Foundations of Data Science

Nathaniel Coulter

Due: May 13, 2025

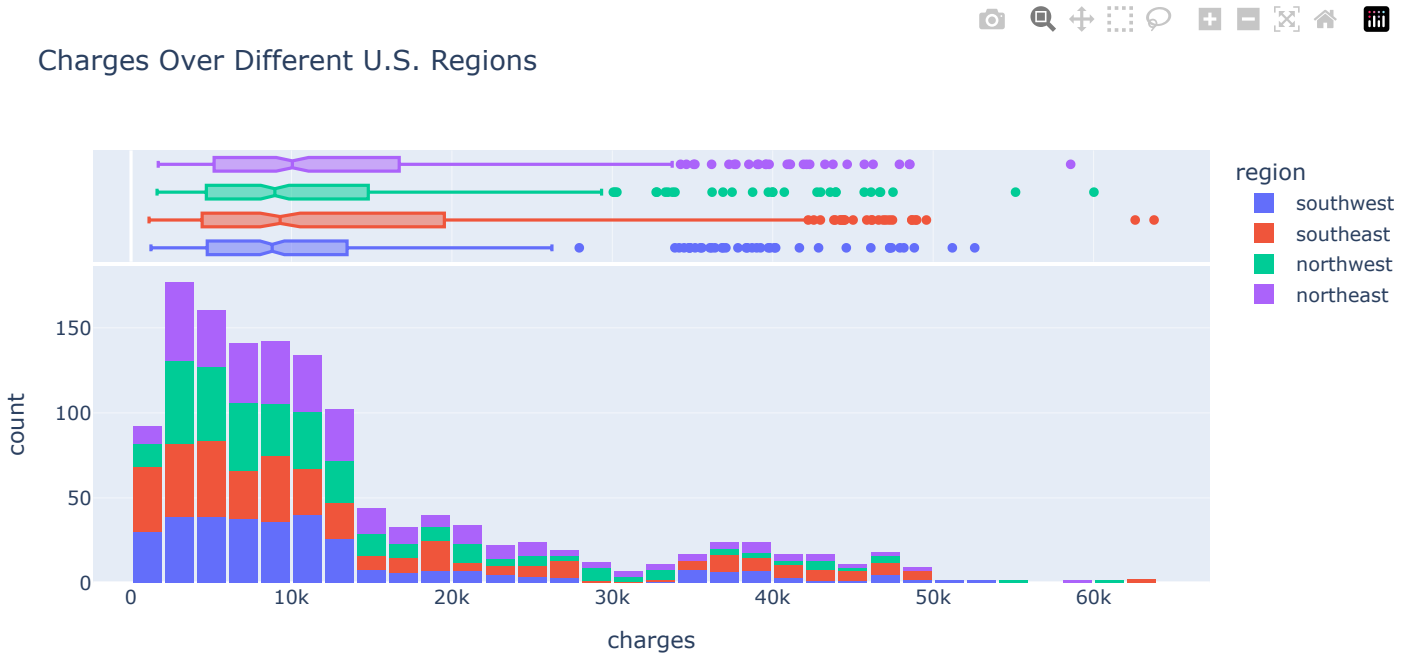
How Health Insurance Premiums vary based on individual demographics:

We chose to use the “US Health Insurance Dataset” from Kaggle. https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data (https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data)

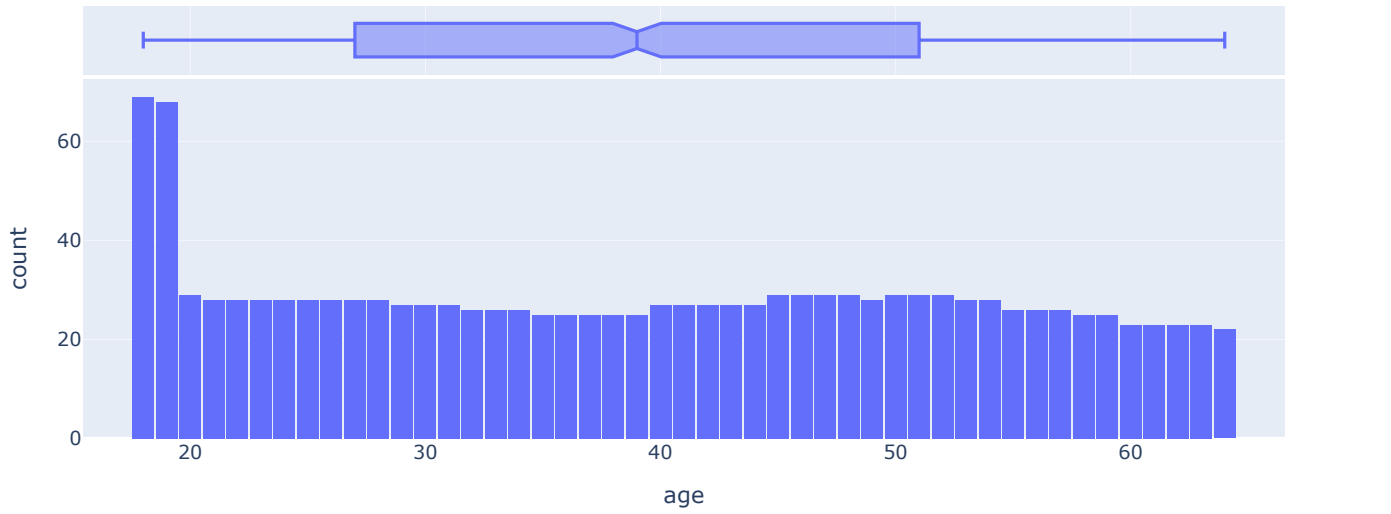
Our dataset includes multiple variables relating to a person’s health and demographic that we used to discover trends and relationships insurance companies are using to price annual premium’s (relative to specific variablea). Variables in the dataset include, age, sex, BMI (Body Mass Index), how many children they have, if a person Smokes and geographic region.

```
# Packages we will use.
library(reticulate)
use_python("C:/Users/hocke/AppData/Local/Programs/Python/Python313/python.exe", required = TRUE)

library(ggplot2)
library(dplyr)
library(reshape2)
library(factoextra)

insurance <- read.csv("C:/Users/hocke/OneDrive/Documents/R Programs (Intro to Data Science Class)/Data Sets/insurance5yrs.
csv")
```

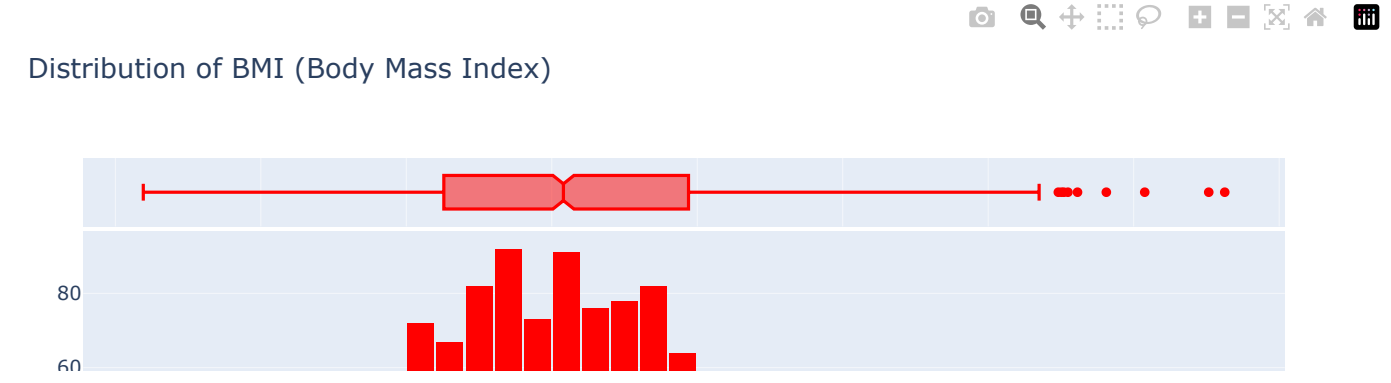
```
import pandas as pd
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import os
from contextlib import redirect_stdout


sns.set_style('darkgrid')
plt.rcParams['font.size'] = 14
plt.rcParams['figure.figsize'] = (10, 6)
plt.rcParams['figure.facecolor'] = '#00000000'

medical_df = pd.read_csv("C:/Users/hocke/OneDrive/Documents/R Programs (Intro to Data Science Class)/Data Sets/insurance5y
rs.csv")

with redirect_stdout(open(os.devnull, 'w')):
    fig_age = px.histogram(medical_df, x='age', marginal='box', nbins=47, title='Distribution of Age')
    fig_age.update_layout(bargap=0.1)

    fig_bmi = px.histogram(medical_df, x='bmi', marginal='box', color_discrete_sequence=['red'], title='Distribution of BM
I (Body Mass Index)')
    fig_bmi.update_layout(bargap=0.1)

    fig_charges_smoker = px.histogram(medical_df, x='charges', marginal='box', color='smoker',
                                      color_discrete_sequence=['red', 'grey'], title='Annual Medical Charges by Smoking St
atus')
    fig_charges_smoker.update_layout(bargap=0.1)

    fig_charges_gender = px.histogram(medical_df, x='charges', color='sex',
                                      color_discrete_sequence=['blue', 'red'], title='Different Charges Over Genders')
    fig_charges_gender.update_layout(bargap=0.1)

    fig_charges_region = px.histogram(medical_df, x='charges', marginal='box', color='region',
                                      title='Charges Over Different U.S. Regions')
    fig_charges_region.update_layout(bargap=0.1)
```

