House Sales Data Analysis

sirius_ife

2024-03-06

Importing the Dataset

```
# Read the Excel file
HouseData <- read_excel("HouseData.xlsx")

# Convert to tibble
house <- as_tibble(HouseData)
```

ETL (Extract, Transform & Load)

```
# Extracting and transforming date column
house$date <- substr(house$date, 1, 8)
house$date <- ymd(house$date)

# Converting waterfront to factor
house$waterfront <- as_factor(house$waterfront)
```

Exploratory Data Analysis (EDA)

```
# Summary statistics
summary(house)
```

```
##        id                 date                price            bedromms
##  Min.   :1.000e+06   Min.   :2014-05-02   Min.   :  75000   Min.   :0.000
##  1st Qu.:2.125e+09   1st Qu.:2014-07-22   1st Qu.: 319950   1st Qu.:3.000
##  Median :3.905e+09   Median :2014-10-16   Median : 445000   Median :3.000
##  Mean   :4.591e+09   Mean   :2014-10-29   Mean   : 500270   Mean   :3.343
##  3rd Qu.:7.335e+09   3rd Qu.:2015-02-17   3rd Qu.: 625000   3rd Qu.:4.000
##  Max.   :9.900e+09   Max.   :2015-05-27   Max.   :1495000   Max.   :7.000
##    bathrooms      sqft_living      sqft_lot          floors      waterfront
##  Min.   :0.000   Min.   : 290   Min.   :    520   Min.   :1.000   No :20976
##  1st Qu.:1.500   1st Qu.:1410   1st Qu.:   5001   1st Qu.:1.000   Yes:   84
##  Median :2.250   Median :1890   Median :   7554   Median :1.000
##  Mean   :2.078   Mean   :2019   Mean   :  14743   Mean   :1.485
##  3rd Qu.:2.500   3rd Qu.:2490   3rd Qu.:  10454   3rd Qu.:2.000
##  Max.   :6.750   Max.   :7480   Max.   :1651359   Max.   :3.500
##      view            condition             grade          yr_built
##  Length:21060       Length:21060       Min.   : 1.00   Min.   :1900
##  Class :character   Class :character   1st Qu.: 7.00   1st Qu.:1951
##  Mode  :character   Mode  :character   Median : 7.00   Median :1975
##                                        Mean   : 7.59   Mean   :1971
##                                        3rd Qu.: 8.00   3rd Qu.:1996
##                                        Max.   :12.00   Max.   :2015
##     zipcode           lat             long
##  Min.   :98001   Min.   :47.16   Min.   :-122.5
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
##  Median :98065   Median :47.57   Median :-122.2
##  Mean   :98078   Mean   :47.56   Mean   :-122.2
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :98199   Max.   :47.78   Max.   :-121.3
```

```
# Descriptive statistics
describe(house)
```

