Sp23_Midterm_Exam

Areej Mulla

3/2/2023

Question 1: Reading the File & Printing the Summary

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
housing_prices <- read_csv(paste0(getwd(), "/melbourne_housing_prices.csv"))
```

```
## Rows: 13580 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (8): Suburb, Address, Type, Method, SellerG, Date, CouncilArea, Regionname
## dbl (13): Rooms, Price, Distance, Postcode, Bedroom2, Bathroom, Car, Landsiz...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(housing_prices)
```

```
##     Suburb            Address              Rooms            Type
##  Length:13580       Length:13580       Min.   : 1.000   Length:13580
##  Class :character   Class :character   1st Qu.: 2.000   Class :character
##  Mode  :character   Mode  :character   Median : 3.000   Mode  :character
##                                        Mean   : 2.938
##                                        3rd Qu.: 3.000
##                                        Max.   :10.000
##
##      Price            Method            SellerG             Date
##  Min.   :  85000   Length:13580       Length:13580       Length:13580
##  1st Qu.: 650000   Class :character   Class :character   Class :character
##  Median : 903000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1075684
##  3rd Qu.:1330000
##  Max.   :9000000
##
##     Distance        Postcode       Bedroom2         Bathroom
##  Min.   : 0.00   Min.   :3000   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 6.10   1st Qu.:3044   1st Qu.: 2.000   1st Qu.:1.000
```