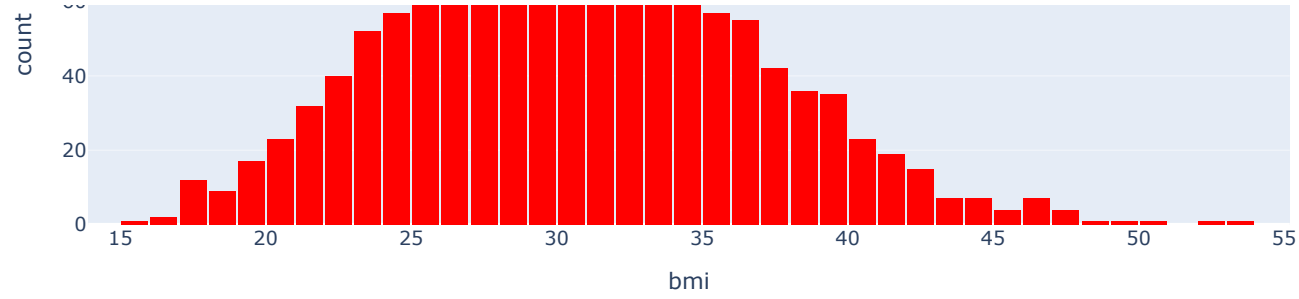



```
fig_age.show()
```



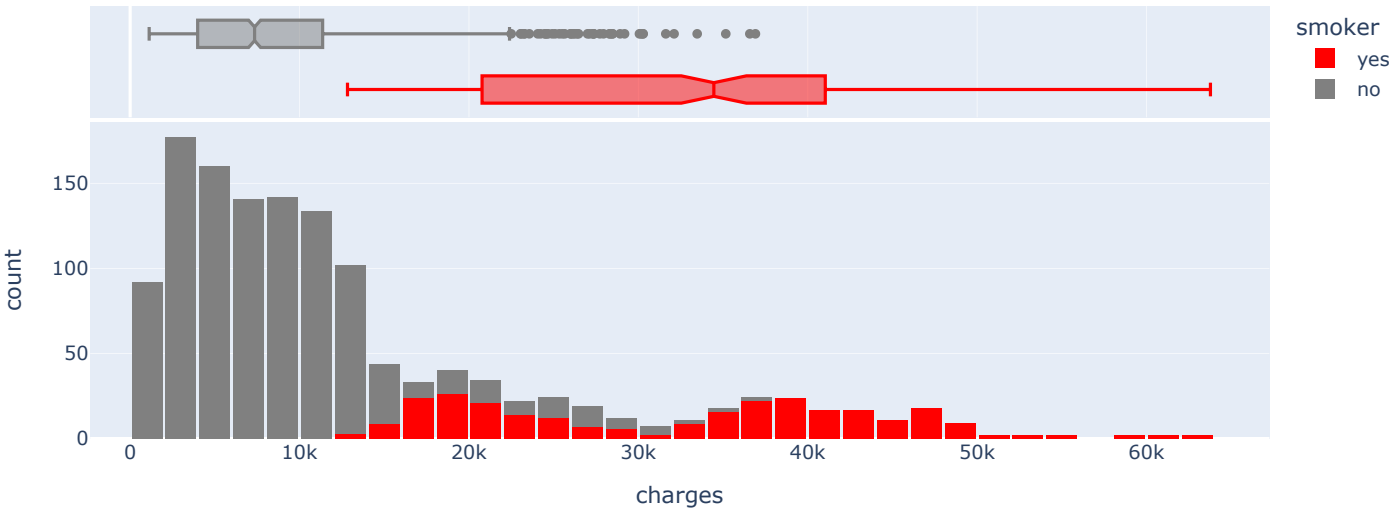

```
fig_bmi.show()
```





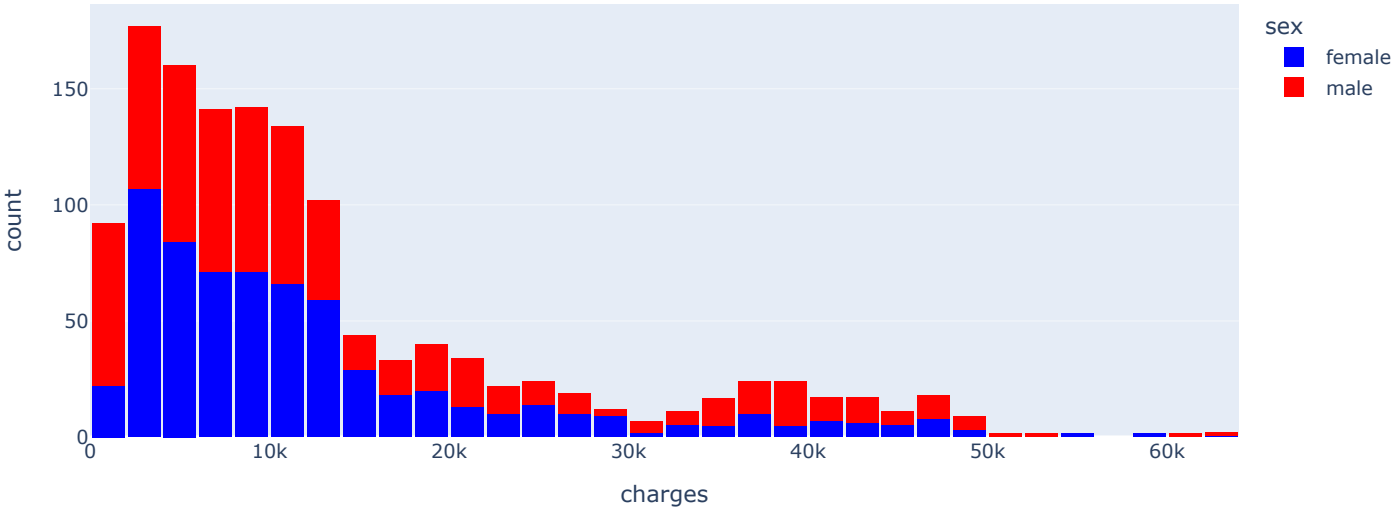

```
fig_charges_smoker.show()
```

Annual Medical Charges by Smoking Status

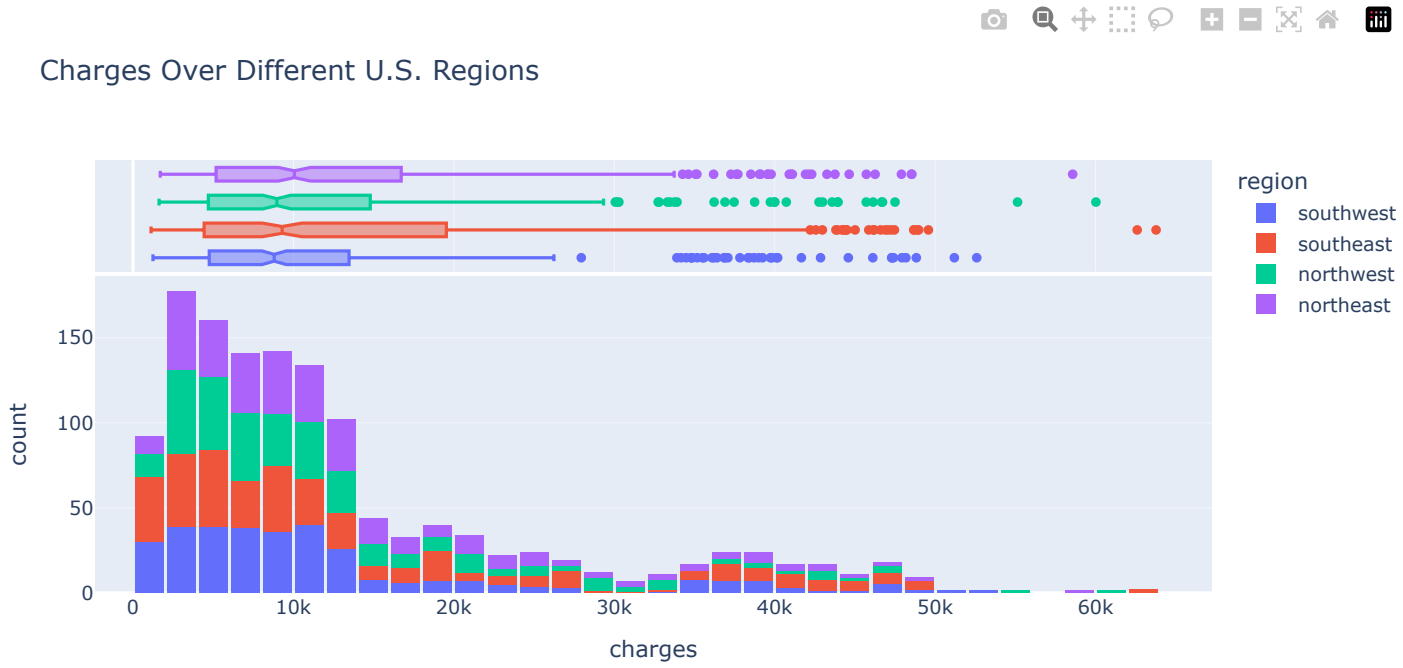



```
fig_charges_gender.show()
```

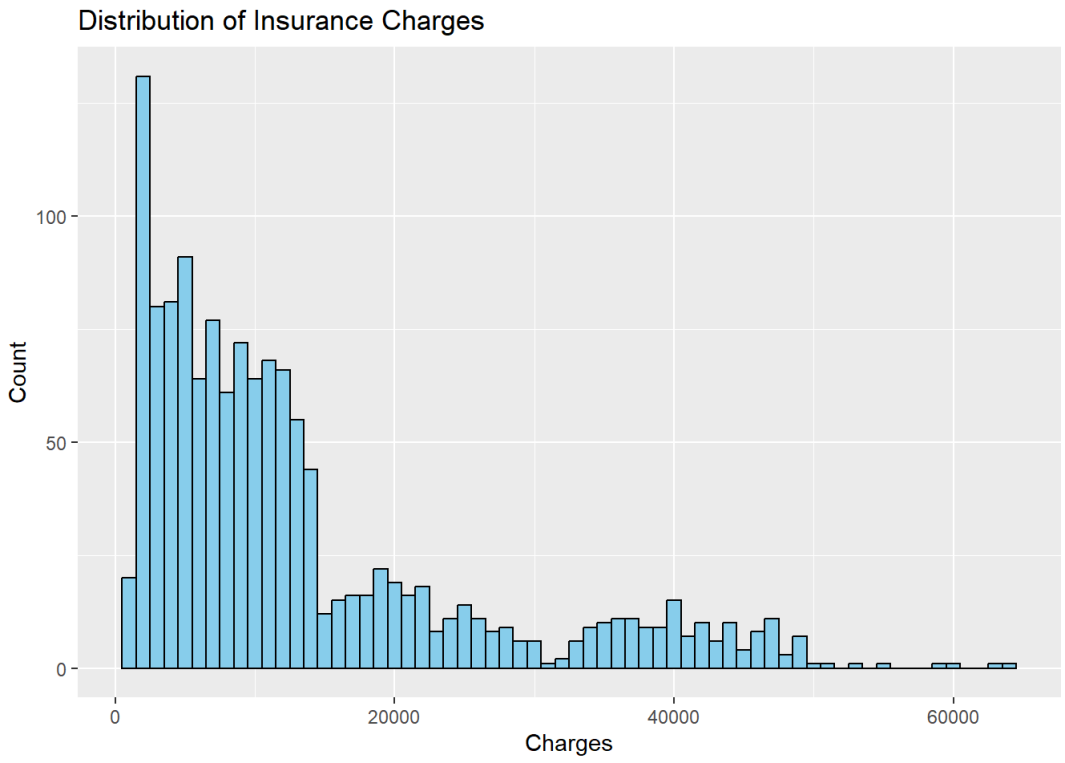
Different Charges Over Genders



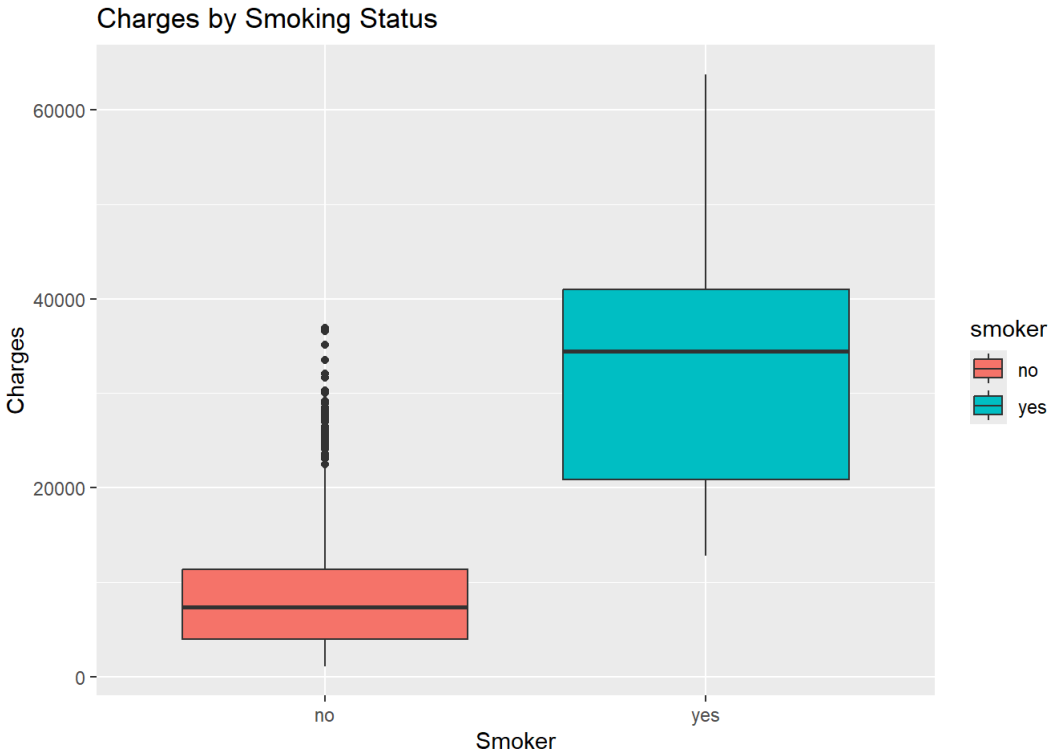
```
fig_charges_region.show()
```



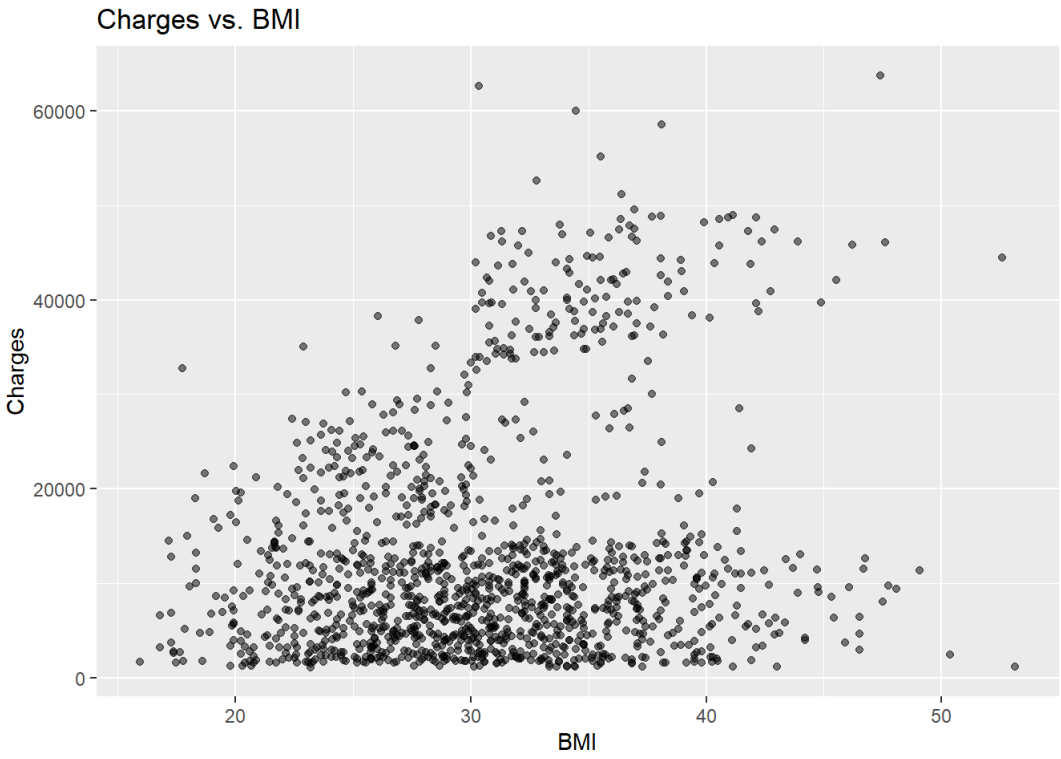

```
library(ggplot2)
library(dplyr)

# 1. Distribution of Charges
ggplot(insurance, aes(x = charges)) +
  geom_histogram(binwidth = 1000, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Insurance Charges", x = "Charges", y = "Count")
```



