Data preprocessing

In this sections, we convert the categorical variables `sex`, `smoker`, and `region` into one-hot vectors, using the mutate funtion by dplyr

```
install.packages("car")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘cowplot’, ‘Deriv’, ‘microbenchmark’, ‘numDeriv’, ‘doBy’, ‘Sparsel
```

```
    # Load necessary libraries
library(dplyr)

# Load the dataset
insurance_data <- read.csv("/content/insurance.csv")

# Convert categorical variables into dummy variables
insurance_data_with_dummies <- insurance_data %>%
  mutate(
    sex_male = ifelse(sex == "male", 1, 0),
    smoker_yes = ifelse(smoker == "yes", 1, 0),
    region_northwest = ifelse(region == "northwest", 1, 0),
    region_southeast = ifelse(region == "southeast", 1, 0),
    region_southwest = ifelse(region == "southwest", 1, 0)
  ) %>%
  select(-sex, -smoker, -region)

# Display the first few rows of the new dataset with dummy variables
head(insurance_data_with_dummies)

```

A data.frame: 6 × 9

	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	re
	<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	
1	19	27.900	0	16884.924	0	1	0	
2	18	33.770	1	1725.552	1	0	0	
3	28	33.000	3	4449.462	1	0	0	
4	33	22.705	0	21984.471	1	0	1	
5	32	28.880	0	3866.855	1	0	1	
6	31	25.740	0	3756.622	0	0	0	

Double-click (or enter) to edit

Linear Regression

```
initial_model <- lm(charges ~ age + bmi + children + sex_male + smoker_yes + region_northwest + reg
summary(initial_model)
```

```
Call:
lm(formula = charges ~ age + bmi + children + sex_male + smoker_yes +
    region_northwest + region_southeast + region_southwest, data =
insurance_data_with_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -11938.5      987.8 -12.086  < 2e-16 ***
age                 256.9       11.9  21.587  < 2e-16 ***
bmi                 339.2       28.6  11.860  < 2e-16 ***
children            475.5      137.8   3.451 0.000577 ***
sex_male           -131.3      332.9  -0.394 0.693348
smoker_yes        23848.5      413.1  57.723  < 2e-16 ***
region_northwest   -353.0      476.3  -0.741 0.458769
region_southeast  -1035.0      478.7  -2.162 0.030782 *
region_southwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
```

Residuals

The residuals reflect the difference between the observed and predicted values of `charges`. The distribution of residuals is as follows:

- **Min**: -11304.9
- **1st Quartile (1Q)**: -2848.1
- **Median**: -982.1
- **3rd Quartile (3Q)**: 1393.9
- **Max**: 29992.8

Coefficients

The model coefficients represent the estimated effect of each predictor variable on `charges`:

- **(Intercept)**: -11938.5
  - This is the expected value of `charges` when all predictors are zero. However, since it is not realistic for all predictors to be zero (e.g., age cannot be zero), this intercept acts as a baseline reference.
- **age**: 256.9
  - Each additional year of age increases the charges by approximately 256.9.
  - p-value (< 2e-16) -> highly significant
- **bmi**: 339.2
  - Each unit increase in BMI raises the charges by approximately 339.2.
  - p-value (< 2e-16) -> highly significant
- **children**: 475.5
  - Each additional child increases the charges by approximately 475.5.
  - p-value (0.000577) -> significant
- **sex_male**: -131.3
  - Being male decreases the charges by approximately 131.3.
  - p-value (0.693348) -> not significant

- **smoker_yes**: 23848.5
    - Being a smoker increases the charges by approximately 23848.5.
    - p-value (< 2e-16) -> highly significant
- **region_northwest**: -353.0
    - Living in the Northwest decreases the charges by approximately 353.0 compared to the baseline region (likely the Northeast or another omitted category).
    - p-value (0.458769) -> not significant
- **region_southeast**: -1035.0
    - Living in the Southeast decreases the charges by approximately 1035.0 compared to the baseline region.
    - p-value (0.030782) -> significant
- **region_southwest**: -960.0
    - Living in the Southwest decreases the charges by approximately 960.0 compared to the baseline region.
    - p-value (0.044765) -> significant

Model Performance

- **Residual Standard Error**: 6062 on 1329 degrees of freedom
    - This measures the average deviation of the observed charges from the predicted charges.
- **Multiple R-squared**: 0.7509
    - Approximately 75.09% of the variability in charges is explained by the model.
- **Adjusted R-squared**: 0.749
    - This adjusted version of R-squared accounts for the number of predictors in the model, indicating that approximately 74.9% of the variability in charges is explained by the model.

