Lab 5

Code

AUTHOR
Maxim Dokukin

Remember, **follow the instructions below and use R Markdown to create a pdf document with your code and answers to the following questions on Gradescope.** You may find a template file by clicking “Code” in the top right corner of this page.

0. Cook County Assessor’s Office

For this lab, you will work with data from Cook County Assessor’s Office from Illinois, which inspects properties across Chicago and its suburbs to **assess** the value of each property to determine the amount in property taxes owed by each property owner. All property owners are required to pay taxes which are used to fund public services at the state level. These assessments are based on property values which are often estimated via statistical models that account for variables like the size and location of a property.

Since these models determine how much property owners must pay each year, it is desirable to have assessments that are fair. However, in 2017, the office of the former Cook County Assessor Joseph Berrios was sued by two Chicago nonprofits who alleged that Berrios’ office “disproportionately put the burden of residential property taxes on minority homeowners,” so that wealthy property owners paid proportionally less in taxes compared with lower-income, and often minority, property owners. The Chicago Tribune investigated property assessments from 2003 to 2015, arguing that assessments had indeed been discrimnatory. Their four-part investigation can be found here.

Since this investigation, the Cook County Assessor’s Office has strived to be more transparent in disclosing their methods and data for property valuation. In this assignment, we will look at data they have released on property valuation from 2013-2019. The office has also published open-source code for their models, which is written in R! This assignment is based on a module developed by instructors at UC Berkeley.

A. Residential Sales Data

1. Download the data from this link. How many rows are there in this dataset? What does each row represent? (Hint: be precise here).

```
sales_data <- read.csv("/Users/xewe/Documents/Education/BS SJSU/

dim(sales_data)
```

```
[1] 583370     83
```

583370 rows, 83 cols; each row is a sale report for a property

2. Examine the `Site Desirability` variable. What do each of the levels of this variable represent? You may need to refer to the codebook to learn about this variable. Is it explained how this variable is determined?

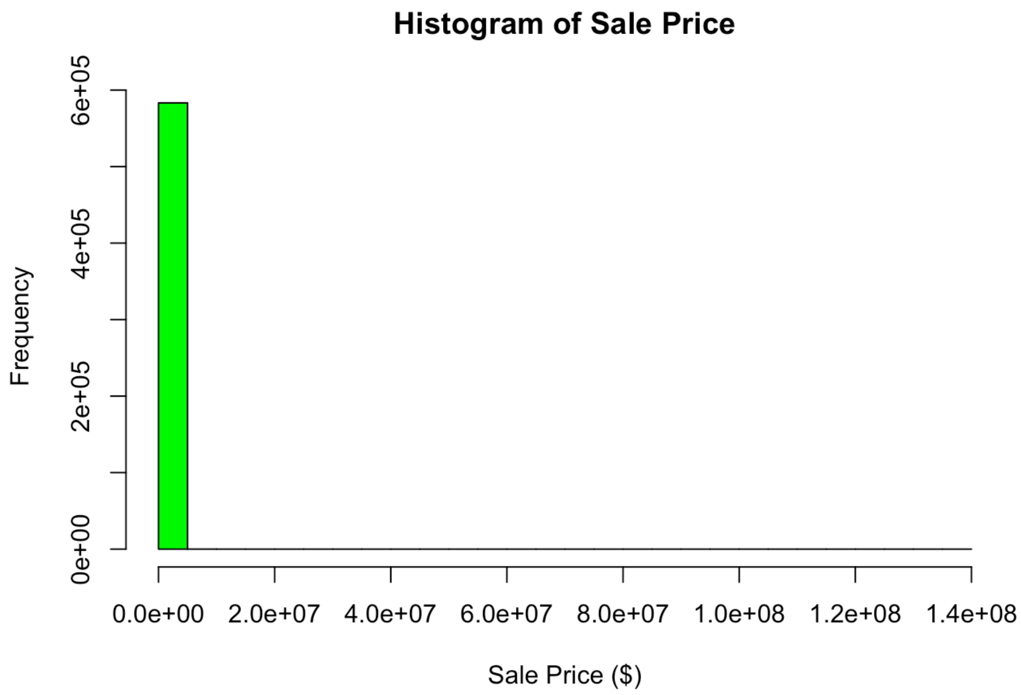
Site desirability - 1 = Beneficial to Value, 2 = Not relevant to Value, 3 = Detracts from Value. This field lack sufficient variation to be useful for modeling. No specific method behind how this value is obtained.