```
##  Median : 9.20   Median :3084   Median : 3.000   Median :1.000
##  Mean   :10.14   Mean   :3105   Mean   : 2.915   Mean   :1.534
##  3rd Qu.:13.00   3rd Qu.:3148   3rd Qu.: 3.000   3rd Qu.:2.000
##  Max.   :48.10   Max.   :3977   Max.   :20.000   Max.   :8.000
##
##       Car           Landsize         BuildingArea     YearBuilt
##  Min.   : 0.00   Min.   :     0.0   Min.   :    0   Min.   :1196
##  1st Qu.: 1.00   1st Qu.:   177.0   1st Qu.:   93   1st Qu.:1940
##  Median : 2.00   Median :   440.0   Median :  126   Median :1970
##  Mean   : 1.61   Mean   :   558.4   Mean   :  152   Mean   :1965
##  3rd Qu.: 2.00   3rd Qu.:   651.0   3rd Qu.:  174   3rd Qu.:1999
##  Max.   :10.00   Max.   :433014.0   Max.   :44515   Max.   :2018
##  NA's   :62                         NA's   :6450    NA's   :5375
##  CouncilArea          Lattitude        Longtitude     Regionname
##  Length:13580       Min.   :-38.18   Min.   :144.4   Length:13580
##  Class :character   1st Qu.:-37.86   1st Qu.:144.9   Class :character
##  Mode  :character   Median :-37.80   Median :145.0   Mode  :character
##                     Mean   :-37.81   Mean   :145.0
##                     3rd Qu.:-37.76   3rd Qu.:145.1
##                     Max.   :-37.41   Max.   :145.5
##
##  Propertycount
##  Min.   :  249
##  1st Qu.: 4380
##  Median : 6555
##  Mean   : 7454
##  3rd Qu.:10331
##  Max.   :21650
##
```

Question 2: Computing the Sum of Missing Records for Each Column

The summary demonstrates that the following columns having missing values:

- "Car" = 62 missing values
- "BuildingArea" = 6450 missing values
- "YearBuilt" = 5375 missing values
- "CouncilArea" = 1369 missing values

Below is the analysis on whether it is acceptable to eliminate the missing values and the reasoning:

- "Car": 62 records is not a significant number compared to the total number of records of 13580, and thus the corresponding entries can be eliminated.
- "CouncilArea" does not include a significant number of missing values, and thus the corresponding entries can be eliminated.
- "BuildingArea" and "YearBuilt" have a significant number of missing values and are crucial indicators of a house price, and thus the corresponding entries should not be eliminated, but rather should be imputed.

Question 3: Creating a Histogram to Illustrate Price Distribution

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

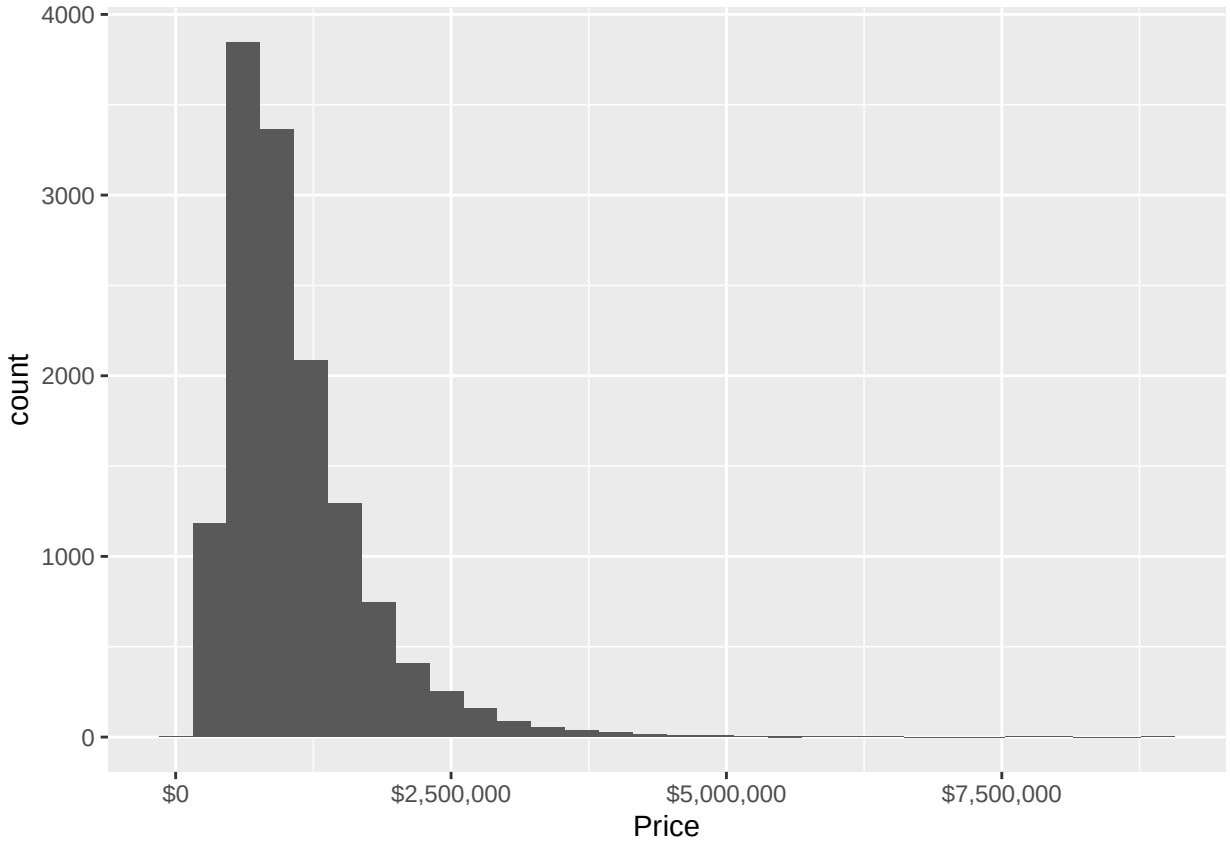
```
layer1 <- ggplot( data = housing_prices)
```

Plot #1:

```
layer1 +
  geom_histogram(aes(x=Price)) +
  scale_x_continuous(labels = scales::dollar_format())
```

