CSCI 550 - Project 2

Eliot Liucci, Eric Folsom, Nick Clausen, Christal O'Connell

2024-10-25

1 Executive Summary

2 Data Preprocessing and Exploration

2.1 Data Cleaning

The data received had a few issues that needed to be dealt with. First of all, the `Description` variable contained pieces of key information that were extracted (number of bathrooms, number of bedrooms, total number of rooms, and sell date).

```
# Extracting important information from descriptions
sell_date = as.numeric(0)
rooms = as.numeric(0)
bedrooms = as.numeric(0)
baths = as.numeric(0)

# Takes a minute to run, not too bad though
for(i in 1:nrow(data)){
    sell_date[i] = str_split(
        str_split(data$Description[i], "sold on ")[[1]][2],
         ", is a")[[1]][1]
    rooms[i] = str_extract_all(
        str_split(data$Description[i], "total of ")[[1]][2],
         "\\d")[[1]][1]
    bedrooms[i] = str_extract_all(
        str_split(data$Description[i], "total of ")[[1]][2],
         "\\d")[[1]][2]
    baths[i] = str_split(
        str_split(data$Description[i], "bedrooms, and ")[[1]][2],
         " of which are bathrooms")[[1]][1]
}
```

Once these variables were extracted, the `Description` variable was dropped from the data set. We noticed some houses with strange recording like "42 bathrooms and 7 rooms", so we dropped any listings where the number of bedrooms and number of bathrooms was greater than the total recorded number of rooms.

```
# Adding features of descriptions, removing descriptions
data = data %>%
    mutate(
      Sell_Date = mdy(sell_date),
      Rooms = as.numeric(rooms),
      Bedrooms = as.numeric(bedrooms),
      Baths = as.numeric(baths)
    ) %>%
    select(-Description)
```

```r
# Removing rows where num bathrooms/bedrooms exceeds number of rooms
data = data %>%
    filter(Bedrooms < Rooms | Baths < Rooms)
```

Next, we dropped any listings with a missing value in at least one of the variables.

```r
# Removing rows with at least 1 missing value
data = data %>%
    drop_na()
```

We also wanted to deal with outliers, so a function was written that would identify a listing as an outlier if it was greater than 3 standard errors away from the mean value and used this to remove outliers for variables where the maximum value was significantly higher than the 3rd quartile.

```r
# Filtering extreme observations
is_outlier = function(x){
    result = abs(x - mean(x)) > 3*sd(x)
    return(result)
}

# Removing outliers for variables where max() is greater than q3()
data = data %>%
    filter(!is_outlier(`Sale Price`),
           !is_outlier(`Land Square Feet`),
           !is_outlier(Baths),
           !is_outlier(`Lot Size`),
           !is_outlier(`Town and Neighborhood`),
           !is_outlier(`Age Decade`),
           !is_outlier(`Age`),
           !is_outlier(`Estimate (Land)`),
           !is_outlier(`Estimate (Building)`),
           !is_outlier(`Building Square Feet`),
           !is_outlier(`Other Improvements`))
```

Finally, we removed all spaces from variable names and replaced them with underscores.

```r
# Removing Spaces
data = data %>%
  rename(Property_Class = `Property Class`,
         Neighborhood_Code = `Neighborhood Code`,
         Land_Square_Feet = `Land Square Feet`,
         ...
         Age_Decade = `Age Decade`,
```

```
        Neighborhood_Code_Mapping = `Neigborhood Code (mapping)`,
        Town_and_Neighborhood = `Town and Neighborhood`,
        )
```

This cleaned data set was written as `data_cleaned.csv` so it could easily be reloaded for the remainder of the analysis.

```
# Save cleaned data
write_csv(data, "data_cleaned.csv")
```

2.2 Exploration of Data

Within the data, latitude and longitude coordinates were provided for each listing. A map of the sale price of listings is overlaid on a satellite map of the region (Figure 1). Here, it can be seen that a lot of the higher price listings are on the water, with sale price generally decreasing the more in-land the listing is.

```
# Spatial Map of Sale Price
ggmap(Map, darken = c(0.1, "white")) +
    geom_polygon(data = cook_map, aes(x = long, y = lat),
        fill = NA, color = "orange") +
    geom_point(data = data,
                aes(x = Longitude,
                    y = Latitude,
                    color = Sale_Price),
                size = 0.02,
                alpha = 0.75) +
    scale_color_gradient(low = "#6fe7f7", high = "#890000") +
    labs(x = "Longitude", y = "Latitude", color = "Sale Price ($)")
```

