DataScience and R

2023

INSURANCE AND VARIOUS FACTORS

Team Leader - Pramana Rio(2023318129)
Kang Seoyoung (2022312663)
Jongmin Park(2022312352)

TEAM 1

Contents

1.
Introduction

Insurance Dataset

- It includes various factors that can influence medical costs and premiums for health insurance.
- It has 10 variables.

Age	BMI	Smoking Status	Income	Occupation
Gender	Number of Children	Region	Education	Insurance Plan

- It generated randomly, ensuring it represented the population in US.
- It was created to explore the relationship between diffrent factors and medical costs.

2.
Preliminary Findings

1

Subset Data

1) young adults (18-35)

2) the number of children is '0'

2

Remove Outliers

BMI and charges have outliers, then remove it before visiualization

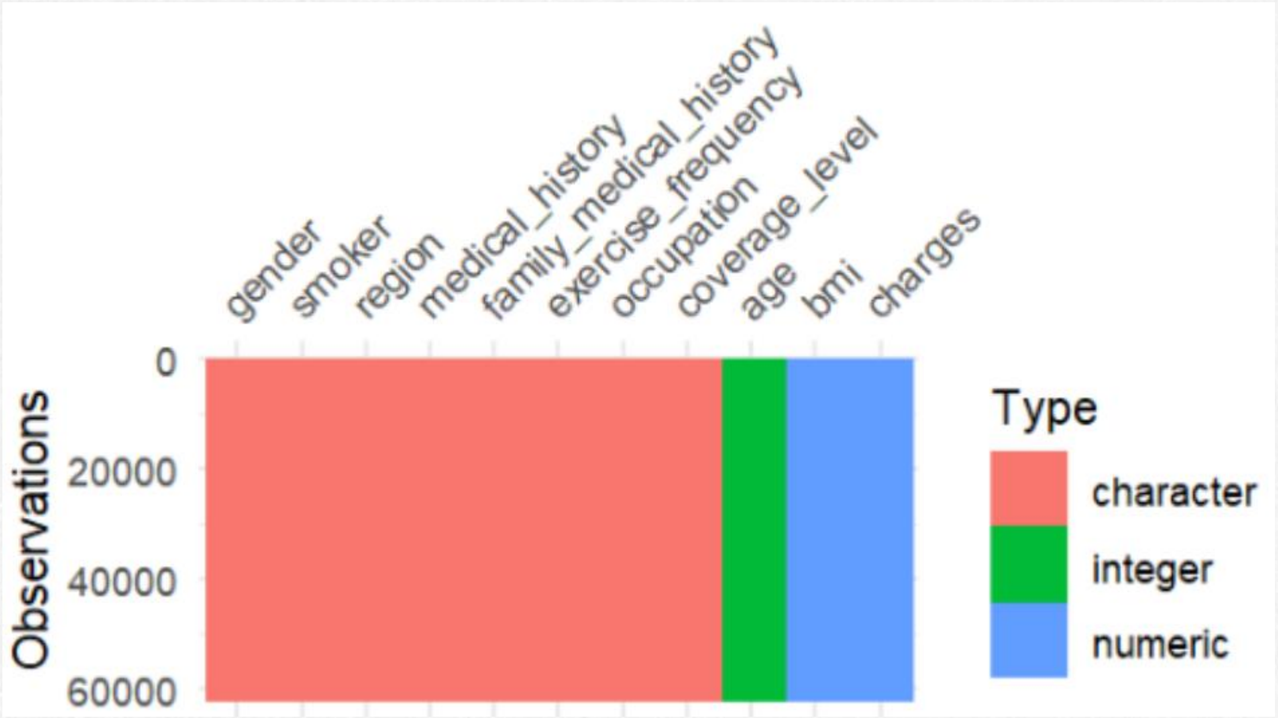
3

Investigate Dataset

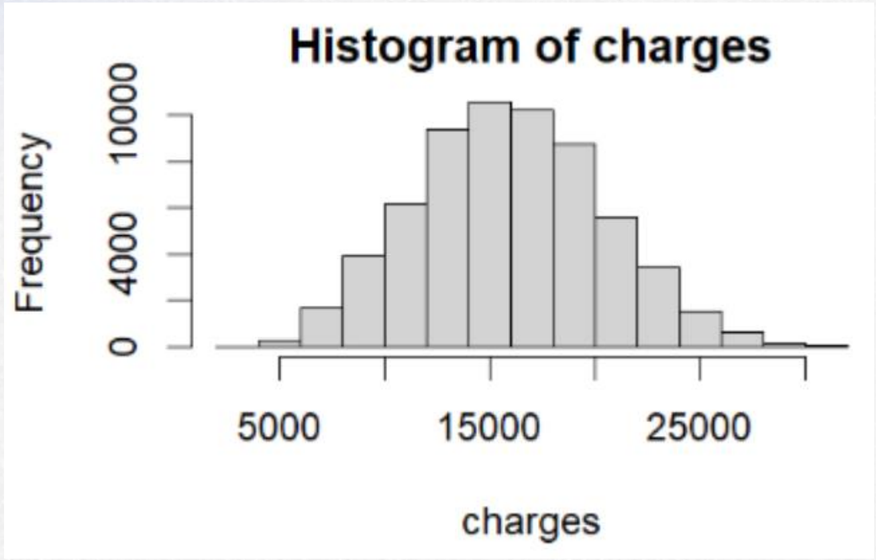
Our data is extremely balanced.

7. Subset Data

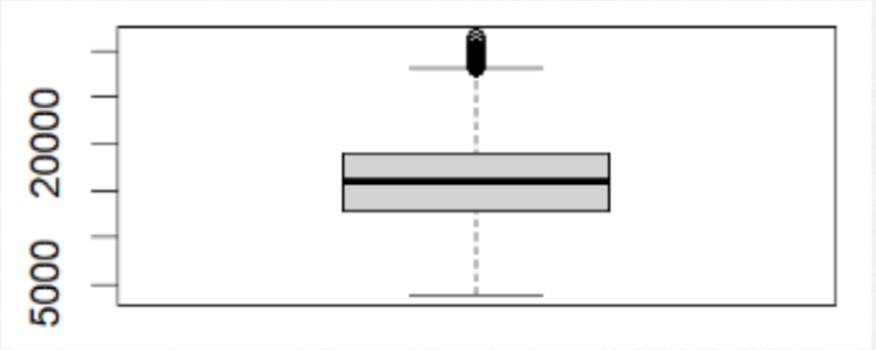
Subset data into young adults & children == 0 (remove children column)
Total data is 62250.



