C2}, reveal some prices exceeding 6 million across various bedroom counts, particularly in the 1-5 bedroom range. These extreme values may skew the data. Moreover, the high prices in the 0-bedroom category suggest possible data issues, such as excessive zeros or missing values, which could affect the overall analysis.

Can Sale Price be predicted?

Based on the data from graphs A - C, sale price can be predicted to a certain extent using variables such as house square footage, year built, and number of bedrooms.

```
House Square Footage: There is a moderate correlation (Pearson correlation ~0.53) between house
square footage and sale price. As square footage increases, the sale price generally rises, alth
ough there are outliers, particularly at higher square footage values-that slightly reduce this
correlation.

Year Built: The correlation between year built and sale price is weaker (Pearson correlation ~0.
23). However, certain years (such as homes built in the 1880s and 1930s) tend to have higher sal
e prices compared to other periods. This suggests that while newer homes may have some impact, s
pecific historical periods or architectural styles may also influence prices.

Number of Bedrooms: There is a weak correlation between the number of bedrooms and sale price (P
earson correlation ~0.21). While more bedrooms generally correlate with a higher price, there ar
e notable outliers, such as homes with 0 or a very high number of bedrooms, which distort the tr
end.
```

Thus, while it is possible to predict sale price based on these factors, outliers and additional variables (e.g., location, historical significance, or amenities) likely play a significant role in the final sale price.

Part 2:

Is missingness informative? Create a binary variable that indicates whether the view quality is missing. Missing values in this data file are represented as an empty character string; that is, ``. Choose three variables. (They may, but do not need to be, the same variables from part 1.) Explore the association between these variables and the binary variable that you created. Follow the same instructions as in Part 1.

Correlation of Variables for Has View Quality

	Has View Quality	Sale Price	House Square Feet	Has Waterfront
Correlation	1	0.2391971	0.1353988	-0.0340727

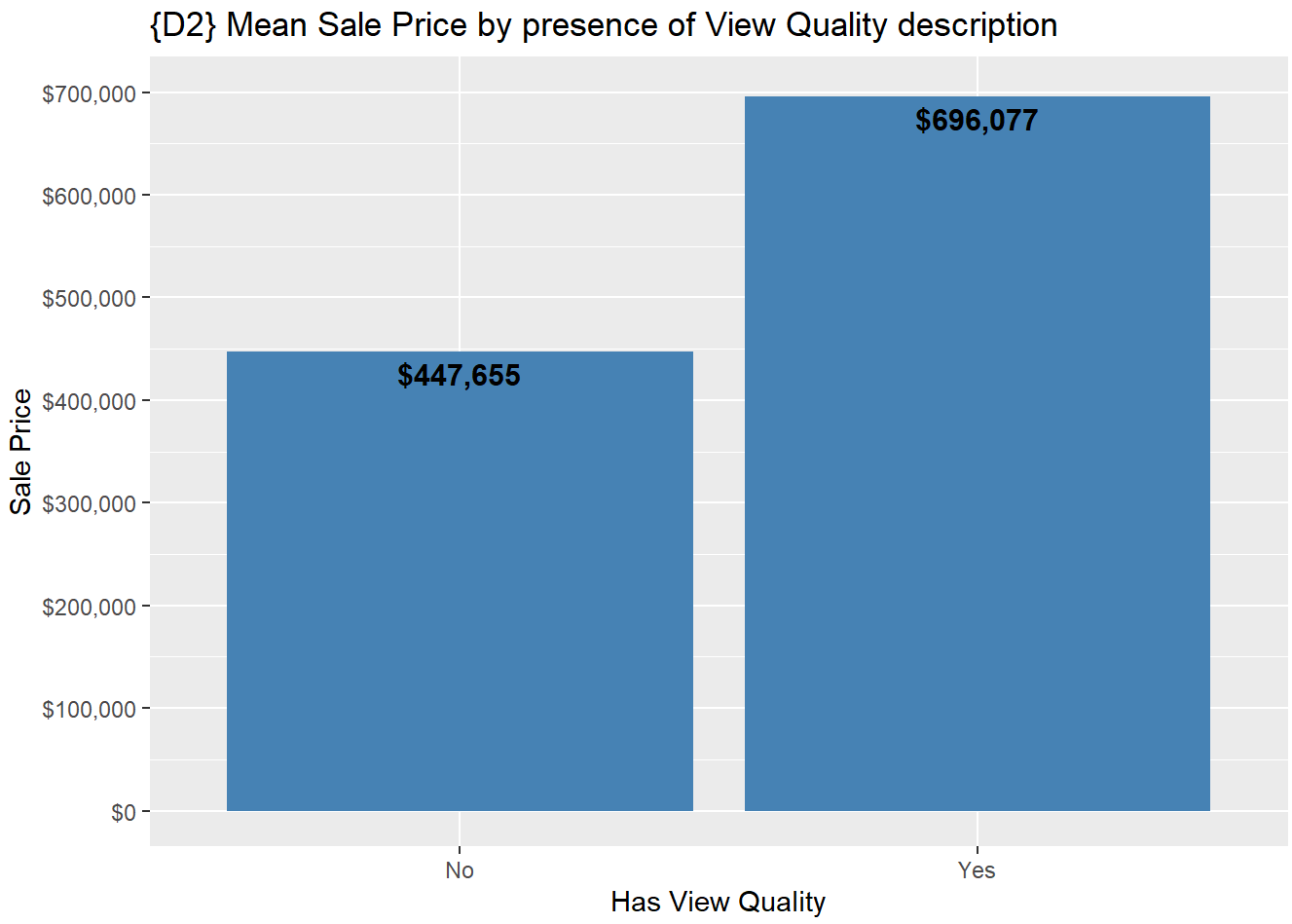
- The table above presents a breakdown of three variables from the Pierce County House Sales dataset: Sale Price, House Square Feet, and Has Waterfront, in relation to the indicator for View Quality.
- In our analysis, we conducted a correlation assessment to evaluate the relationships between these variables and View Quality. This indicates that Sale Price and House Square Feet have moderate positive relationships with View Quality, while the Has Waterfront shows a weak negative correlation.

