Medical Insurance Costs

Exploratory Data Analysis & Insights

Noor Syed

Dataset Overview

- Dataset Name: Medical Insurance Cost Prediction
- Dataset Source: Kaggle
- Link: Medical Insurance Cost Dataset
- Total Entries: 2,700

Why This Dataset?

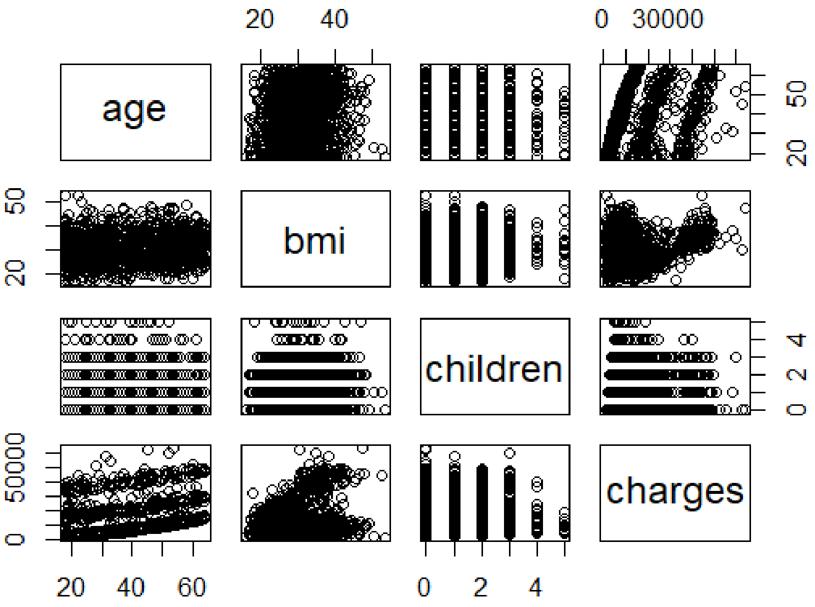
Multifaceted Data

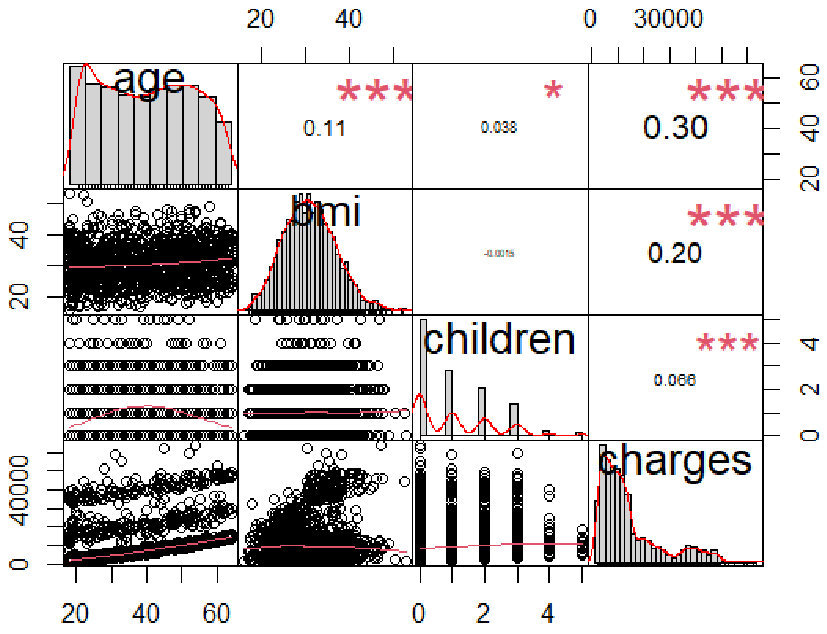
Real-World Relevance

Analytical Depth

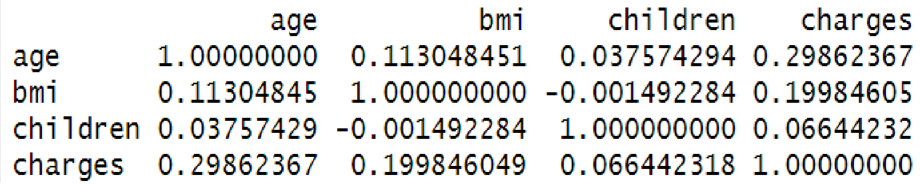
Rich Variables → **Age, Sex, BMI, Children, Smoker, Region, Charges**

EDA (Scatterplot & Correlation Matrix)





