**Final Project**

Jack Galloway

Analysis of Home Cost In Ames, Iowa

It is well known that during different parts of the year, more people tend to purchase homes. Generate a visualization which shows whether there is evidence of this in Ames, Iowa. What should your boss take away from this?

Almost 700 of the 1460 total data points are sold with three months! According to this visualization, we can see that there is a specific season where sales are high. This May-July. My boss should take this into account when they consider when to sell properties that have been renovated and are ready to flip. Since almost 50% of homes are sold at this time, they are likely to be able to sell at a higher price than normal. They should also consider the opposite season (the winter months December-February) as a good time to buy homes at a lower price.

A graph of a bar chart

Description automatically generated

**B.)**

With a deeper understanding of the sales per month, you dig a little deeper. You are curious to know, were certain home styles (**HouseStyle**) selling better per month. Generate a table which shows the average sale price of **each home style** (HouseStyle) **per month**.

A screenshot of a computer

Description automatically generated

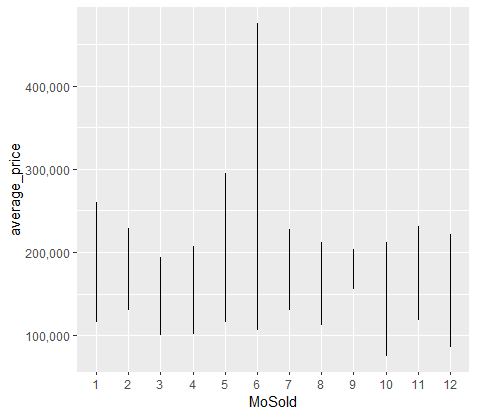
**C.)**

Generate a line plot for the data in **B** using this code:

line\_plot <- ggplot(data = pipeline\_data, mapping = aes(x = MoSold, y = Calculated Column))

line\_plot+geom\_line()

Please describe why the figure is being plotted as this. What is happening? Be clear on why this figure is being plotted this way. What type of variables are being plotted? Why are we not getting separate lines across the months? **Place screenshot below**

****

In this visual, we can see the average prices of all homes sold per month in Ames, Iowa. This data includes a range of average housing price. For example, if we take Month 1, average prices range from $115,610 to $260,279. $115,610 is the average price for a 1.5Fin house style (1 and a half story, 2nd level finished). $260,279 is the average price for the 2Story house style (two story). The lines do not move across months sold because we did not specify grouping by housetype. In this visual, we see the average range for all housetypes sold, rather than grouping by specific type.

**D.)**

Fix the figure above such that we get a clear visualization of your table in **B**. (Hint: the lines alone are not informative). Please describe what you did to improve this figure. Describe what you would want your boss to take away from this figure, what does it tell you about the average sale price of each home style (HouseStyle) per month. Are there any identifiable trends or anomalies? **Place screenshot below**

A graph of a house sale

Description automatically generated with medium confidence

To improve the previous figure, I made sure to group by housestyle in the plot. This created 8 lines, rather than 1, for each month, so that a change in average sale price per housestyle could be examined across the 12 months. I also added a title and changed the x and y axis to make it clearer that we were looking at the variance in average sale price in each home style across months. I changed the line type and color to show the change in average sale price by housestyle. By looking at this graph, it is much easier to see that, for example, in June, house style 2.5Fin had a significantly higher average sale price than any other housestyle in the dataset. Another example, 1.5Unf had a significantly lower sale price in October than any other housestyle in the dataset. Another noticeable trend that can be seen in this graph is that average sale price increases across housestyle during the summer months. An exception or anomaly to this would be the 1.5Fin housestyle, in which the average sale price jumped during February.

