# Activity 2

Luke Kerwin, Nick Malloy, Dinni Bhardwaj

## Introduction / Background

Hatfield Insurance Company, a giant in the healthcare industry, seeks to optimize its pricing strategies for health insurance policies. Determining the cost of these policies is crucial for the company's sustainability and ensuring value for its policyholders. To address this, our team will employ linear regression, a basic yet effective statistical method, to predict potential policy costs. By analyzing how various factors influence these costs, we aim to provide Hatfield with a foundational model to guide their pricing decisions.

## Data Exploration

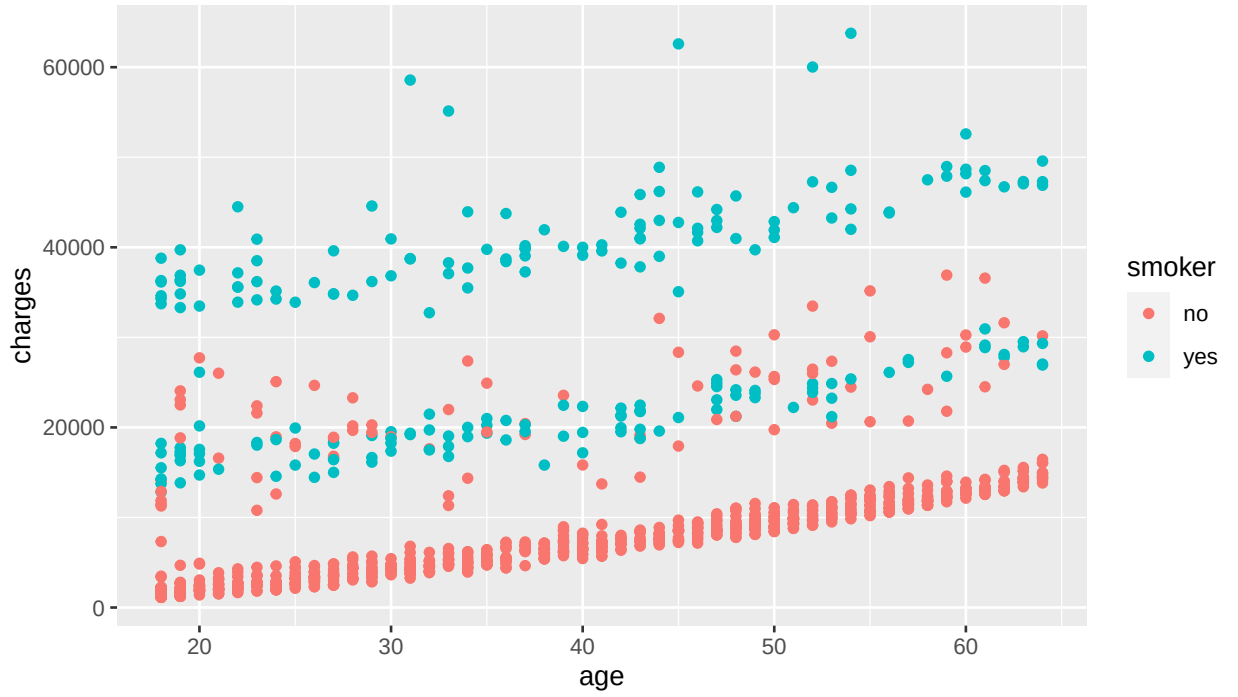
First we will get a look at the raw data itself:

#### Raw Data

| caseID | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 1 | 19 | female | 27.900 | 0 | yes | southwest | 16884.924 |
| 4 | 33 | male | 22.705 | 0 | no | northwest | 21984.471 |
| 5 | 32 | male | 28.880 | 0 | no | northwest | 3866.855 |
| 6 | 31 | female | 25.740 | 0 | no | southeast | 3756.622 |
| 8 | 37 | female | 27.740 | 3 | no | northwest | 7281.506 |
| 9 | 37 | male | 29.830 | 2 | no | northeast | 6406.411 |
| 10 | 60 | female | 25.840 | 0 | no | northwest | 28923.137 |
| 11 | 25 | male | 26.220 | 0 | no | northeast | 2721.321 |
| 12 | 62 | female | 26.290 | 0 | yes | southeast | 27808.725 |
| 13 | 23 | male | 34.400 | 0 | no | southwest | 1826.843 |

