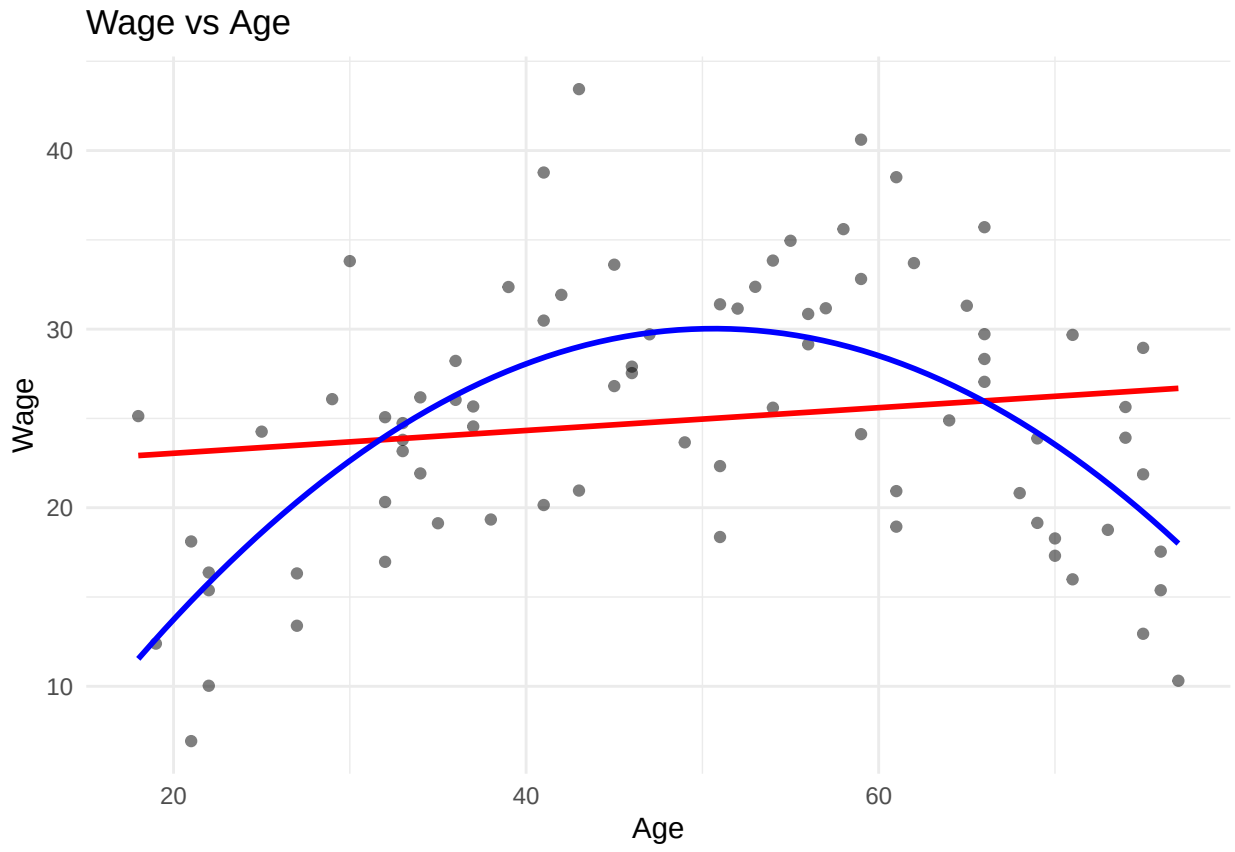
Regression

2025-03-31

```
# 1. Plot Wage against Age and evaluate whether a linear or quadratic model would better capture the re
ggplot(data, aes(x = Age, y = Wage)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = FALSE) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), col = "blue", se = FALSE) +
  labs(title = "Wage vs Age",
       x = "Age",
       y = "Wage") +
  theme_minimal()
```



