Project 1

Brandon Keck

2024-10-09

```
knitr::opts_chunk$set(echo = TRUE)
```

```
#install.packages("corrplot")
#install.packages("scales")
```

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.2.3
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
train <- read_csv("~/Desktop/house-prices-advanced-regression-techniques/train.csv")
```

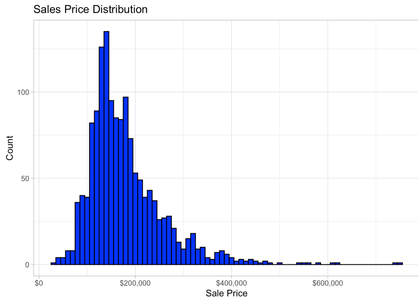
```
## Rows: 1460 Columns: 81
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## 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.
```

```
train2 <- select(train, SalePrice, LotArea, GrLivArea, YearBuilt, OverallQual, OverallCond, Neighborhood)
```

```
summary(train2)
```

```
##    SalePrice         LotArea         GrLivArea      YearBuilt
##  Min.   : 34900   Min.   :  1300   Min.   : 334   Min.   :1872
##  1st Qu.:129975   1st Qu.:  7554   1st Qu.:1130   1st Qu.:1954
##  Median :163000   Median :  9478   Median :1464   Median :1973
##  Mean   :180921   Mean   : 10517   Mean   :1515   Mean   :1971
##  3rd Qu.:214000   3rd Qu.: 11602   3rd Qu.:1777   3rd Qu.:2000
##  Max.   :755000   Max.   :215245   Max.   :5642   Max.   :2010
##   OverallQual      OverallCond    Neighborhood
##  Min.   : 1.000   Min.   :1.000   Length:1460
##  1st Qu.: 5.000   1st Qu.:5.000   Class :character
##  Median : 6.000   Median :5.000   Mode  :character
##  Mean   : 6.099   Mean   :5.575
##  3rd Qu.: 7.000   3rd Qu.:6.000
##  Max.   :10.000   Max.   :9.000
```

```
ggplot(train2, aes(x = SalePrice)) +
  geom_histogram(binwidth = 10000, fill = "blue", color = "black") +
  labs(title = "Sales Price Distribution", x = "Sale Price", y = "Count") +
  scale_x_continuous(labels = dollar) +
  theme_light()
```




```
ggplot(train2, aes(x = GrLivArea, y = SalePrice)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Living Area vs. Sale Price", x = "Living Area (sq ft)", y = "Sales Price ($)") +
  scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = dollar) +
  theme_light()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


```
model <- lm(SalePrice ~ GrLivArea + OverallQual + YearBuilt, data = train2)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea + OverallQual + YearBuilt,
##     data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -393773  -22639   -2424   18437  290554
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.053e+06  8.376e+04  -12.57   <2e-16 ***
## GrLivArea    6.209e+01  2.581e+00   24.06   <2e-16 ***
## OverallQual  2.520e+04  1.172e+03   21.50   <2e-16 ***
## YearBuilt    5.001e+02  4.409e+01   11.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40750 on 1456 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7368
## F-statistic:  1363 on 3 and 1456 DF,  p-value: < 2.2e-16
```

```
ggplot(train2, aes(x = Neighborhood, y = SalePrice, fill = Neighborhood)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Sale Price by Neighborhood", x = "Neighborhood", y = "Sale Price") +
  scale_y_continuous(labels = dollar_format()) +
  theme(legend.position = "none")
```