- Age has a moderate positive correlation with charges (0.299)
- BMI has a lower positive correlation with charges (0.200)
- Smoker_numeric variable shows a strong positive correlation with charges (0.789)

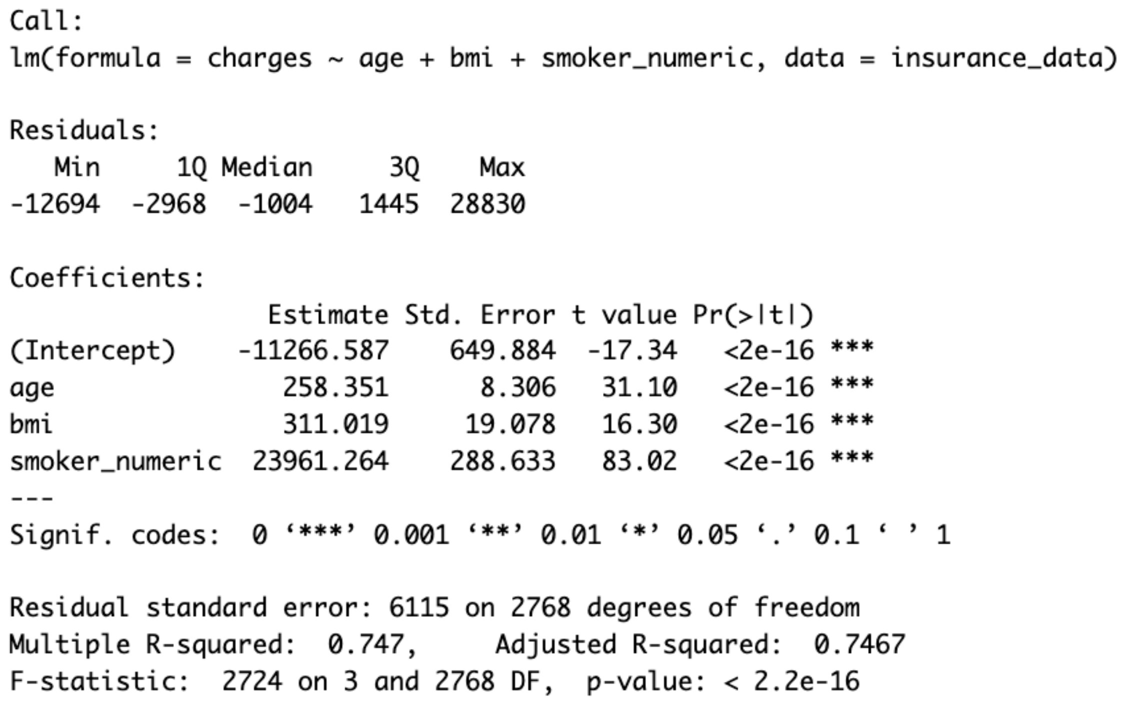


```
                age          bmi     children    charges
age      1.00000000  0.113048451  0.037574294 0.29862367
bmi      0.11304845  1.000000000 -0.001492284 0.19984605
children 0.03757429 -0.001492284  1.000000000 0.06644232
charges  0.29862367  0.199846049  0.066442318 1.00000000
```

Multiple Linear Regression (Model 1)