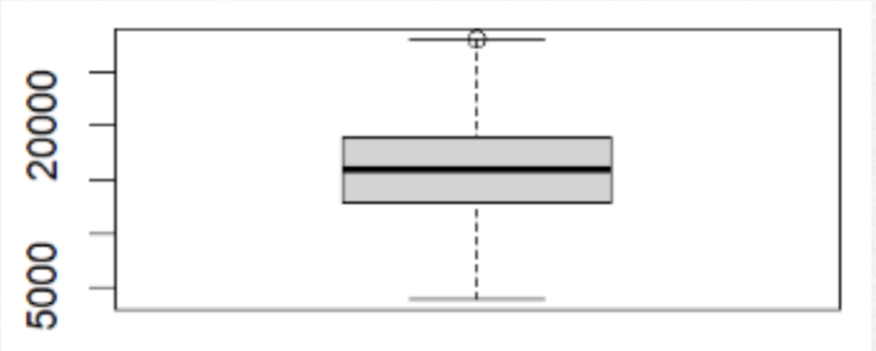
2. Remove Outliers



Normal distribution of charges



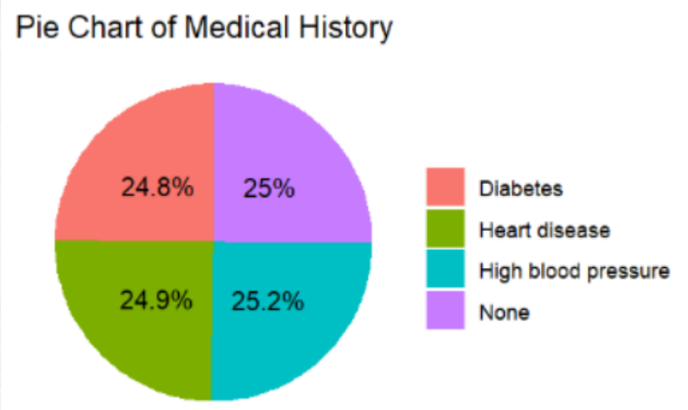
boxplot before removal

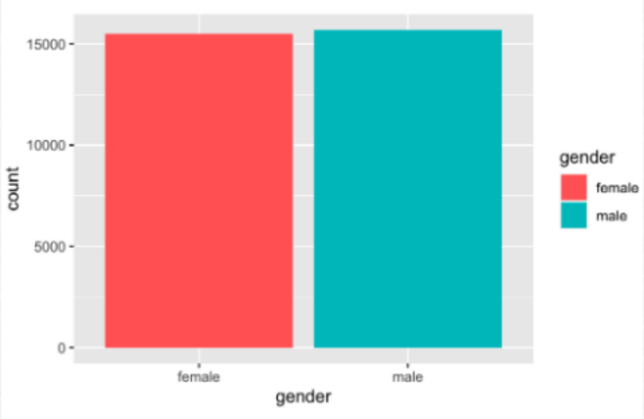


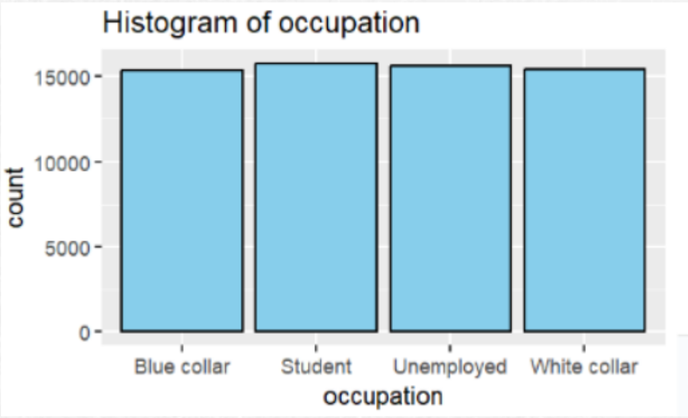
boxplot after removal

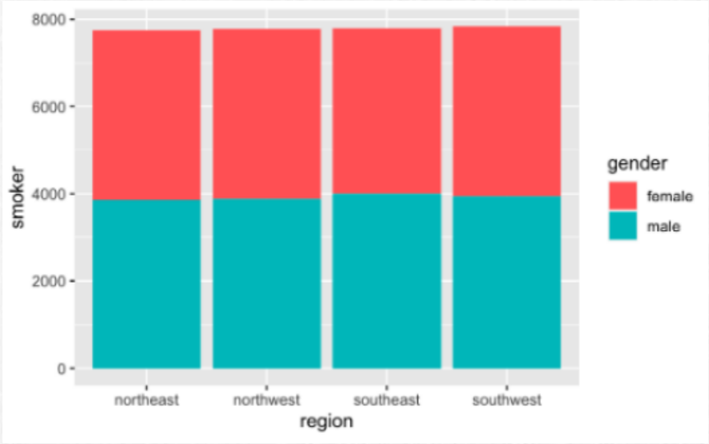
3. Investigate Dataset

Most of our data is perfectly balanced.









Let's analyze the relationship between different factors and charges!

3-1.
EDA

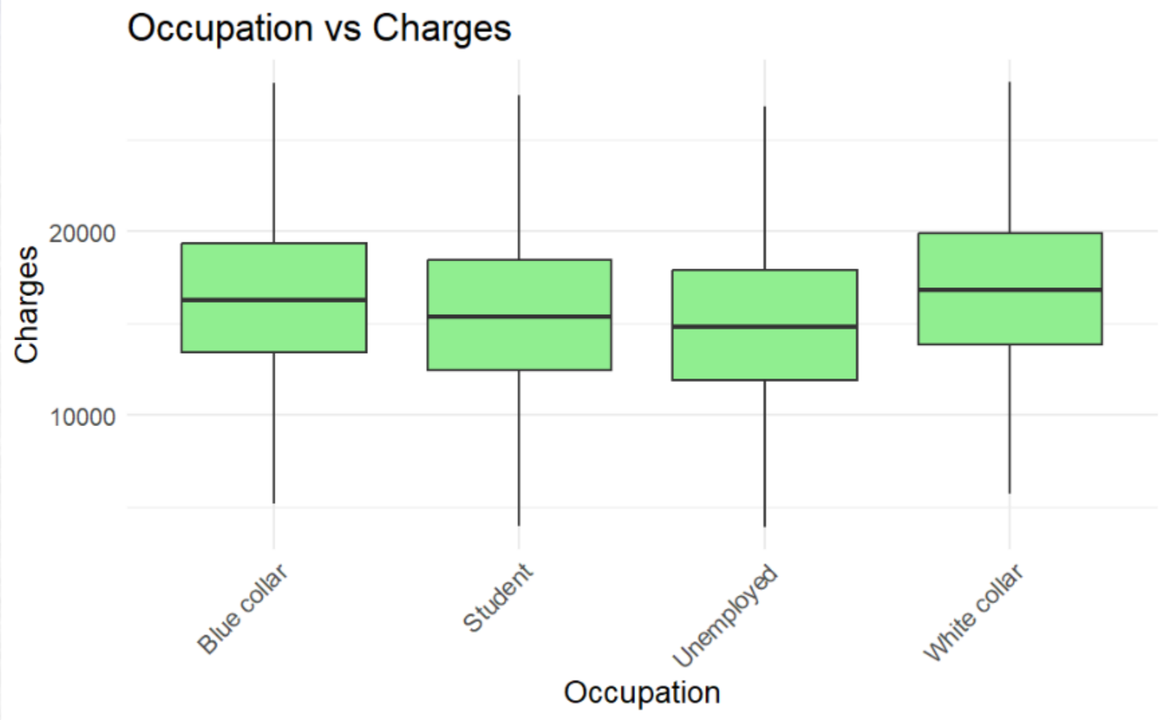
: How companies set insurance premium?

