Week4_Assignment_Housing

Henry Pham

2023-04-08

```
# Set the working directory to the root of your DSC 520 directory

getwd()
```

```
## [1] "C:/R/DSC520"
```

```
dir("C://R//DSC520//data")
```

```
## [1] "acs-14-1yr-s0201.csv"           "G04ResultsDetail2004-11-02.xls"
## [3] "r4ds"                           "scores.csv"
## [5] "tidynomicon"                    "week-6-housing.csv"
```

```
if (!file.exists("data"))
{
  # set data working directory
  setwd("C://R//DSC520//data")
}

# Load the data
house_data <- read.csv("week-6-housing.csv")
head(house_data)
```

```
##   Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1  1/3/2006     698000           1               3                    R1
## 2  1/3/2006     649990           1               3                    R1
## 3  1/3/2006     572500           1               3                    R1
## 4  1/3/2006     420000           1               3                    R1
## 5  1/3/2006     369900           1               3           15       R1
## 6  1/3/2006     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat
building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139
9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731
9
## 3 13315 174TH AVE NE 98052            REDMOND -122.1085 47.71986
8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914
8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748
7
## 6   8101 229TH DR NE 98053            REDMOND -122.0341 47.67545
7
```

```
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot
prop_type
## 1               0       2003              0             R4      6635
R
## 2               1       2006              0             R4      5570
R
## 3               1       1987              0             R6      8444
R
## 4               1       1968              0             R4      9600
R
## 5               1       1980              0             R6      7526
R
## 6               1       2005              0          URPSO      7280
R
##   present_use
## 1           2
## 2           2
## 3           2
## 4           2
## 5           2
## 6           2
```

```r
# Use the apply function on a variable in your dataset
# Compute the average sale price of the houses
avg_price <- mean(house_data$Sale.Price)
cat("Average sale price:", avg_price, "\n")
```

```
## Average sale price: 660737.7
```

```r
# Use the aggregate function on a variable in your dataset
# Compute the average sale price by year of sale
yearly_avg_price <- aggregate(Sale.Price ~ year_built, data = house_data, FUN
= mean)
cat("Yearly average sale price:\n")
```

```
## Yearly average sale price:
```

```r
print(yearly_avg_price)
```

```
##     year_built Sale.Price
## 1         1900   394499.7
## 2         1903   430000.0
## 3         1905   620000.0
## 4         1906   550000.0
```

```
## 5         1909     1070.0
## 6         1910   150000.0
## 7         1912   619666.7
## 8         1913   457500.0
## 9         1914   835000.0
## 10        1915   228150.0
## 11        1916   350000.0
## 12        1918  1033833.3
## 13        1919   476800.0
## 14        1920   509083.3
## 15        1922   424587.5
## 16        1923   300000.0
## 17        1924   649500.0
## 18        1925   387250.0
## 19        1926   318333.3
## 20        1927  1173750.0
## 21        1928   520000.0
## 22        1929  1242500.0
## 23        1930   402191.7
## 24        1931   168828.5
## 25        1932   588146.2
## 26        1933   440500.0
## 27        1934   750000.0
## 28        1935  1616333.3
## 29        1936   485182.3
## 30        1937   846594.3
## 31        1938  1675500.0
## 32        1939   520000.0
## 33        1940   681411.1
## 34        1941   348517.2
## 35        1942   343561.0
## 36        1943   501200.0
## 37        1944   335626.5
## 38        1945   354330.9
## 39        1946   626875.0
## 40        1947   390378.7
## 41        1948   713522.6
## 42        1949   485525.4
## 43        1950   360315.0
## 44        1951   583972.0
## 45        1952   786191.7
## 46        1953   463553.7
## 47        1954   657591.3
## 48        1955   563706.3
## 49        1956   625561.5
## 50        1957   511411.5
## 51        1958   428233.8
## 52        1959   468616.6
## 53        1960   451005.4
## 54        1961   581580.0
```

```
## 55        1962   515826.5
## 56        1963   508518.7
## 57        1964   566355.5
## 58        1965   484418.3
## 59        1966   478482.7
## 60        1967   497566.3
## 61        1968   446930.1
## 62        1969   444439.2
## 63        1970   419788.3
## 64        1971   442688.5
## 65        1972   552177.1
## 66        1973   556947.5
## 67        1974   591669.8
## 68        1975   535944.1
## 69        1976   502248.9
## 70        1977   494102.5
## 71        1978   512763.1
## 72        1979   545454.4
## 73        1980   546471.3
## 74        1981   539075.9
## 75        1982   586006.0
## 76        1983   527091.5
## 77        1984   561059.2
## 78        1985   599990.3
## 79        1986   583642.8
## 80        1987   662669.3
## 81        1988   774747.3
## 82        1989   762350.0
## 83        1990   837696.4
## 84        1991   807708.3
## 85        1992   630408.5
## 86        1993   700939.1
## 87        1994   752529.6
## 88        1995   694532.9
## 89        1996   689408.3
## 90        1997   738764.9
## 91        1998   791991.1
## 92        1999  1016032.6
## 93        2000   829172.7
## 94        2001   695094.1
## 95        2002   599826.2
## 96        2003   645323.4
## 97        2004   632882.3
## 98        2005   647728.2
## 99        2006   692548.0
## 100       2007   664465.2
## 101       2008   866785.5
## 102       2009   756906.6
## 103       2010   649072.9
## 104       2011   677745.2
```

```
## 105       2012   922800.5
## 106       2013   912130.4
## 107       2014   825761.6
## 108       2015   888559.7
## 109       2016   893875.0
```