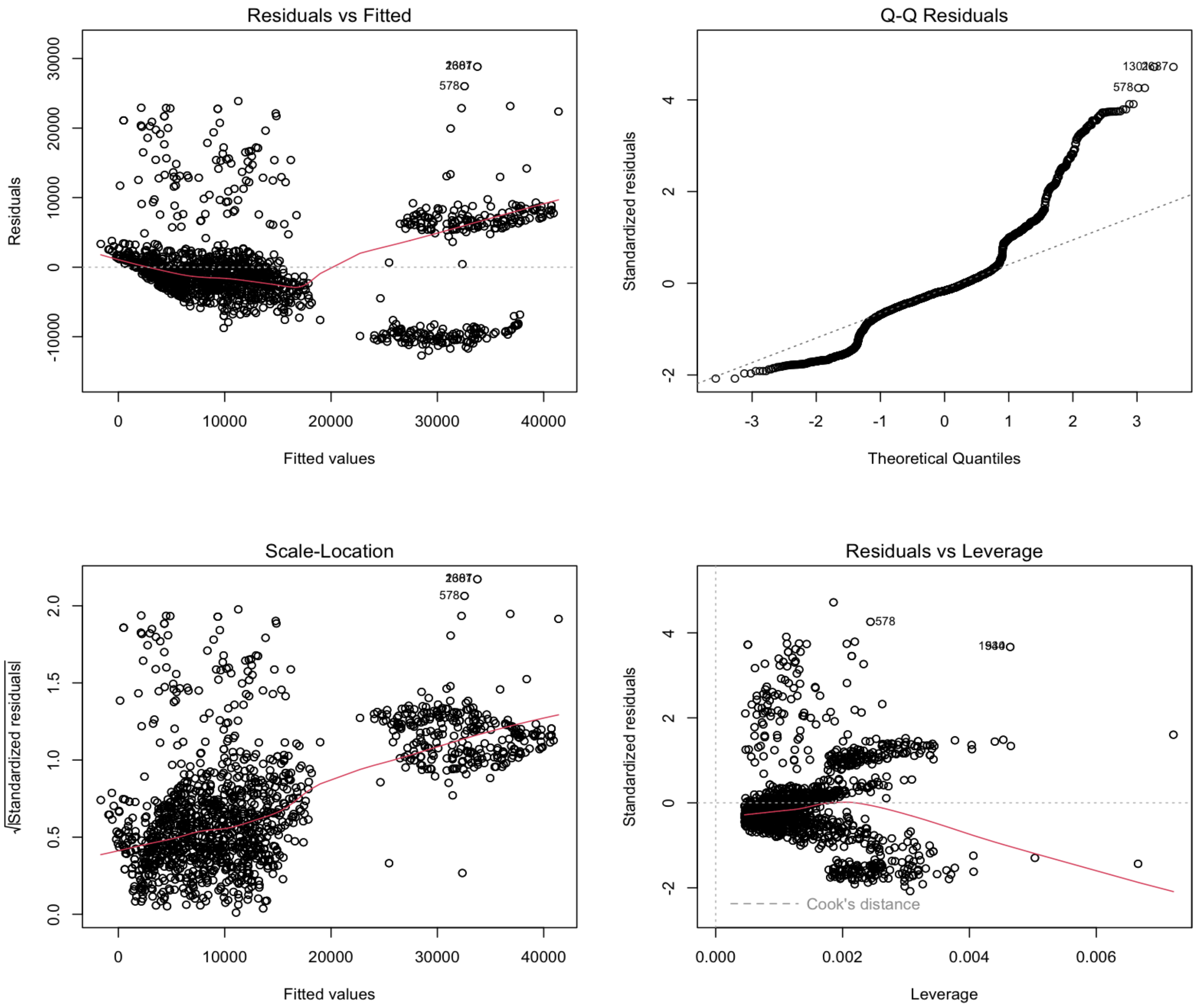
Coefficients for **age, BMI, and smoker** status are all **significant**:

- **Age:** For each additional year of age, the insurance charge increases by about $258.
- **BMI:** For each unit increase in BMI, insurance charges increase by approximately $311.
- **Smoker Status**: Being a smoker is associated with an increase of about $23,961 in charges compared to a non-smoker.
- **Adjusted R-squared** value is approximately 0.747.
- **AIC** and **BIC** values are (AIC: 56208.35, BIC: 56237.98).