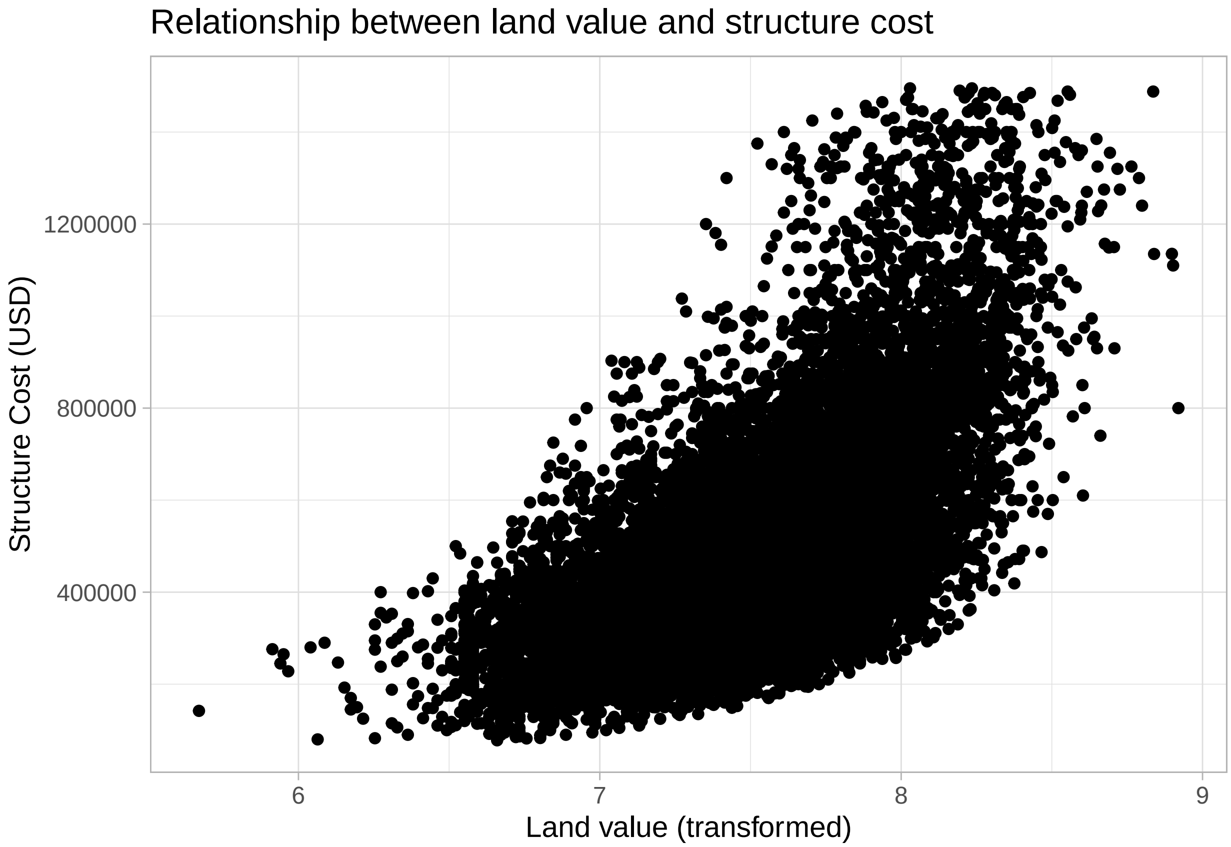
```
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
```

```
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
```

```
##             vars     n          mean           sd       median       trimmed
## id             1 21060 4591326900.25 2.877902e+09  3.90503e+09 4514025585.07
## date           2 21060           NaN           NA           NA           NaN
## price          3 21060     500269.62 2.465775e+05  4.45000e+05     470079.13
## bedrooms       4 21060          3.34 8.800000e-01  3.00000e+00          3.32
## bathrooms      5 21060          2.08 7.300000e-01  2.25000e+00          2.05
## sqft_living    6 21060       2019.49 8.220200e+02  1.89000e+03       1947.61
## sqft_lot       7 21060      14743.01 3.995697e+04  7.55350e+03       8104.70
## floors         8 21060          1.48 5.400000e-01  1.00000e+00          1.44
## waterfront*    9 21060          1.00 6.000000e-02  1.00000e+00          1.00
## view*         10 21060          4.75 8.800000e-01  5.00000e+00          5.00
## condition*    11 21060          1.85 1.260000e+00  1.00000e+00          1.62
## grade         12 21060          7.59 1.100000e+00  7.00000e+00          7.54
## yr_built      13 21060       1970.92 2.927000e+01  1.97500e+03       1972.99
## zipcode       14 21060      98078.38 5.341000e+01  9.80650e+04      98075.23
## lat           15 21060         47.56 1.400000e-01  4.75700e+01         47.57
## long          16 21060       -122.21 1.400000e-01 -1.22230e+02       -122.23
##                      mad        min           max        range  skew kurtosis
## id          3.578967e+09 1000102.00 9900000190.00 9.899000e+09  0.24    -1.27
## date                  NA        Inf          -Inf         -Inf    NA       NA
## price       2.149770e+05   75000.00    1495000.00 1.420000e+06  1.22     1.62
## bedrooms    1.480000e+00       0.00          7.00 7.000000e+00  0.32     0.68
## bathrooms   7.400000e-01       0.00          6.75 6.750000e+00  0.26     0.24
## sqft_living 7.857800e+02     290.00       7480.00 7.190000e+03  0.96     1.43
## sqft_lot    3.812510e+03     520.00    1651359.00 1.650839e+06 13.23   304.05
## floors      0.000000e+00       1.00          3.50 2.500000e+00  0.65    -0.44
## waterfront* 0.000000e+00       1.00          2.00 1.000000e+00 15.74   245.69
## view*       0.000000e+00       1.00          5.00 4.000000e+00 -3.63    11.89
## condition*  0.000000e+00       1.00          5.00 4.000000e+00  1.22     0.34
## grade       1.480000e+00      1.00        12.00 1.100000e+01  0.60     0.87
## yr_built    3.410000e+01   1900.00      2015.00 1.150000e+02 -0.47    -0.65
## zipcode     6.227000e+01  98001.00     98199.00 1.980000e+02  0.40    -0.86
## lat         1.600000e-01     47.16        47.78 6.200000e-01 -0.46    -0.72
## long        1.500000e-01   -122.52      -121.32 1.200000e+00  0.88     1.01
##                      se
## id          19831102.24
## date                 NA
## price           1699.12
## bedrooms           0.01
## bathrooms          0.01
## sqft_living        5.66
## sqft_lot         275.34
## floors             0.00
## waterfront*        0.00
## view*              0.01
## condition*         0.01
## grade              0.01
## yr_built           0.20
## zipcode            0.37
## lat                0.00
## long               0.00
```