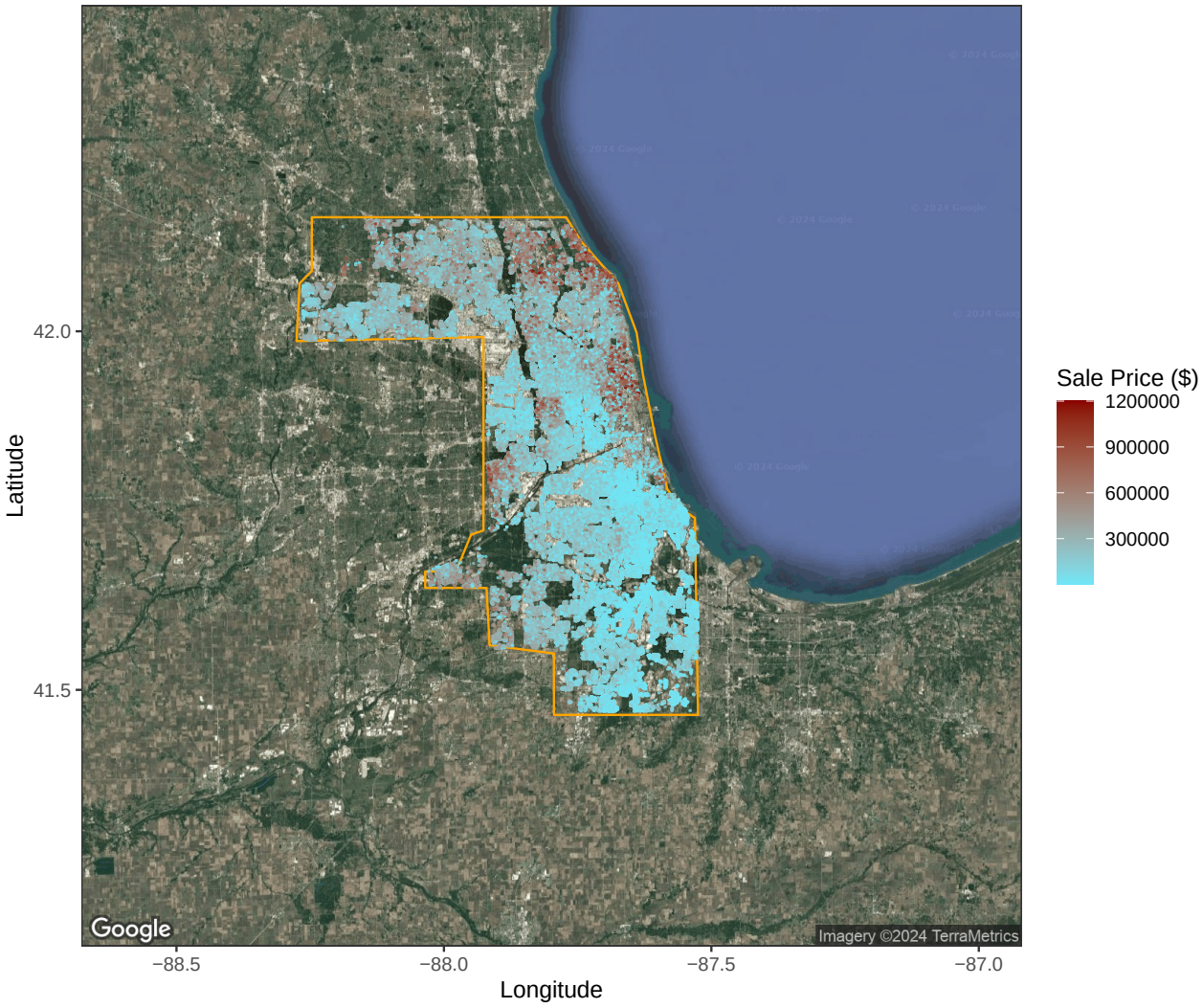

Figure 1: A spatial map of the sale prices over the region.