3. Give an example of a variable that is **not** included in this dataset that could be useful in determining property value.

Neighborhood crime rate per 1000 residents could be beneficial in assessing property value.

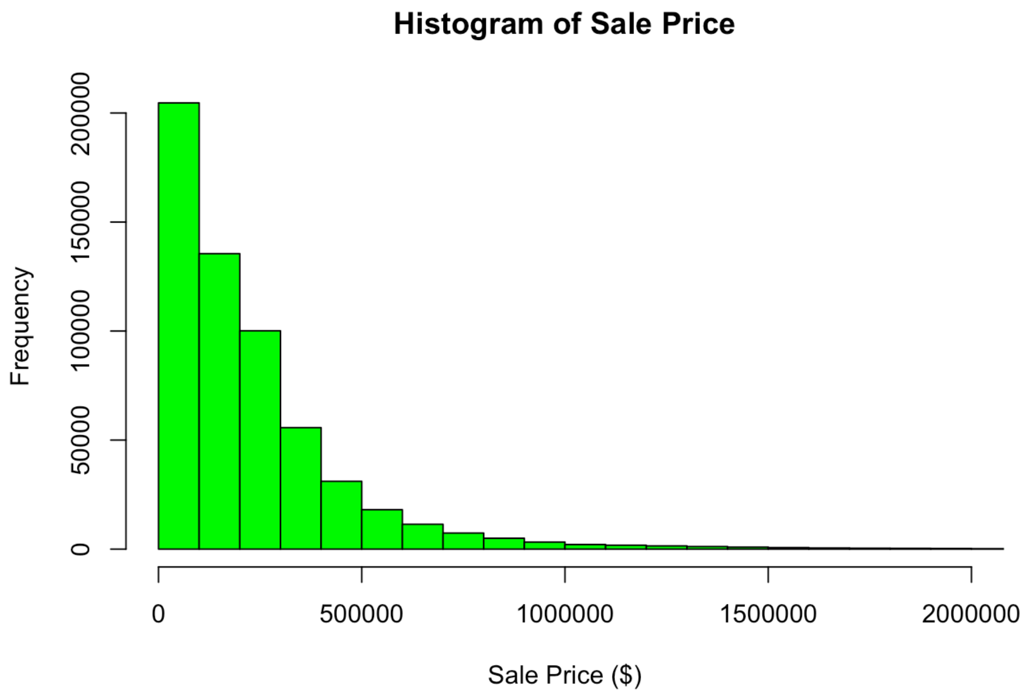
4. Create a histogram of `Sale Price` for this dataset. Identify one issue with this visualization and attempt to address this issue.

```
hist(sales_data$Sale.Price,
     main = "Histogram of Sale Price",
     xlab = "Sale Price ($)",
     col = "green")
```



The x scale is off. There must be a very expensive outlier in the dataset. I am adjusting x axis scale.

```r
hist(sales_data$Sale.Price,
     main = "Histogram of Sale Price",
     xlab = "Sale Price ($)",
     col = "green",
     xlim = c(0, 2000000),
     breaks = 1000)
```