A quadratic model would better capture the relationship since the points are going in a downward U-shape and not in a linear shape

```r
#Estimate a multiple regression model of Wage using Age and Education as independent (X) variables; ass
linear_model <- lm(Wage ~ Age + Educ, data = data)

# Summary of the model
summary(linear_model)
```

```
##
## Call:
## lm(formula = Wage ~ Age + Educ, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2598  -1.9734   0.3785   2.7700   9.9533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.63808    2.36649   1.115    0.268
## Age          0.04717    0.03062   1.541    0.127
## Educ         1.44101    0.13123  10.981   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.678 on 77 degrees of freedom
## Multiple R-squared:  0.6187, Adjusted R-squared:  0.6088
## F-statistic: 62.47 on 2 and 77 DF,  p-value: < 2.2e-16
```

#we get a R-squared of 0.6088 and the most significant variable being education more than age, and age being impacted by education years the most

```r
#Estimate another multiple regression model of Wage using Age and Education as independent (X) variable

quad_model <- lm(Wage ~ Age + I(Age^2) + Educ, data = data)
summary(quad_model)
```

```
##
## Call:
## lm(formula = Wage ~ Age + I(Age^2) + Educ, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7285 -1.7124 -0.3596  1.9203  7.8048
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.721936   3.022859  -7.517 9.22e-11 ***
## Age           1.350002   0.133973  10.077 1.19e-15 ***
## I(Age^2)     -0.013322   0.001354  -9.840 3.34e-15 ***
## Educ          1.253959   0.089631  13.990  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.123 on 76 degrees of freedom
```

```
## Multiple R-squared:  0.8323, Adjusted R-squared:  0.8257
## F-statistic: 125.7 on 3 and 76 DF,  p-value: < 2.2e-16
```

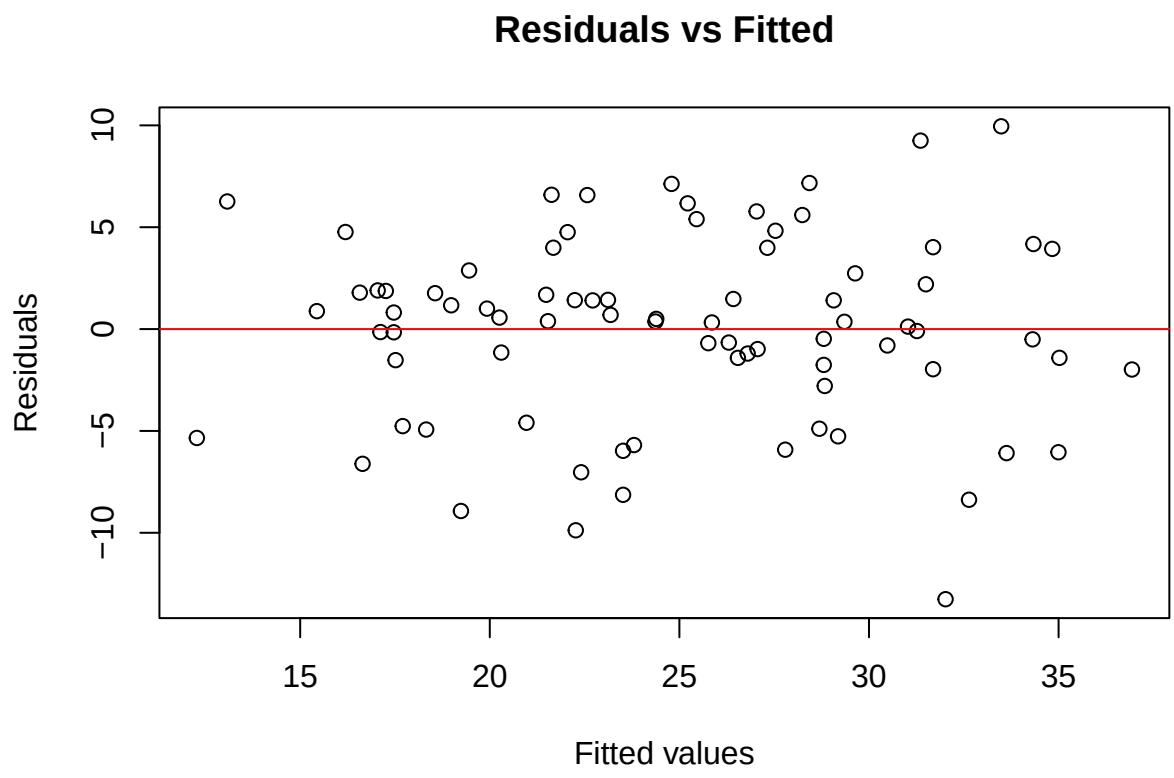
#we can see that now with age having a quadratic relationship, the R-squared value went up in accuracy.

```
#Use the appropriate model to predict hourly wages for someone with 16 years of education and age equal

new_data <- data.frame(Age = c(30, 50, 70), Educ = rep(16, 3))

predicted_wages <- predict(quad_model, newdata = new_data)
predicted_wages
```

