San Francisco Housing

Bradley West

2022-12-06

Conducting simple regression on San Francisco housing data.

Data Provided by scrapfishies on Kaggle

$$H_o : Sqft\ is\ not\ an\ accurate\ predictor\ of\ Price$$

$$H_a : Sqft\ is\ significant\ predictor\ on\ Housing\ Costs$$

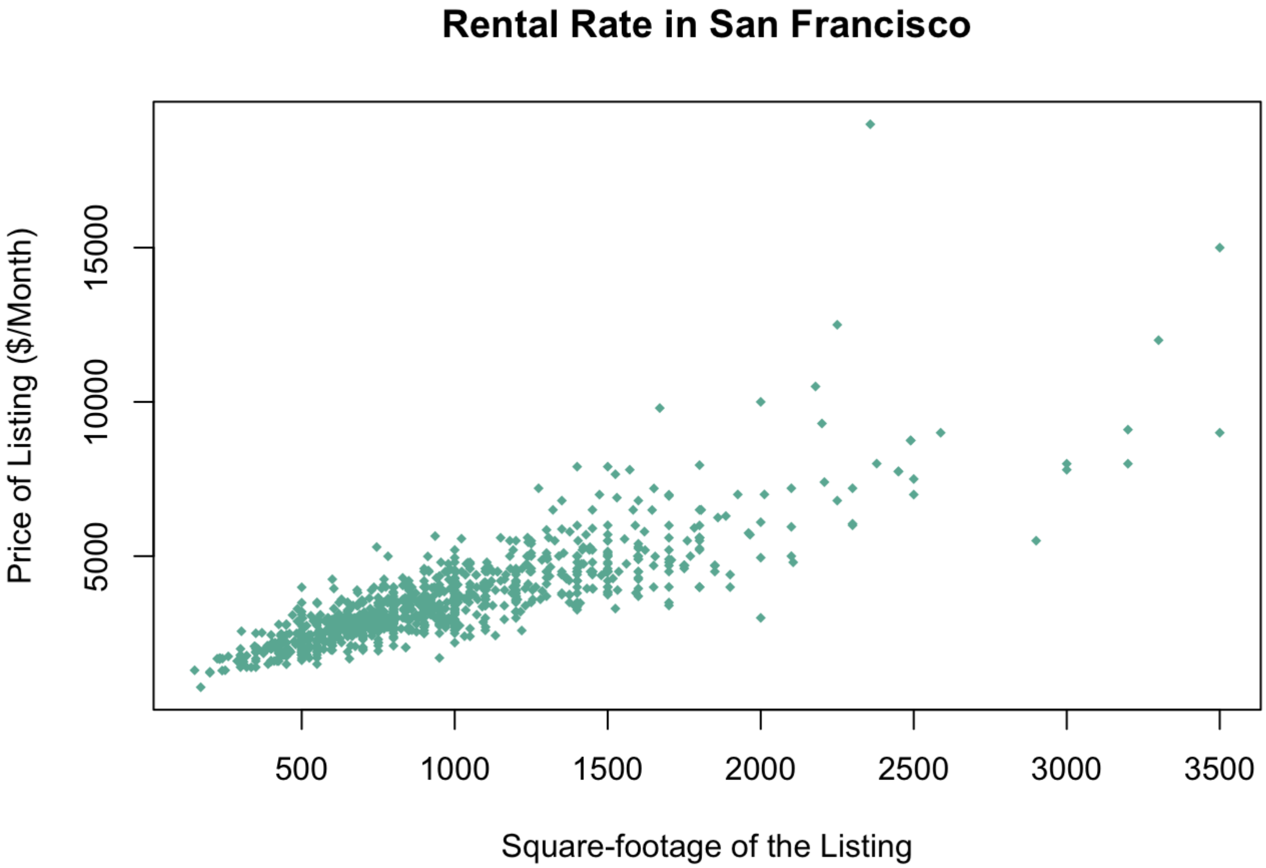
$$\alpha = 0.05$$

```
# Read in and View Data
df <- read.csv('sf_clean.csv')
head(df)
```

	price	sqft	b…	b…	laundry	pets	housing_type	parking	hood_di
	<int>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<
1	6800	1600	2	2.0	(a) in-unit	(d) no pets	(c) multi	(b) protected	
2	3500	550	1	1.0	(a) in-unit	(a) both	(c) multi	(b) protected	
3	5100	1300	2	1.0	(a) in-unit	(a) both	(c) multi	(d) no parking	
4	9000	3500	3	2.5	(a) in-unit	(d) no pets	(c) multi	(b) protected	
5	3100	561	1	1.0	(c) no laundry	(a) both	(c) multi	(d) no parking	
6	3800	800	2	1.0	(b) on-site	(c) cats	(c) multi	(b) protected	