Multiple boxplots were created for all of the variables pertaining to a garage against the response (Figure 2). As the first garage increases in size, so too does the sale price. Having a garage attached to the building (`Garage_#_Attachement` = 1) is also associated with a higher sale price for both garage 1 and garage 2. Higher quality materials (larger values of `Garage_#_Material`) is associated with higher sale prices too.

```
# Plotting Against Garage Variables
data %>%
    select(Sale_Price,
           Garage_1_Area,
           Garage_1_Size,
           Garage_1_Attachment,
           Garage_1_Material,
           Garage_2_Area,
           Garage_2_Size,
```

```
            Garage_2_Attachment,
            Garage_2_Material) %>%
    pivot_longer(cols = 2:9, names_to = "Variable", values_to = "Value") %>%
    ggplot(aes(x = factor(Value), y = Sale_Price, group = Value)) +
    geom_boxplot() +
    facet_wrap(~Variable, scales = "free_x", nrow = 2) +
    labs(x = " ")
```

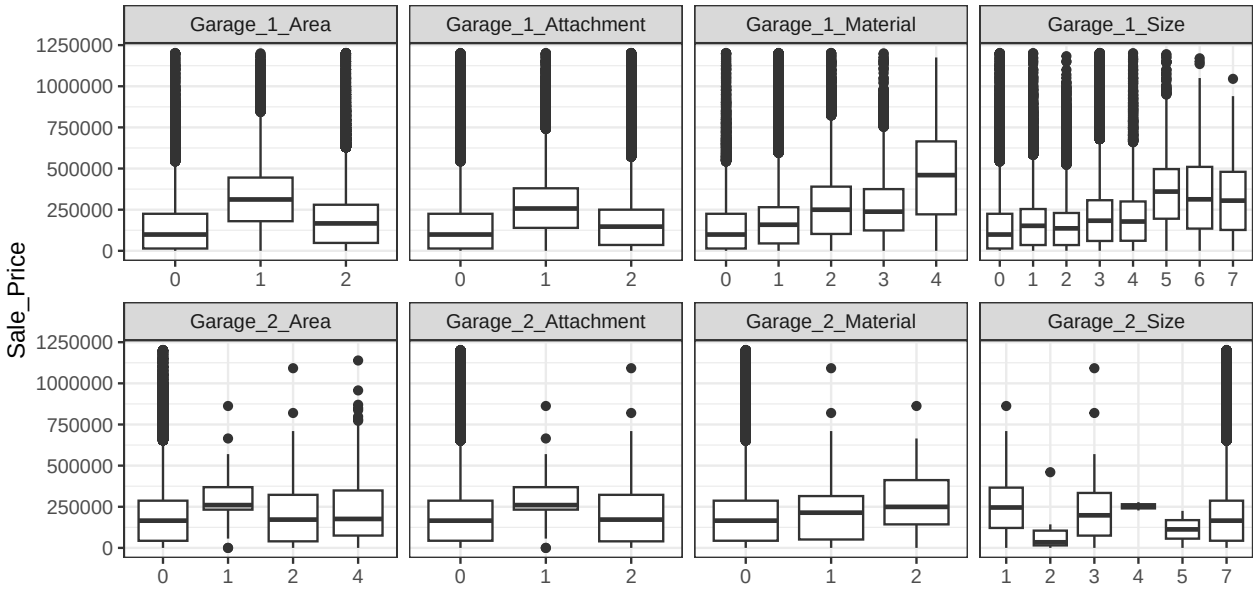

Figure 2: A series of boxplots for all garage variables against sale price.

When comparing sale price to variables related to the physical property, it can be noted that for the `Apartments` variable, there is large varibility in the "0" group, which may be due to non-apartment buildings being more expensive (Figure 3). Sale price is generally the same across attic types, porch groups, and design plans. Sale price appears to increase as the number of fireplaces increases. Additionally, sale price is higher for listings with cathedral ceilings. The `Proprty_Class` variable apperas to have equal variability in sale price for all classes except "209".

```
data %>%
    select(`Sale_Price`,
           `Property_Class`,
           `Apartments`,
           `Basement`,
           `Attic_Type`,
           `Design_Plan`,
           `Cathedral_Ceiling`,
           Fireplaces,
```

```
            Porch) %>%
    pivot_longer(cols = 2:9, names_to = "Variable", values_to = "Value") %>%
    ggplot(aes(x = factor(Value), y = Sale_Price, group = Value)) +
    geom_boxplot() +
    facet_wrap(~Variable, scales = "free_x", nrow = 2) +
    labs(x = " ")
```