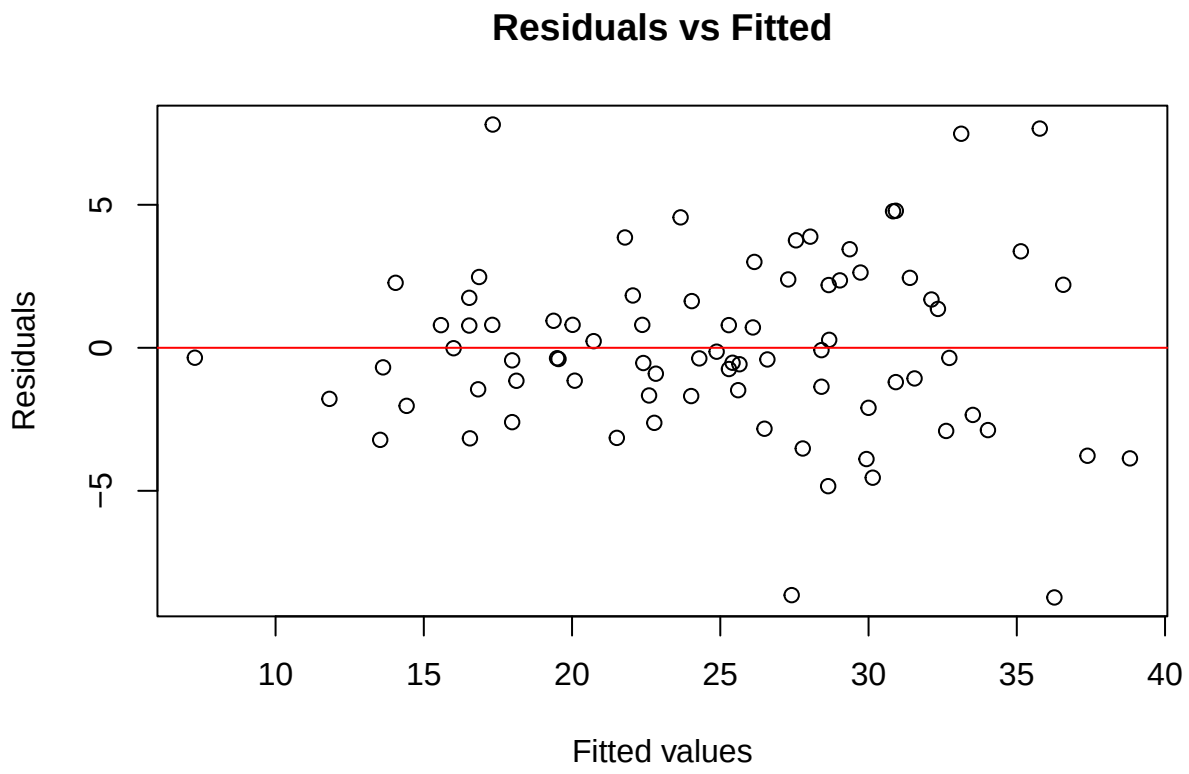
```
##        1        2        3
## 25.85187 31.53709 26.56490
```

#someone with 50 years of age will obtain the highest wages

```
#residuals and fitted values with our models
plot(fitted(linear_model), residuals(linear_model),
     main="Residuals vs Fitted",
     xlab="Fitted values", ylab="Residuals")
abline(h=0, col="red")
```