6 rows

Beginning Simple Linear Regression



It seems evident that Square Footage plays a large in the Rental Rates in San Francisco.

```
slr <- lm(price ~ sqft, data = df)
summary(slr)
```

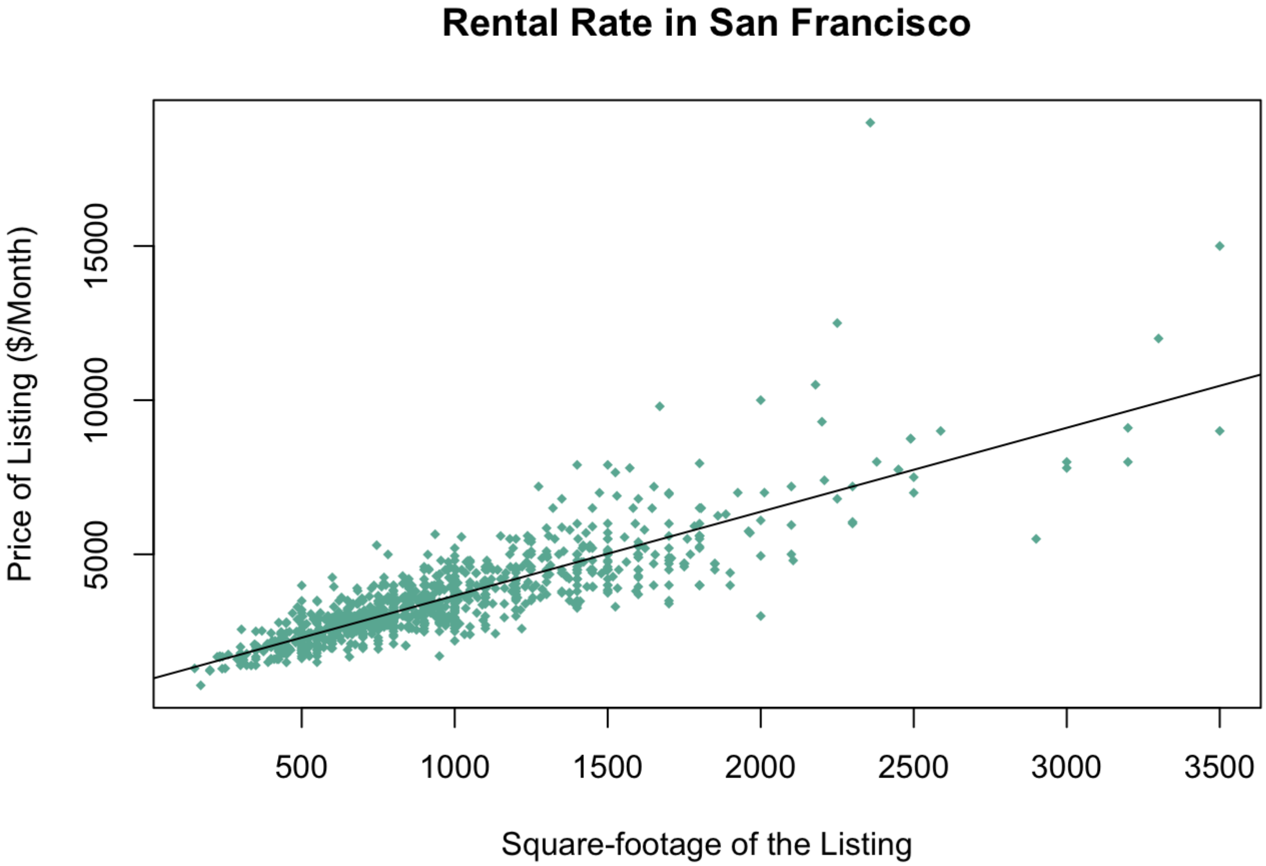
```
## 
## Call:
## lm(formula = price ~ sqft, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3382.2  -402.9   -58.3   341.7 11644.0 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 935.36946   61.81526   15.13   <2e-16 ***
## sqft          2.72293    0.05693   47.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 849.3 on 987 degrees of freedom
## Multiple R-squared:  0.6986, Adjusted R-squared:  0.6983 
## F-statistic:  2288 on 1 and 987 DF,  p-value: < 2.2e-16
```

Based on the p-value produced from our linear model, we can reject the null hypothesis at the 5% significance level. There is sufficient evidence to suggest that sqft is a significant predictor in the housing prices in San Francisco.