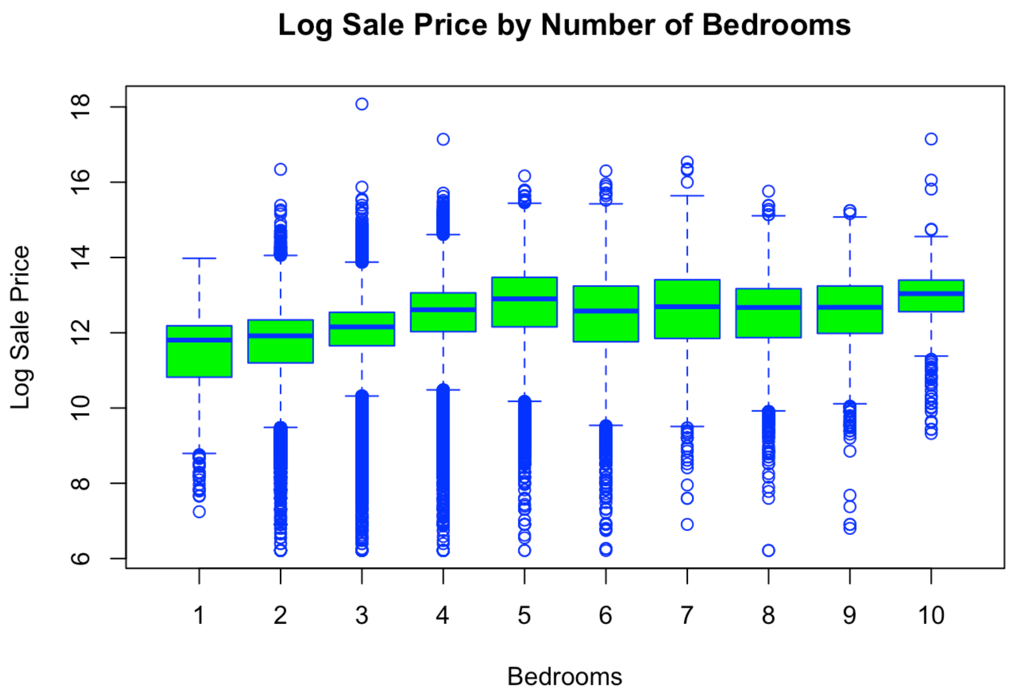


5. For the rest of the assignment, we will focus on a subset of properties. Provide code that creates a new `data.frame` called `clean_data` that contains only properties whose sale price is at least $500. Create a new column in this data frame called `log_sale_price` that contains the log-transformed `Sale Price` values.

```
clean_data <- subset(sales_data, Sale.Price >= 500)
clean_data$log_sale_price <- log(clean_data$Sale.Price)
```

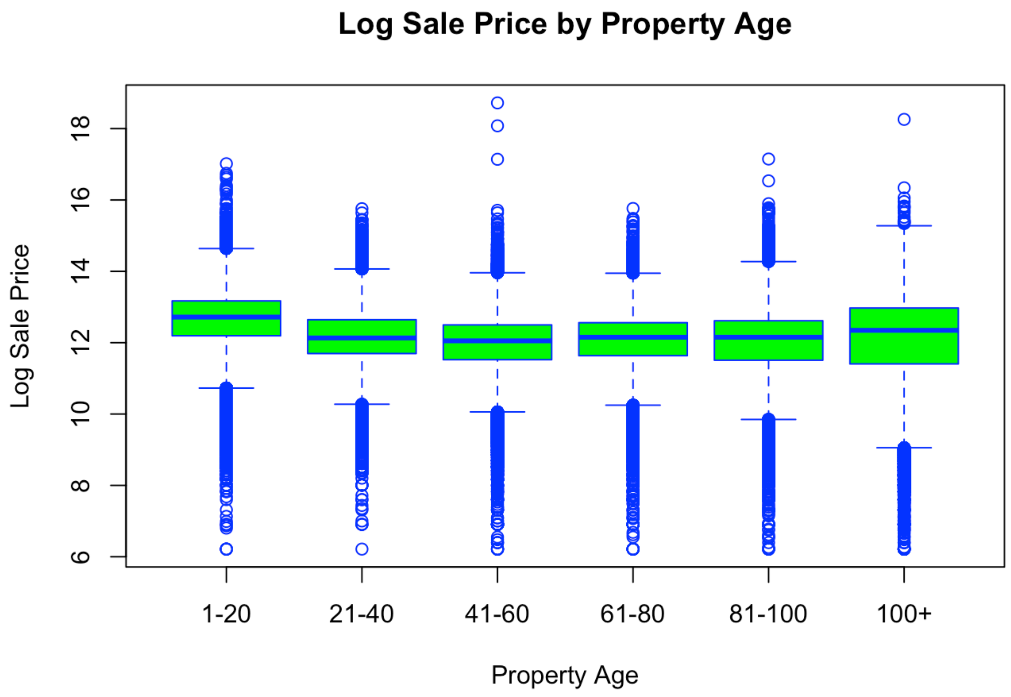
6. Visualize the association between number of bedrooms and `log_sale_price` using parallel box plots. You may need to convert `Bedrooms` to a factor before you are able to construct the parallel box plots. For clarity, only include properties with 10 or fewer bedrooms. Interpret your results.

```
price_bed <- subset(clean_data, Bedrooms <= 10)
price_bed$Bedrooms <- as.factor(price_bed$Bedrooms)

boxplot(log_sale_price ~ Bedrooms, data = price_bed,
        xlab = "Bedrooms", ylab = "Log Sale Price",
        main = "Log Sale Price by Number of Bedrooms",
        col = "green", border = "blue")
```