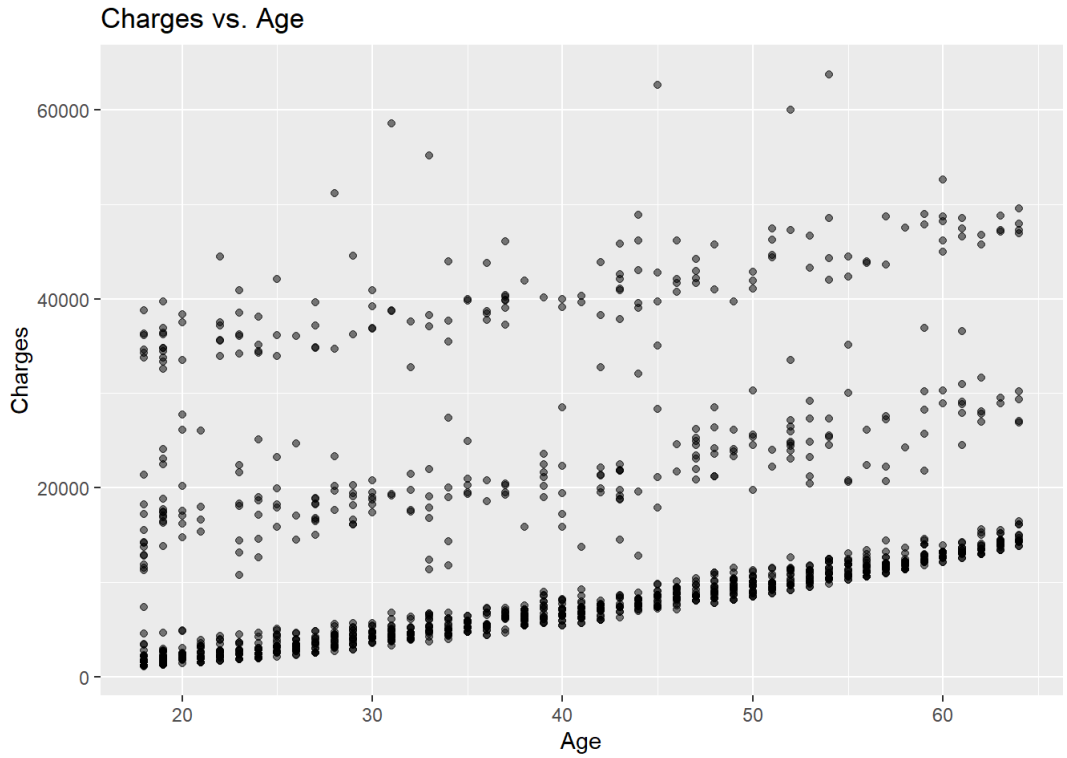
```
# 2. Charges by Smoking Status
ggplot(insurance, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot() +
  labs(title = "Charges by Smoking Status", x = "Smoker", y = "Charges")
```



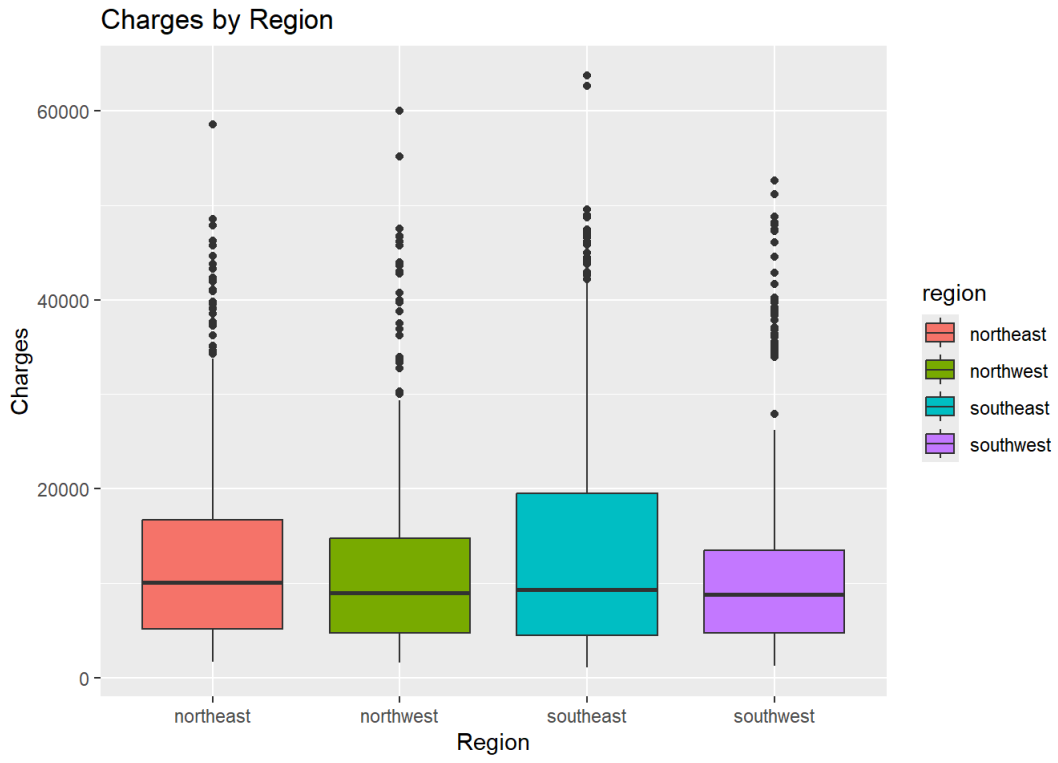

```
# 3. Charges by BMI
ggplot(insurance, aes(x = bmi, y = charges)) +
  geom_point(alpha = 0.5) +
  labs(title = "Charges vs. BMI", x = "BMI", y = "Charges")
```



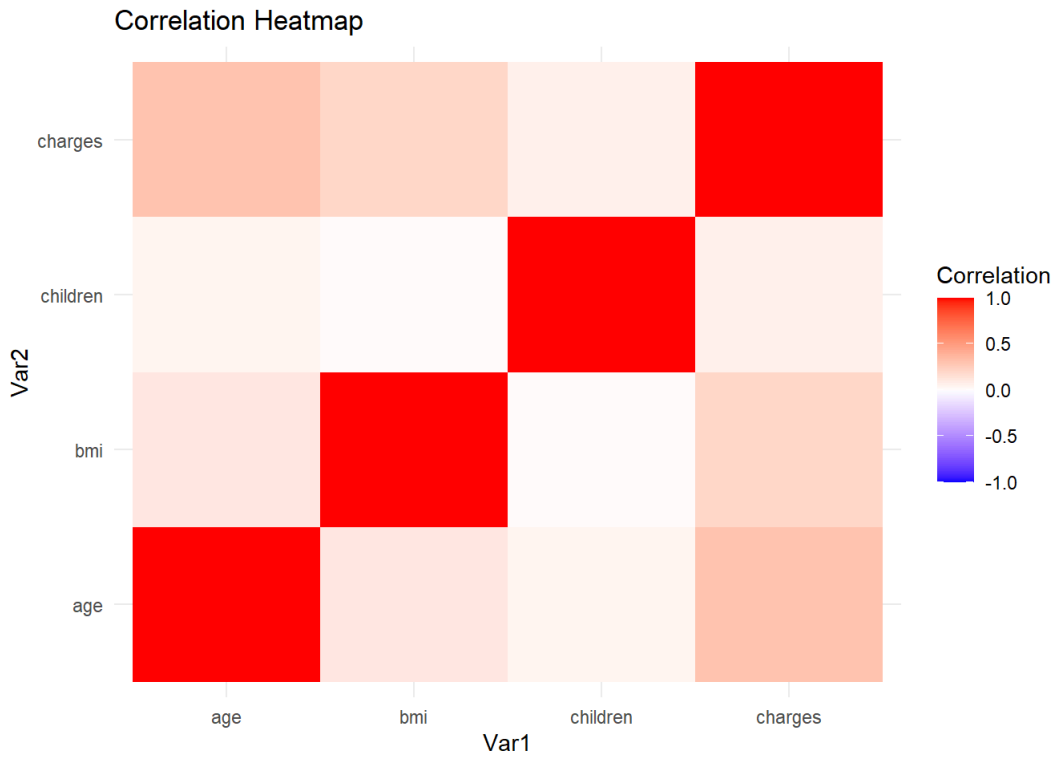
```
# 4. Charges by Age
ggplot(insurance, aes(x = age, y = charges)) +
  geom_point(alpha = 0.5) +
  labs(title = "Charges vs. Age", x = "Age", y = "Charges")
```




```
# 5. Charges by Region
ggplot(insurance, aes(x = region, y = charges, fill = region)) +
  geom_boxplot() +
  labs(title = "Charges by Region", x = "Region", y = "Charges")
```



```
# 6. Correlation Heatmap
# Calculate correlation matrix
numeric_vars <- insurance %>% select(age, bmi, children, charges)
cor_matrix <- cor(numeric_vars)

library(reshape2)
melted_cor <- melt(cor_matrix)

# Heatmap
ggplot(data = melted_cor, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Correlation") +
  theme_minimal() +
  labs(title = "Correlation Heatmap")
```



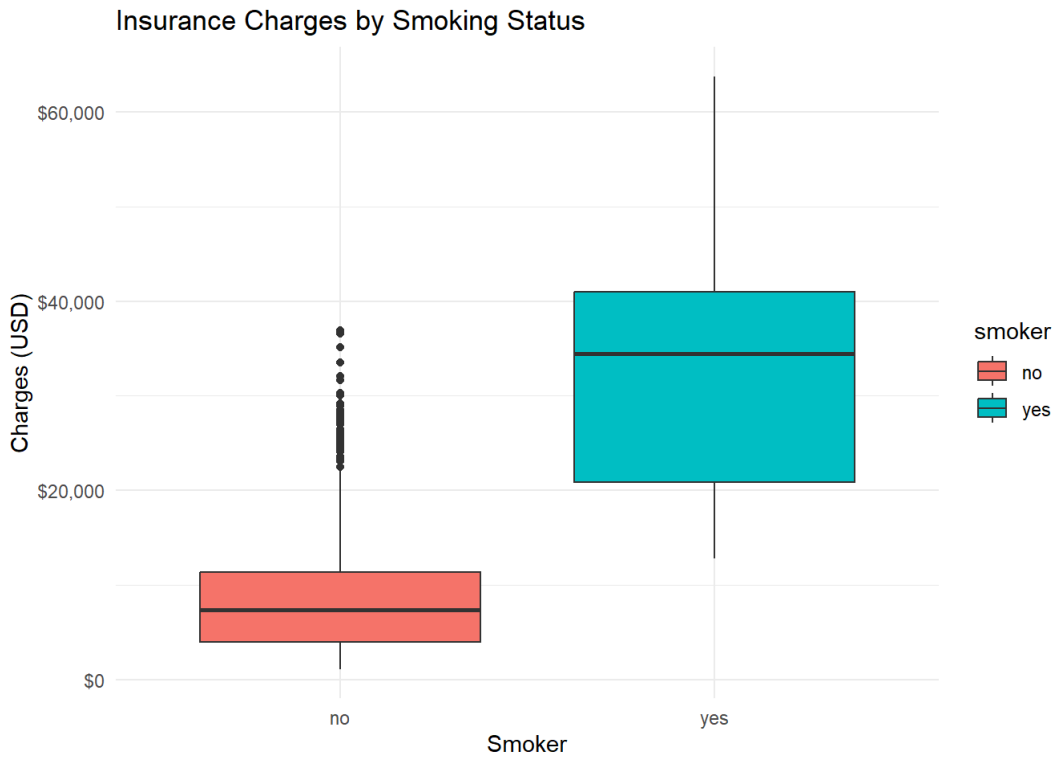
Is smoking a stronger predictor of premium than BMI or age? - Multiple Linear Regression Model

```
insurance <- read.csv("C:/Users/hocke/OneDrive/Documents/R Programs (Intro to Data Science Class)/Data Sets/insurance5yrs.
csv")

insurance$smoker <- as.factor(insurance$smoker)
insurance$region <- as.factor(insurance$region)
insurance$sex <- as.factor(insurance$sex)

lm_model <- lm(charges ~ age + bmi + smoker + region + children + sex, data = insurance)
summary(lm_model)
```