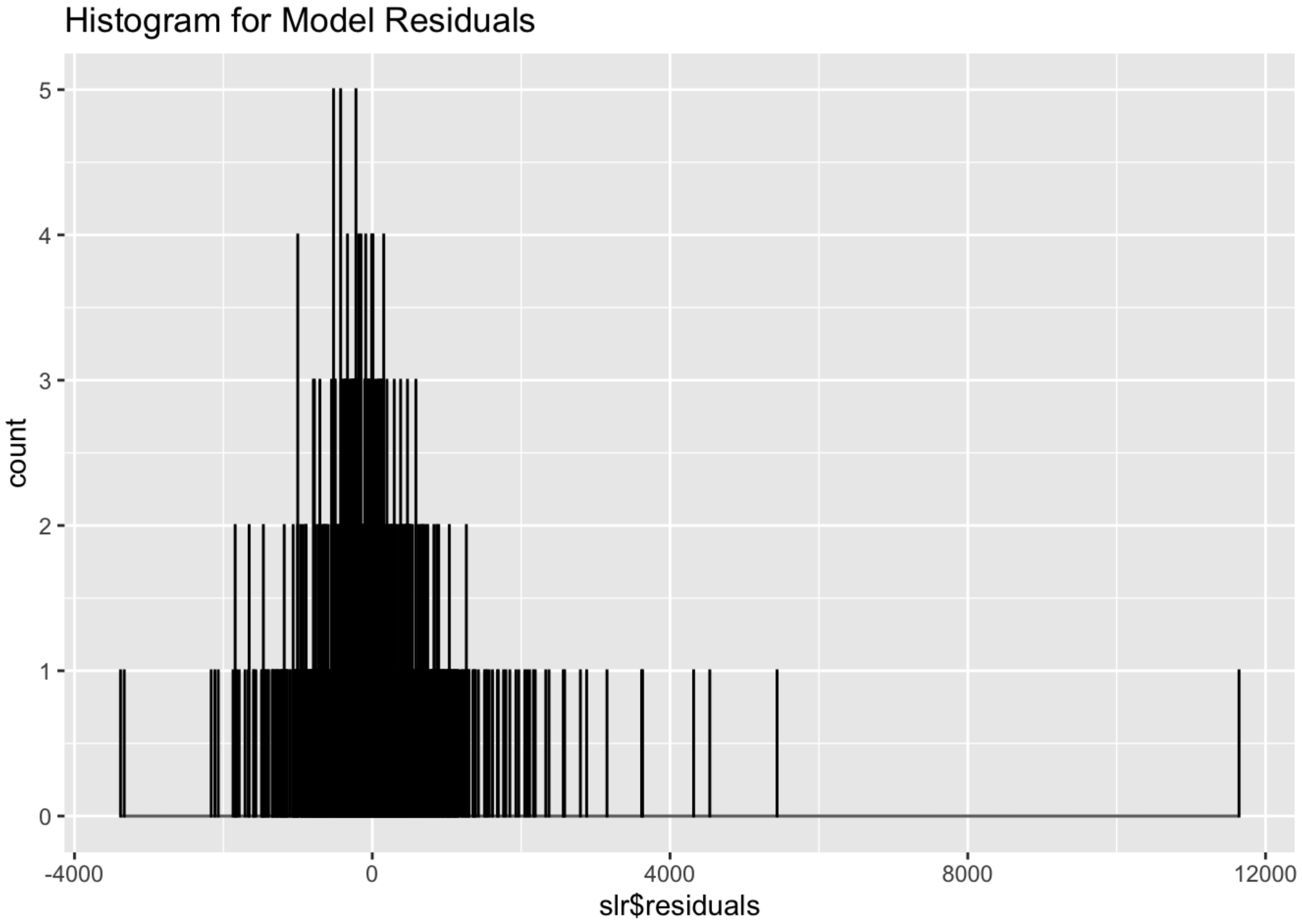
Least Squares =

$$\hat{y} = 2.72293x + 935.36946$$

Plot the linear Model



Plot histogram of the Residuals, showing normal distribution around 0

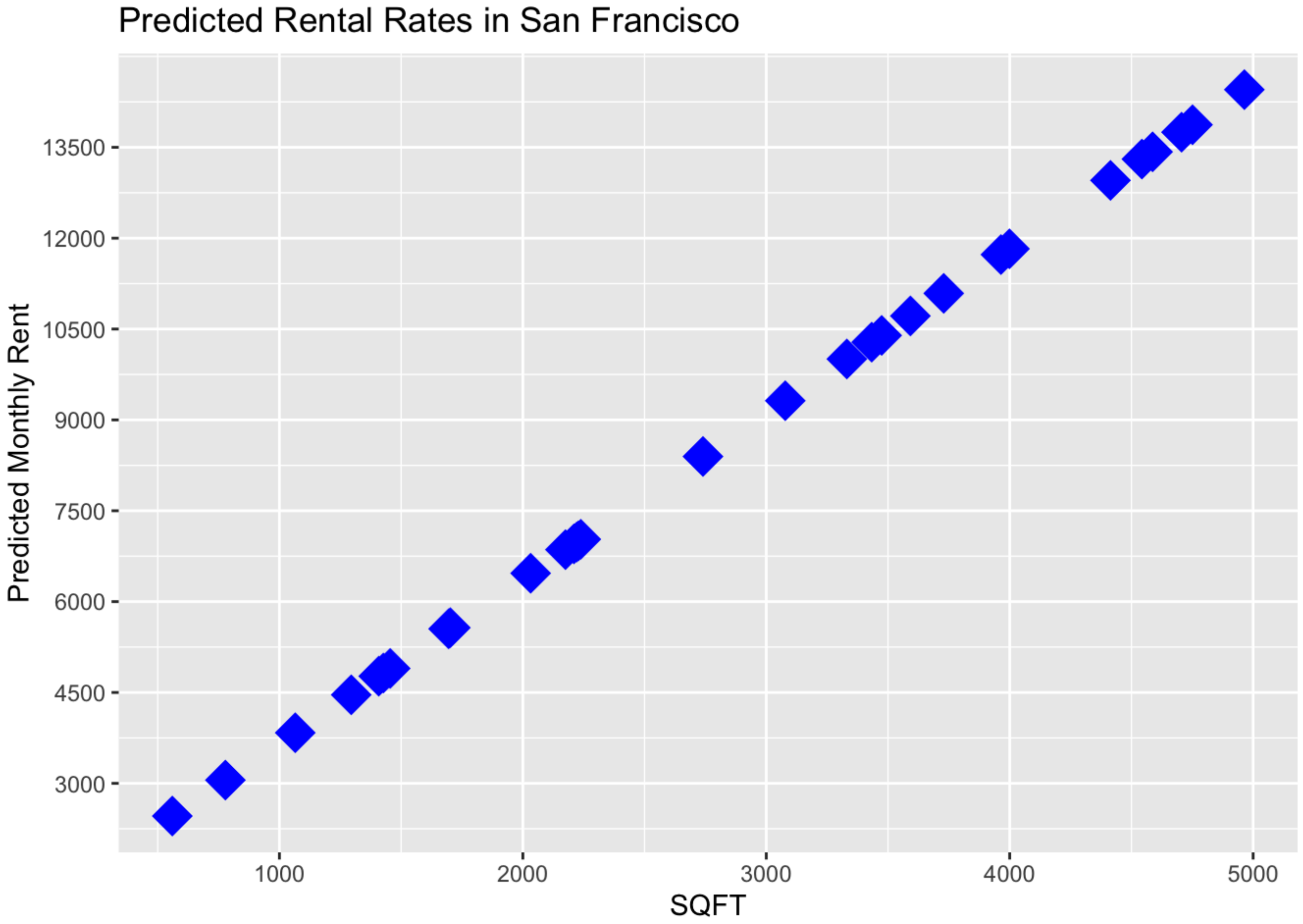

Histogram for Model Residuals
count
5
4
3
2
1
0
-4000
0
4000
8000
12000
slr$residuals

Create price prediction model, given random sqft

```
set.seed(1)
random_sqft <- as.data.frame(
    matrix(
    round(
      runif(n = 30,
            min = 500,
            max = 5000)
    )
  )
)
colnames(random_sqft) <- c('sqft')

predicted_price <- predict(slr, newdata = random_sqft)
predicted_model_df <- cbind(predicted_price, random_sqft)
summary(predicted_model_df)
```

```
##  predicted_price      sqft
##  Min.   : 2460   Min.   : 560
##  1st Qu.: 5556   1st Qu.:1697
##  Median : 7713   Median :2489
##  Mean   : 8541   Mean   :2793
##  3rd Qu.:11569   3rd Qu.:3905
##  Max.   :14452   Max.   :4964
```