**Question 2.)**

**A.)**

Next, you are trying to identify what may be a good predictor of SalePrice. It is known that square footage leads to more space and oftentimes higher sales price. So, you examine what the relationship is between GrLivArea and SalePrice. Generate a figure which shows this relationship. What do you observe? Do there appear to be outliers? What should your boss take away from this plot? **Place screenshot below**

In this visual, we can see a strong positive correlation between GrLivArea (square footage) and SalePrice. This means that as square footage increases, SalePrice increases likewise at an almost consistent rate. There are outliers however: At around 4600 square feet we can see a datapoint with a sale price below $200,000, and at around 5600 square feet we can see a data point with a sale price below $200,000. According to the graph we would have expected these outliers to sell for $700,000+, and we can hypothesize that they sold at a low price because of the impact of a different variable(s). Besides these outliers, we can tell our boss with a high degree of confidence, that they will be able to sell homes with a higher square footage at a higher price. However, most of the data is concentrated between 1000 and 2500 feet, so I would avoid suggesting to the boss to buy houses any larger than that.

A graph with black dots

Description automatically generated

**B.)**

Utilize annotate() to highlight outliers. (Hint: you can have multiple areas highlighted by having two layers of annotate() functions) Place your figure below. Refer to page 124 in the Data Visualization book.

A graph with black dots

Description automatically generated

**C.)**

To help add clarity, you color code your figure in **A** to show the homes which had the largest square footage (GrLivArea >2000) **and** had **excellent** kitchen quality (these are 2 important variables with homes). This could help to give insights into the home purchase prices. **Place the screenshot below.**

A graph with red and blue dots

Description automatically generated

Additionally, generate the same figure in **the previous screenshot** and utilize geom\_text\_repel() to label the homes that had the largest square footage (GrLivArea >2000) **and** had **excellent** kitchen quality with how many car garage the homes have (i.e. GarageCars). **Place screenshot below**

A graph with red and blue dots

Description automatically generated

Please describe what you observe in both plots and discuss which is more effective (and why). What should your boss take away from these plots?

When considering visuals, it is important to remember simplicity. In the first plot in 2C, we already see that in order for the conditions to be met, the home must have large square footage and excellent kitchen quality. I would tell my boss that if he wishes to examine the number of cars that will fit into the garage, that should be done in a separate plot. However, he should note that homes with a large square footage and excellent kitchen quality often sell at a higher price, at an average of $400,000 to $600,000. And while it would be my advice to not consider the garage, it can be noted that these homes also typically have garages that can fit more cars.

EXTRA CREDIT:

You will obtain extra points if you choose your own set of colors manually or with a pre-built set of colors.

**Question 3.)**

**A.)**

Next, you are curious to understand what style of buildings are being sold in Ames, Iowa. Generate a figure which shows the number of each **BldgType** that has been sold. What do you observe? How many TwnhsE and 1Fam have been sold? What should your boss take away from this? **Place screenshot below**

More 1 family homes have been sold than any other type of home. Other home types aren’t even close. TwnhsE or townhouse end unit, is a type of home that sold a little over a hundred units, less than 10% of how many 1 family units were sold. My boss should consider this when flipping houses! The majority of the market is selling 1-family homes. Around 80% of all homes being sold are 1-family homes. This is important to consider when choosing which homes to buy, and how to market the home once it has been renovated and is ready to flip.

TwnhsE- 114

1Fam- 1220

A graph with numbers and a bar

Description automatically generated

**B.)**

Confirm your amounts stated in **A** by developing a table which shows those amounts. (click on the object in the Global Environment to pull up the table (click on the object name **not** the blue circle)). Place a screen shot of the table below. (use a pipeline to generate a table showing the count of each building type)

A screenshot of a computer

Description automatically generated

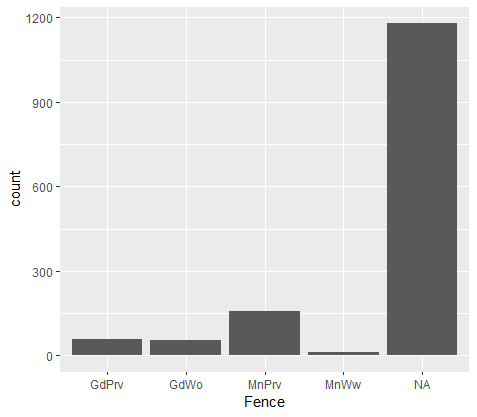
**C.)**

Next, you are curious to understand if the homes sold in Ames, Iowa, most often have a fence or not. Additionally, of the homes with a fence, what type of fence is most common? Generate a visualization that can answer those two questions. What do you observe? If your investor were to invest in upgrading the homes via a fence, what would you recommend and why (use your visualization to support your argument). **Place screenshot below**