D) Presence of View Quality and Sale Price

- First, we will analyze Sale Price, the variable with the highest correlation (~0.24) to View Quality

{D1} Summary Statistics for Has View Quality by Sale Price

Has View Quality	Count	Mean Sale Price	Median Sale Price	SD of Sale Price	Min Sale Price	Max Sale Price
0	15895	447655	410000	225581	2000	6130000
1	919	696077	665000	285186	12752	2600000


{D2} Mean Sale Price by presence of View Quality description
$700,000
$600,000
$500,000
$400,000
$300,000
$200,000
$100,000
$0
Sale Price
$447,655
$696,077
No
Yes
Has View Quality


{D3} Sale Price Distribution by presence of View Quality description
$6,000,000
$4,000,000
$2,000,000
$0
Sale Price
No
Yes
Has View Quality

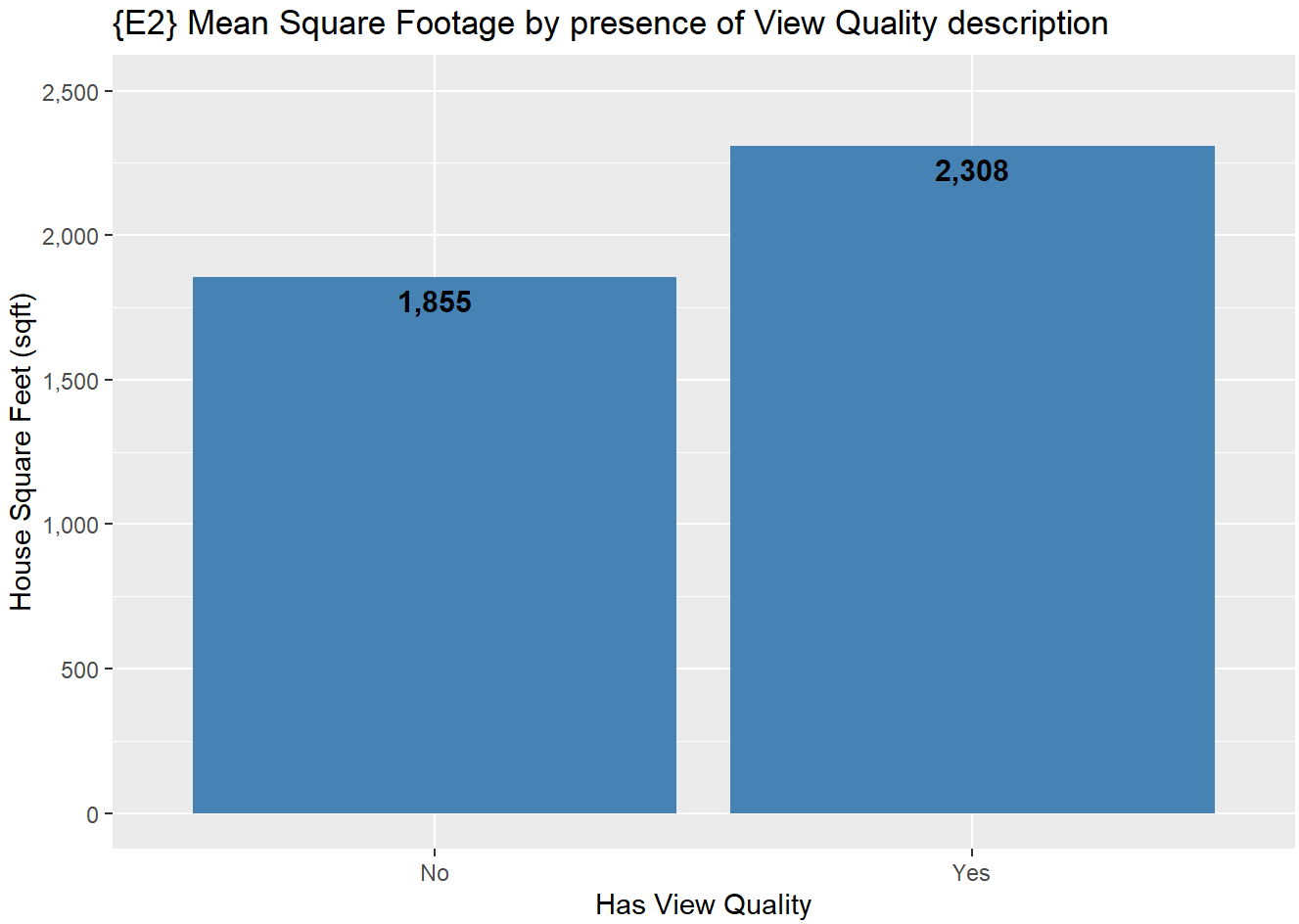
- The summary statistics {D1} show that the mean sale price for homes with a view quality is $696,077, which is 1.55 times greater than the mean sale price of $447,655 for those without. The difference in median sale price is even more substantial, with homes having a view quality achieving a median of $665,000 compared to $410,000 for those without. This suggests that a view quality description is a premium feature that can increase the value of a property, as confirmed by our Pearson Correlation Coefficient, calculated to be 0.2391971.
- Outliers can be seen in graph {D3} for homes without view quality, as some sale prices exceeded $6 million. However, due to the existence of 8,960 other housing sales without view quality, the mean sale price remains at a much lower value and does not skew our overall analysis.