Other than just visually interpreting the correlation/relationship between each independent variables with 'charges', we are also looking to quantify these relationships using their adjusted R-squared value (for numerical variable) or their eta squared value (for categorical variables)

Visualization of factors with Charges

*All p-value is smaller than 0.05

OCCUPATION



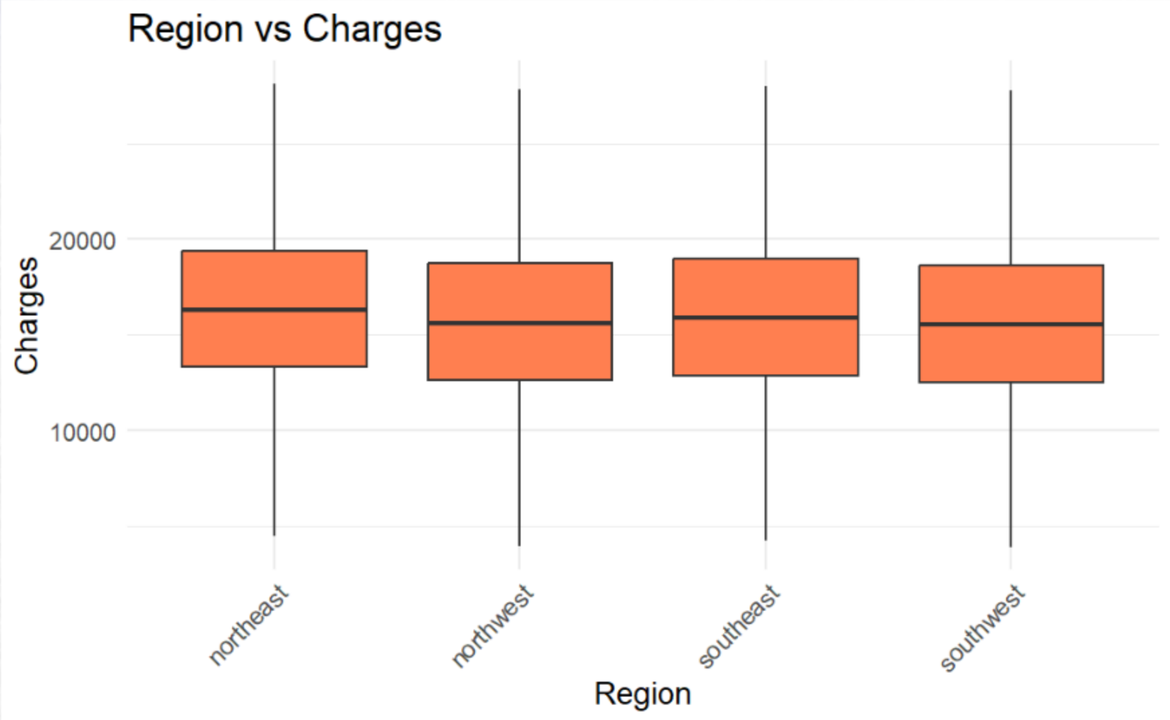
<Linear Regression Statistics>

Eta-squared for occupation:
0.0313582225624953

Visualization of factors with Charges

*All p-value is smaller than 0.05

REGION



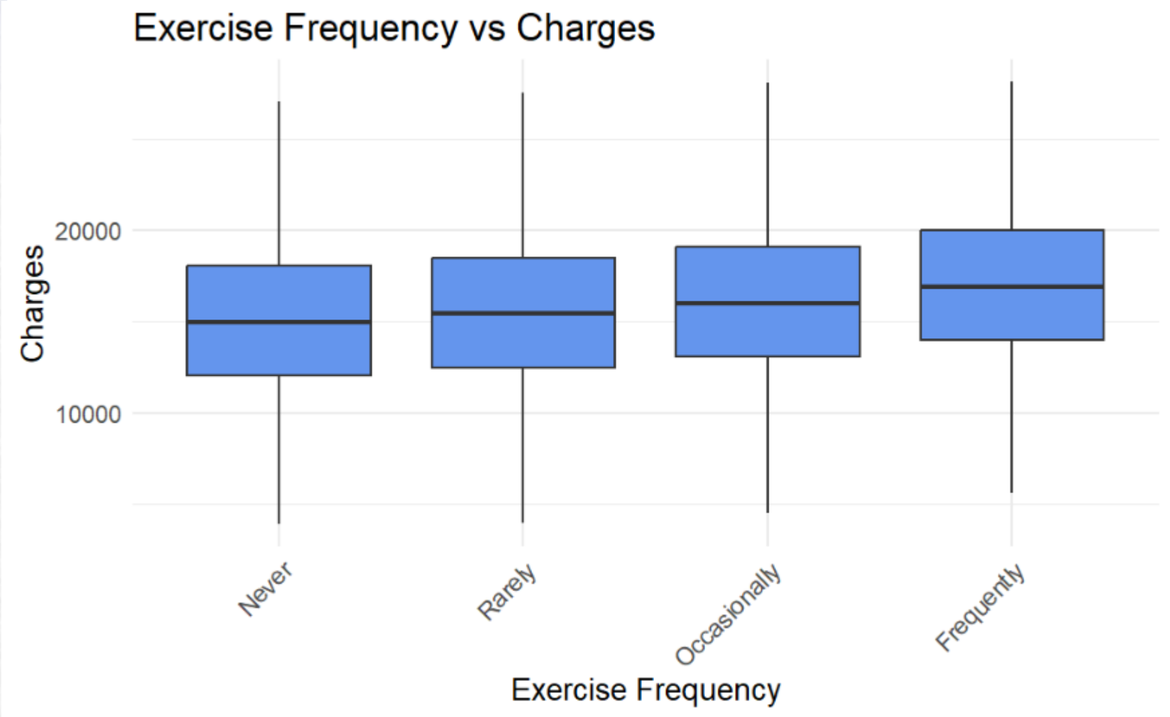
<Linear Regression Statistics>

Eta-squared for occupation:
0.0044615364913614 5