Conclusion

The initial model shows that age, BMI, number of children, and smoking status are significant predictors of insurance charges. Additionally, living in the Southeast and Southwest regions significantly decreases the charges compared to the baseline region. However, being male and living in the Northwest region are not significant predictors.

```
library(car)

vif_values <- vif(initial_model)
print(vif_values)
```

```
              age              bmi         children         sex_male 
         1.016822         1.106630         1.004011         1.008900 
      smoker_yes region_northwest region_southeast region_southwest 
         1.012074         1.518823         1.652230         1.529411 
```

```r
# Function to calculate VIF and remove high VIF predictors
remove_high_vif <- function(model, threshold = 5) {
  # Calculate VIF
  vif_values <- vif(model)
  print(vif_values)

  # Check if any VIF values are above the threshold
  while(any(vif_values > threshold)) {
    # Identify the predictor with the highest VIF
    max_vif_predictor <- names(which.max(vif_values))

    cat("Removing predictor with high VIF:", max_vif_predictor, "\n")

    model_formula <- as.formula(paste("charges ~", paste(setdiff(names(coef(model)), c("(Intercept)

    # Fit the new model
    model <- lm(model_formula, data = data)

    # Recalculate VIF
    vif_values <- vif(model)
    print(vif_values)
  }

  return(model)
}

# Apply the function to the initial model
final_model <- remove_high_vif(initial_model)
summary(final_model)
```

```
             age              bmi         children         sex_male
        1.016822         1.106630         1.004011         1.008900
      smoker_yes region_northwest region_southeast region_southwest
        1.012074         1.518823         1.652230         1.529411

Call:
lm(formula = charges ~ age + bmi + children + sex_male + smoker_yes +
    region_northwest + region_southeast + region_southwest, data =
insurance_data_with_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -11938.5      987.8 -12.086  < 2e-16 ***
age                 256.9       11.9  21.587  < 2e-16 ***
bmi                 339.2       28.6  11.860  < 2e-16 ***
children            475.5      137.8   3.451 0.000577 ***
sex_male           -131.3      332.9  -0.394 0.693348
smoker_yes        23848.5      413.1  57.723  < 2e-16 ***
region_northwest   -353.0      476.3  -0.741 0.458769
region_southeast  -1035.0      478.7  -2.162 0.030782 *
region_southwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
```

Use step function to get final model

```r
# Use step function for variable selection
final_model <- step(initial_model, direction = "both")

# Summarize the final model
summary(final_model)
```

```
Start:  AIC=23316.43
charges ~ age + bmi + children + sex_male + smoker_yes + region_northwest +
    region_southeast + region_southwest

                   Df  Sum of Sq        RSS   AIC
- sex_male          1 5.7164e+06 4.8845e+10 23315
- region_northwest  1 2.0183e+07 4.8860e+10 23315
<none>                           4.8840e+10 23316
- region_southwest  1 1.4829e+08 4.8988e+10 23318
- region_southeast  1 1.7180e+08 4.9011e+10 23319
- children          1 4.3755e+08 4.9277e+10 23326
- bmi               1 5.1692e+09 5.4009e+10 23449
- age               1 1.7124e+10 6.5964e+10 23717
- smoker_yes        1 1.2245e+11 1.7129e+11 24993

Step:  AIC=23314.58
charges ~ age + bmi + children + smoker_yes + region_northwest +
    region_southeast + region_southwest

                   Df  Sum of Sq        RSS   AIC
- region_northwest  1 2.0094e+07 4.8865e+10 23313
<none>                           4.8845e+10 23315
+ sex_male          1 5.7164e+06 4.8840e+10 23316
- region_southwest  1 1.4808e+08 4.8993e+10 23317
- region_southeast  1 1.7159e+08 4.9017e+10 23317
- children          1 4.3596e+08 4.9281e+10 23324
- bmi               1 5.1645e+09 5.4010e+10 23447
- age               1 1.7151e+10 6.5996e+10 23715
- smoker_yes        1 1.2301e+11 1.7186e+11 24996

Step:  AIC=23313.13
charges ~ age + bmi + children + smoker_yes + region_southeast +
    region_southwest

                   Df  Sum of Sq        RSS   AIC
<none>                           4.8865e+10 23313
+ region_northwest  1 2.0094e+07 4.8845e+10 23315
- region_southwest  1 1.3139e+08 4.8997e+10 23315
+ sex_male          1 5.6275e+06 4.8860e+10 23315
- region_southeast  1 1.5694e+08 4.9022e+10 23315
- children          1 4.3080e+08 4.9296e+10 23323
- bmi               1 5.1638e+09 5.4029e+10 23446
- age               1 1.7155e+10 6.6021e+10 23714
- smoker_yes        1 1.2317e+11 1.7203e+11 24995

Call:
lm(formula = charges ~ age + bmi + children + smoker_yes + region_southeast
+
    region_southwest, data = insurance_data_with_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-11377.1  -2855.0   -973.6   1349.5  29926.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -12165.38     949.54 -12.812  < 2e-16 ***
age                 257.01      11.89  21.617  < 2e-16 ***
bmi                 338.64      28.55  11.860  < 2e-16 ***
children            471.54     137.66   3.426 0.000632 ***
smoker_yes        23843.87     411.66  57.921  < 2e-16 ***
region_southeast   -858.47     415.21  -2.068 0.038873 *
region_southwest   -782.75     413.76  -1.892 0.058734 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6059 on 1331 degrees of freedom
Multiple R-squared:  0.7508,     Adjusted R-squared:  0.7497
```

```r
# Calculate VIF for the final model
final_vif_values <- vif(final_model)
print(final_vif_values)
```

```
              age              bmi         children       smoker_yes 
         1.016174         1.104196         1.002831         1.005747 
region_southeast region_southwest 
         1.244250         1.147367 
```