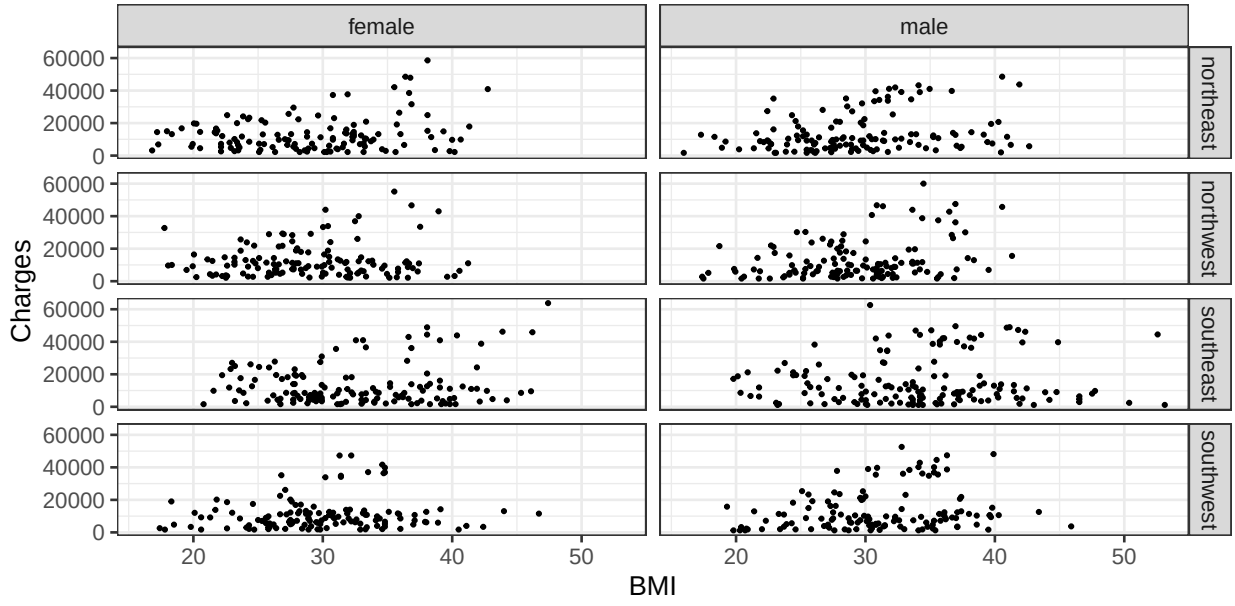
From the sample above, we can see that there are eight columns. Six of the eight columns have a category characteristic to them. `caseID` is ordinal and likely is irrelevent to predicting charges, but we will leave it in for now. `age`, `sex`, `children`, `smoker`, and `region` are the remaining "categorical" columns. Each one likely is somewhat correlated to our target column; `charges`. We also notice that since we are using linear regression techniques, we will have to manipulate the categorical columns so that the model can easily read them.

**Demographic Exploration**




Here we can see some of the customers with the highest charge amounts are smokers and that age isn't a huge indicator of high charges.

Next we will take a look at the different demographics using `bmi` and `charges`:




Here we can see that there likely isn't much of a different between males and females due to the plots all looking relatively the same. It can also be said for region as well. The plots arent the same, but they have roughly the same shape. Now lets take a look at the correlation between the continuous variables as well as with the target variable.

**Correlation Matrix (Continuous Variables)**

| | caseID | age | bmi | children | charges |
|---|---|---|---|---|---|
| caseID | 1.0000 | -0.0330 | -0.0123 | 0.0222 | 0.0136 |
| age | -0.0330 | 1.0000 | 0.1150 | 0.0683 | 0.3053 |
| bmi | -0.0123 | 0.1150 | 1.0000 | 0.0158 | 0.1830 |
| children | 0.0222 | 0.0683 | 0.0158 | 1.0000 | 0.0850 |
| charges | 0.0136 | 0.3053 | 0.1830 | 0.0850 | 1.0000 |

The table above shows that we dont have any major co-linear columns which is a good thing. It means we wont have to remove any columns. We also can see that the column that correlates the most with `charges` is `age`. The second and third most correlated variables are `bmi` and `children`. As hypothesized before, `caseID` essentially has no correlation to charges which means we can get exclude it from the model.