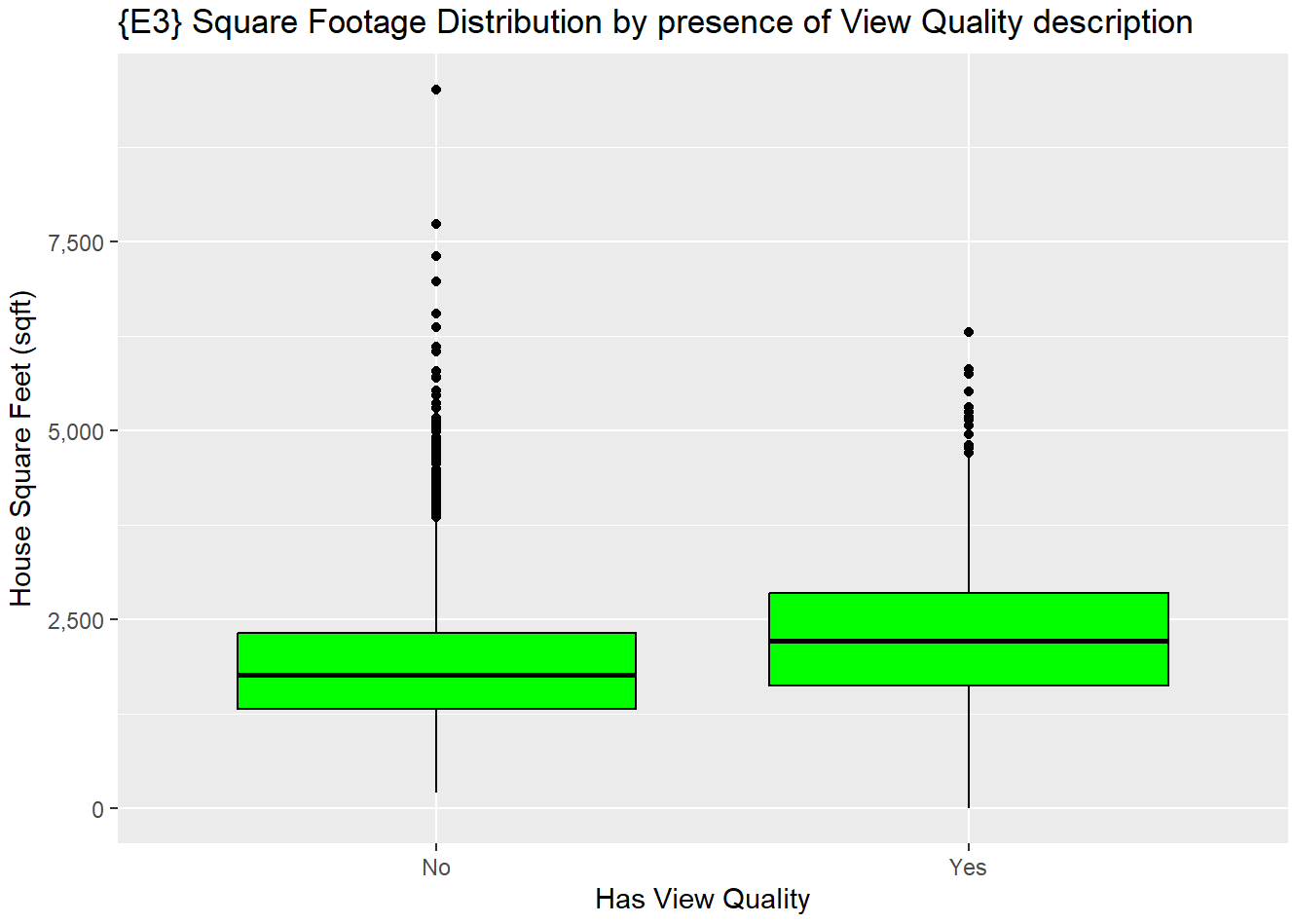
E) Presence of View Quality and House Square Footage

- Next is the analysis of association between House Square Footage and presence of View Quality description

{E1} Summary Statistics for Has View Quality by Square Feet

Has View Quality	Count	Mean Square Feet	Median Square Feet	SD of Square Feet	Min Square Feet	Max Square Feet
0	15895	1855	1756	743	200	9510
1	919	2308	2213	912	1	6295


{E2} Mean Square Footage by presence of View Quality description
2,500
2,000
1,500
1,000
500
0
House Square Feet (sqft)
1,855
2,308
No
Yes
Has View Quality


{E3} Square Footage Distribution by presence of View Quality description
7,500
5,000
2,500
0
House Square Feet (sqft)
No
Yes
Has View Quality