Analysis of the Final Linear Regression Model

The final linear regression model has been fitted to predict `charges` based on `age`, `bmi`, `children`, `smoker_yes`, `region_southeast`, and `region_southwest`. Below is a detailed analysis of the model:

Residuals

The residuals reflect the difference between the observed and predicted values of `charges`. The distribution of residuals is as follows:

- **Min**: -11377.1
- **1st Quartile (1Q)**: -2855.0
- **Median**: -973.6
- **3rd Quartile (3Q)**: 1349.5
- **Max**: 29926.4

Coefficients

The model coefficients represent the estimated effect of each predictor variable on `charges`:

- **(Intercept)**: -12165.38
  - This is the expected value of `charges` when all predictors are zero. While having all predictors at zero is not realistic (e.g., age cannot be zero), this intercept serves as a baseline reference.
- **age**: 257.01
  - Each additional year of age increases the charges by approximately 257.01.
  - p-value (< 2e-16) -> highly significant
- **bmi**: 338.64
  - Each unit increase in BMI raises the charges by approximately 338.64.
  - p-value (< 2e-16) -> highly significant
- **children**: 471.54
  - Each additional child increases the charges by approximately 471.54.
  - p-value (0.000632) -> significant
- **smoker_yes**: 23843.87
  - Being a smoker increases the charges by approximately 23843.87.
  - p-value (< 2e-16) -> highly significant
- **region_southeast**: -858.47
  - Living in the Southeast decreases the charges by approximately 858.47 compared to the baseline region (likely the Northeast or another omitted category).
  - p-value (0.038873) -> significant
- **region_southwest**: -782.75
  - Living in the Southwest decreases the charges by approximately 782.75 compared to the baseline region.
  - p-value (0.058734) -> marginally significant (at the 0.1 level)