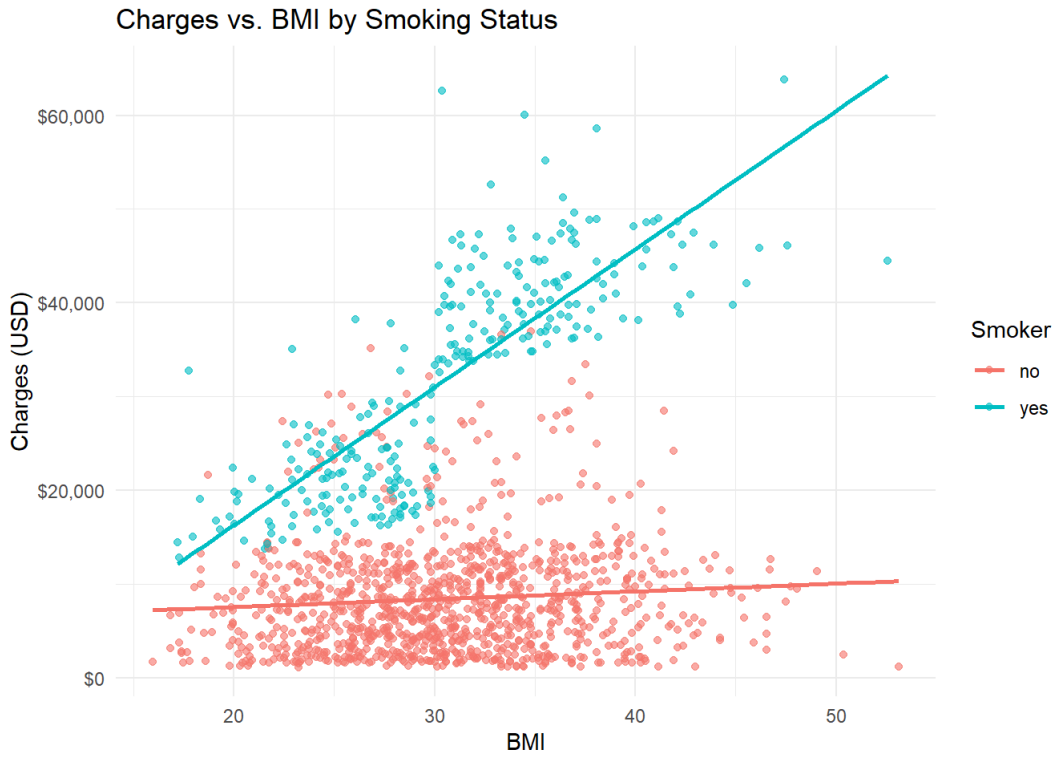
```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker + region + children +
##     sex, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (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 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## children           475.5      137.8   3.451 0.000577 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## ---
## 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
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Visual Representations

```
library(ggplot2)

ggplot(insurance, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(
    title = "Insurance Charges by Smoking Status",
    x = "Smoker",
    y = "Charges (USD)"
  ) +
  theme_minimal()
```



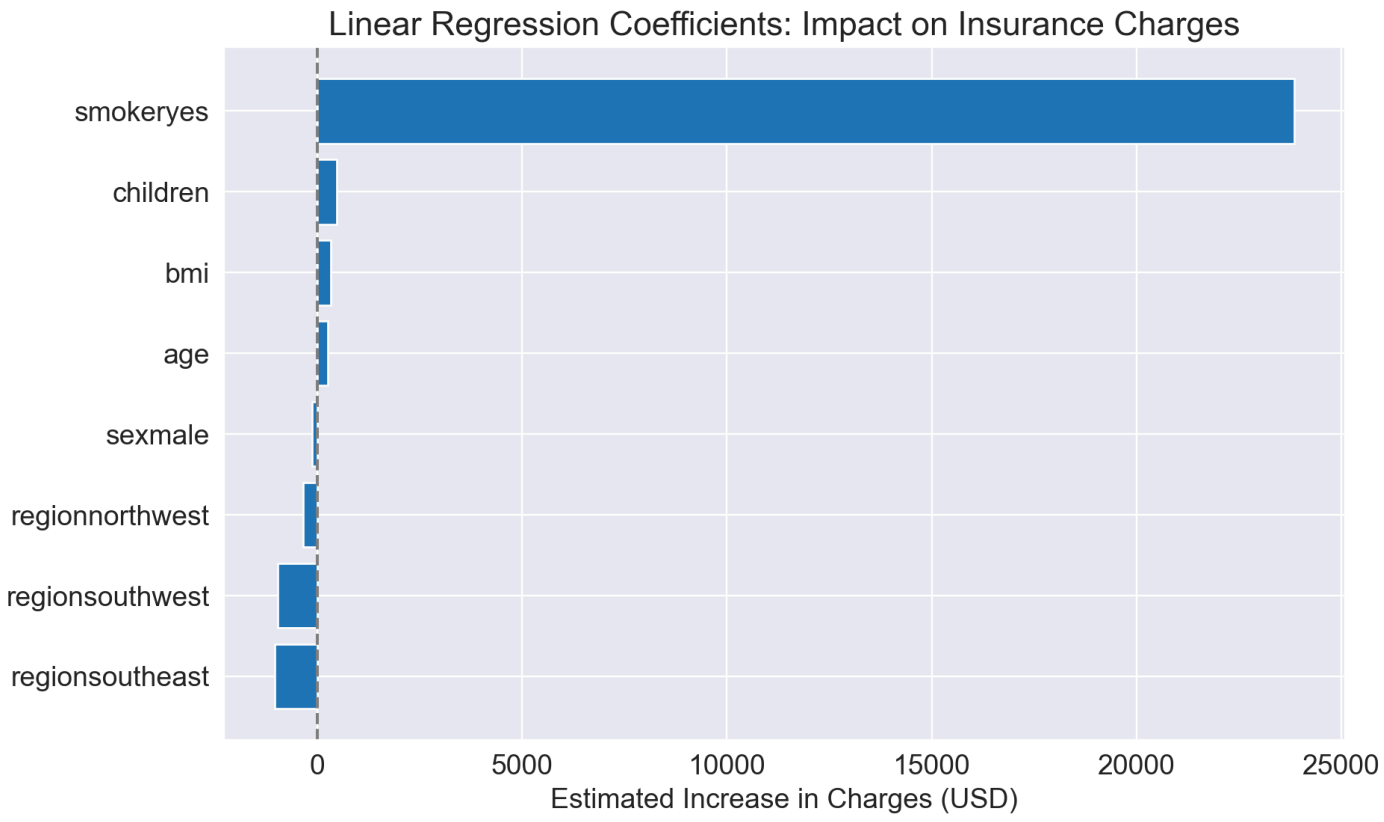

```
ggplot(insurance, aes(x = bmi, y = charges, color = smoker)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(
    title = "Charges vs. BMI by Smoking Status",
    x = "BMI",
    y = "Charges (USD)",
    color = "Smoker"
  ) +
  theme_minimal()
```



Boxplot: Smokers predictably face significantly higher insurance premiums than non-smokers. With a much wider spread and higher median cost.

Scatterplot: The charges will generally increase with BMI, and still, smokers are charged substantially more than non-smokers at every BMI level. Note: his is indicating a compounding effect (increased age + smoker = 2x higer risk variables)

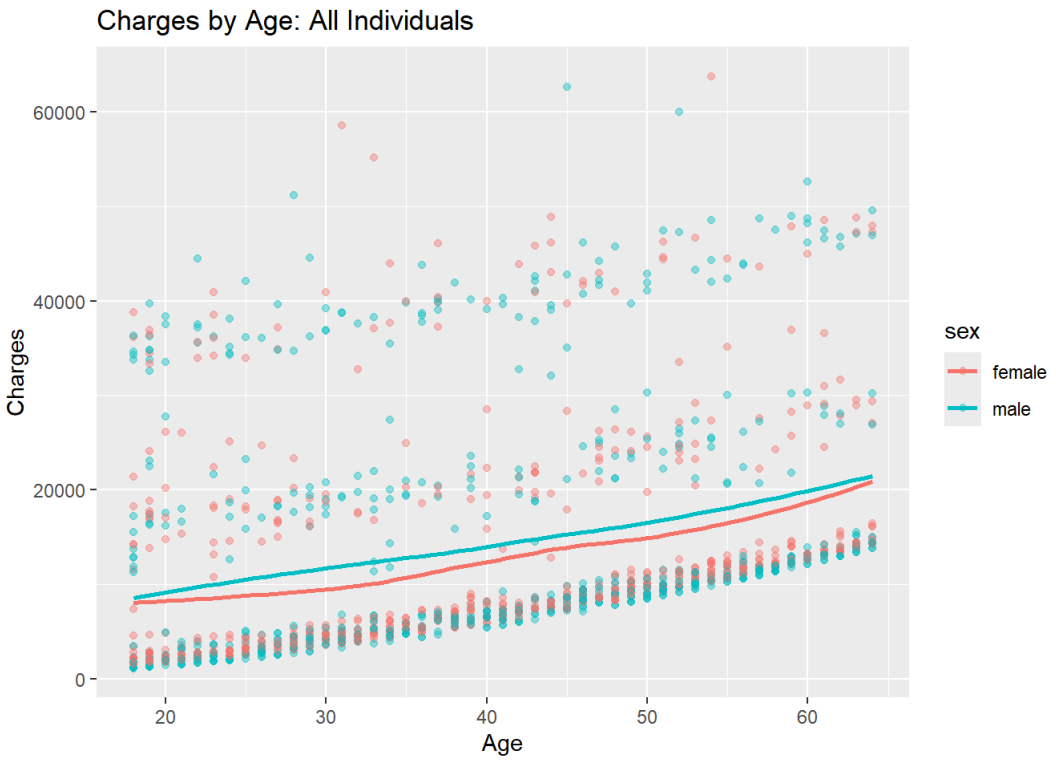
```
import pandas as pd
import matplotlib.pyplot as plt

coefficients = {
    'smokeryes': 23848.5,
    'age': 256.9,
    'bmi': 339.2,
    'children': 475.5,
    'regionsoutheast': -1035.0,
    'regionsouthwest': -960.0,
    'regionnorthwest': -353.0,
    'sexmale': -131.3
}

coef_df = pd.DataFrame(coefficients.items(), columns=['Variable', 'Estimate'])
coef_df = coef_df.sort_values(by='Estimate', ascending=False)