```
##
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

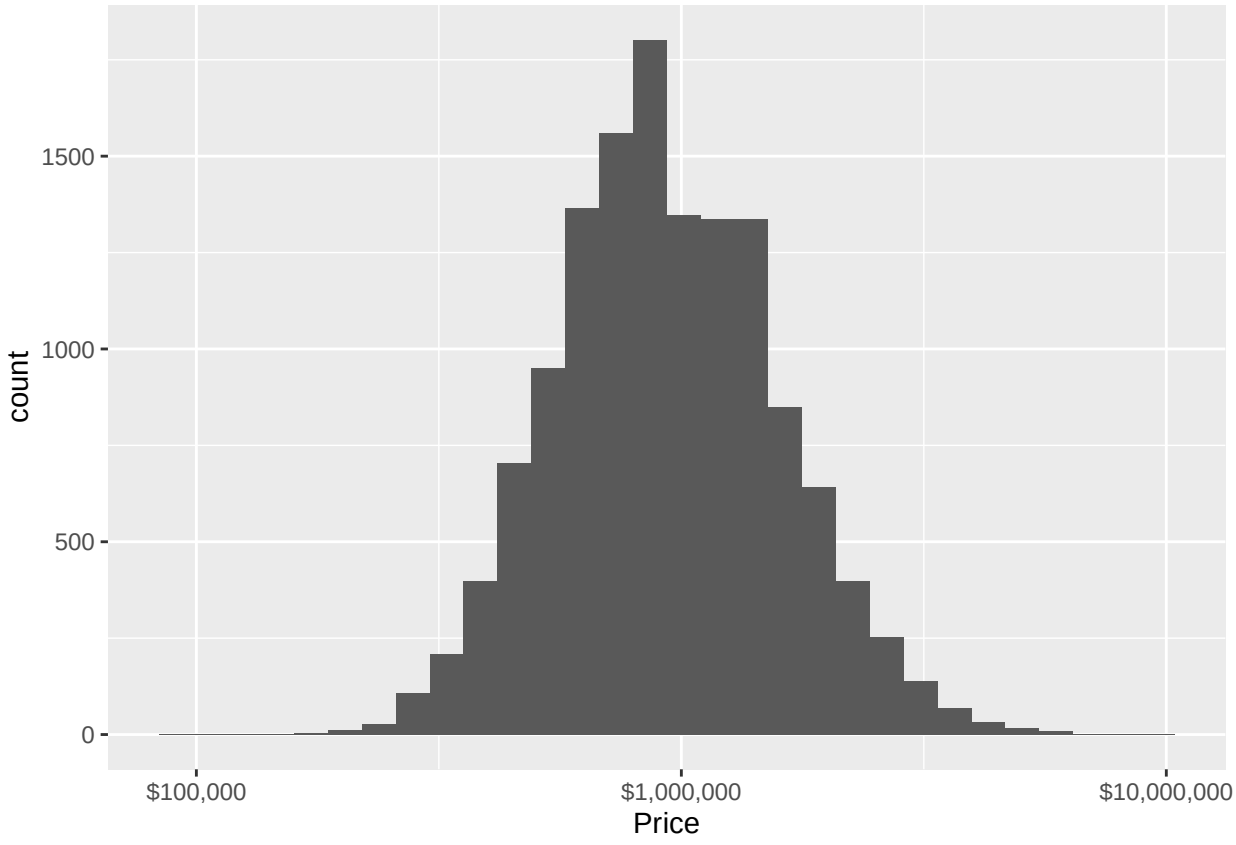


Plot #2 (With log10 Scale):

```
layer1 +
  geom_histogram(aes(x=Price)) +
    scale_x_log10(labels = scales::dollar_format())
```

```
##
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



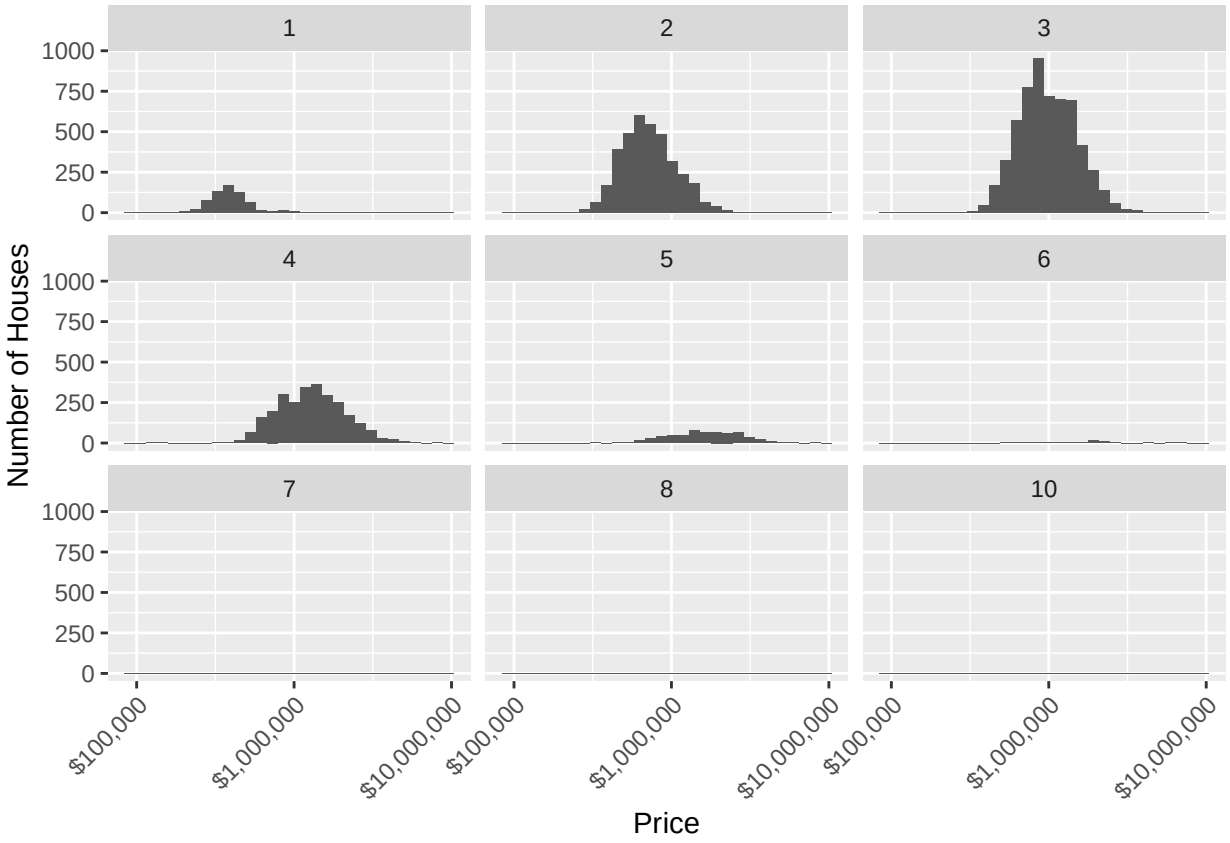
Question 4: Creating a Facet Histogram

Plot #1:

```
ggplot(housing_prices[!is.na(housing_prices$Price),]) +
  geom_histogram(aes(x = Price)) +
  scale_x_log10(labels = scales::dollar_format(), name = "Price") +
  facet_wrap(~Rooms) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
  ylab("Number of Houses")
```