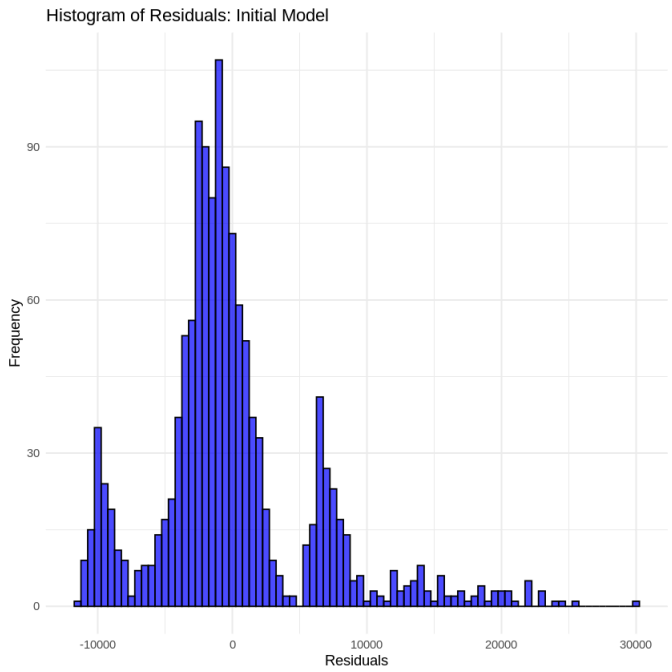
Model Performance

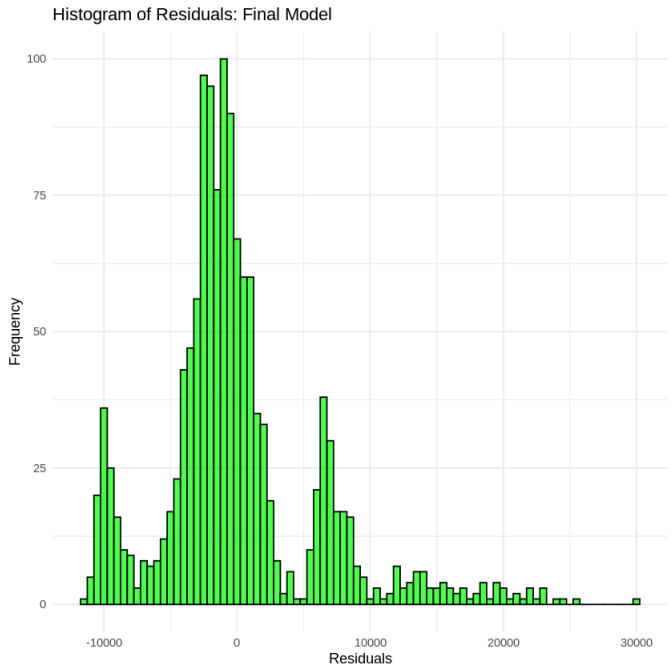
- **Residual Standard Error**: 6059 on 1331 degrees of freedom
    - This measures the average deviation of the observed charges from the predicted charges.
- **Multiple R-squared**: 0.7508
    - Approximately 75.08% of the variability in charges is explained by the model.
- **Adjusted R-squared**: 0.7497
    - This adjusted version of R-squared accounts for the number of predictors in the model, indicating that approximately 74.97% of the variability in charges is explained by the model.

Conclusion

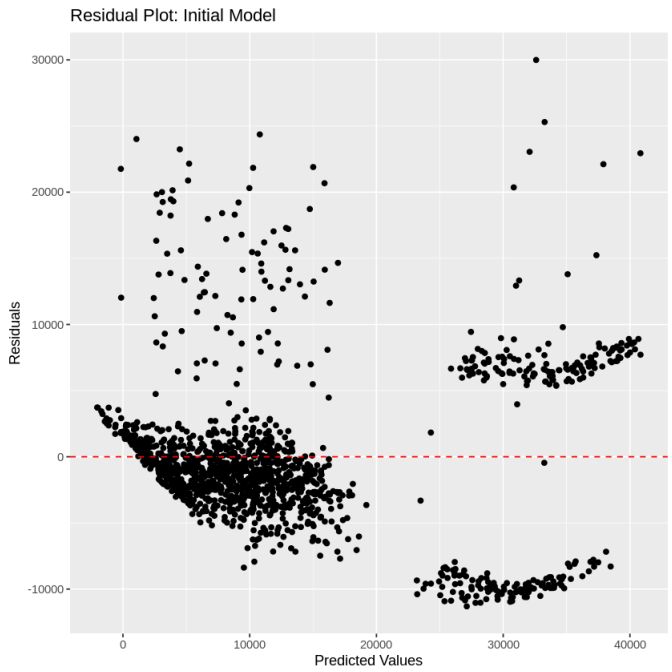
The final model shows that age, BMI, number of children, and smoking status are significant predictors of insurance charges. Additionally, living in the Southeast and Southwest regions significantly decreases the charges compared to the baseline region. The high R-squared value suggests that the model fits the data well. The use of the `step` function helped in removing unnecessary variables, leading to a more parsimonious model.