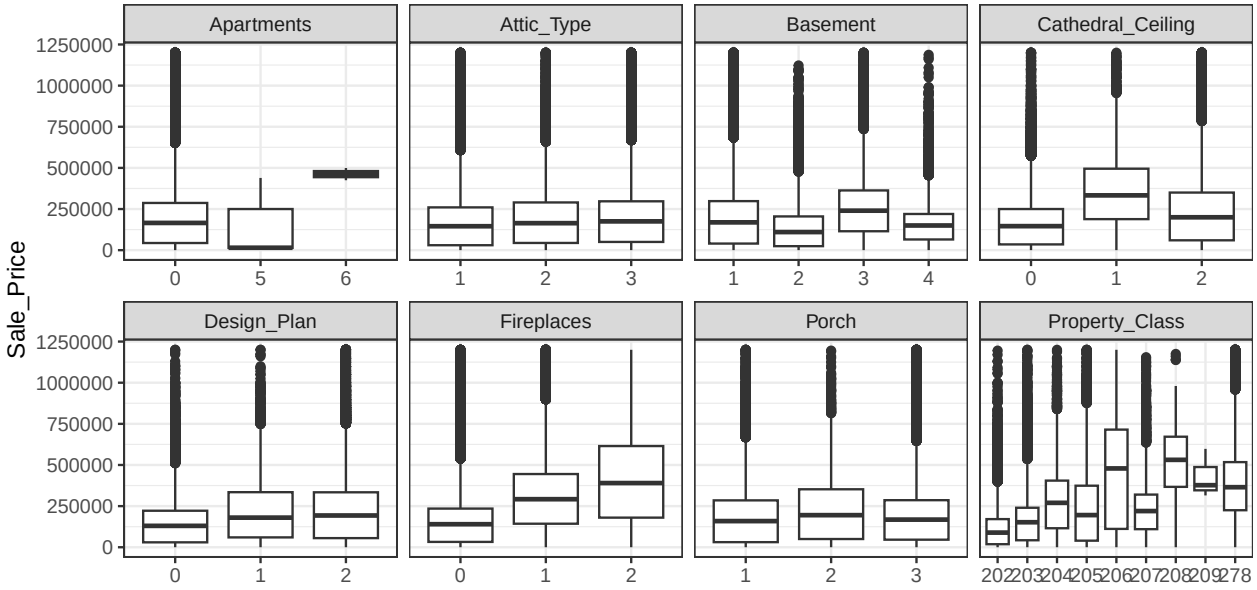

Figure 3: A series of boxplots for variables related to the property itself.

Additional boxplots were created for the number of rooms in each listing (Figure 4). For all room variables, there is generally an increase in sale price as the number of rooms increases. However, the sale price for listings with 1 room is higher, on average, than all other room groups. The same goes for listings with 0 bedrooms.

```
# Plotting Against Room Variables
data %>%
    select(`Sale_Price`,
           `Bedrooms`,
           `Baths`,
           `Rooms`) %>%
    pivot_longer(cols = 2:4, names_to = "Variable", values_to = "Value") %>%
    ggplot(aes(x = factor(Value), y = Sale_Price, group = Value)) +
    geom_boxplot() +
    facet_wrap(~Variable, scales = "free_x") +
    labs(x = " ")
```