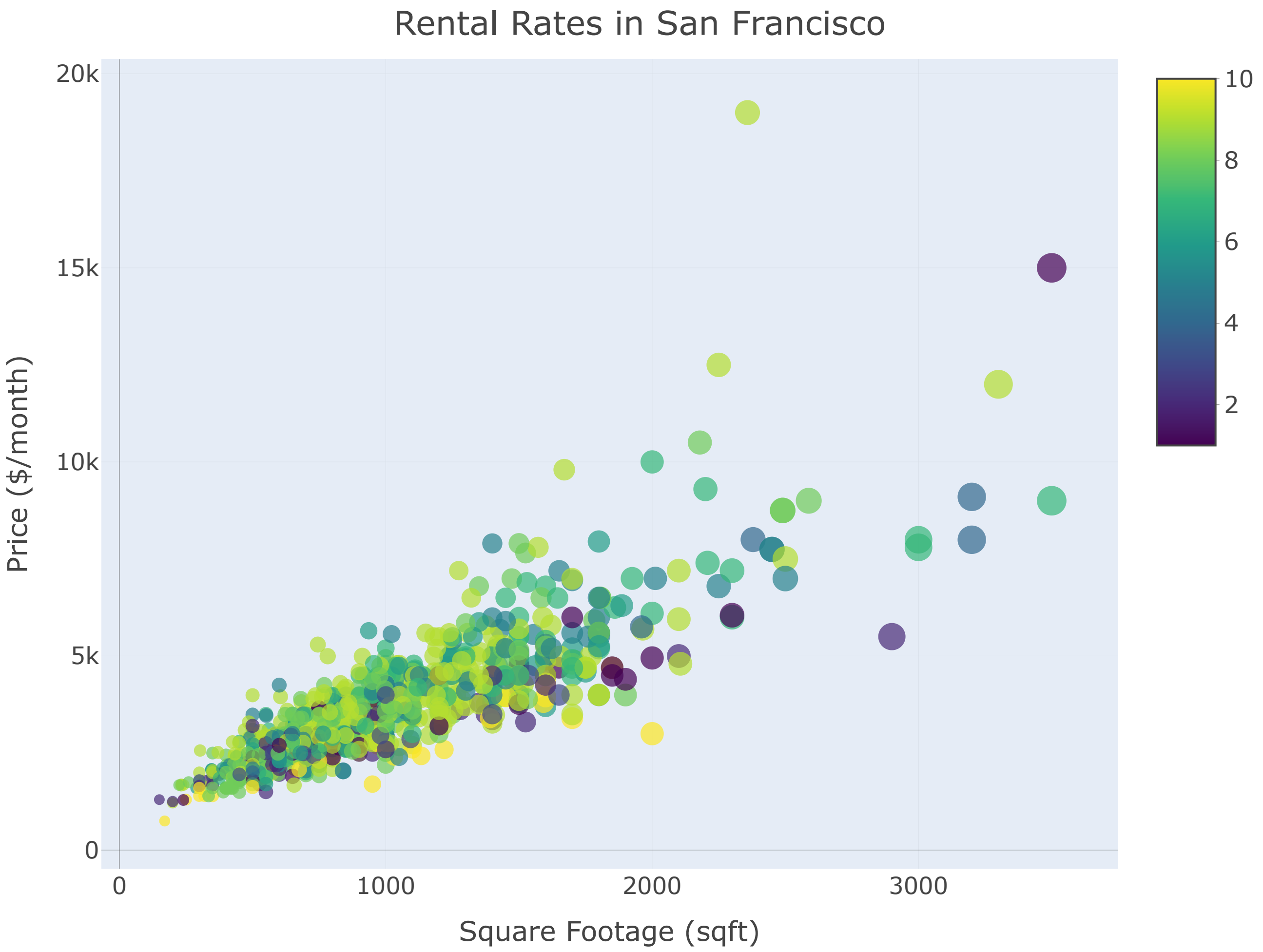

Predicted Rental Rates in San Francisco
Predicted Monthly Rent
13500
12000
10500
9000
7500
6000
4500
3000
1000
2000
3000
4000
5000
SQFT

Create Multiple Linear Regression Model, adding Neighborhood District to the model

```
mlr <- lm(price ~ sqft + hood_district, data = df)
summary(mlr)
```

```
##
## Call:
## lm(formula = price ~ sqft + hood_district, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3473.7  -408.6   -60.0   351.6 11579.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   724.9008   101.9471   7.111 2.22e-12 ***
## sqft            2.7286     0.0568  48.035  < 2e-16 ***
## hood_district  29.0593    11.2118   2.592  0.00969 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 846.8 on 986 degrees of freedom
## Multiple R-squared:  0.7007, Adjusted R-squared:    0.7
## F-statistic:  1154 on 2 and 986 DF,  p-value: < 2.2e-16
```

Plot the new MLR Model


Rental Rates in San Francisco
Price ($/month)
20k
15k
10k
5k
Square Footage (sqft)