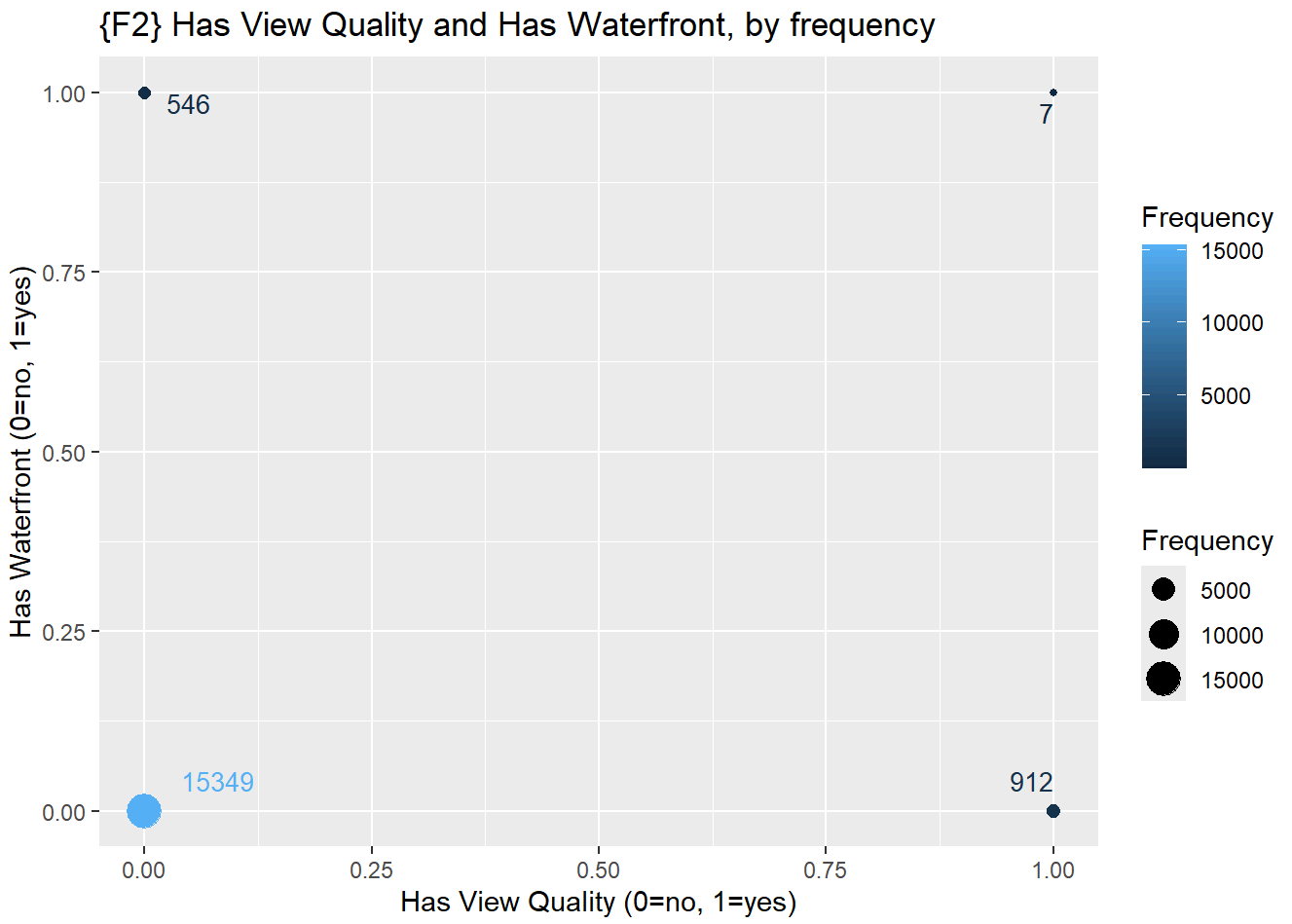
- From our graph {E2}, we can observe a very slight correlation between the presence of view quality and square footage, with homes containing view quality data being approximately 1.24 times larger than those without a view quality description. This is reflected in our Pearson Correlation Coefficient, calculated to be 0.1353988.
- Outliers can be detected in the data. For homes without view quality data, the mean square footage is 1,855 sqft; however, this group also contains properties as large as 9,510 sqft {E1}. Similarly, outliers are present in the group of homes with view quality data, which has a mean size of 2,308 sqft but includes values as low as 1 sqft {E1}. Both of these outliers affect our analysis by lowering the Pearson Correlation Coefficient; however, due to the size of both groups, these outliers are less impactful.

F) Presence of View Quality and Presence of Waterfront

- Lastly, we will view the association between View Quality description and the Presence of a Waterfront

{F1} Cross Table of Frequency for View Quality and existance of a Waterfront

Has View Quality (0=no, 1=yes)	Has Waterfront (0=no, 1=yes)	Frequency
0	0	15349
0	1	546
1	0	912
1	1	7