# Model Construction

We decided to make two models; one simple linear regression model and one multiple linear regression model.

## Model 1: Simple Linear Regression

Simple linear regression (SLR) models use one variable for X. In this case we have chosen `age` as it is the most correlated to our target `charges`. Here is how the SLR model using `age` performs:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 2923.578 | 1035.910 | 2.822 | 0.005 |
| Age (years) | 262.748 | 25.003 | 10.509 | 0.000 |

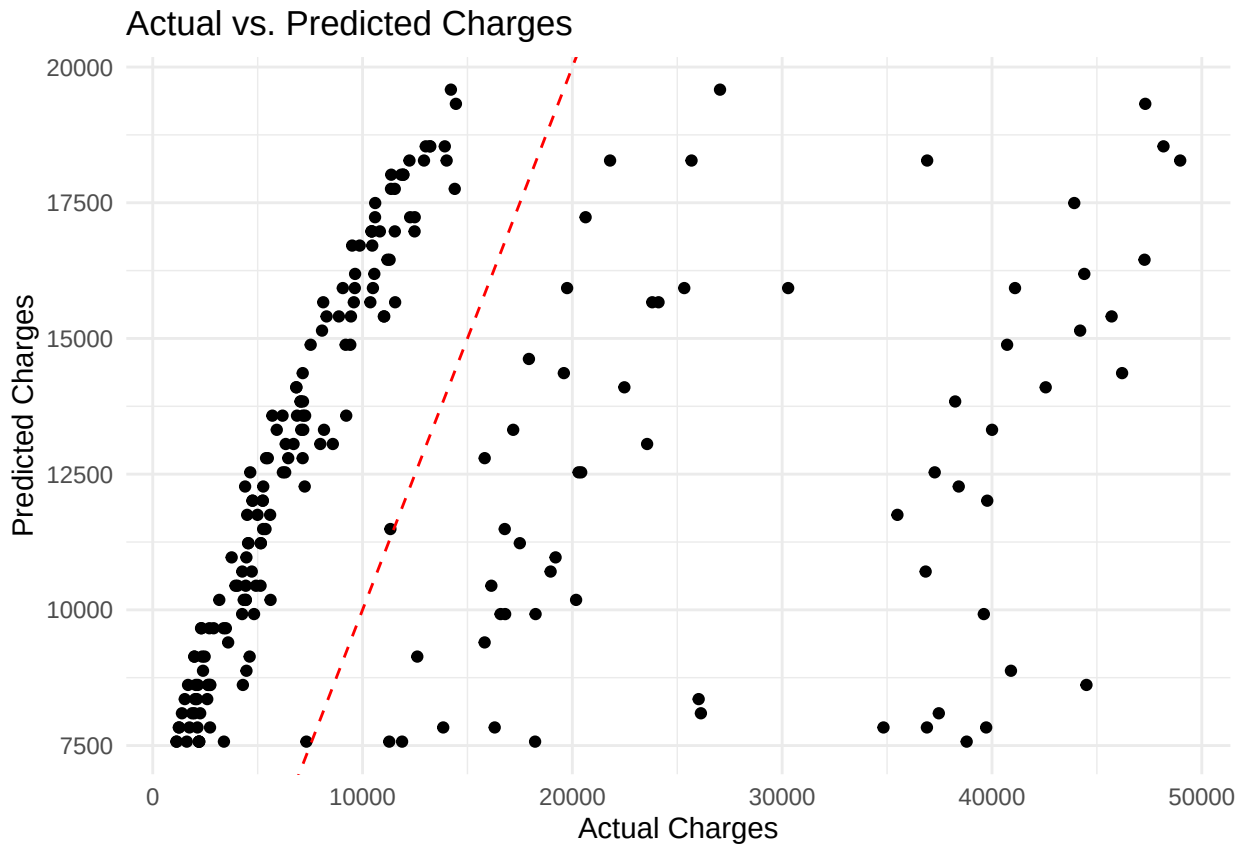
| R_Squared | Adjusted_R_Squared |
|---|---|
| 0.093 | 0.092 |

Trying to predict `charges` from one variable is going to be tough. The model tells us that essentially for every 1 year increase in age, the predicted charges increase by 262.75 dollars. We also see that the R^2 value is `0.093`. This means that 9.3% of the variability in `charges` can be explained by the `age` in our model. Not great, but also not horrible considering we are only using one variable. Now lets split the data into a training and test set so we can properly asses the model.