Checking Correlation

```
# Plotting relationship between price and log(sqft_living)
house %>%
  subset(year(date) == 2014) %>%
  ggplot(aes(y = price, x = log(sqft_living))) +
  geom_point() +
  theme_light() +
  labs(x = "Land value (transformed)",
       y = "Structure Cost (USD)",
       title = "Relationship between land value and structure cost")
```



```r
# Correlation matrix
cor(house[, c('price', 'sqft_living', 'sqft_lot')])
```

```
##                  price sqft_living   sqft_lot
## price       1.00000000   0.6558679 0.08708296
## sqft_living 0.65586791   1.0000000 0.16223459
## sqft_lot    0.08708296   0.1622346 1.00000000
```

Both sqft_living and sqft_lot are positively correlated with price.

Simple Regression Model: Price / sqft_living

```r
lm1 <- lm(price ~ sqft_living, data = house)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -774560 -131573  -18035  100573  955205
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102959.86    3402.27   30.26   <2e-16 ***
## sqft_living    196.74       1.56  126.08   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 186100 on 21058 degrees of freedom
## Multiple R-squared:  0.4302, Adjusted R-squared:  0.4301
## F-statistic: 1.59e+04 on 1 and 21058 DF,  p-value: < 2.2e-16
```

Multiple Regression Model: Price / sqft_living + sqft_lot

```
lm2 <- lm(price ~ sqft_living + sqft_lot, data = house)
summary(lm2)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -776536 -131430  -18116  100337  954997
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.028e+05  3.401e+03  30.227  < 2e-16 ***
## sqft_living  1.977e+02  1.581e+00 125.060  < 2e-16 ***
## sqft_lot    -1.225e-01  3.252e-02  -3.765 0.000167 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 186100 on 21057 degrees of freedom
## Multiple R-squared:  0.4305, Adjusted R-squared:  0.4305
## F-statistic:  7960 on 2 and 21057 DF,  p-value: < 2.2e-16
```

lm2 has a higher R-squared value of 0.4305, indicating that it is a better model for predicting price.