```
##
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

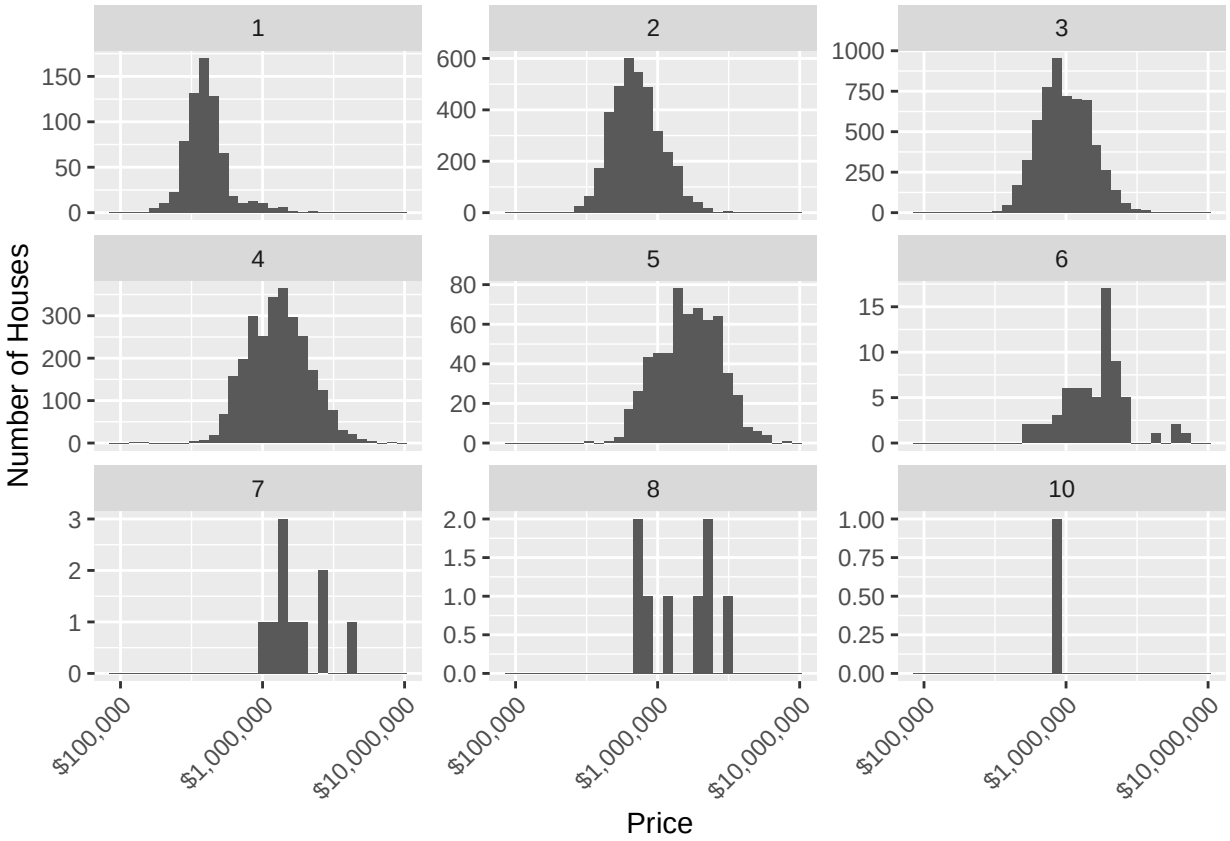


Plot #2 (With free_y):

```
ggplot(housing_prices[!is.na(housing_prices$Price),]) +
  geom_histogram(aes(x = Price)) +
  scale_x_log10(labels = scales::dollar_format(), name = "Price")+
  facet_wrap(~Rooms, scale = "free_y") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
  ylab("Number of Houses")
```

```
##
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 5: Creating a Correlation Matrix

```
options(digits = 2)
numeric_column_names <- unlist(sapply(housing_prices, is.numeric))
cor_matrix <- cor(housing_prices[,numeric_column_names], use="pairwise.complete.obs")
knitr::kable(cor_matrix)
```

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude	Propertycount
Rooms	1.00	0.50	0.29	0.06	0.94	0.59	0.41	0.03	0.12	-0.07	0.02	0.10	-0.08
Price	0.50	1.00	-0.16	0.11	0.48	0.47	0.24	0.04	0.09	-0.32	-0.21	0.20	-0.04
Distance	0.29	-0.16	1.00	0.43	0.30	0.13	0.26	0.03	0.10	0.25	-0.13	0.24	-0.05
Postcode	0.06	0.11	0.43	1.00	0.06	0.11	0.05	0.02	0.06	0.03	-0.41	0.45	0.06
Bedroom2	0.94	0.48	0.30	0.06	1.00	0.58	0.41	0.03	0.12	-0.05	0.02	0.10	-0.08
Bathroom	0.59	0.47	0.13	0.11	0.58	1.00	0.32	0.04	0.11	0.15	-0.07	0.12	-0.05
Car	0.41	0.24	0.26	0.05	0.41	0.32	1.00	0.03	0.10	0.10	0.00	0.06	-0.02
Landsize	0.03	0.04	0.03	0.02	0.03	0.04	0.03	1.00	0.50	0.04	0.01	0.01	-0.01
BuildingArea	0.12	0.09	0.10	0.06	0.12	0.11	0.10	0.50	1.00	0.02	0.04	-0.02	-0.03
YearBuilt	-0.07	-0.32	0.25	0.03	-0.05	0.15	0.10	0.04	0.02	1.00	0.06	0.00	0.01
Lattitude	0.02	-0.21	-0.13	-0.41	0.02	-0.07	0.00	0.01	0.04	0.06	1.00	-0.36	0.05

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude	Propertycount
Longtitude	0.10	0.20	0.24	0.45	0.10	0.12	0.06	0.01	-0.02	0.00	-0.36	1.00	0.07
Propertycount	-0.08	-0.04	-0.05	0.06	-0.08	-0.05	-0.02	-0.01	-0.03	0.01	0.05	0.07	1.00

Honors Pledge:

As a student of the Dr. Robert B. Pamplin Jr. School of Business I have read and strive to uphold the University's Code of Academic Integrity and promote ethical behavior. In doing so, I pledge on my honor that I have not given, received, or used any unauthorized materials or assistance on this examination or assignment. I further pledge that I have not engaged in cheating, forgery, or plagiarism and I have cited all appropriate sources.

Student Signature: Areej Mulla