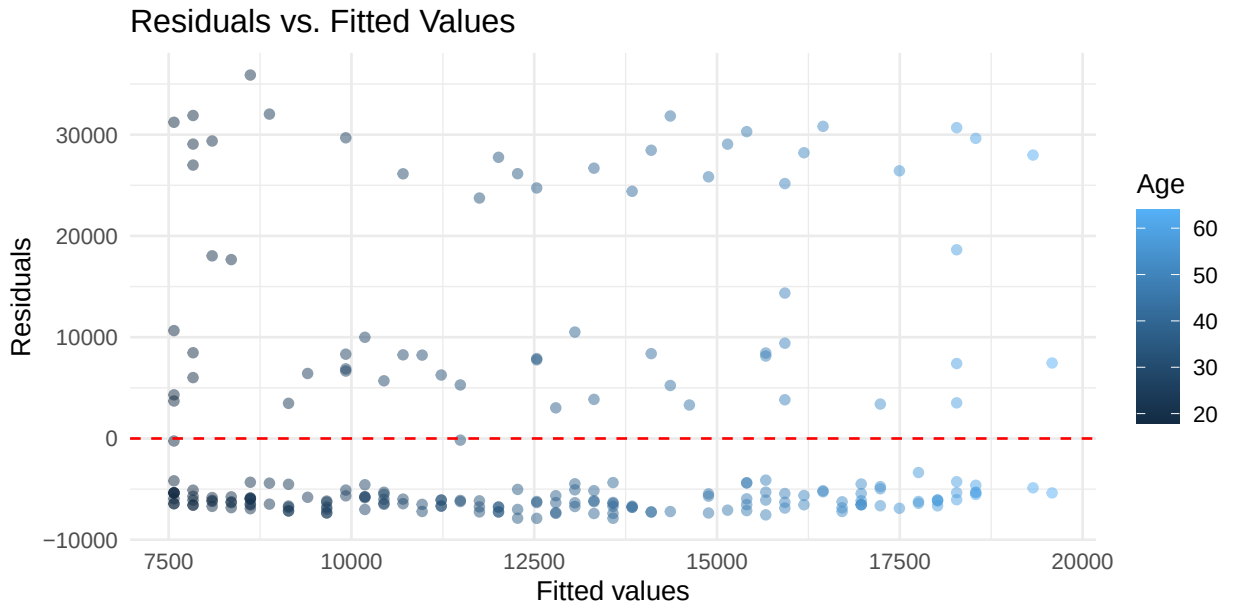
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 2872.278 | 1151.534 | 2.494 | 0.013 |
| Age (years) | 261.112 | 27.496 | 9.496 | 0.000 |

| R_Squared | Adjusted_R_Squared |
|---|---|
| 0.095 | 0.094 |

**Results**




Using a test/train split of `0.8` we get a training data set with 860 rows and a testing data set with 216 rows. Using the training set we roughly the same results as when we used the entire data set, which was expected. When we look at the RMSE, we get a value of 11920.55 which tells us that there is an average error of 11920.55 between the predictions and the actuals. Here is a plot of the residuals:



## Model 2: Multiple Linear Regression

Unlike SLR models, Multiple Linear Regression models use more than one variable for X. For fun we are going to use all of the variables possible to see what the model looks like compared to the SLR model made earlier. But before we can do that, we need to make our data friendly for the `lm()` function to use. This means we have to manipulate our data so that the categorical variables are in matrix form. After the manipulation, our data now looks like this:

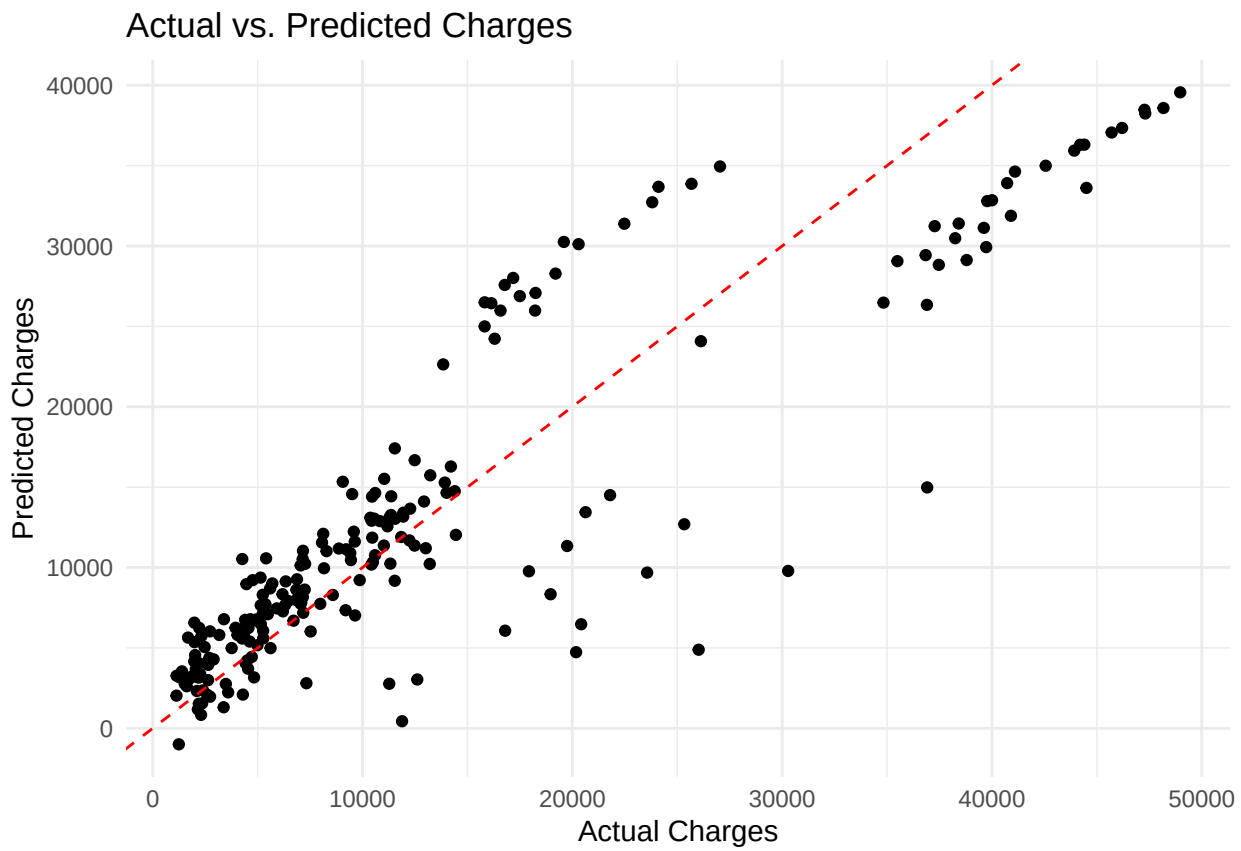
| caseID | age | bmi | children | charges | is_male | is_smoker | is_southwest | is_northwest | is_northeast |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 19 | 27.900 | 0 | 16884.924 | 0 | 1 | 1 | 0 | 0 |
| 4 | 33 | 22.705 | 0 | 21984.471 | 1 | 0 | 0 | 0 | 0 |
| 5 | 32 | 28.880 | 0 | 3866.855 | 1 | 0 | 0 | 0 | 0 |
| 6 | 31 | 25.740 | 0 | 3756.622 | 0 | 0 | 0 | 0 | 0 |
| 8 | 37 | 27.740 | 3 | 7281.506 | 0 | 0 | 0 | 0 | 0 |
| 9 | 37 | 29.830 | 2 | 6406.411 | 1 | 0 | 0 | 0 | 1 |
| 10 | 60 | 25.840 | 0 | 28923.137 | 0 | 0 | 0 | 0 | 0 |
| 11 | 25 | 26.220 | 0 | 2721.321 | 1 | 0 | 0 | 0 | 1 |
| 12 | 62 | 26.290 | 0 | 27808.725 | 0 | 1 | 0 | 0 | 0 |
| 13 | 23 | 34.400 | 0 | 1826.843 | 1 | 0 | 1 | 0 | 0 |