Visualization of factors with Charges

*All p-value is smaller than 0.05

EXERCISE FREQUENCY



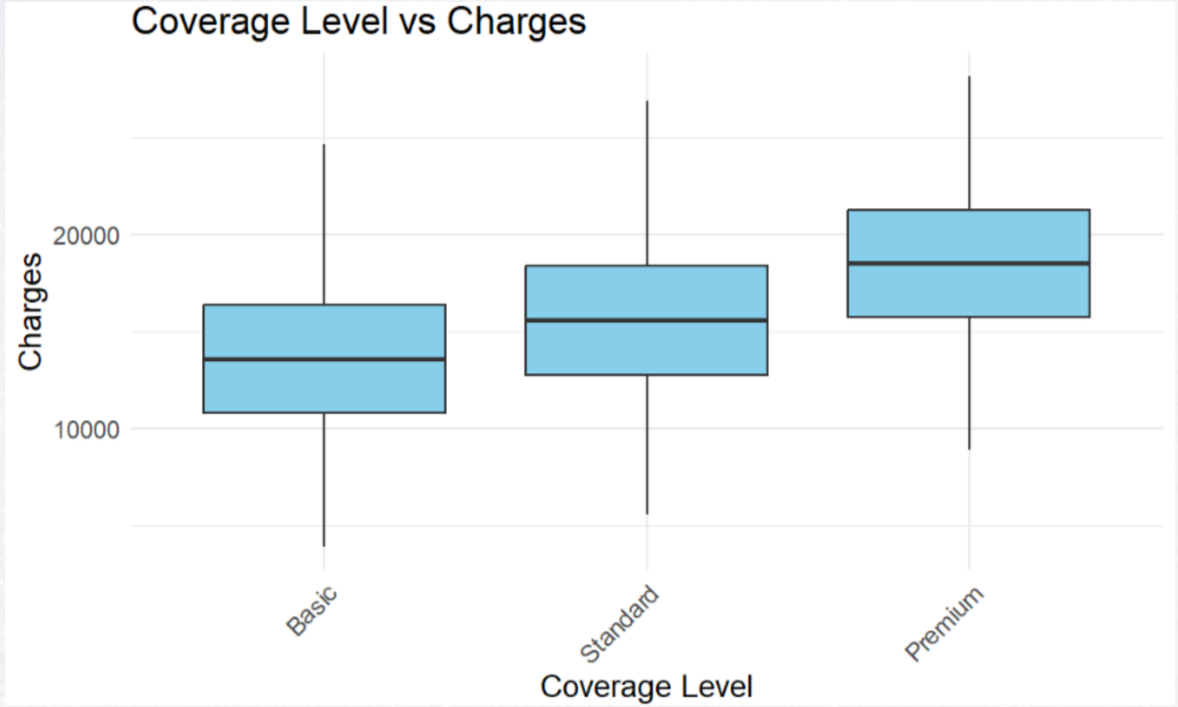
<Linear Regression Statistics>

Eta-squared for occupation:
0.0273637928887102

Visualization of factors with Charges

*All p-value is smaller than 0.05

COVERAGE_LEVEL



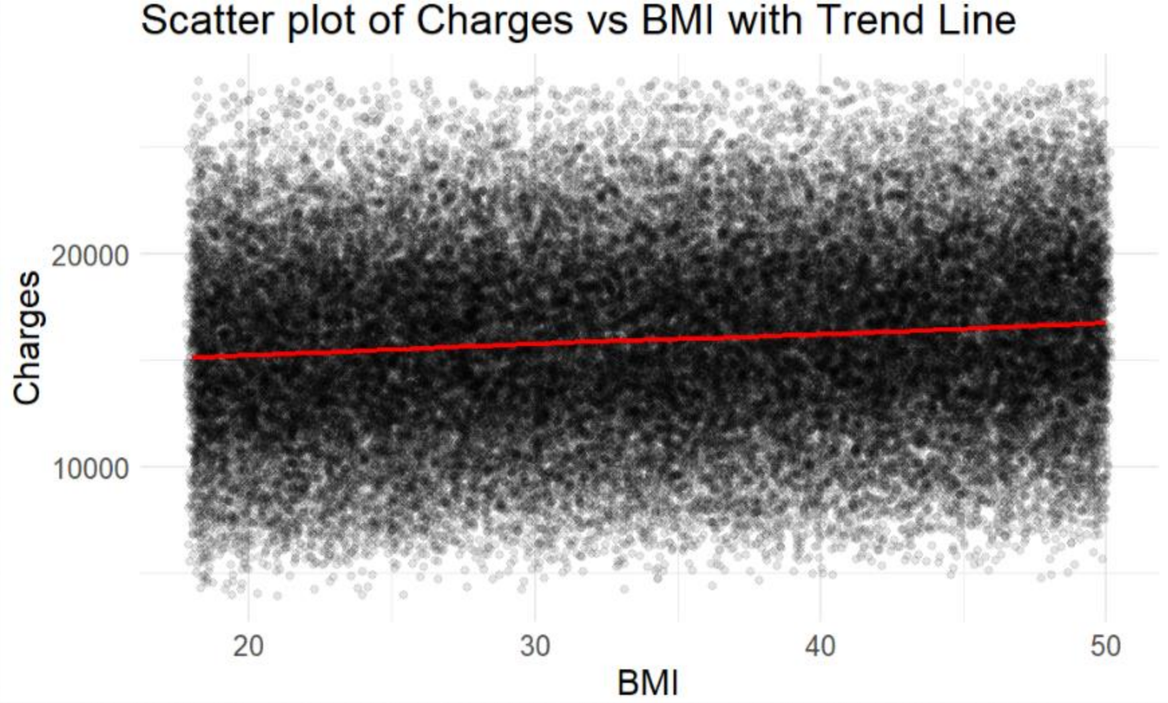
<Linear Regression Statistics>

Eta-squared for coverage_level:
0.214920282615371

Visualization of factors with Charges

*All p-value is smaller than 0.05

BMI



<Linear Regression Statistics>

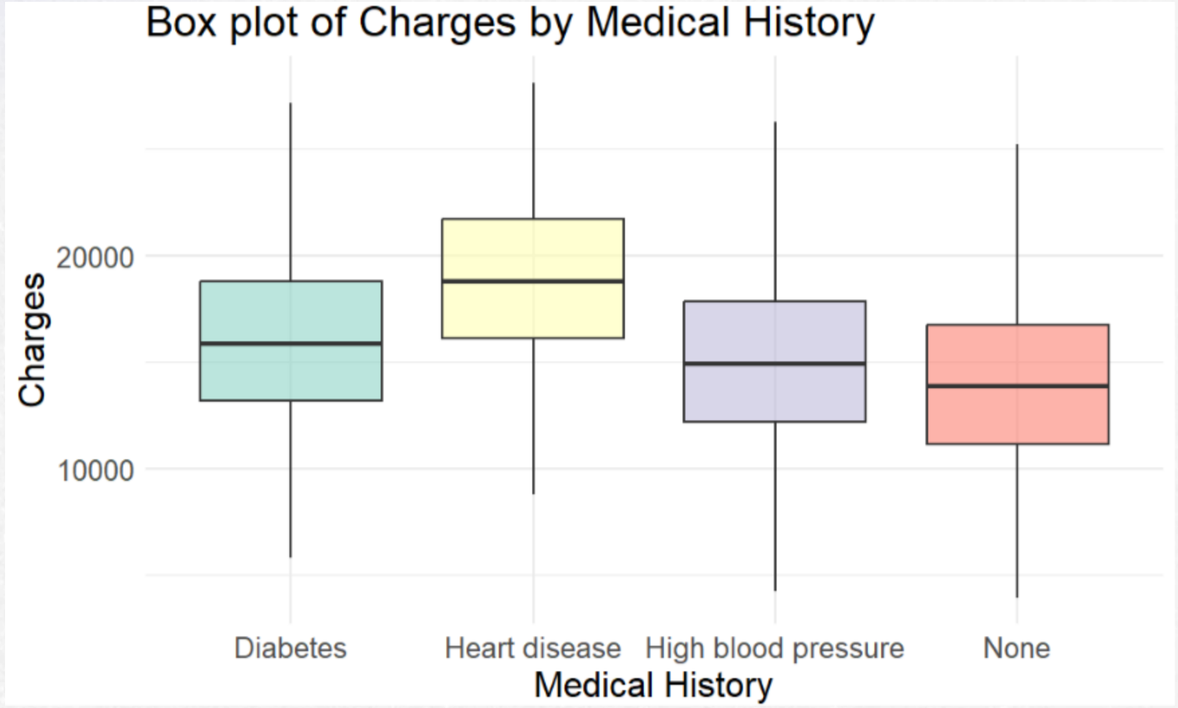
Pearson's product-moment correlation: 0.1054016

Adjusted R-squared: 0.01109 (linear regression)