plt.figure(figsize=(10, 6))
plt.barh(coef_df['Variable'], coef_df['Estimate'])
plt.title('Linear Regression Coefficients: Impact on Insurance Charges')
plt.xlabel('Estimated Increase in Charges (USD)')
plt.axvline(0, color='gray', linestyle='--')
plt.tight_layout()
plt.gca().invert_yaxis()
plt.grid(True, axis='x')
plt.show()
```



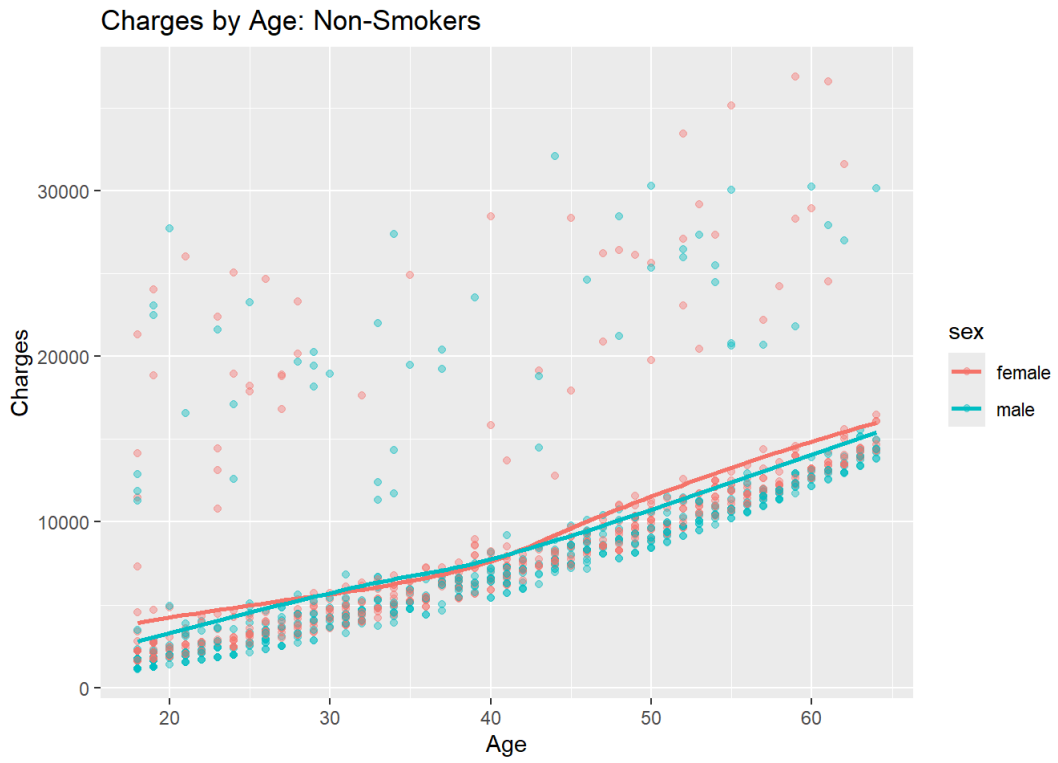
Conclusion: Our multiple linear regression model shows that smoking has the greatest impact on insurance premiums. As shown, the coefficient for smokers is approximately $23,848. Meaning, on average when holding all other variables constant, that is how much smoking will increase a person's Health Insurance Premium.

Also I'd like to state, the effect becomes SIGNIFICANTLY larger versus the same $ increases with age $257 per year or BMI $339 per unit. Which makes smoking the "best" (strongest) predictor of health insurance premiums present in our data set. Visuals included below.

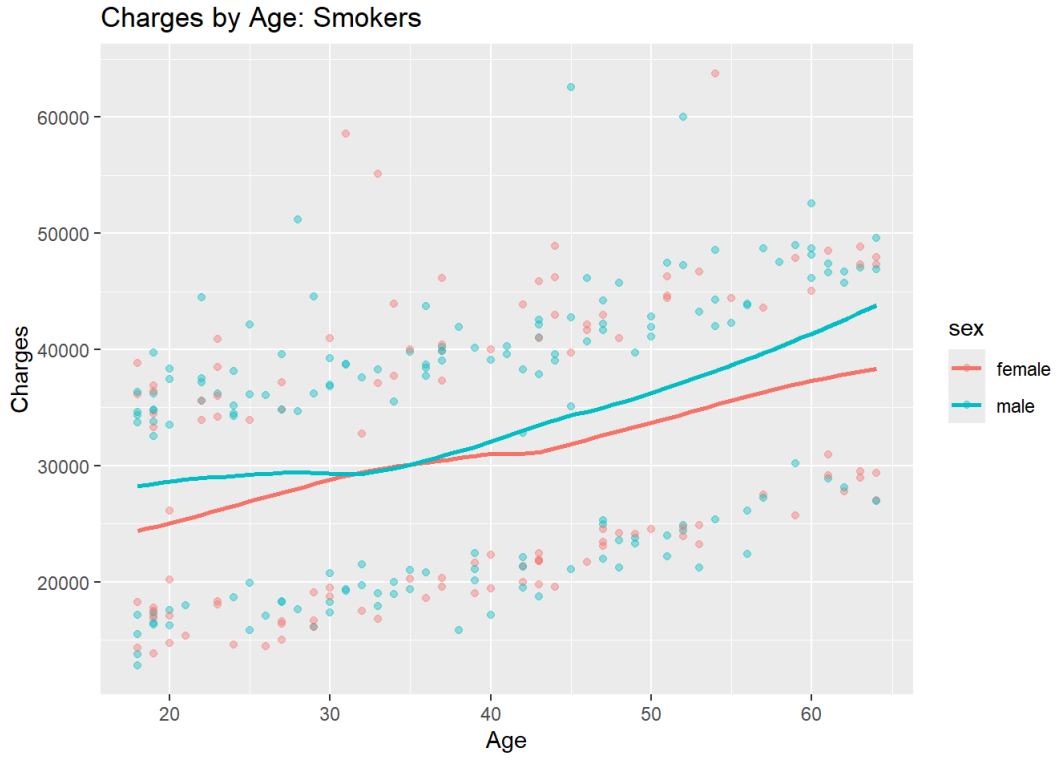
```
library(ggplot2)
library(dplyr)

data_all <- insurance

# 1. All
ggplot(data_all, aes(x = age, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by Age: All Individuals", x = "Age", y = "Charges")
```



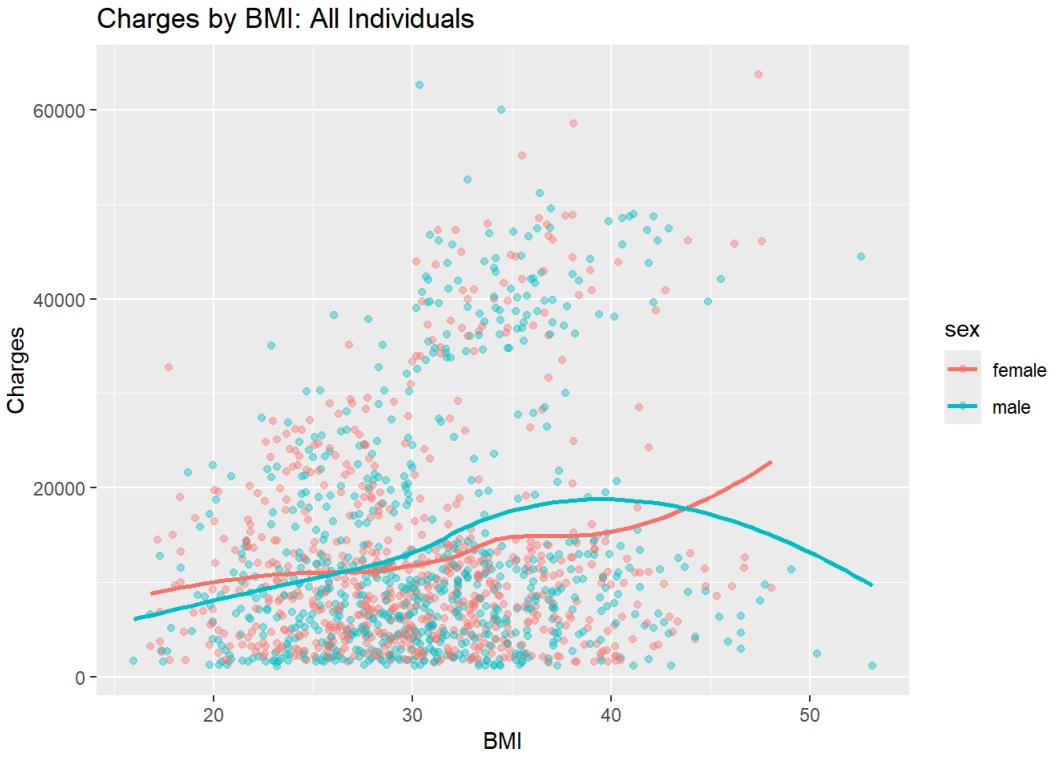

```
# 2. Non-smokers (only)
data_nonsmokers <- filter(insurance, smoker == "no")

ggplot(data_nonsmokers, aes(x = age, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by Age: Non-Smokers", x = "Age", y = "Charges")
```



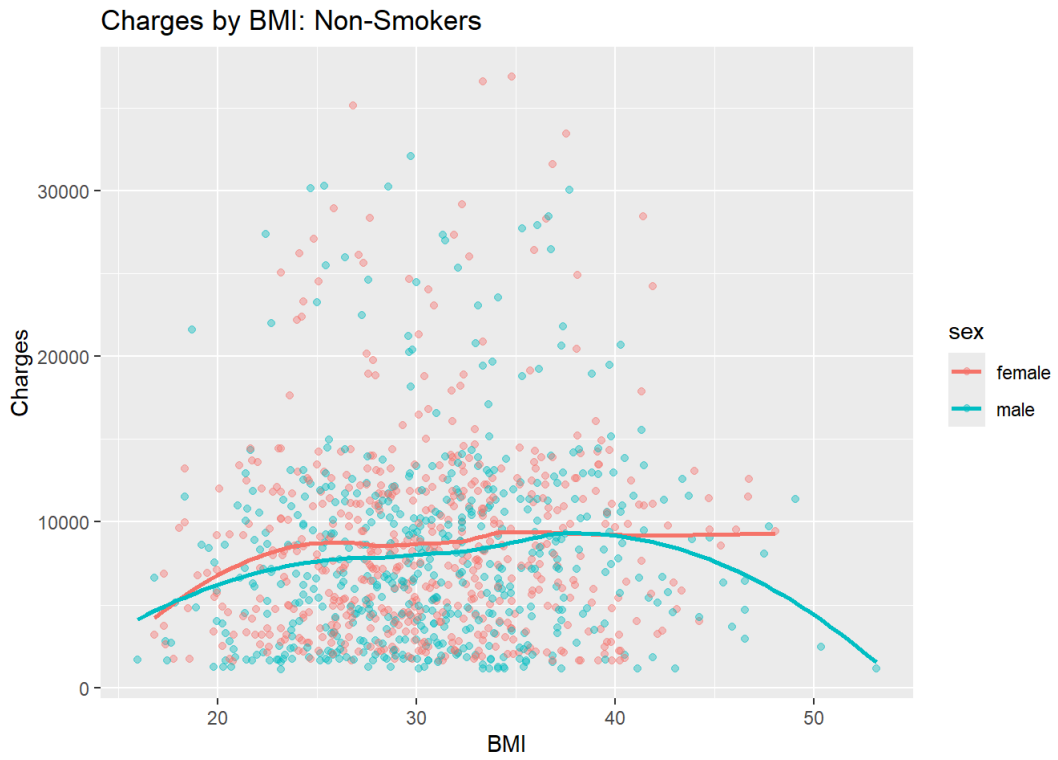