7. Create a new factor variable called `age_bin` that has levels `1-20`, `21-40`, `41-60`, `61-80`, `81-100`, and `100+`. Visualize the association between `age_bin` and `log_sale_price` using parallel box plots. Interpret your results.

```
clean_data$age_bin <- cut(clean_data$Age,
                breaks = c(-Inf, 20, 40, 60, 80, 100, Inf),
                labels = c("1-20", "21-40", "41-60", "61-80", "8
                right = FALSE)

boxplot(log_sale_price ~ age_bin, data = clean_data,
        xlab = "Property Age", ylab = "Log Sale Price",
        main = "Log Sale Price by Property Age",
        col = "green", border = "blue")
```



B. Assessor First Pass Values

8. Not all of the properties in the above dataset have public assessment values. You can download another dataset containing “First Pass Values” representing the Assessor’s initial valuations for a set of properties in 2019 here. How many rows are in this dataset?

```
values_data <- read.csv("/Users/xewe/Documents/Education/BS SJSU

dim(values_data)
```

```
[1] 31163    17
```

31163 rows

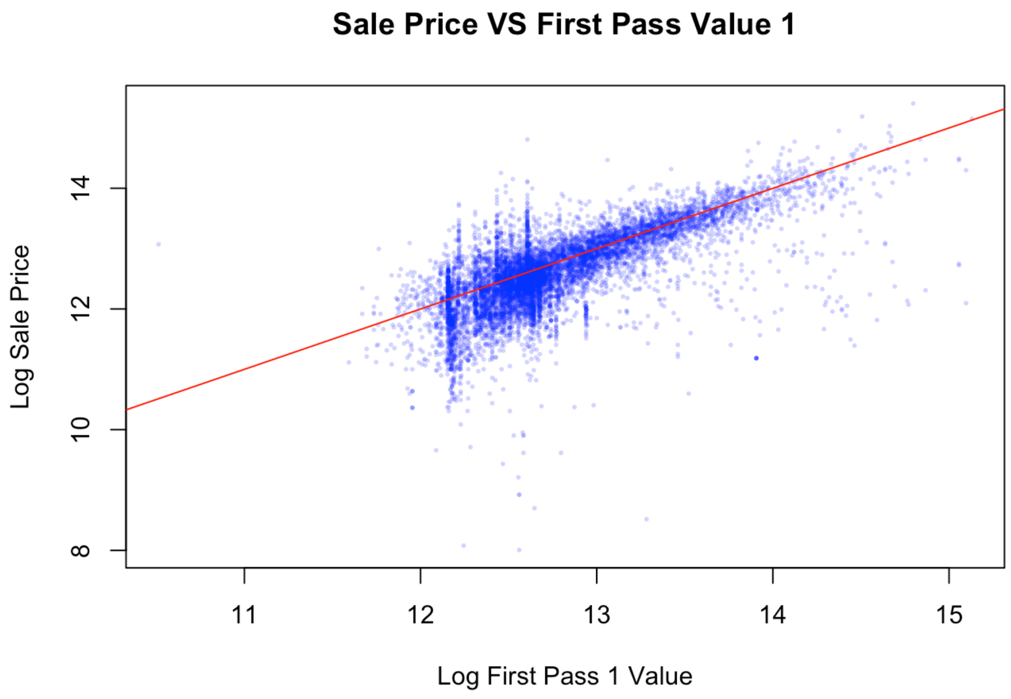
9. Use an appropriate function to combine the first pass values data with the `clean_data` from Part A. You should keep only rows that have both `log_sale_price` (from `clean_data`) and `First Pass Value 1` from the first pass values data. How many rows are in this combined dataset?

```
combined_data <- merge(clean_data, values_data, by = "PIN")
combined_data <- combined_data[!is.na(combined_data$log_sale_pr
nrow(combined_data)
```

```
[1] 9437
```