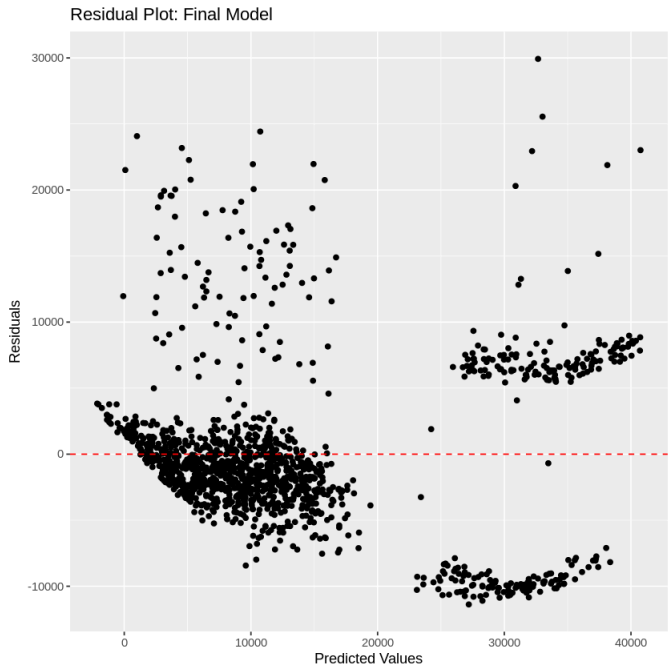
```
# Calculate residuals for initial and final models
initial_residuals <- residuals(initial_model)
final_residuals <- residuals(final_model)

# Plot histogram for initial model residuals
hist_initial <- ggplot(data.frame(initial_residuals), aes(x = initial_residuals)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Residuals: Initial Model", x = "Residuals", y = "Frequency") +
  theme_minimal()

# Plot histogram for final model residuals
hist_final <- ggplot(data.frame(final_residuals), aes(x = final_residuals)) +
  geom_histogram(binwidth = 500, fill = "green", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Residuals: Final Model", x = "Residuals", y = "Frequency") +
  theme_minimal()

# Print the histograms
print(hist_initial)
print(hist_final)