NA or No fence-1179 out of 1460

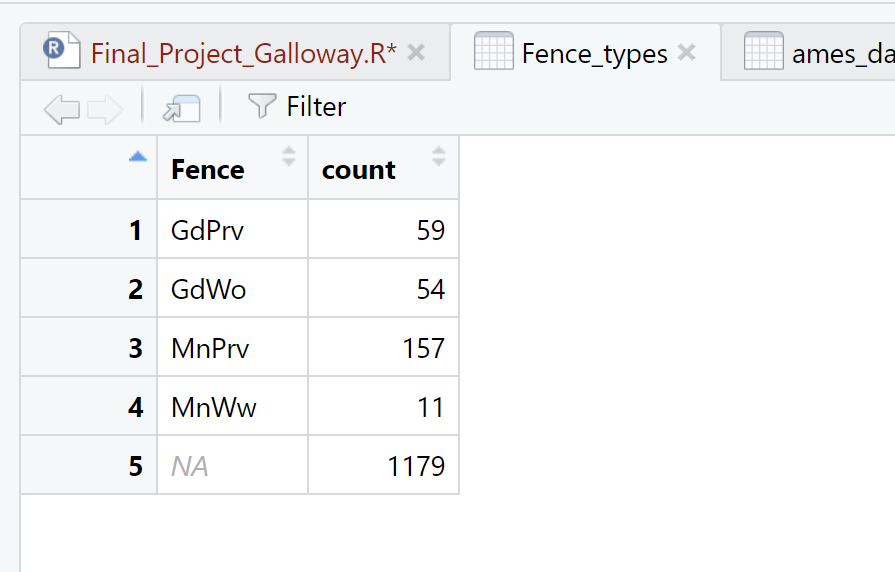
MnPrv or Minimum Privacy-157 out of 1460

If my investor was to invest in upgrade homes via a fence, I would recommend to them to build minimum privacy fences. Out of 1460 data points, only 281 houses had a fence. However, out of the 281 homes with a fence, 157 had a minimum privacy fence. Supported by the following visual, we can see that over 50% of fences built were minimum privacy.



**D.)**

Confirm your amounts stated in **C** by developing a table which shows those amounts. (click on the object in the Global Environment to pull up the table (click on the object name **not** the blue circle)). Place a screenshot of the table below. (use a pipeline to generate a table showing the count of each **fence**)



**Problem 4.)**

**A.)**

Next you are curious about the variation of the sale price across the different GarageType. Generate a figure which effectively visualizes this distribution. What do you observe? What should your boss take away from this plot? **Place screenshot below**

In this box chart, we can see the highlights of the variation of sale prices. We can see the median sale price for each home, as well as the interquartile range of prices. The median is an important statistic in the realm of home buying because it is less affected by outliers. As opposed to the average, which can change significantly due to outliers. Our boss should take away that the greatest variation in prices in is Attchd and Builtin garagetypes, which also have the highest median sale price. Even though there is a large variation, I would still recommend to my boss to buy homes with these two garage types.

A graph with black and white lines and dots

Description automatically generated

**B.)**

Generate a Cleveland plot to show the average sale price across each **GarageType.** Discuss the advantage/disadvantage to this chart as compared to the plot in **A**. **Place screenshot below**

As I mentioned previously, mean/average can be dramatically changed due to outliers in a dataset. Here we can see that Attchd and BuiltIn still have the highest sale prices, however I would caution the boss against thinking that these are “normal” prices. As we can see in the box plot graph, these style of garage have a lot of outlier sale prices. I would advise the boss to look at the median prices for those garage types, which should provide a more “normal” sale price.

A graph of numbers and points

Description automatically generated with medium confidence

**Problem 5.)**

**A.)**

Lastly, you are hoping to gain more insight into what is predictive of SalePrice. Pick a variable you want to use to predict SalePrice, generate the scatterplot. Additionally, within the figure, plot your linear model fit to the data. What do you observe? What should your boss take away from this plot? **Place screenshot below**

In this plot we used the variable YearBuilt to predict the SalePrice. When we plotted a linear model on top of the scatter plot, we can quickly notice a positive linear relationship between YearBuilt and SalePrice. This shows that as the YearBuilt increase, the SalePrice also increases. Our boss should consider that when purchasing new homes, homes that were built more recently are going to most likely have a higher Sale Price.