Visualization of factors with Charges

*All p-value is smaller than 0.05

MEDICAL HISTORY



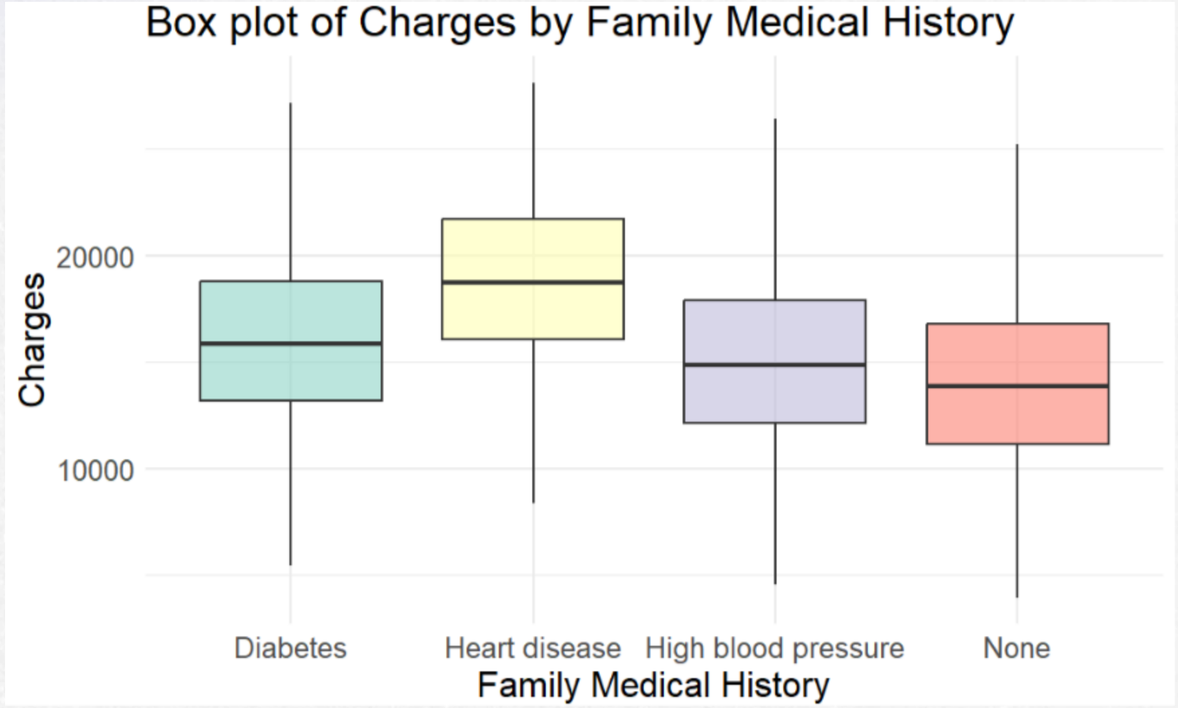
<Linear Regression Statistics>

Eta-squared of ANOVA:
0.1766539

Visualization of factors with Charges

*All p-value is smaller than 0.05

FAMILY MEDICAL HISTORY



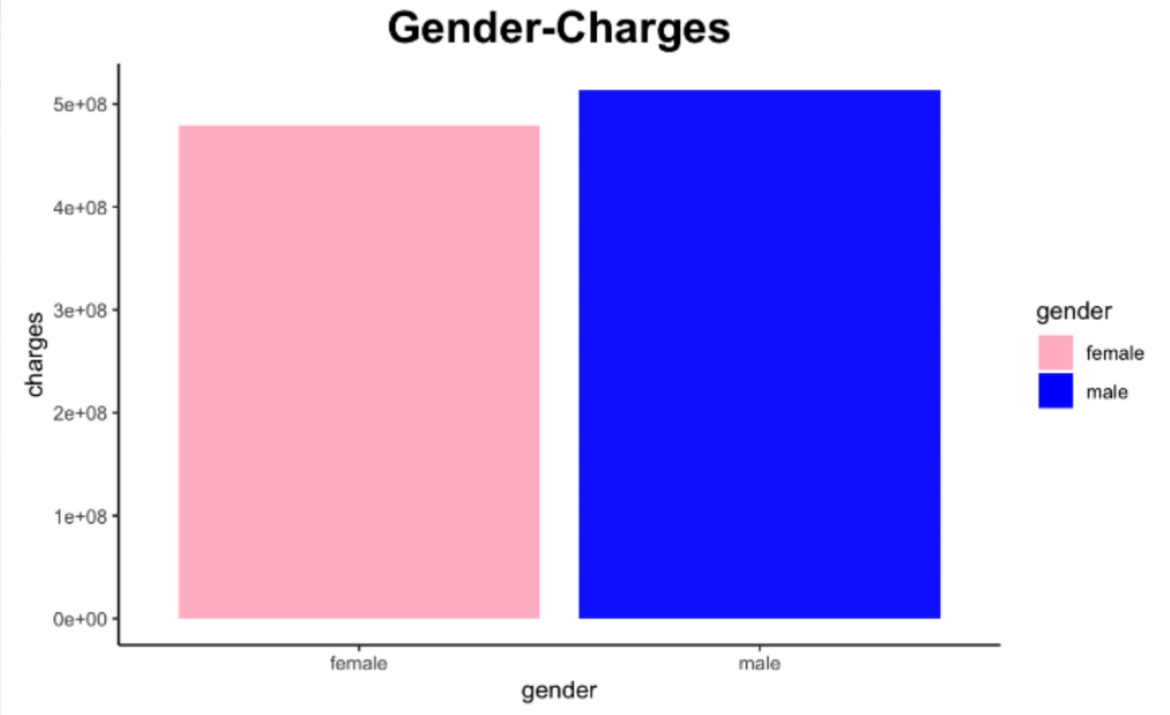
<Linear Regression Statistics>

Eta-squared of ANOVA:
0.1754327

Visualization of factors with Charges

*All p-value is smaller than 0.05

GENDER



<Linear Regression Statistics>

Multiple R-squared: 0.01332

Adjusted R-squared: 0.01331

p-value: < 2.2e-16

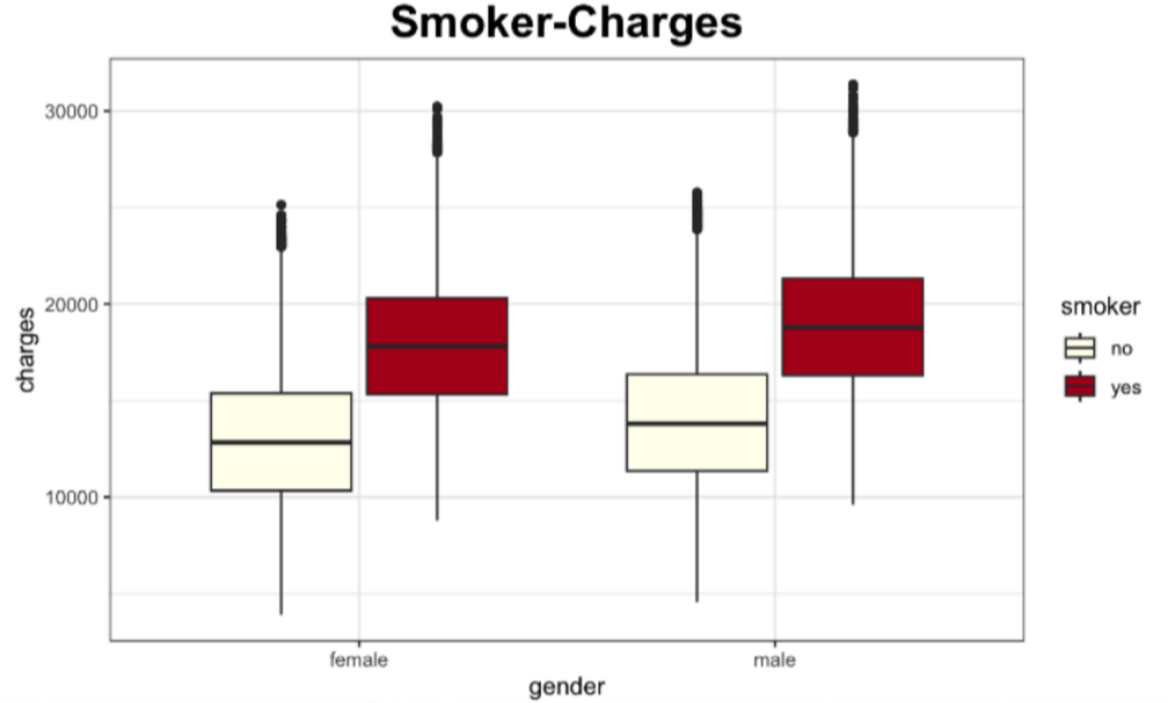
F-statistic: 840.6

charges = 15423.25 + 985.50 * gendermale

Visualization of factors with Charges

*All p-value is smaller than 0.05

SMOKER



<Linear Regression Statistics>

Multiple R-squared: 0.3234

Adjusted R-squared: 0.3234

p-value: < 2.2e-16

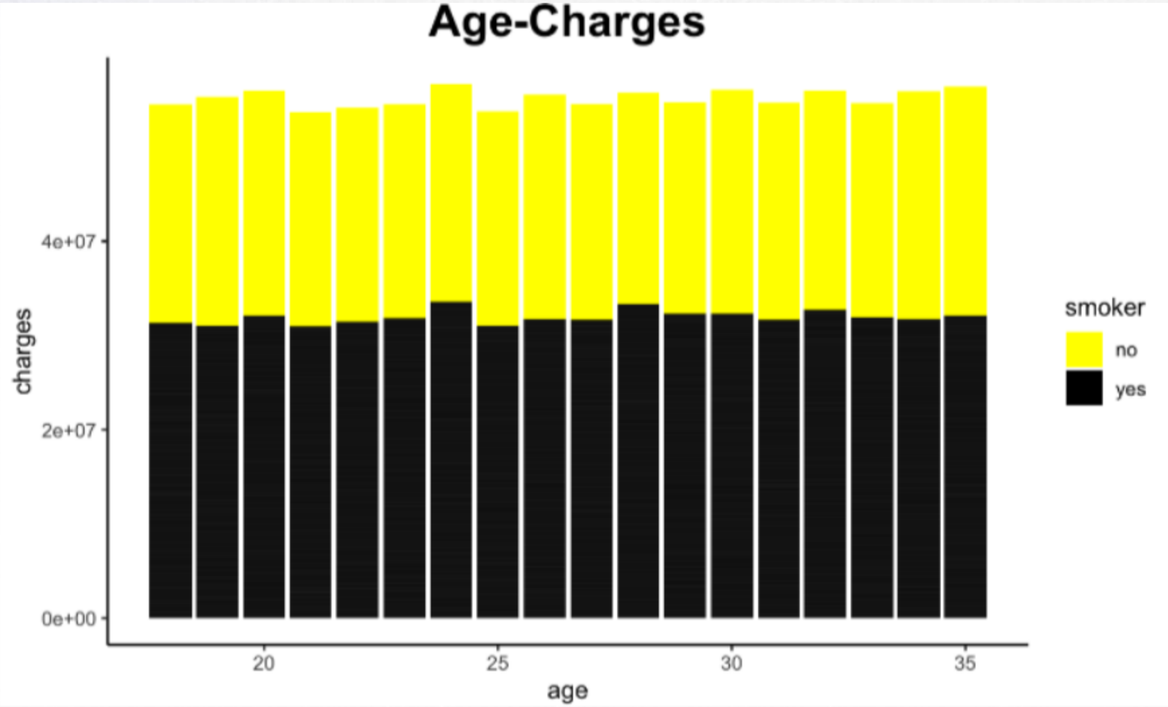
F-statistic: 2.976e+04

charges = 13447.34 + 4943.32 * smokeryes

Visualization of factors with Charges

*All p-value is smaller than 0.05

AGE



<Linear Regression Statistics>

Multiple R-squared: 0.0007603

Adjusted R-squared: 0.0007442

p-value: 5.955e-12

F-statistic: 47.36

charges = 15321.814 + 22.471 * age

3-2.
Additional EDA
: Use Machine Learning

RandomForest to compare with ours

This model is almost consistent with our analysis.

Fit

```r
#model training
forest_m <- randomForest(charges ~ ., data=train, importance = T)


#predict charges
y_pred <- predict(forest_m, test, type='response')

# Adjusted-R-squared
n <- nrow(test)
p <- ncol(reduced_insurance_data) - 1
adjusted_R_squared <- 1 - (1 - r_squared) * (n - 1) / (n - p - 1)
adjusted_R_squared
```

RandomForest to compare with ours

This model is almost consistent with our analysis.

```
> adjusted_R_squared
[1] 0.9751178
```

feature importance(%)

age	gender	bmi	children
0.03232093	1.13760236	0.85157270	0.00000000
smoker	region	medical_history	family_medical_history
35.02418411	0.28949714	17.83256418	17.53350795
exercise_frequency	occupation	coverage_level	
2.35415307	2.50240162	22.44219593	

These feature importance from the Random Forest model matches what we got from the previous analysis which involves adjusted R-squared value or eta squared value that describes/quantify the univariate relationships

3-2.
Additional EDA

: Are the most important variables enough to determine insurance charges?