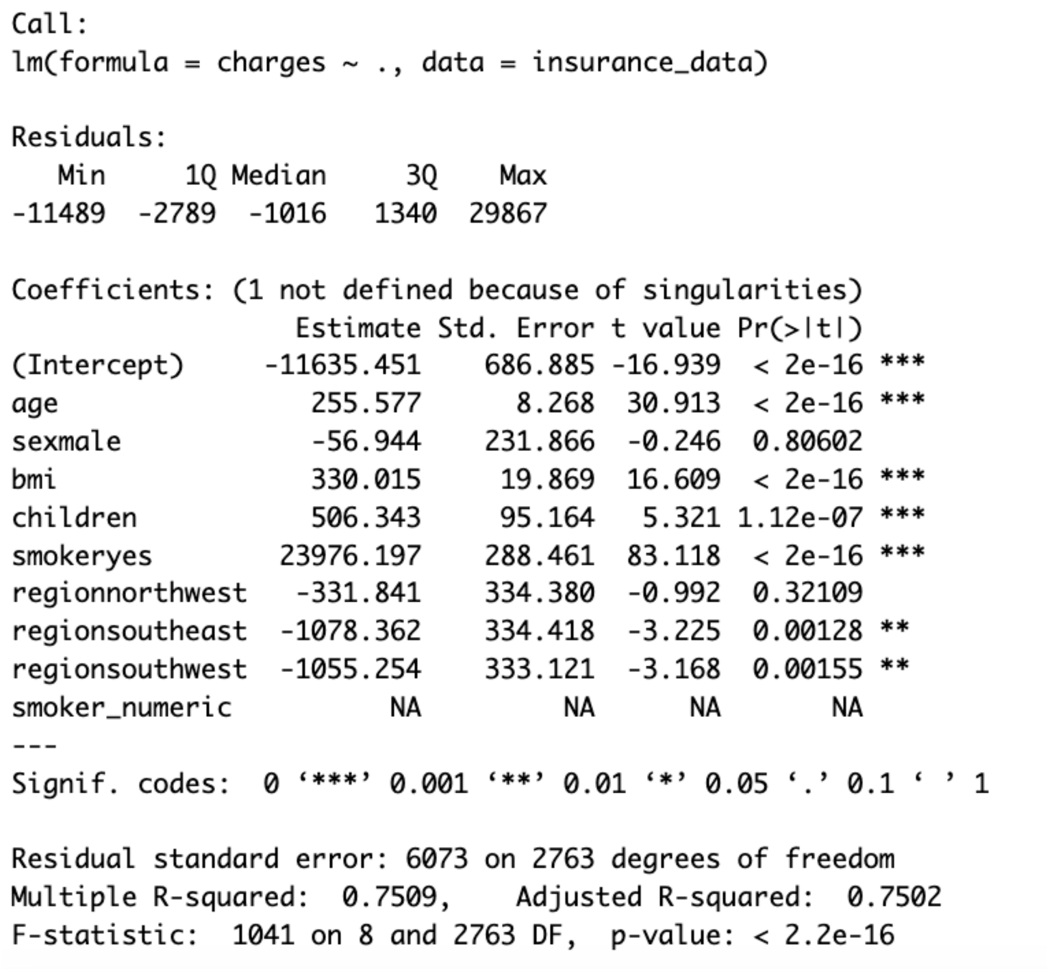
```
Call:
lm(formula = charges ~ age + bmi + smoker_numeric, data = insurance_data)

Residuals:
   Min     1Q Median     3Q    Max 
-12694  -2968  -1004   1445  28830 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -11266.587    649.884  -17.34   <2e-16 ***
age              258.351      8.306   31.10   <2e-16 ***
bmi              311.019     19.078   16.30   <2e-16 ***
smoker_numeric 23961.264    288.633   83.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6115 on 2768 degrees of freedom
Multiple R-squared:  0.747,	Adjusted R-squared:  0.7467 
F-statistic:  2724 on 3 and 2768 DF,  p-value: < 2.2e-16
```

Model 1 Diagnostic Plots



Multiple Linear Regression (Model 2)