- The association between the presence of view quality and waterfront is minimal. From table {F1}, we see that the vast majority of homes without waterfront access also lack a view quality designation (15,349 homes). Only a small number of properties have both waterfront and view quality (7 homes). This suggests that while there is some overlap, view quality and waterfront presence are largely independent features. This is further confirmed by our Pearson Correlation Coefficient of -0.0340727.

Is missingness informative?

Missingness in “view_quality” proves to be informative based on data from tables and graphs D - F.

```
Sale Price: Homes with recorded "view_quality" tend to have higher sale prices, with a mean of
$696,077 compared to $447,655 for homes where view quality is missing. This suggests that missin
g "view_quality" is associated with lower-value properties. Therefore, missingness in this featu
re could indicate homes without premium attributes, such as a scenic view.

Square Footage: Properties with "view_quality" data also have a higher average square footage
(2,308 sq ft compared to 1,855 sq ft). This further emphasizes that missing view quality data ma
y correspond to smaller, less expensive homes.

Waterfront: The association between having "view_quality" and being on the waterfront is minima
l, with only seven homes showing both features. Most homes either lack both features or have onl
y one, indicating that these characteristics do not strongly co-occur.
```

Thus, the missingness of “view_quality” data provides valuable information, potentially signaling homes with fewer premium features.