10. Create a scatter plot with `log(First Pass Value 1)` on the x-axis and `log_sale_price` on the y-axis. Add a line to your plot indicating the line where `y=x`. Interpret your results. What do points above the line represent? What do points below the line represent?

```
combined_data$log_first_pass_value_1 = log(combined_data$`First

plot(combined_data$log_first_pass_value_1, combined_data$log_sa
     xlab = "Log First Pass 1 Value",
     ylab = "Log Sale Price",
     main = "Sale Price VS First Pass Value 1",
     col = rgb(0, 0, 1, 0.2), pch = 16, cex = 0.4)
abline(a = 0, b = 1, col = "red")
```

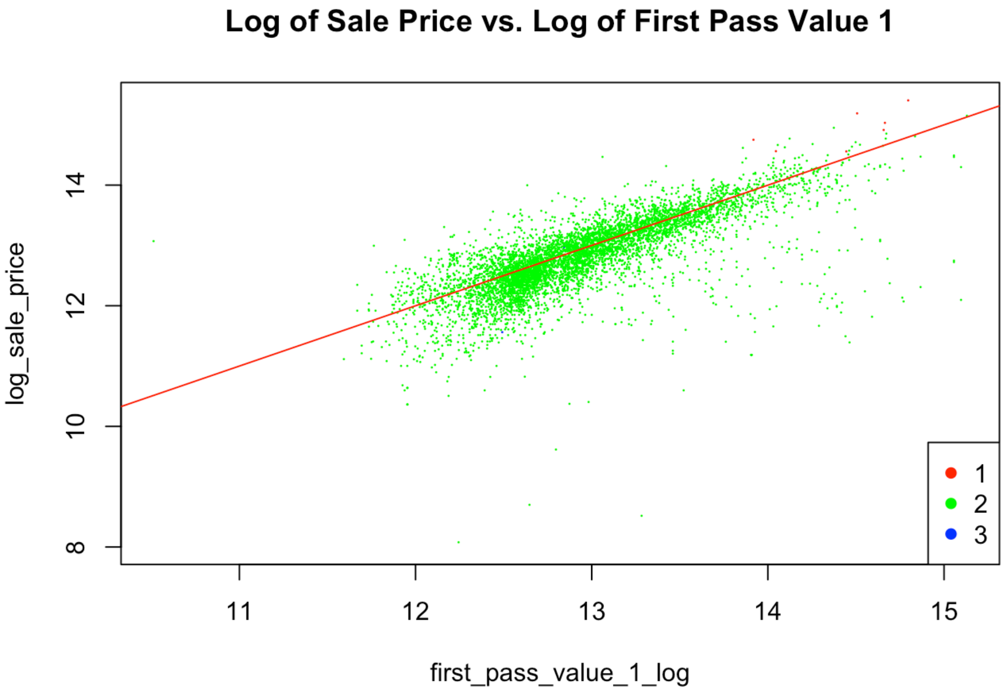


```
#PS
#Chat GPT wrote code below
```

```
combined_data$Site.Desirability <- as.factor(combined_data$Site.
levels(combined_data$Site.Desirability)
```

```
[1] "1" "2" "3"
```

```
colors <- rainbow(length(levels(combined_data$Site.Desirability)

plot(combined_data$log_first_pass_value_1,
     combined_data$log_sale_price,
     col = colors[as.numeric(combined_data$Site.Desirability)],
     xlab = "first_pass_value_1_log",
     ylab = "log_sale_price",
     main = "Log of Sale Price vs. Log of First Pass Value 1",
     pch = 16,  # Choose a plotting character
     cex = 0.2) # Choose size of points

abline(a = 0, b = 1, col = "red")

legend("bottomright",
       legend = levels(combined_data$Site.Desirability),
       col = colors,
       pch = 16)
```



Nothing too interesting. Some of the more desirable properties are sold for higher than evaluation price (expected).