Predictive Model for Charges

INITIAL_MODEL

smoker

coverage_level

medical_history

family_medical_history

SECOND_MODEL

smoker

coverage_level

medical_history

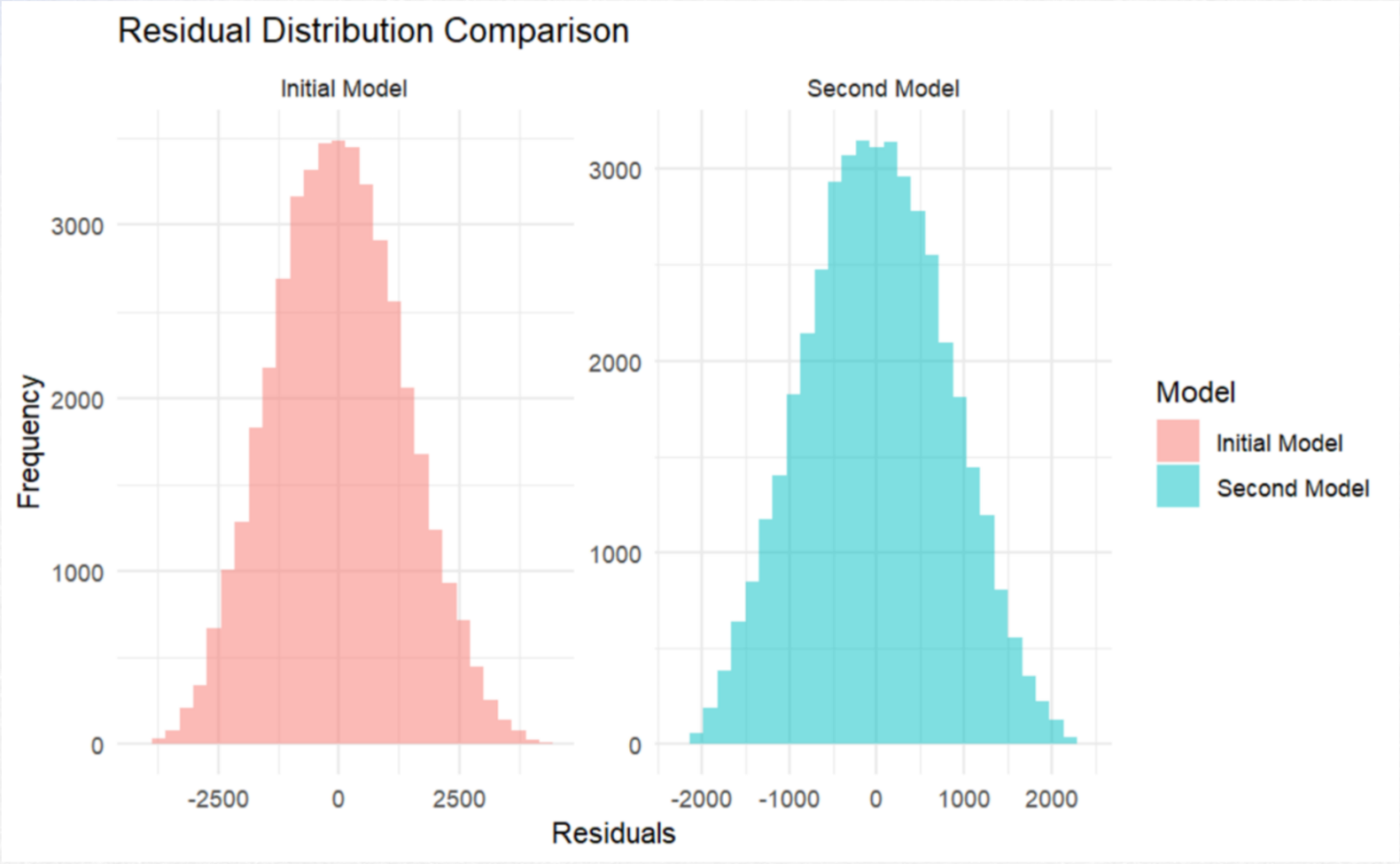
family_medical_history

+ Occupation

+Exercise_frequency

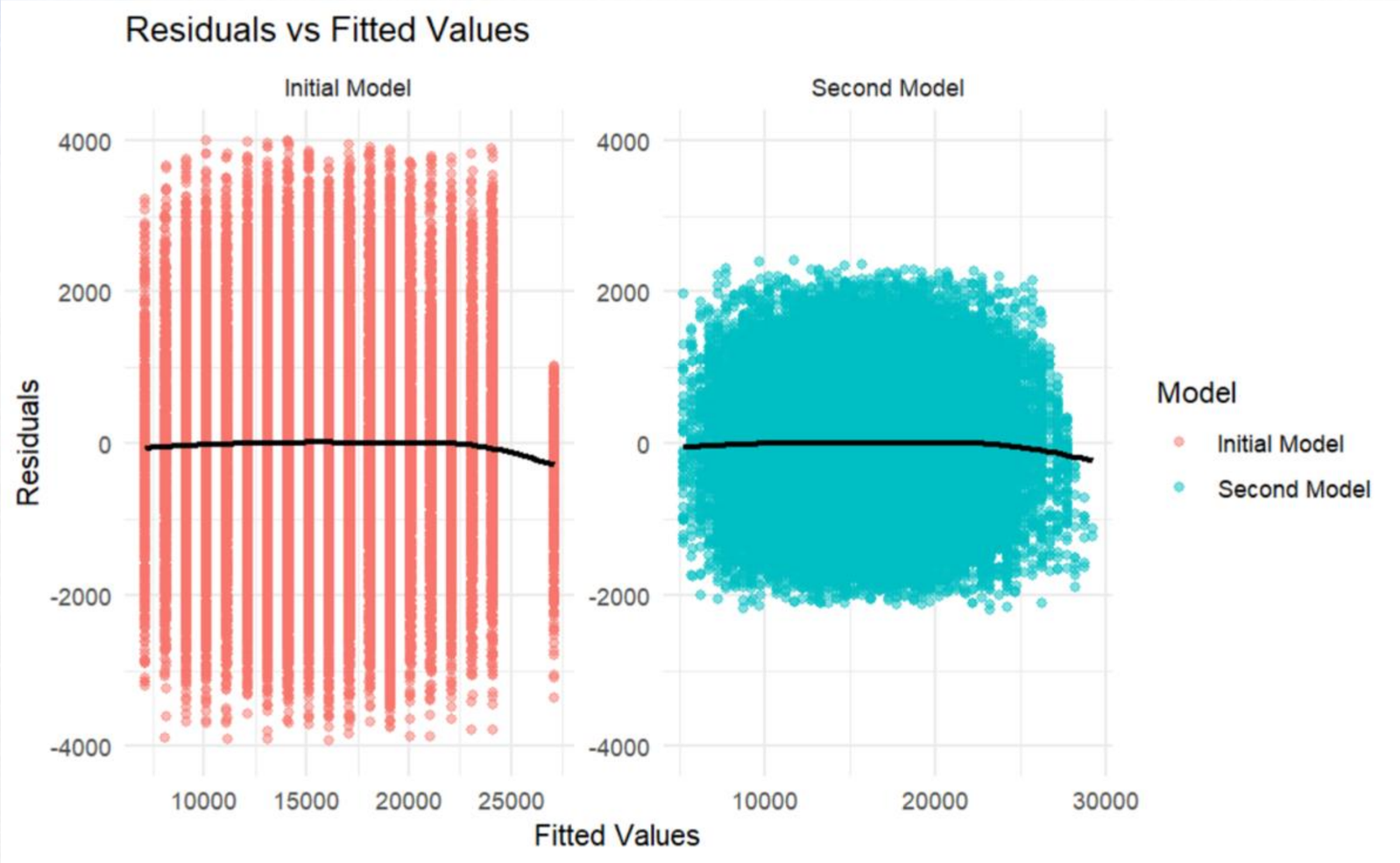
The initial_model uses 4 most important variables (sorted based on their feature importance/adjusted R-squared value/eta squared value), while the second_model uses 6 most important variables.
We used the second model as a comparison because the initial model already accounts for over 90% of charges' variability, but it does not mean that it will perform well. The second model uses occupation and exercise frequency as addition and they both only combine for 5% of charges' variability

Normality of Residuals



We want a normal distribution of the residuals to ensure a reliable model. Both models show normal distribution

Prediction Errors Plot



What we want to see is a random pattern (the data points are evenly distributed above and below the line). We also want to see the data being as close as possible to 0.
From the graph, we can see that the second model fits this description better