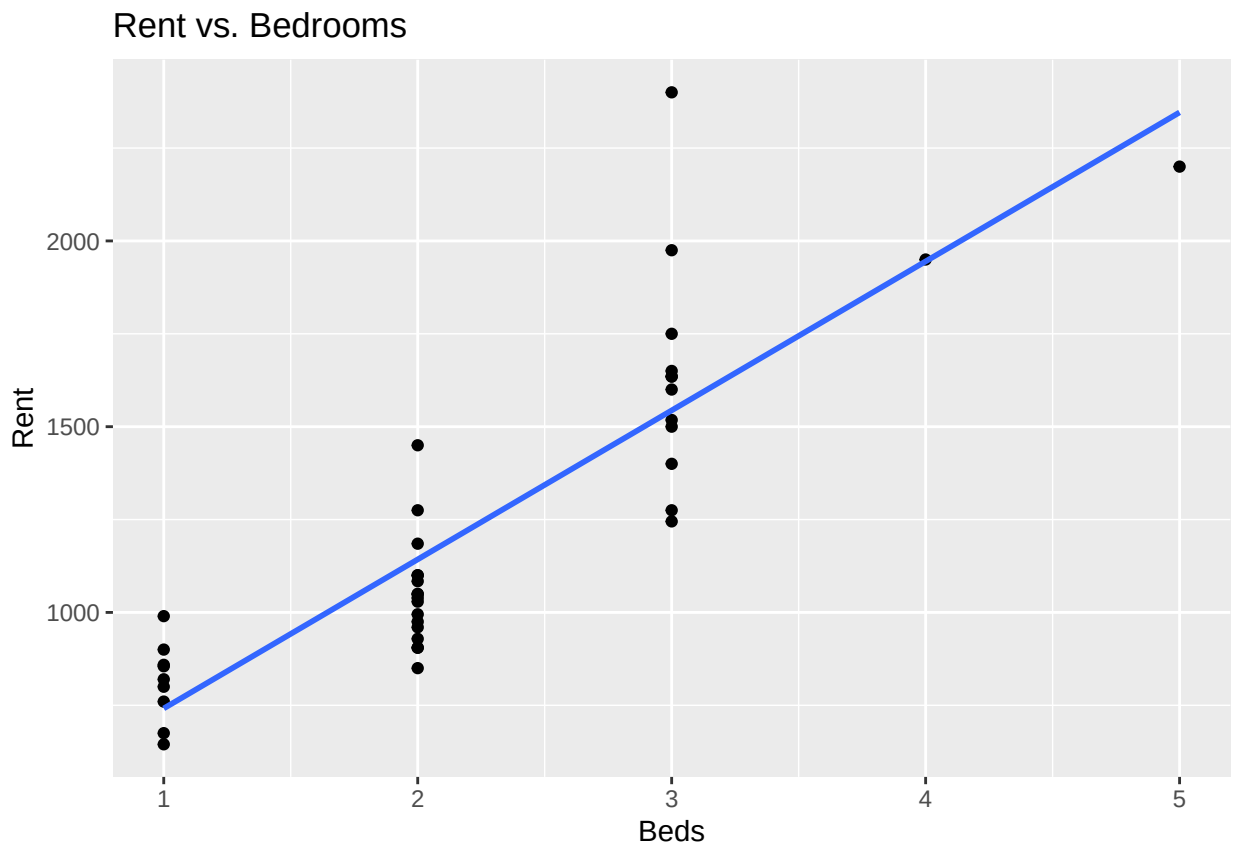
```r
plot(fitted(quad_model), residuals(quad_model),
     main="Residuals vs Fitted",
     xlab="Fitted values", ylab="Residuals")
abline(h=0, col="red")
```



#after plotting residuals and fitted values we can confirm the quadratic model has more points closer to the line than on the outside like the linear model