**Results**

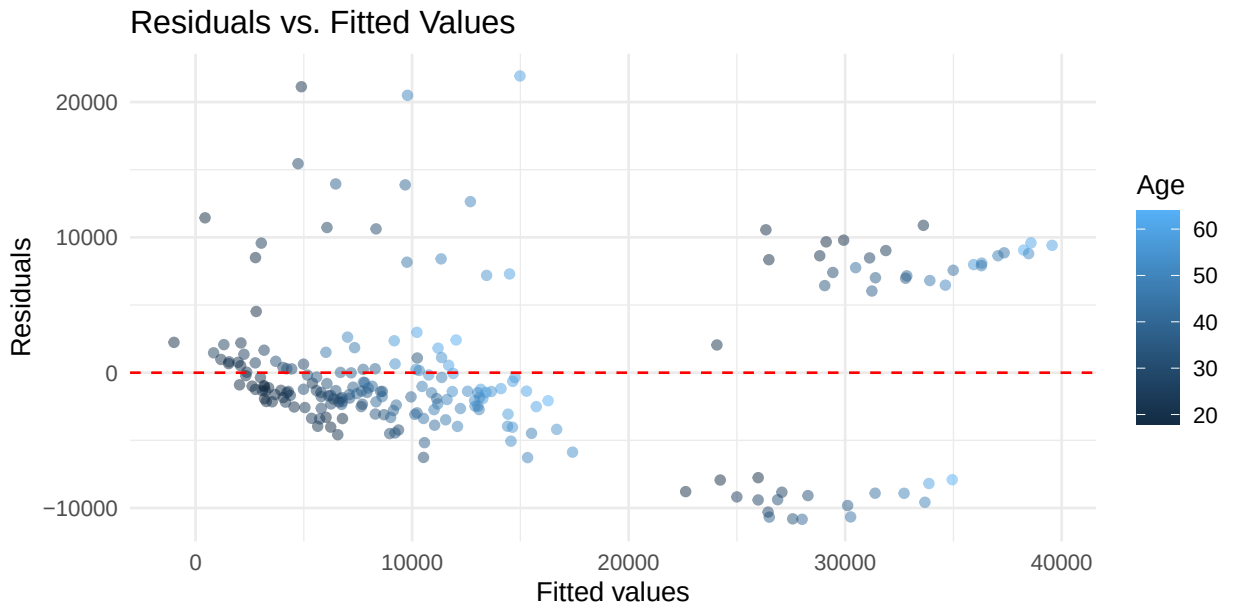
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -11728.481 | 1287.117 | -9.112 | 0.000 |
| age | 257.154 | 15.439 | 16.656 | 0.000 |
| bmi | 312.051 | 36.766 | 8.488 | 0.000 |
| children | 508.826 | 182.676 | 2.785 | 0.005 |
| is_male | -275.379 | 431.415 | -0.638 | 0.523 |
| is_smoker | 23043.122 | 535.900 | 42.999 | 0.000 |
| is_southwest | -337.982 | 530.782 | -0.637 | 0.524 |
| is_northeast | 800.547 | 531.378 | 1.507 | 0.132 |

| R_Squared | Adjusted_R_Squared |
|---|---|
| 0.723 | 0.72 |

Using the same training and testing sets as the SLR model, we see a massive improvement in model performance. Our intercept doesnt make a ton of sense, but that was expected with the fact we are using seven variables. We can also see that being a smoker makes a big different in the predicted charges. If someone is a smoker, it adds on average `$22,802` to their predicted charges. Our R^2 value is now `0.723`, over a 800% increase in performance. Even the Actual vs Predicted plot looks much better than the previous:




Looking at these residuals, we can see that the gaps are much closer for the smaller charges, but we still have pretty large residuals for charges greater than $20,000.