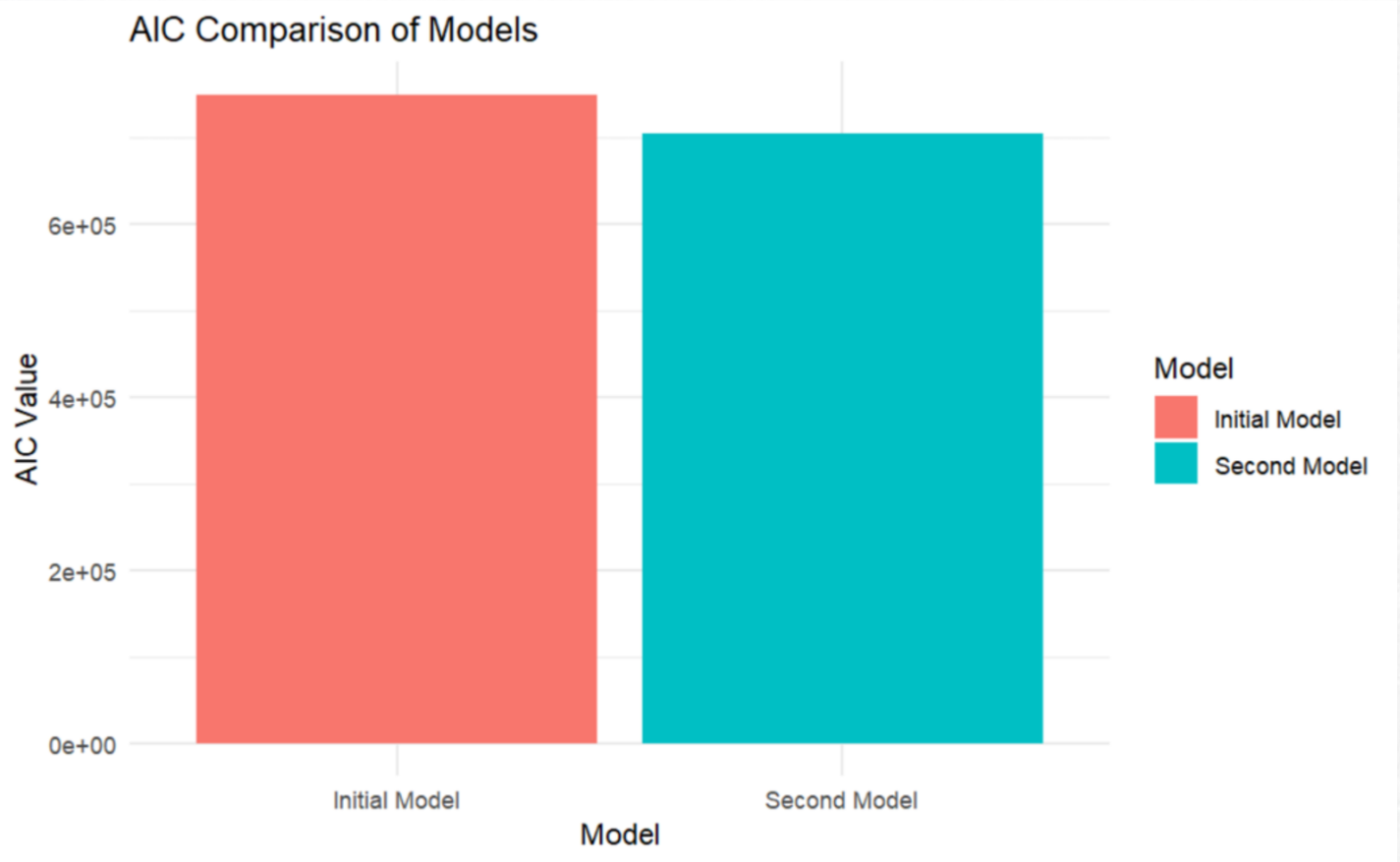
No Multicollinearity

```
                          GVIF Df GVIF^(1/(2*Df))
smoker                1.000182  1        1.000091
coverage_level        1.000376  2        1.000094
medical_history       1.000470  3        1.000078
family_medical_history 1.000397  3        1.000066
                          GVIF Df GVIF^(1/(2*Df))
smoker                1.000413  1        1.000207
coverage_level        1.000535  2        1.000134
medical_history       1.000635  3        1.000106
family_medical_history 1.000959  3        1.000160
occupation            1.000582  3        1.000097
exercise_frequency    1.000725  3        1.000121
```

We do not want high correlation between any of the independent variables. The fact that the GVIF values of all variables on both models are very close to 1, shows that there is no multicollinearity

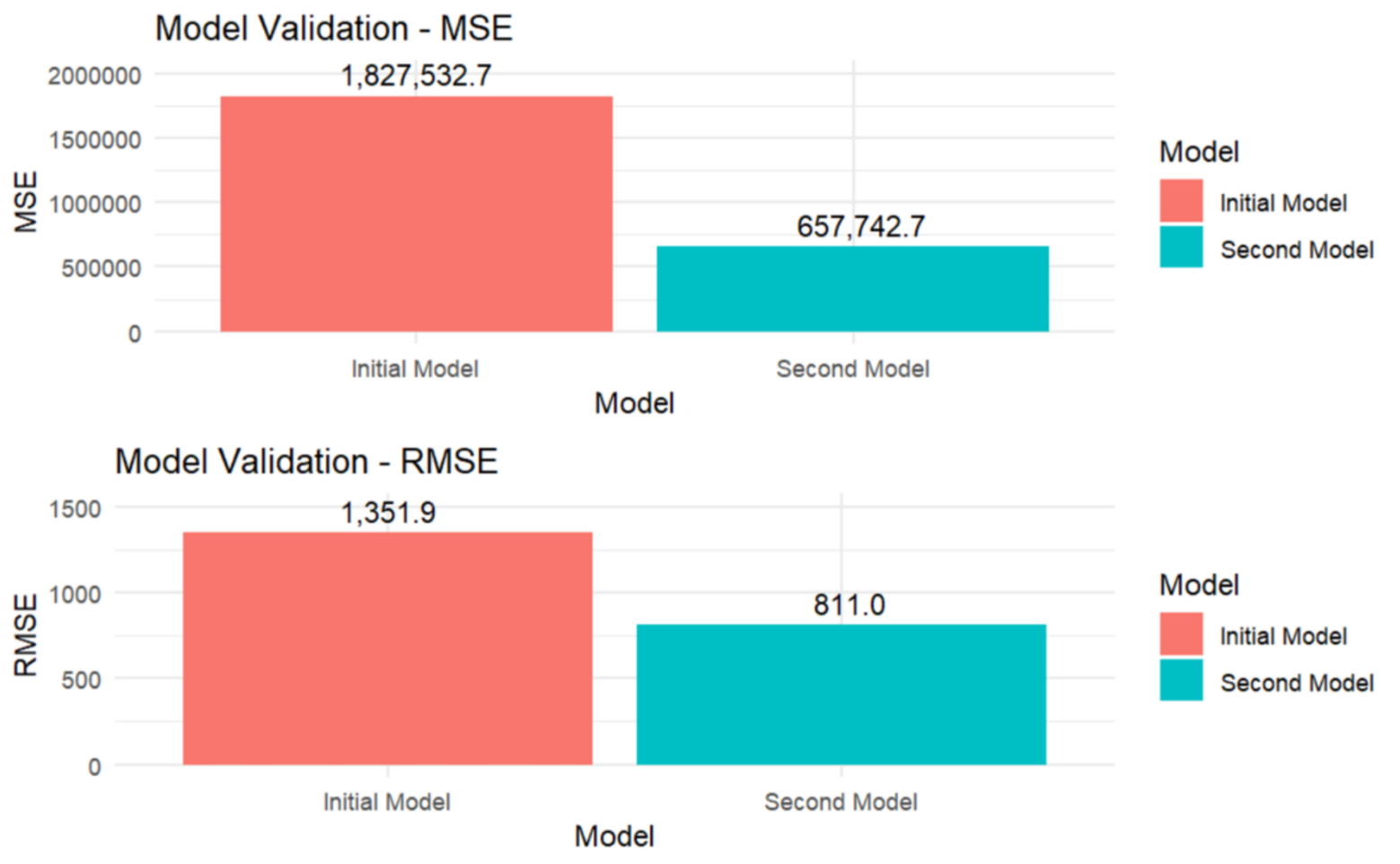
Metric 1: AIC

Initial_model: 626650.8 / Second_model: 582145.4



The lower the AIC, the better the model is because it indicates that the model has less overfitting

Metric 2 & 3: MSE & RMSE



MSE and RMSE focuses on showing the errors and how big they are. Lower MSE and RMSE indicates a better performing model.

From 3 of the previous metrics, all of them shows that the second model (using 6 most important variables) performs much better than the initial model (using only 4 most important variables)

4.
Challenges

Problem of Our Data

TOO MUCH BALANCED DATA

Hard to dig for insights & trends
because dataset is perfectly balanced

UNFAMILIAR STATISTICAL TERM

To answer the questions,
we had to learn a large number of new statistical terms

5.
Conclusions

1. The insurance dataset (young adults with no children) is perfectly balanced

2. "4 key variables (coverage_level, smoker, medical_history, family_medical_history) can already explain 90% of insurance charges' variability

3. Occupation and exercise frequency are crucial to determine insurance charges accurately

Team 1

Analyze Insurance Data

Thank You

DATASCIENCE AND R