Assignment 8

Surenther

2024-10-18

Load XLSX file

```
#Load XLSX
library(readxl)
mydata <- read_excel("week-6-housing.xlsx", 1)
```

Data Cleanup

```
library(dplyr,warn.conflicts = FALSE)
# Remove unwanted Columns
housing_data <- select(mydata,-c(sale_reason,sale_instrument,sale_warning,sitetype,
                                 ctyname,present_use))
```

Explanation
The columns such as Sale Reason, Sale Instrument, Sale Warning, and Present Use contain references without any accompanying explanation, making their inclusion unclear. Additionally, columns like Sitetype and City Name are duplicates of Postalcty and Prop Type, respectively. Therefore, these columns have been removed for clarity and to avoid redundancy.

```
# Rename Columns
housing_data <- rename(housing_data,c(sale_date='Sale Date',sale_price='Sale Price',
                      address=addr_full,zip=zip5,city=postalctyn,longtitude=lon,
                      lattitude=lat))
```

Explanation
Rename the columns by converting all characters to lowercase and assigning more meaningful names for clarity and consistency.

```
# Add Year Column
housing_data <- housing_data %>%
  mutate('sale_year' = format(housing_data$'sale_date','%Y'))
```

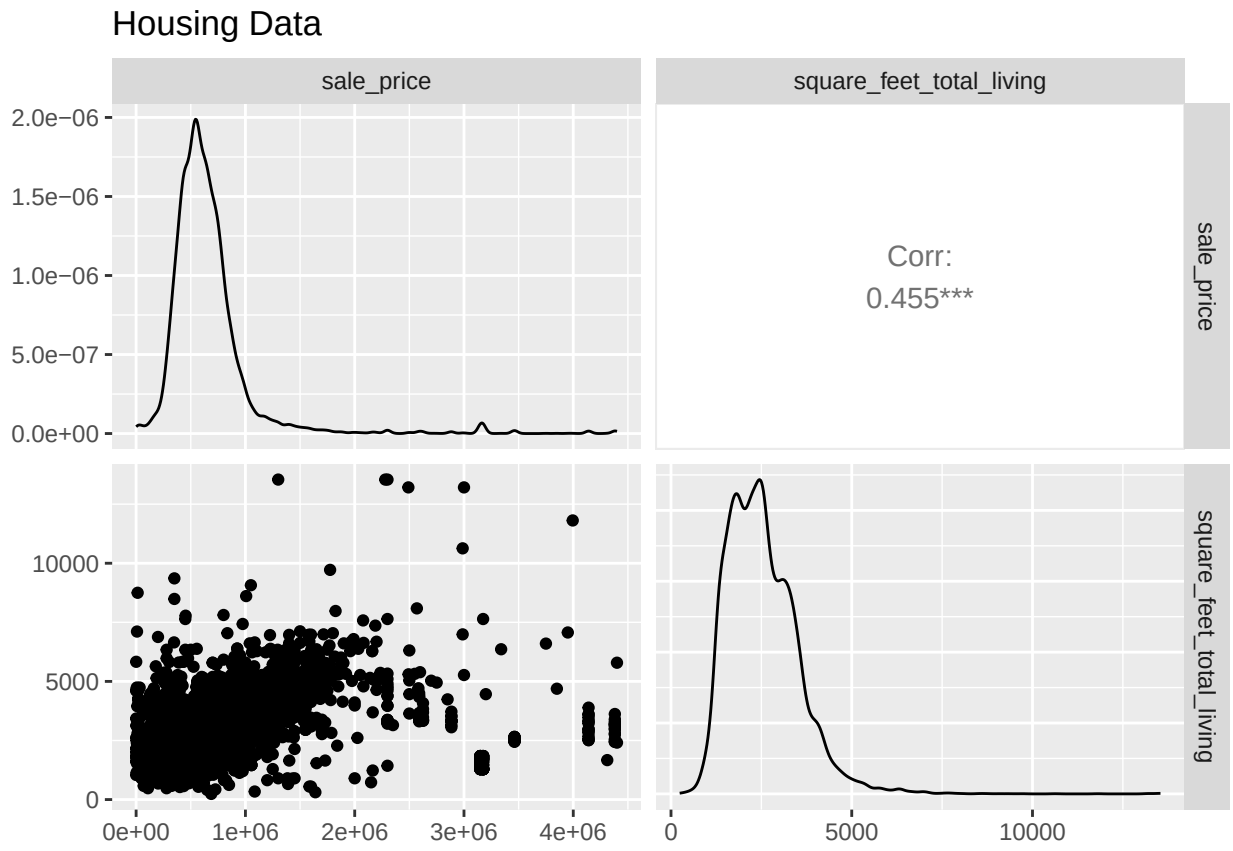
Explanation
A new "year" column was generated from the sale date to facilitate year-based calculations.

Linear regression model

```r
#Identify Relationship between Sale Price and Sq_ft
library("ggplot2")
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(housing_data[,c(2,9)], title="Housing Data")
```



Findings

While a relationship exists between square footage and sale price, the correlation coefficient of 0.455 suggests a relatively weak association.

```r
#Linear Regression Model
lm_sale_sqft <- lm(sale_price ~ square_feet_total_living, data = housing_data)
#Summary of Lm
summary(lm_sale_sqft)
```

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living, data = housing_data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1800136  -120257   -41547    44028  3811745
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.891e+05  8.745e+03   21.62   <2e-16 ***
## square_feet_total_living 1.857e+02  3.208e+00   57.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360200 on 12863 degrees of freedom
## Multiple R-squared:  0.2066, Adjusted R-squared:  0.2066
## F-statistic:  3351 on 1 and 12863 DF,  p-value: < 2.2e-16
```

Findings

Coefficients

(Intercept): This is the predicted sale_price when square_feet_total_living is 0. In this case, it's $189,100.

square_feet_total_living: This represents the change in sale_price for every one-unit increase in square_feet_total_living. So, for each additional square foot of living area, the sale_price increases by $185.70. Both coefficients are statistically significant (indicated by the Pr(>|t|) values being much smaller than 0.05).

Model Fit

Residual standard error: This is the average distance between the actual sale_price and the predicted sale_price by the model. In this case, it's $360,200.

Multiple R-squared: This measures the proportion of variance in sale_price explained by the model. Here, it's 0.2066, meaning that about 20.66% of the variation in sale_price can be explained by square_feet_total_living.

Adjusted R-squared: This is a similar metric to R-squared but penalizes for the number of predictors in the model. In this case, it's also 0.2066, indicating that adding more predictors might not significantly improve the model's fit.

F-statistic and p-value: These test the overall significance of the model. A p-value of less than 0.05 suggests that the model is statistically significant, meaning that at least one predictor (in this case, square_feet_total_living) is significantly related to the sale_price.

In Summary

The model indicates that there is a significant positive relationship between sale_price and square_feet_total_living. However, only about 20% of the variation in sale_price can be explained by this variable. This suggests that other factors may also influence the sale_price.