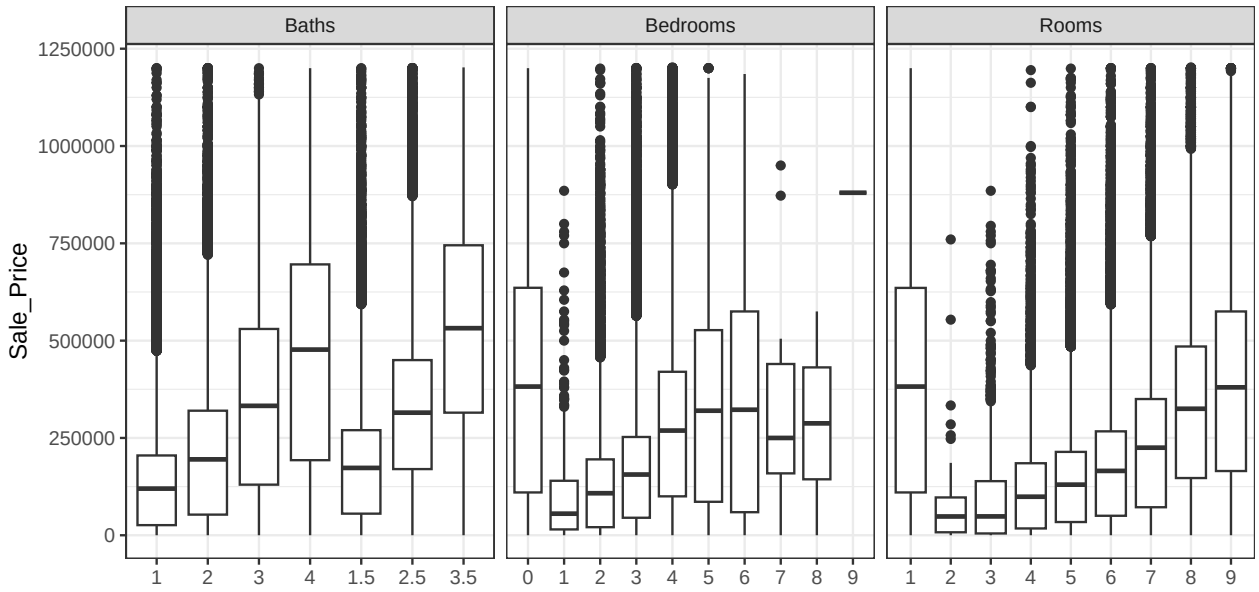

Figure 4: A series of boxplots for the variables relating to number of rooms present.

The final visualization of interest is a scatterplot of sale price against both building square footage and land square footage (Figure 5). Here, it can be seen that for listings with high building square footage, we see higher sale price. The relationship is the same for land square footage, although the highest sale prices occur at high building square footage and average land square footage.

```
data %>%
    ggplot(aes(x = `Building_Square_Feet`,
               y = `Land_Square_Feet`,
               color = `Sale_Price`)) +
    geom_point(size = 2) +
    labs(x = "Building Square Footage",
        y = "Land Square Footage",
        color = "Sale Price ($)")
```

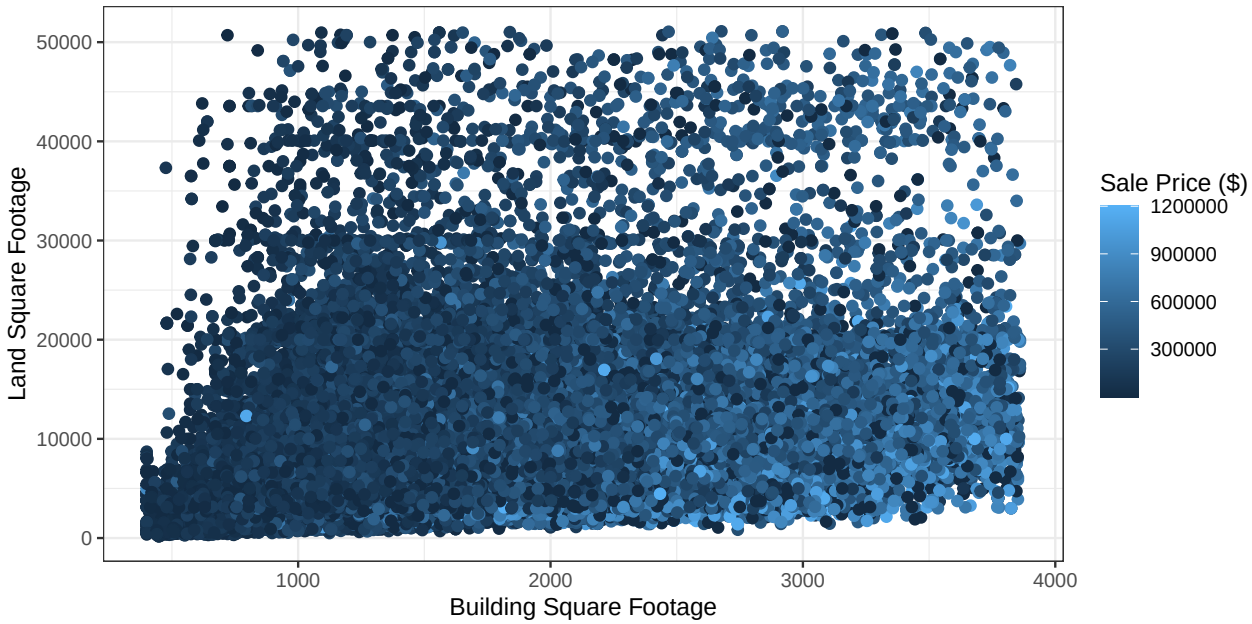

Figure 5: A scatterplot of the relationship between sale price and building/land square footage.

2.3 Hypothesis Development

From the exploration performed above, we hypothesize that building square footage, land square footage, the number of rooms, and location are likely going to have the highest impact on sale price.

3 Model Development and Performance Evaluation