```
# 3. Smokers (only)
data_smokers <- filter(insurance, smoker == "yes")

ggplot(data_smokers, aes(x = age, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by Age: Smokers", x = "Age", y = "Charges")
```



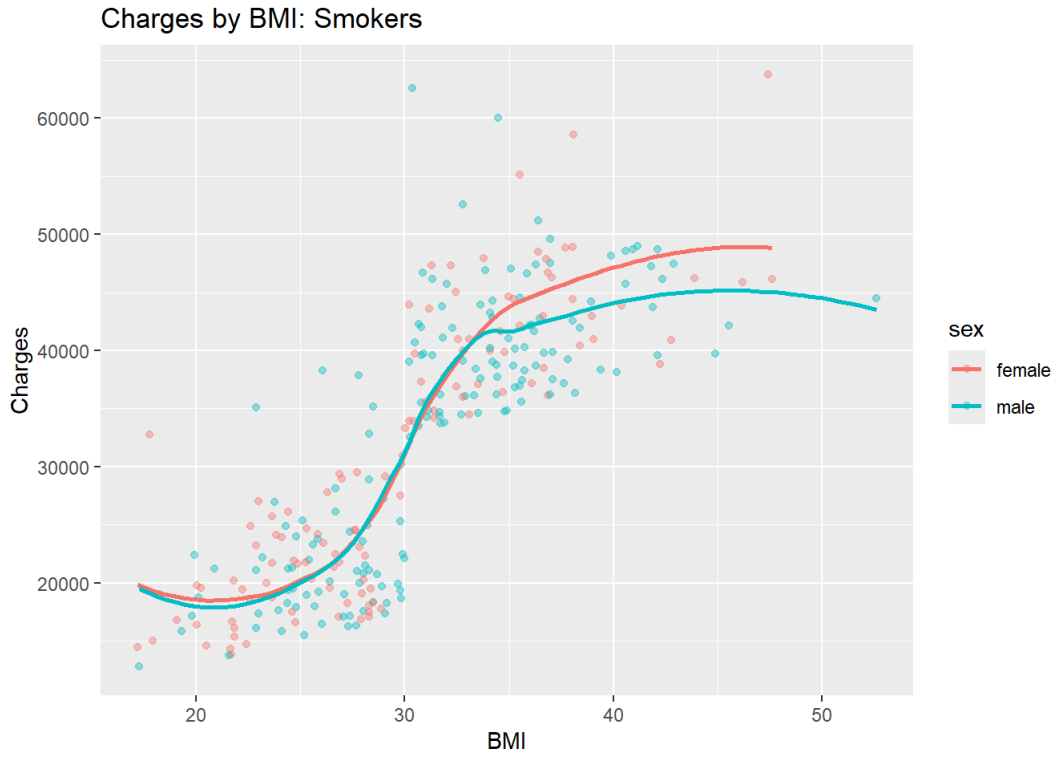
```r
library(ggplot2)
library(dplyr)

# 1. All
ggplot(insurance, aes(x = bmi, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by BMI: All Individuals", x = "BMI", y = "Charges")
```




```r
# 2. Non-smokers only
ggplot(filter(insurance, smoker == "no"), aes(x = bmi, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by BMI: Non-Smokers", x = "BMI", y = "Charges")
```




```
# 3. Smokers only
ggplot(filter(insurance, smoker == "yes"), aes(x = bmi, y = charges, color = sex)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Charges by BMI: Smokers", x = "BMI", y = "Charges")
```



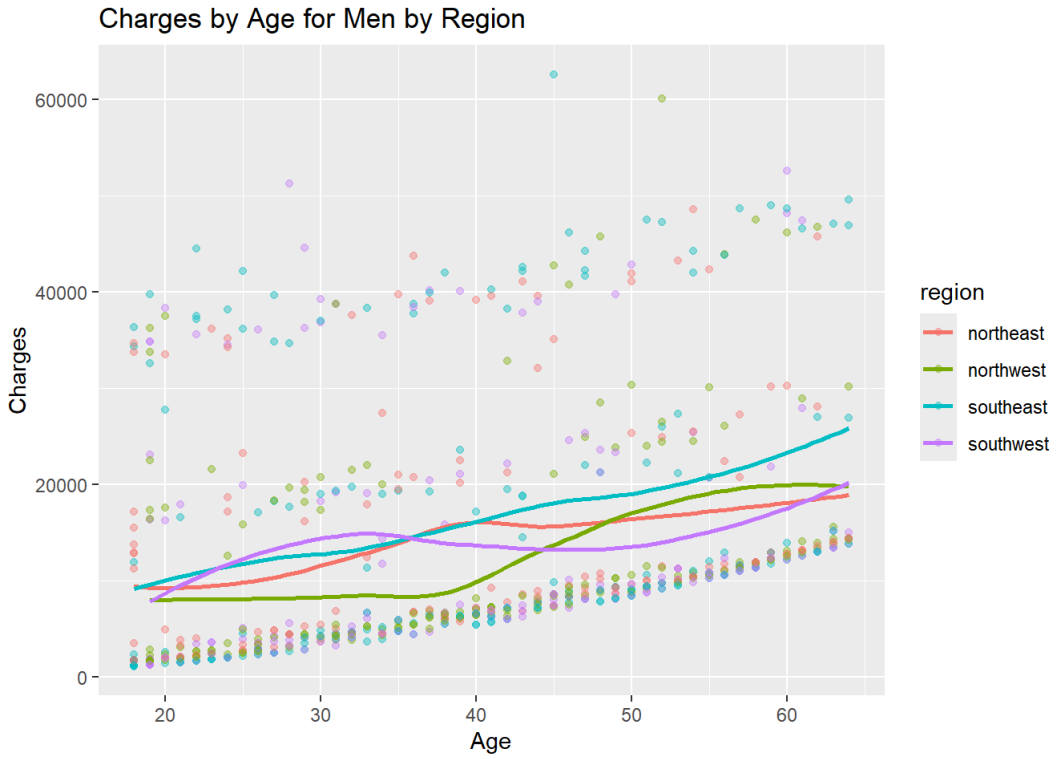
#Trend: I noticed was the higher charges for woman which becomes increasingly noticeable on the second halves of each graph.

Possible Explanations: Women in the dataset may be slightly older or have more health-related charges. (We tested this even with bmi held constant) 1.) Middle aged women may have higher medical expenses related to reproductive health, hormonal care, or chronic condition management. 2.) If health events are costly and occur frequently in the dataset, their respective charges will reflect that.

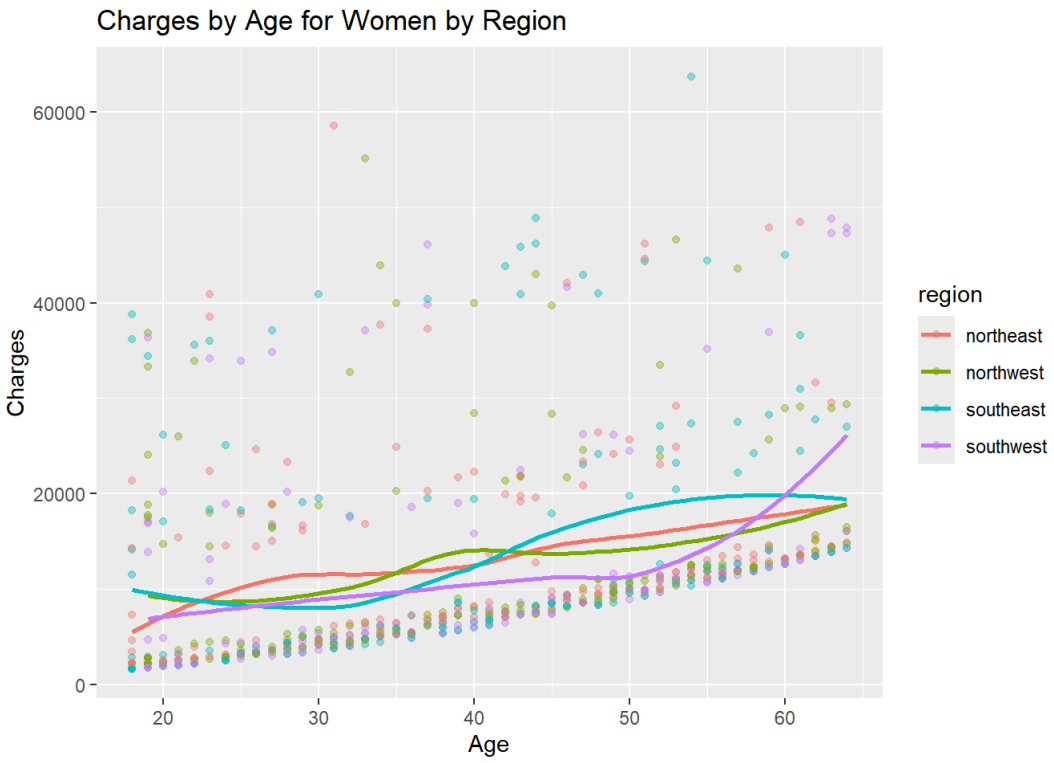
I also ran a summary() on age and charges grouped by sex via: insurance %>% group_by(sex) %>% summarise(avg_age = mean(age), avg_charge = mean(charges)) thinking there might be fewer high-BMI men in the same age range…

Since that isn't the case, I concluded that the data was well collected with each group being fairly well represented even when I started eliminating variables. So since health conditions are ommitted from the data, my best guess is pregnancy is the likely cause of insurance being higher for women around age 30, even though that doesn't explain the fact that this trend is sustained past typical child rearing age.

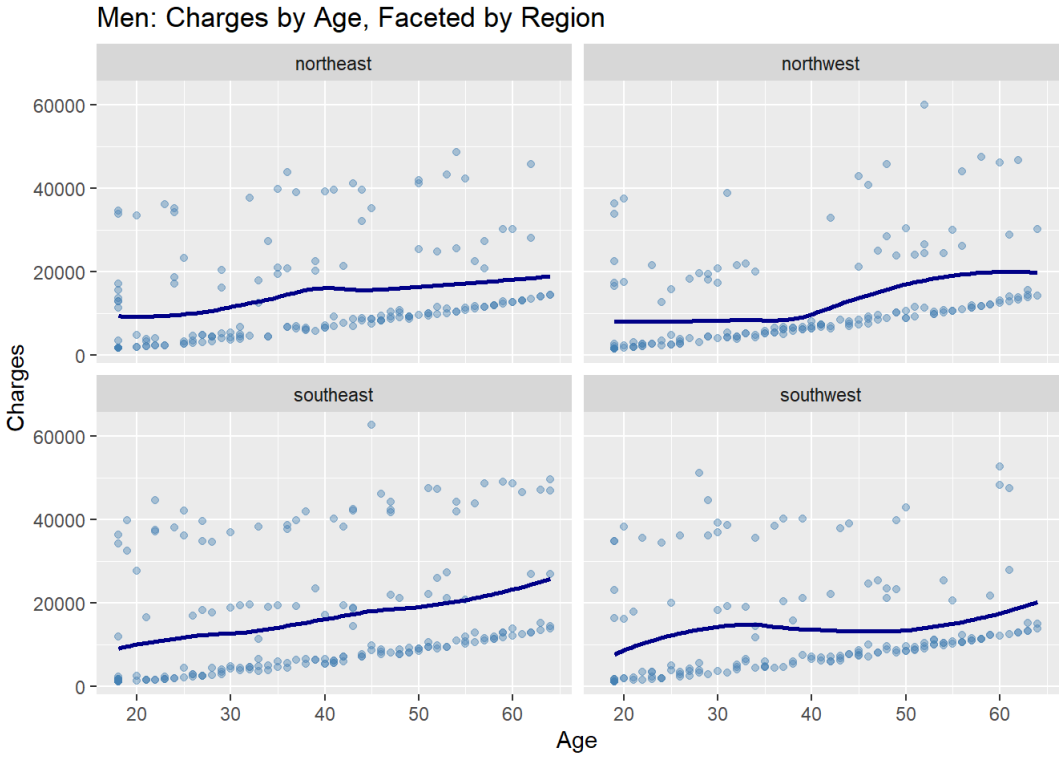
```
library(ggplot2)
library(dplyr)

men_data <- filter(insurance, sex == "male")
women_data <- filter(insurance, sex == "female")

# 1. Charges vs. Age for Men by Region
ggplot(men_data, aes(x = age, y = charges, color = region)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(
    title = "Charges by Age for Men by Region",
    x = "Age",
    y = "Charges"
  )
```



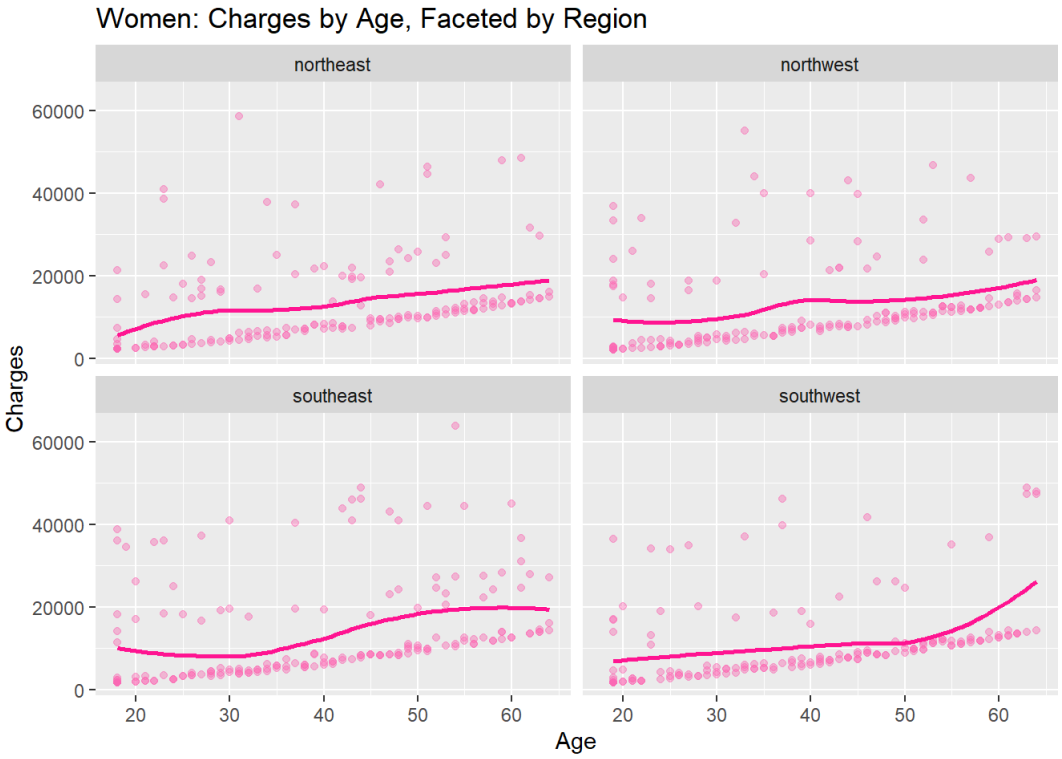