### Conclusion

As previously mentioned, we saw a huge increase in performance going from a simple LR model to a multiple LR model. In conclusion, it makes the most sense to use MLR in this scenario as there are not any single variables that correlate enough to the target to use SLR. Theres also room to build on our findings and use feature selection to possibly optimize the model even further. The counter point would be that it isnt great for predicting charges over `$20,000` so there is room for growth there.

### Final Reccomendation

We reccomend using a multiple linear regression model that uses a standardized dataset consisting of the columns `age`, `bmi`, `children`, `is_male`, `is_smoker`, `is_southwest`, `is_northwest`, and `is_northeast`. The model will do a solid job of predicting charges less than `$20,000`. Hatfield Insurance Company may want to consider collecting more information on each case to help improve the predictions on cases over `$20,000`. The exact parameters of the model can be seen below:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -11728.481 | 1287.117 | -9.112 | 0.000 |
| age | 257.154 | 15.439 | 16.656 | 0.000 |
| bmi | 312.051 | 36.766 | 8.488 | 0.000 |
| children | 508.826 | 182.676 | 2.785 | 0.005 |
| is_male | -275.379 | 431.415 | -0.638 | 0.523 |
| is_smoker | 23043.122 | 535.900 | 42.999 | 0.000 |
| is_southwest | -337.982 | 530.782 | -0.637 | 0.524 |
| is_northeast | 800.547 | 531.378 | 1.507 | 0.132 |

## Author Contributions

- Luke Kerwin – Modeling and Model Evaluation
- Nick Malloy – EDA and EDA Visualizations
- Dinni Bhardwaj – Report creation and formatting