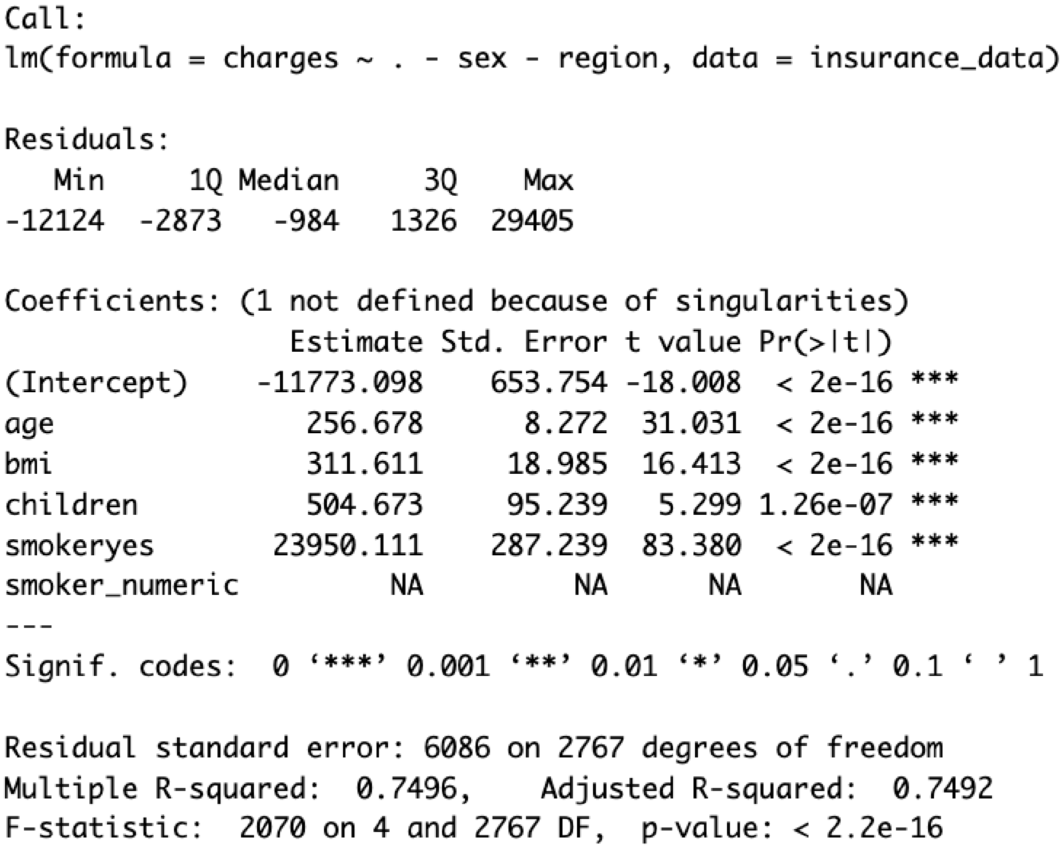
- **Age:** For each additional year, there is an increase of approximately $255.58 in insurance charges.
- **Sex (male):** The coefficient for 'sexmale' is not statistically significant (p-value = 0.806).
- **BMI:** For each unit increase in BMI, the insurance charges are expected to increase by about $330.01.
- **Children:** Each additional child is associated with an increase of about $506.34 in insurance charges.
- **Smoker Status (yes):** Being a smoker is associated with an increase of roughly $23,976.20 in insurance charges.
- **Region:** 'regionnorthwest' is not a significant predictor (p-value = 0.321).
- **Adjusted R-squared** value is 0.7502.
- **AIC** and **BIC** values are (AIC: 56175.02, BIC: 56234.29).



```
Call:
lm(formula = charges ~ ., data = insurance_data)

Residuals:
   Min     1Q Median     3Q    Max
-11489  -2789  -1016   1340  29867

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11635.451    686.885 -16.939  < 2e-16 ***
age                255.577      8.268  30.913  < 2e-16 ***
sexmale            -56.944    231.866  -0.246  0.80602
bmi                330.015     19.869  16.609  < 2e-16 ***
children           506.343     95.164   5.321 1.12e-07 ***
smokeryes        23976.197    288.461  83.118  < 2e-16 ***
regionnorthwest   -331.841    334.380  -0.992  0.32109
regionsoutheast  -1078.362    334.418  -3.225  0.00128 **
regionsouthwest  -1055.254    333.121  -3.168  0.00155 **
smoker_numeric          NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6073 on 2763 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7502
F-statistic:  1041 on 8 and 2763 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression (Model 3)

- **Age:** Slightly increased to $256.68 per year.
- **BMI:** The coefficient has decreased to $311.61.
- **Children:** Each additional child correlates with an increase in insurance charges by $504.67.
- **Smoker Status:** Being a smoker raises insurance charges by about $23,950.11.
- **Adjusted R-squared** value is 0.7492.
- **AIC** and **BIC** values are (AIC: 56182.36, BIC: 56217.92).



```
Call:
lm(formula = charges ~ . - sex - region, data = insurance_data)

Residuals:
   Min     1Q Median     3Q    Max 
-12124  -2873   -984   1326  29405 

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -11773.098    653.754 -18.008  < 2e-16 ***
age              256.678      8.272  31.031  < 2e-16 ***
bmi              311.611     18.985  16.413  < 2e-16 ***
children         504.673     95.239   5.299 1.26e-07 ***
smokeryes      23950.111    287.239  83.380  < 2e-16 ***
smoker_numeric        NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6086 on 2767 degrees of freedom
Multiple R-squared:  0.7496,	Adjusted R-squared:  0.7492 
F-statistic:  2070 on 4 and 2767 DF,  p-value: < 2.2e-16
```

Model Comparison

AIC and BIC

AIC(model_1) = 56208.35

BIC(model_1) = 56237.98

AIC(model_2) = 56175.02

BIC(model_2) = 56234.29

AIC(model_3) = 56182.36

BIC(model_3) = 56217.92

R-squared

summary_model_1$adj.r.squared = 0.7467351

summary_model_2$adj.r.squared = 0.750212

summary_model_2$adj.r.squared = 0.7491888

Mean Squared Error

mse1 <- mean(resid(model_1)^2) = 37344519

mse2 <- mean(resid(model_2)^2) = 36765310

mse3 <- mean(resid(model_3)^2) = 36969351

Model 2 is appears to be the best model.