A graph with black dots and green line

Description automatically generated

**B.)**

Generate the linear model from **A** and print out the summary statistics. What interpretations can you make and present to your boss? What impact does your variable have on the SalePrice? What's the R2 and your interpretation? What should your boss take away from this information? **Place screenshot below**

YearBuilt: The coefficient for YearBuilt is 1,375. This indicates that, on average, for each one-unit increase in YearBuilt, the estimated SalePrice increases by 1,375 units.

R-squared: 27.34% of the variance in SalePrice is explained by the linear regression model with YearBuilt. This is a low/moderate amount of explanatory power.

My boss should consider that the most practical use of this information is that for every 1-year increase in the date the house was built, there is an expected price increase of $1,375. This should lead his focus towards buying newer homes, for they are likely to sell at a higher price. An R-squared of 27.34% suggests that the model explains a low/moderate amount of variability in sale prices based on the YearBuilt variable. While this can be useful for considering which houses to buy, it also indicates that there are other factors influencing house values that are not accounted for in the model. In other words, while there is a positive correlation between the year built and price, my boss must continue to consider other variables as important predictors as well.

A screenshot of a computer

Description automatically generated

**PART 2:**

**Tableau**

**1.)**

Generate the figures/table from questions **1D, 2A, 3A, 3C, 3D, 4A, 5A** in tableau. Place them below.

A graph with colorful lines

Description automatically generated

A graph showing a number of blue dots

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generatedA graph with numbers and a bar

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA screen shot of a graph

Description automatically generated

**2.)**

Develop a dashboard. You can choose any of the figures you have developed to put in the dashboard or you can generate new figures. Justify why you chose the figures you did and discuss why you designed your dashboard that way. (i.e. What in the IDD book did you consider, why did you order it the way you did? Why did you color it the way you did? Did you use any preattentive attributes to draw attention to certain parts of the visualizations? Be detailed.) The more descriptive the more points.

Remember, your overall goal is to provide insights into the Ames, Iowa, housing market and influence the types of investments your company should make to turn a profit flipping homes. Discuss what your boss should take away from this dashboard and what actions they should take based on your dashboard.

**Answer**:

Selling homes is a competitive market. When considering the best approach to enter that market, I thought of two questions. What types of homes are currently selling in Ames, Iowa? And, what homes are selling at a higher price? After all, selling a highly-priced home is only worth it if there is a demand for it.

When designing my dashboard, I kept these two questions in mind. I chose 3 tables to help answer what types of homes are selling, and I chose 3 graphs to help answer what homes are selling at a higher price. Keeping dashboard design principles in mind, I made sure to keep my dashboard simple. I went with a simple blue color scheme, using complementary colors like orange and green only to highlight important information. I used 3 titles to help provide context to the dashboard. The main titles explains that we are looking at housing data, the second and third subtitles explain that we are trying to answer the two questions I previously mentioned.

Under the “Predicting Sales Price” section of my dashboard, I included square footage, year built, and garage type, all of which showed a correlation to sale price during the research stage of this project. Square footage showed the strongest correlation, with an increase in $104 per square foot. Next year built also showed a strong correlation, with an increase in $1375 per yearly increase on date built. Finally, we saw that with garage type, built in and attached garage sold at the highest prices.

When looking at the types of homes that were sold, we can see that the majority of homes old didn’t have a fence, were a 1 family building style, and majority had an attached garage type.

What can we take away from this? I would recommend to my boss to always try to buy newer homes, as they sell for higher. I would also recommend buying mostly 1-family homes, as these are the majority of homes sold. Once the home is bought, I would recommend building an attached garage if there isn’t already one, because we can see from the dashboard that they are in high demand. I would also recommend that during the renovation process, they definitely should consider increasing the square footage whenever possible, as this will increase the home value significantly.