## Code Appendix

```
# Packages Used
library(ggplot2)
library(tidyr)
library(dplyr)
library(kableExtra)
library(knitr)

# Load in data
data <- read.csv("/Users/lukekerwin/Downloads/activity2_insurance.csv")
data <- na.omit(data)

# Show first 10 rows of the raw data using kable
head(data, 10) %>%
  kable(
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# Simple plot of charges vs age with smoker as a color identifier
data%>%
  select(age,smoker,charges)%>%
  ggplot(aes(x = age, y = charges))+
  geom_point(aes(color = smoker))

# Plot of charges vs bmi with demographics
data %>%
  ggplot(
    mapping = aes(x = bmi, y = charges)
  ) +
  geom_point(size = 0.5) +
  theme_bw() +
  xlab("BMI") +
  ylab("Charges") +
  facet_grid(rows = vars(region), cols = vars(sex))

continuous_data = data %>%
  select(caseID, age, bmi, children, charges)
# Correlation Matrix visual using caseID, age, bmi, children and charges
cor_matrix <- cor(continuous_data)
cor_matrix %>%
  kable(
    digits = 4,
    align = "cccc",
    booktab = TRUE,
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
```

```
    full_width = FALSE
  )

# SLR Model 1 (age)
slr_model_age <- lm(
  formula = charges ~ age,
  data = data,
  na.action = "na.omit"
)
# Coefficient Table (Model 1)
coefTable <- summary(slr_model_age)$coefficients
rownames(coefTable) <- c("Intercept", "Age (years)")
coefTable %>%
  kable(
    digits = 3,
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# R2 Table (Model1)
r2Table <- data.frame(
  R_Squared = summary(slr_model_age)$r.squared,
  Adjusted_R_Squared = summary(slr_model_age)$adj.r.squared
)

r2Table %>%
  kable(
    digits = 3,
    align = "c",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# Form training set ----
set.seed(716)
trainingData <- data %>%
  slice_sample(prop = 0.8)

# Form testing set ----
testingData <- data %>%
  filter(!(caseID %in% trainingData$caseID))

# Train model ----
slr_model_age <- lm(
  formula = charges ~ age,
  data = trainingData,
```

```
  na.action = "na.omit"
)

# Predict ----
predictions <- predict(slr_model_age, newdata=testingData)

# Add predictions to data frame and find residuals ----
testingData$prediction <- predictions
testingData <- testingData %>%
  mutate(
    residuals = charges - prediction
  )

# Output Tables ----
coefTable <- summary(slr_model_age)$coefficients
rownames(coefTable) <- c("Intercept", "Age (years)")
coefTable %>%
  kable(
    digits = 3,
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

r2Table <- data.frame(
  R_Squared = summary(slr_model_age)$r.squared,
  Adjusted_R_Squared = summary(slr_model_age)$adj.r.squared
)

r2Table %>%
  kable(
    digits = 3,
    align = "c",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# Plot actual vs. predicted values
df_mlr <- data.frame(Actual = testingData$charges, Predicted = testingData$prediction)
ggplot(df_mlr, aes(x = Actual, y = Predicted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  theme_minimal() +
  labs(title = "Actual vs. Predicted Charges",
       x = "Actual Charges",
       y = "Predicted Charges")
```

```r
# Scatter plot of residuals
ggplot(testingData, aes(x=prediction, y=residuals)) +
  geom_point(aes(color=age), alpha=0.5) +  # Color points by age for added detail
  geom_hline(yintercept=0, linetype="dashed", color="red") +  # Add a horizontal line at y=0
  theme_minimal() +
  labs(
    title="Residuals vs. Fitted Values",
    x="Fitted values",
    y="Residuals",
    color="Age"
  )

categorical_vars = c("sex","smoker","region")
data$is_male = 0
data$is_male[which(data$sex == "male")] = 1

data$is_smoker = 0
data$is_smoker[which(data$smoker == "yes")] = 1

data$is_southwest = 0
data$is_northwest = 0
data$is_northeast = 0

data$is_southwest[which(data$region == "southwest")] = 1
data$is_northeast[which(data$region == "northeast")] = 1
data$is_northeast[which(data$region == "northeast")] = 1

standardized <- data %>%
  select("caseID","age","bmi","children","charges","is_male","is_smoker","is_southwest","is_northwest",

head(standardized, 10) %>%
  kable(
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# Form training set ----
set.seed(716)
trainingData <- standardized %>%
  slice_sample(prop = 0.8)

# Form testing set ----
testingData <- standardized %>%
  filter(!(caseID %in% trainingData$caseID))

# Train MLR model using all features
formula <- charges ~ age + bmi + children + is_male + is_smoker + is_southwest + is_northwest + is_nortl
model_mlr <- lm(formula, data = trainingData)
```

```r
# Predict on test data
predictions_mlr <- predict(model_mlr, newdata = testingData)

# Add predictions to data frame and find residuals ----
testingData$prediction <- predictions_mlr
testingData <- testingData %>%
  mutate(
    residuals = charges - prediction
  )

# Output Tables ----
coefTable <- summary(model_mlr)$coefficients
coefTable %>%
  kable(
    digits = 3,
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

r2Table <- data.frame(
  R_Squared = summary(model_mlr)$r.squared,
  Adjusted_R_Squared = summary(model_mlr)$adj.r.squared
)

r2Table %>%
  kable(
    digits = 3,
    align = "c",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )

# Plot actual vs. predicted values
df_mlr <- data.frame(Actual = testingData$charges, Predicted = predictions_mlr)
ggplot(df_mlr, aes(x = Actual, y = Predicted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  theme_minimal() +
  labs(title = "Actual vs. Predicted Charges",
       x = "Actual Charges",
       y = "Predicted Charges")

# Scatter plot of residuals
ggplot(testingData, aes(x=prediction, y=residuals)) +
  geom_point(aes(color=age), alpha=0.5) +  # Color points by age for added detail
  geom_hline(yintercept=0, linetype="dashed", color="red") +  # Add a horizontal line at y=0
```

```
  theme_minimal() +
  labs(
    title="Residuals vs. Fitted Values",
    x="Fitted values",
    y="Residuals",
    color="Age"
  )
coefTable <- summary(model_mlr)$coefficients
coefTable %>%
  kable(
    digits = 3,
    align = "cccc",
    booktab = TRUE
  ) %>%
  kableExtra::kable_classic(
    latex_options = c("HOLD_position"),
    full_width = FALSE
  )
```