```



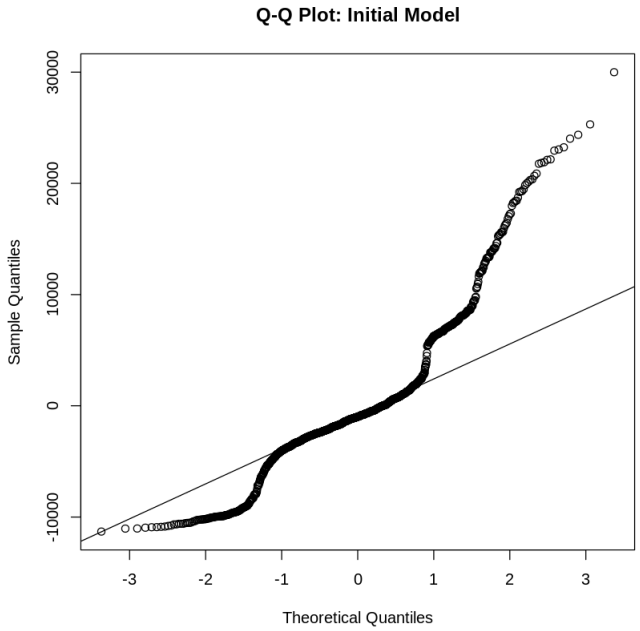


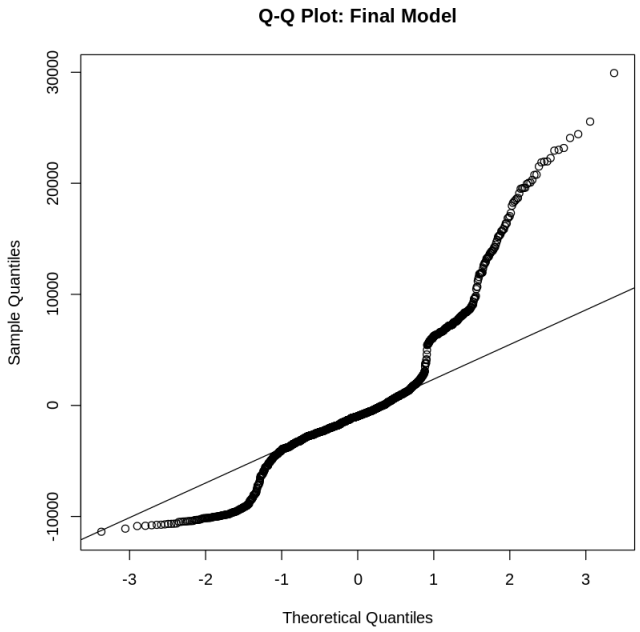

```r
library(ggplot2)

# Residual plot for initial model
ggplot(insurance_data, aes(x = predict(initial_model), y = residuals(initial_model))) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residual Plot: Initial Model", x = "Predicted Values", y = "Residuals")

# Residual plot for final model
ggplot(insurance_data, aes(x = predict(final_model), y = residuals(final_model))) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residual Plot: Final Model", x = "Predicted Values", y = "Residuals")

```



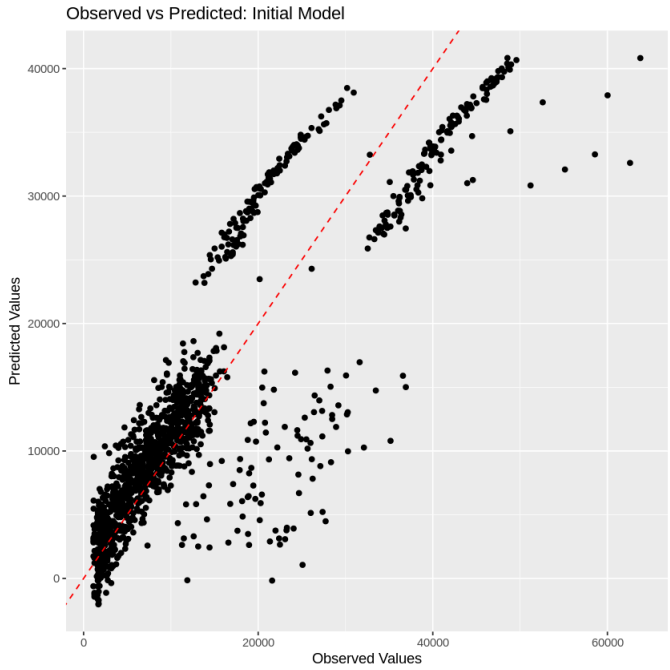


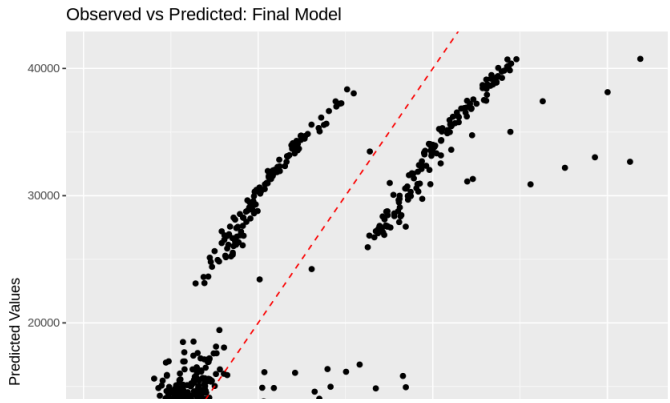

```

# Q-Q plot for initial model
qqnorm(residuals(initial_model), main = "Q-Q Plot: Initial Model")
qqline(residuals(initial_model))

# Q-Q plot for final model
qqnorm(residuals(final_model), main = "Q-Q Plot: Final Model")
qqline(residuals(final_model))
```






```
1
2 # Observed vs Predicted plot for initial model
3 ggplot(insurance_data, aes(x = charges, y = predict(initial_model))) +
4   geom_point() +
5   geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
6   labs(title = "Observed vs Predicted: Initial Model", x = "Observed Values", y = "Predicted Values
7
8 # Observed vs Predicted plot for final model
9 ggplot(insurance_data, aes(x = charges, y = predict(final_model))) +
10   geom_point() +
11   geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
12   labs(title = "Observed vs Predicted: Final Model", x = "Observed Values", y = "Predicted Values")
```






```
1 Start coding or generate with AI.
```

```
Error in data$pred_initial <- predict(initial_model): object of type
'closure' is not subsettable
Traceback:
```

Next steps: Explain error

Observed Values