```
# 2. Charges vs. Age for Women by Region
ggplot(women_data, aes(x = age, y = charges, color = region)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(
    title = "Charges by Age for Women by Region",
    x = "Age",
    y = "Charges"
  )
```




```
# 1. Men: Charges vs. Age, Faceted by Region
ggplot(men_data, aes(x = age, y = charges)) +
  geom_point(alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "loess", se = FALSE, color = "darkblue") +
  facet_wrap(~ region) +
  labs(
    title = "Men: Charges by Age, Faceted by Region",
    x = "Age",
    y = "Charges"
  )
```

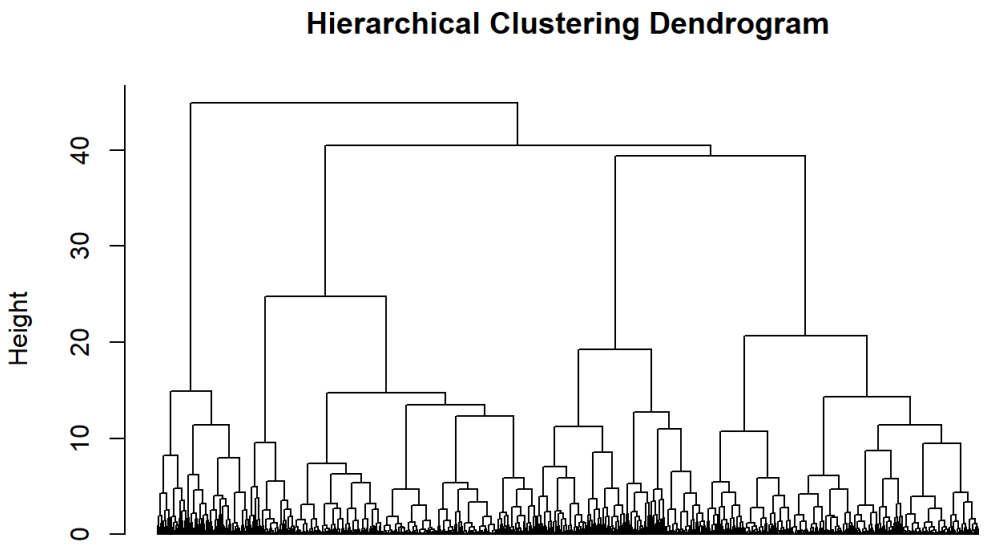


```
# 2. Women: Charges vs. Age, Faceted by Region
ggplot(women_data, aes(x = age, y = charges)) +
  geom_point(alpha = 0.4, color = "hotpink") +
  geom_smooth(method = "loess", se = FALSE, color = "deeppink") +
  facet_wrap(~ region) +
  labs(
    title = "Women: Charges by Age, Faceted by Region",
    x = "Age",
    y = "Charges"
  )
```



Hierarchical Clustering:

```
numeric_data <- insurance[, c("age", "bmi", "children", "charges")]
scaled_data <- scale(numeric_data)
dist_matrix <- dist(scaled_data, method = "euclidean")
hc <- hclust(dist_matrix, method = "ward.D2")

plot(hc, labels = FALSE, hang = -1, main = "Hierarchical Clustering Dendrogram", xlab = "", sub = "")
```



The dendrogram reveals nested groupings among individuals based on age, BMI, children, and charges, highlighting natural similarity clusters that reflect different health risk profiles.

Note I scaled the data: Scaling transforms numeric variables so they're on the same scale by converting them to z-scores (usually mean = 0, standard deviation = 1).

We scaled the variables (age, bmi, children, charges) because they're measured in different units and have different ranges (EX: charges can be tens of thousands, children is single digits). Without scaling, variables with large values (aka charges) would dominate any of our distance calculations thus skewing our clustering results. So basically, scaling ensures everything contributes equally to how similarity is measured in our dendrogram.

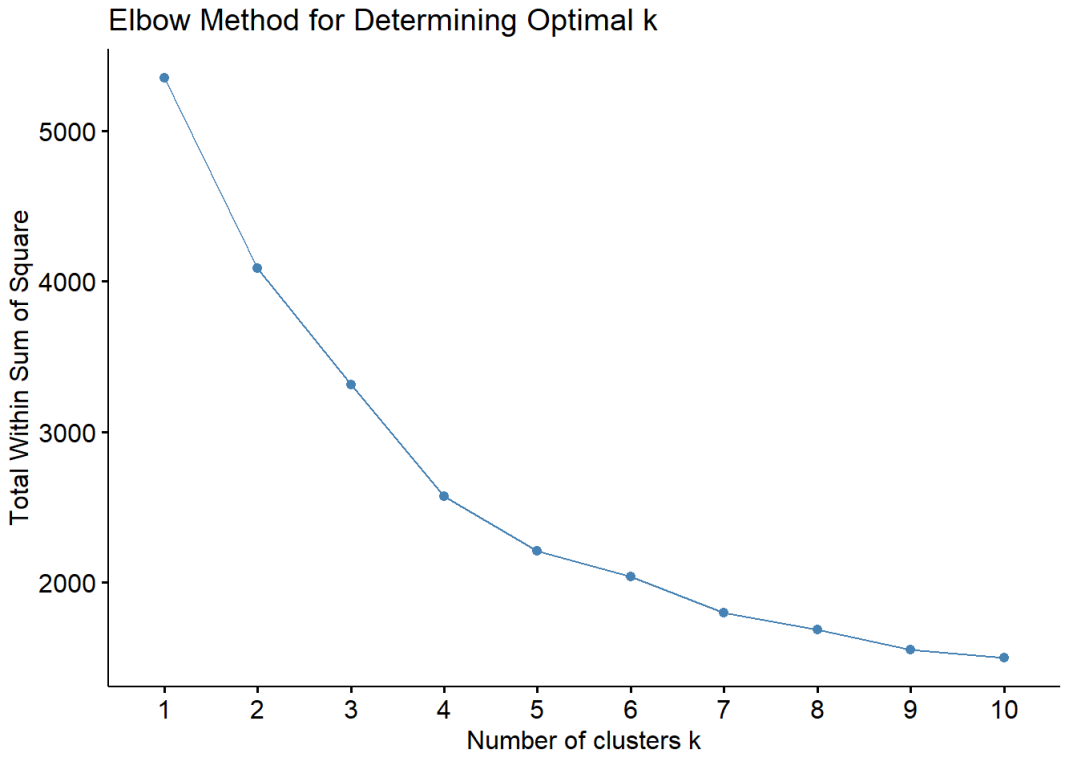
K-Means Clustering / Elbow Method: Scaling is also Essential for k-Means: Why? k-means clustering relies on Euclidean distance between data points. In English: Any variable with a larger numeric range (like charges, that can be $30,000+) dominate the distance calculation again.

Whereas, Smaller-range features (like children, usually being 0–5) will have a negligable impact on clustering unless we scale them.

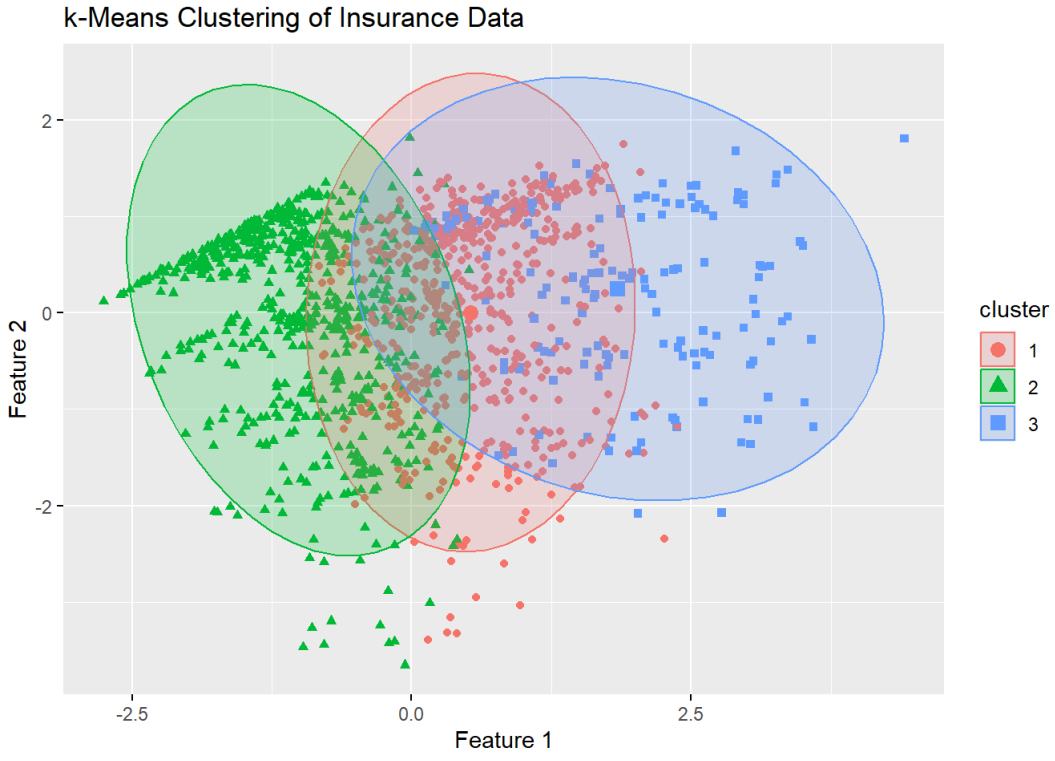
So scaling makes k-means more accurate, since each variable contributes equally to how points are grouped we avoid forming clusters based purely on higher magnitude variables which is especially important if you're using mixed units like we are (USD, years,body fat metrics, counts-children).

Elbow Method: finding the Optimal Number of Clusters The elbow method basically helps us choose the optimal number of clusters by plotting within cluster sum of squares (WSS). We look for the point where adding more clusters stops improving the model significantly (which is where the "elbow" bends).