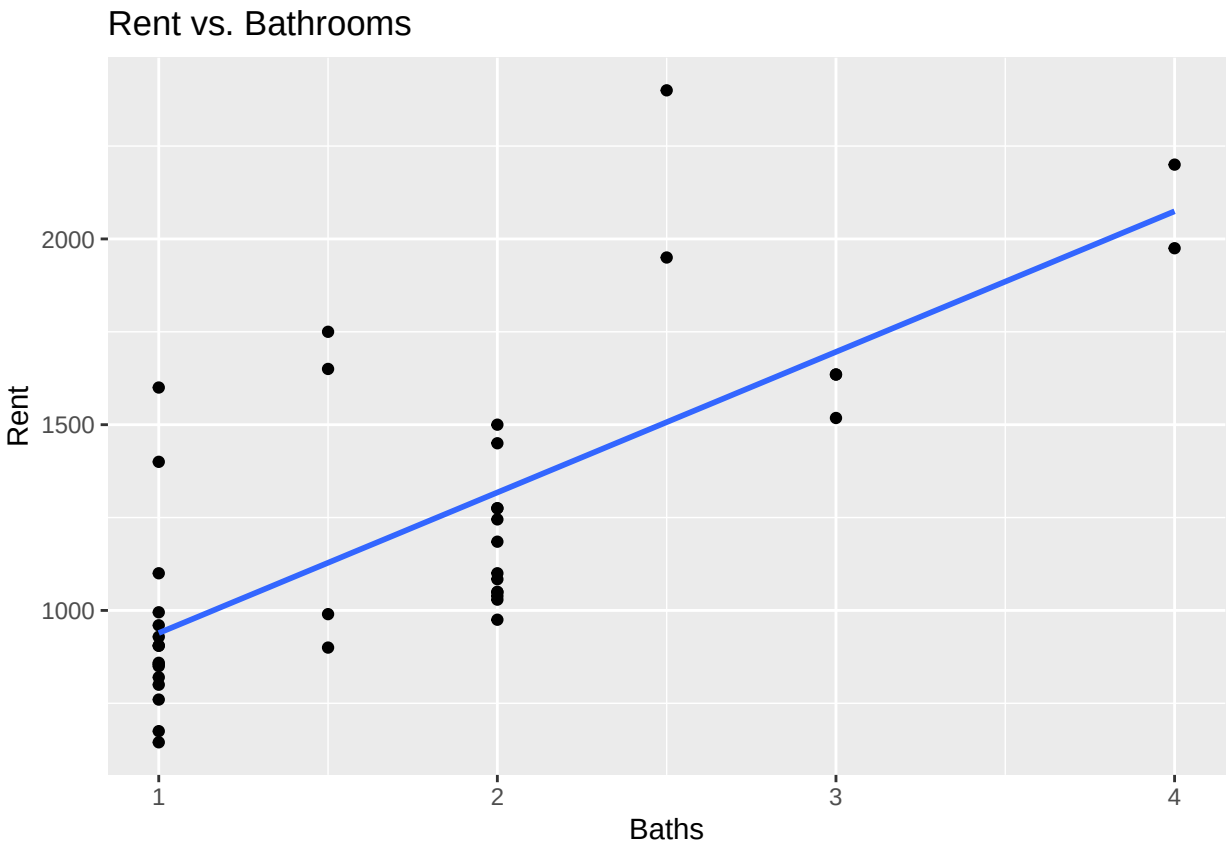
```r
#Plot Rent against each of the three predictor variables and evaluate whether the relationship is best

data <- read_excel("AnnArbor.xlsx")

ggplot(data, aes(x = Beds, y = Rent)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Rent vs. Bedrooms")
```

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




```
ggplot(data, aes(x = Baths, y = Rent)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Rent vs. Bathrooms")
```