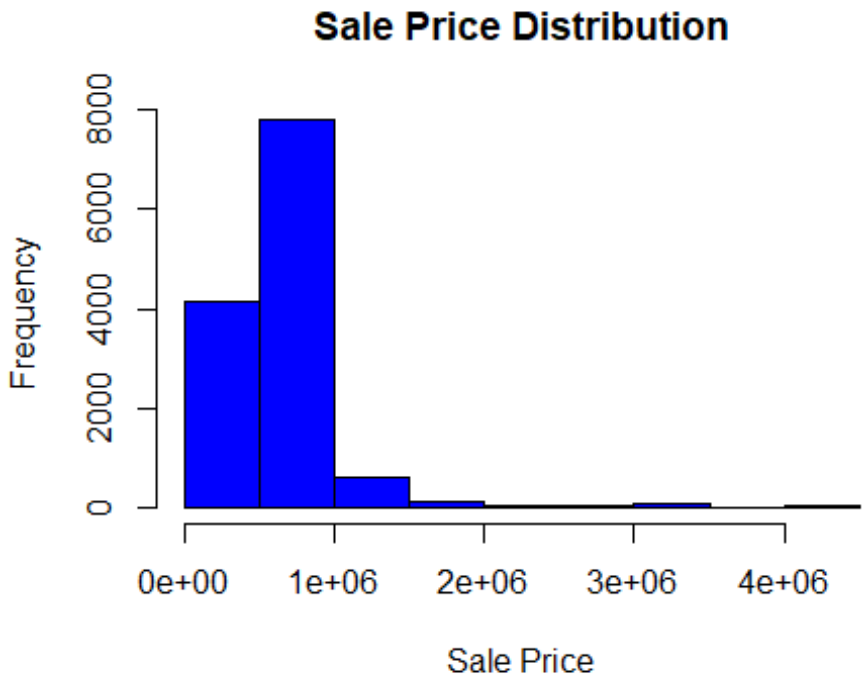
```r
# Use the plyr function on a variable in your dataset
library(plyr)
# Split the data by zip5
# Compute the average sale price and number of houses sold for each zip5
# Combine the results into a new data frame
neighborhood_stats <- ddply(house_data, .(zip5), summarise, AvgPrice =
mean(Sale.Price), NumSales = length(Sale.Price))
cat("Neighborhood statistics:\n")
```

```
## Neighborhood statistics:
```

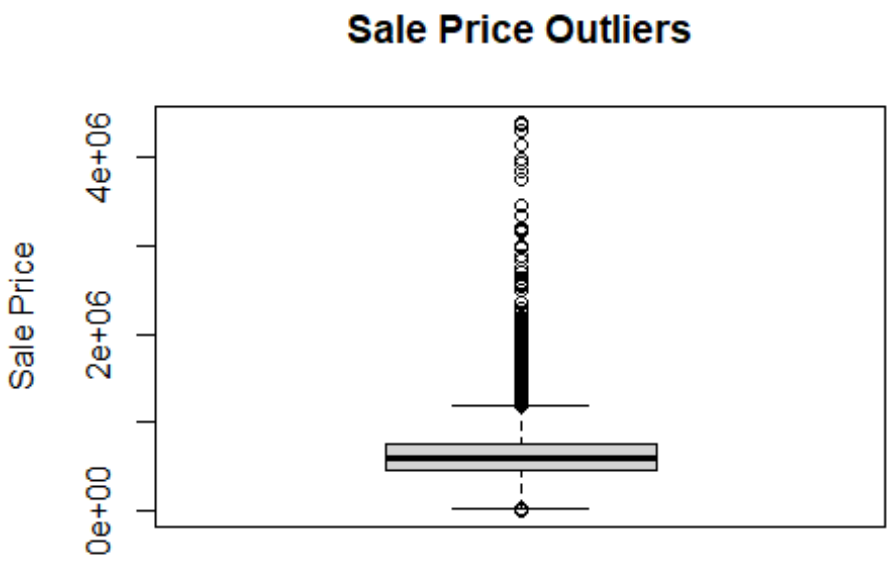
```r
print(neighborhood_stats)
```

```
##    zip5 AvgPrice NumSales
## 1 98052 649375.4     7452
## 2 98053 672623.7     5339
## 3 98059 645000.0        1
## 4 98074 951543.8       73
```

```r
# Check distributions of the data
# Create a histogram of the sale prices
hist(house_data$Sale.Price, main = "Sale Price Distribution", xlab = "Sale
Price", ylab = "Frequency", col = "blue")
```




```r
# Identify if there are any outliers
# Create a boxplot of the sale prices
boxplot(house_data$Sale.Price, main = "Sale Price Outliers", ylab = "Sale Price")
```




```r
# Create at least 2 new variables
# Create a variable for the total rooms of the house
house_data$totalrooms <- house_data$bedrooms + house_data$bath_full_count +
house_data$bath_half_count + house_data$bath_3qtr_coun
# Create a variable for the pct of bedrooms
house_data$Pctbedrooms <- house_data$bedrooms / house_data$totalrooms
```