```
fviz_nbclust(scaled_data, kmeans, method = "wss") +
  labs(title = "Elbow Method for Determining Optimal k")
```




```r
library(factoextra)

set.seed(123) #Note that we used this seed
kmeans_result <- kmeans(scaled_data, centers = 3, nstart = 25)

fviz_cluster(kmeans_result, data = scaled_data,
             geom = "point",
             ellipse.type = "norm",
             main = "k-Means Clustering of Insurance Data",
             xlab = "Feature 1", ylab = "Feature 2")
```

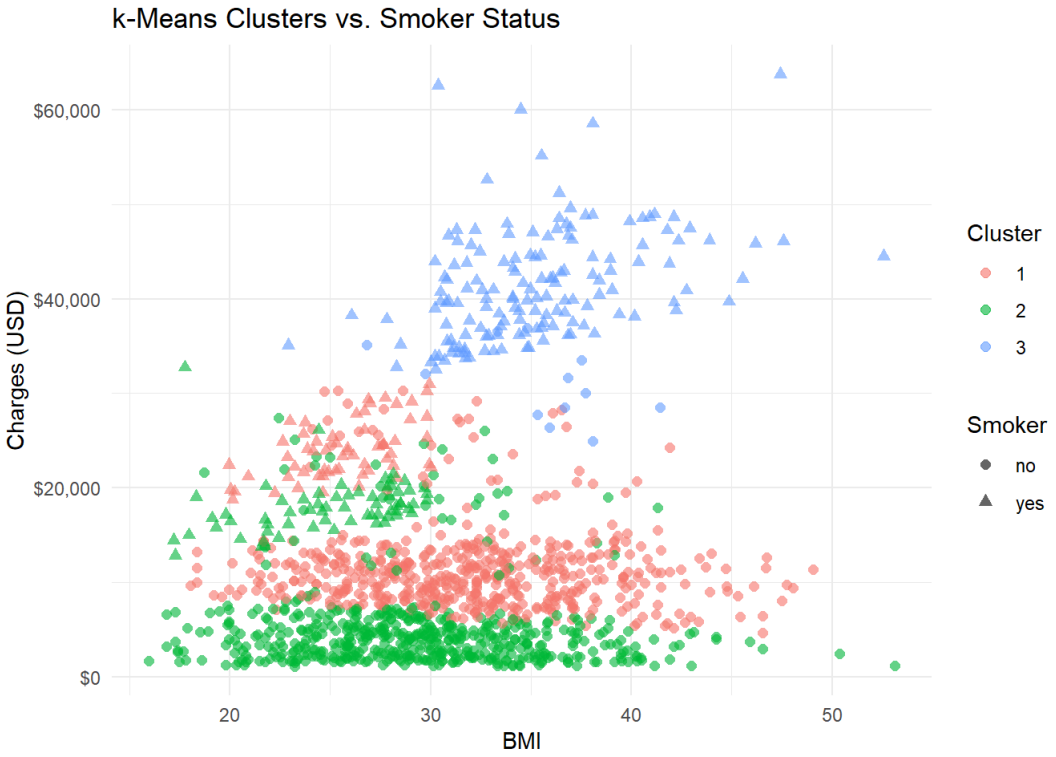


```
insurance$cluster <- as.factor(kmeans_result$cluster)

table(Cluster = insurance$cluster, Smoker = insurance$smoker)
```

```
##        Smoker
## Cluster  no yes
##       1 522  59
##       2 530  66
##       3  12 149
```

Interpretation: k-means clustering identified three distinct groups of individuals based on age, BMI, children, and charges. These clusters may reflect low, moderate, and high-risk insurance profiles based on shared traits.

```
library(ggplot2)

insurance$cluster <- as.factor(kmeans_result$cluster)

ggplot(insurance, aes(x = bmi, y = charges, color = cluster, shape = smoker)) +
  geom_point(alpha = 0.6, size = 2) +
  labs(
    title = "k-Means Clusters vs. Smoker Status",
    x = "BMI",
    y = "Charges (USD)",
    color = "Cluster",
    shape = "Smoker"
  ) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal()
```



Interpretation: The clusters align closely with smoker status, especially in the higher charge ranges, this tells us that k-means was effective in separating high-risk individuals even without smoker information. Most smokers are grouped into the same cluster, highlighting the relationship between smoking, BMI, and higher insurance charges.

#kNN Classification:

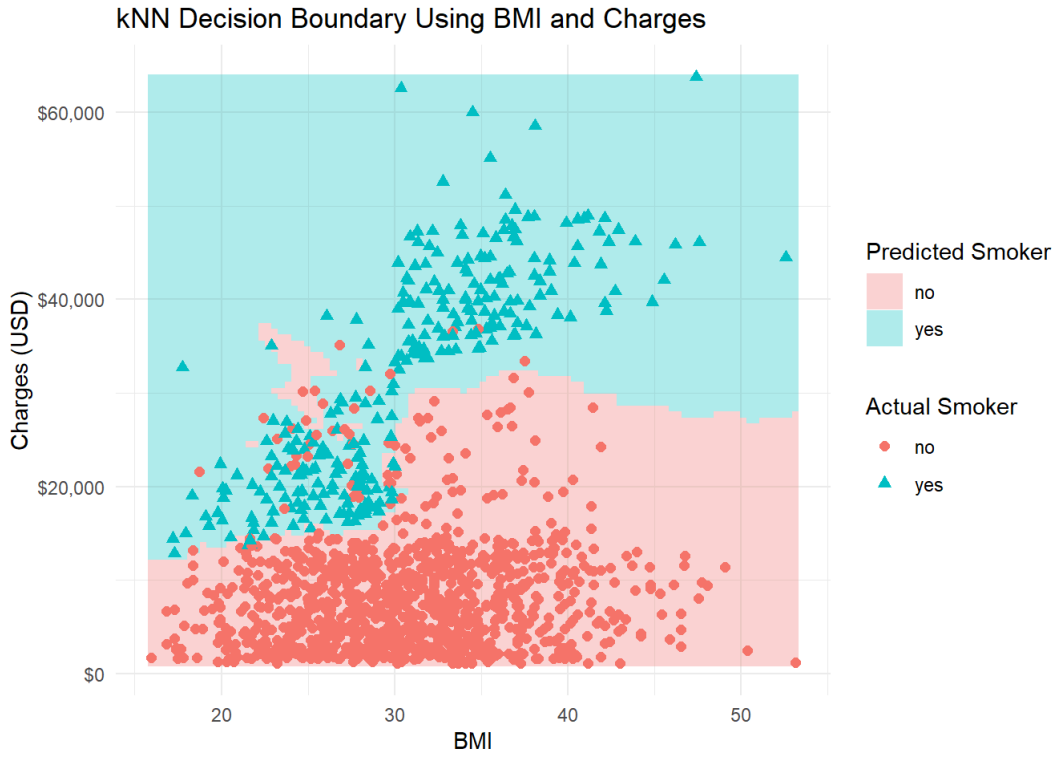
```r
library(class)
library(caret)
library(dplyr)

knn_data <- insurance %>%
  select(age, bmi, children, charges, smoker)

knn_data$smoker <- as.factor(knn_data$smoker)

scaled_features <- scale(knn_data[, c("age", "bmi", "children", "charges")])

set.seed(123)
sample_index <- sample(1:nrow(knn_data), 0.8 * nrow(knn_data))

train_X <- scaled_features[sample_index, ]
test_X  <- scaled_features[-sample_index, ]

train_Y <- knn_data$smoker[sample_index]
test_Y  <- knn_data$smoker[-sample_index]

knn_pred <- knn(train = train_X, test = test_X, cl = train_Y, k = 5)
confusionMatrix(knn_pred, test_Y)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##        no  216   2
##        yes   7  43
##
##                Accuracy : 0.9664
##                  95% CI : (0.9372, 0.9845)
##     No Information Rate : 0.8321
##     P-Value [Acc > NIR] : 4.613e-12
##
##                   Kappa : 0.8849
##
##  Mcnemar's Test P-Value : 0.1824
##
##             Sensitivity : 0.9686
##             Specificity : 0.9556
##          Pos Pred Value : 0.9908
##          Neg Pred Value : 0.8600
##              Prevalence : 0.8321
##          Detection Rate : 0.8060
##    Detection Prevalence : 0.8134
##       Balanced Accuracy : 0.9621
##
##        'Positive' Class : no
##
```

Quite Pleased as our kNN model performed very well. #Accuracy: 96.64% |Sensitivity (the true positive rate for non-smokers): 96.9% |Specificity (the true positive rate for smokers): 95.6% #Kappa: 0.88 => has a strong agreement beyond chance (Only 2 smokers misclassified as non-smokers and 7 non-smokers misclassified as smokers.)

Conclusion: The kNN achieved 96.6% accuracy in predicting smoker status using age, BMI, children, and charges, with a strong balance of sensitivity and specificity. This is suggestive that the chosen health and cost-related variables provide a reliable signal for identifying smokers.

```r
library(ggplot2)

features <- knn_data[, c("bmi", "charges")]
features_scaled <- scale(features)

scaler_center <- attr(features_scaled, "scaled:center")
scaler_scale <- attr(features_scaled, "scaled:scale")

bmi_range <- seq(min(features$bmi), max(features$bmi), length.out = 100)
charges_range <- seq(min(features$charges), max(features$charges), length.out = 100)
grid <- expand.grid(bmi = bmi_range, charges = charges_range)

grid_scaled <- as.data.frame(scale(grid, center = scaler_center, scale = scaler_scale))

knn_grid_pred <- knn(train = features_scaled[sample_index, ],
                     test = grid_scaled,
                     cl = knn_data$smoker[sample_index],
                     k = 5)

grid$smoker_pred <- knn_grid_pred

plot_data <- knn_data
plot_data$smoker <- as.factor(plot_data$smoker)

ggplot() +
  geom_tile(data = grid, aes(x = bmi, y = charges, fill = smoker_pred), alpha = 0.3) +
  geom_point(data = plot_data, aes(x = bmi, y = charges, shape = smoker, color = smoker), size = 2) +
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(
    title = "kNN Decision Boundary Using BMI and Charges",
    x = "BMI",
    y = "Charges (USD)",
    fill = "Predicted Smoker",
    color = "Actual Smoker",
    shape = "Actual Smoker"
  ) +
  theme_minimal()
```



Here we see that the kNN decision boundary shows clear separation between smokers and non-smokers based on BMI and charges, as smokers are primarily concentrated in the higer charge region.

MY APP!

https://github.com/Nathaniel-Coulter/Data-Science-in-R/blob/main/insurance_app.zip (https://github.com/Nathaniel-Coulter/Data-Science-in-R/blob/main/insurance_app.zip)

from flask import Flask, render_template, request import os from flask import Flask, render_template, request import pandas as pd

app = Flask(**name**)

data = pd.read_csv(r"C:Programs (Intro to Data Science Class)Sets2yrs.csv")

@app.route('/', methods=['GET', 'POST']) def index(): if request.method == 'POST': # Get form inputs age = int(request.form['age']) sex = request.form['sex'] feet = int(request.form['feet']) inches = int(request.form['inches']) height = feet * 12 + inches weight = float(request.form['weight']) children_input = request.form['children'] smoker = request.form['smoker'] region = request.form['region']

```
    # Step 1: Calculate BMI
    bmi = round((weight / (height ** 2)) * 703, 1)
    bmi_range = (bmi * 0.9, bmi * 1.1)  # ±10%

    # Step 2: Start filtering
    filtered = data.copy()

    # Age ±1
    filtered = filtered[(filtered['age'] >= age - 1) & (filtered['age'] <= age + 1)]

    # Sex exact match
    filtered = filtered[filtered['sex'].str.lower() == sex.lower()]

    # BMI ±10%
    filtered = filtered[(filtered['bmi'] >= bmi_range[0]) & (filtered['bmi'] <= bmi_range[1])]

    # Children
    if children_input == 'no':
        filtered = filtered[filtered['children'] == 0]
    # if "yes", include all

    # Smoker exact match
    filtered = filtered[filtered['smoker'].str.lower() == smoker.lower()]

    # Region exact match
    filtered = filtered[filtered['region'].str.lower() == region.lower()]

    # Step 3: Calculate average charge
    if not filtered.empty:
        average_charge = round(filtered['charges'].mean(), 2)
        result_message = f"${average_charge} (based on {len(filtered)} matching profiles)"
    else:
        result_message = "No similar profiles found in the dataset."

    return render_template('result.html', estimated_charge=result_message)

return render_template('index.html')
```

if **name** == '**main**': app.run(debug=True)

Final Thoughts

Overall, through our analysis of health insurance premiums we have discovered multiple consistent, and meaningful trends across the various statistical methods we used. Beginning our recap with the multilinear regression model, we found that smoking status was by far the strongest predictor of annual charges, far surpassing other factors like BMI, age, sex, or region. This insight was further reinforced by our exploratory visualizations, where the difference in charges between smokers and non-smokers was both visually stark and statistically robust. Additional modeling that included hierarchical clustering, k-means, and k-nearest neighbors (kNN) — all confirmed the separability of smokers and non-smokers as distinct groups. Specifically, with kNN achieving high accuracy when predicting smoking status from demographic and health

features. We also examined gender-based trends in charges across age and BMI, revealing that women tended to incur slightly higher charges during middle age and in higher BMI ranges, this likely being a result of the combination of healthcare usage patterns, childbirth and unknown dataset-specific factors (like diseseases).Note: after independentlyfiltering for outliers and looking at the national median and average health insurance premium costs for men and women with type 1 & 2 diabetes, I am fairly confident that approximately 20-30 of our outliers fit the profile demographic (parent / smoker /ages) for having said ailment based on their variables and the costs paid via google. Regional and gender comparisons provided further nuance, but across all models and visualizations, smoking consistently emerged as the most powerful driver of cost variation. These results reflect how actuarial models are closely tied not only to statistical patterns BUT to real-world behavioral and health risks, providing a strong foundation for predictive insurance modeling.