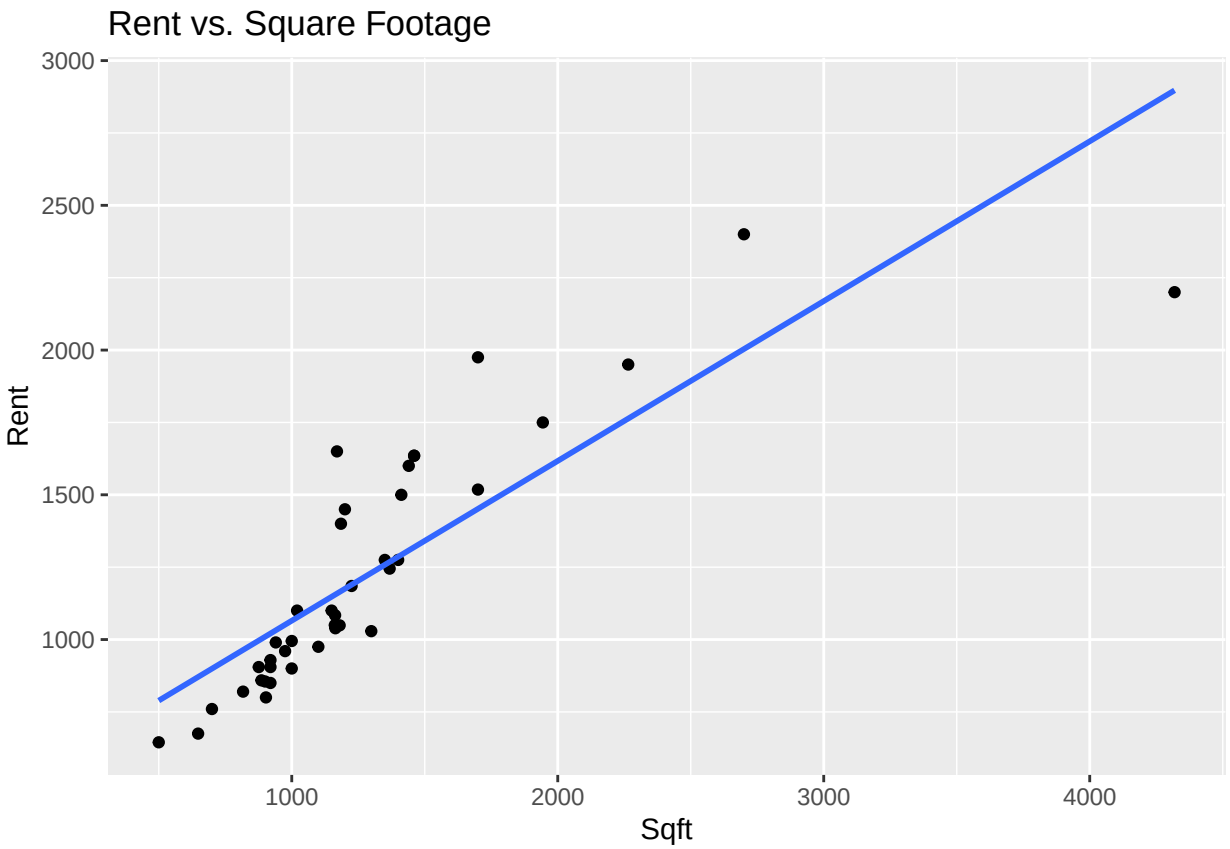
```
## `geom_smooth()` using formula = 'y ~ x'
```



#both bedrooms and bathrooms both are going up and appear somewhat linear but with different jumps but bathrooms have a more nonlinear trends

```
ggplot(data, aes(x = Sqft, y = Rent)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Rent vs. Square Footage")
```

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



#for the square feet it looks like it may a log transformation since it does have a curved pattern

```
data$Log_Sqft <- log(data$Sqft)


model <- lm(Rent ~ Beds + Baths + Log_Sqft, data = data)

# Display model summary
summary(model)
```

```
##
## Call:
## lm(formula = Rent ~ Beds + Baths + Log_Sqft, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -347.34 -102.11  -42.25   91.99  488.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3909.74    1053.79  -3.710 0.000696 ***
## Beds          131.78      61.68   2.136 0.039513 *
## Baths          36.43      52.68   0.691 0.493703
## Log_Sqft      675.26     169.41   3.986 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 172.3 on 36 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8355
## F-statistic: 67.05 on 3 and 36 DF,  p-value: 8.287e-15
```

#model has an r-squared of 0.8355 degrees meaning its a good fit and our t statistic showing it is statistically significant with most square feet being the most significant and beds too

```r
new_data <- data.frame(Beds = 3, Baths = 2, Log_Sqft = log(1600))

# Predict rent
predicted_rent <- predict(model, newdata = new_data)

# Print predicted rent
print(predicted_rent)
```

```
##        1
## 1540.384
```

#the predicted rent is 1540.384 for a 3 beds 2 baths and 1600 squared feet