Introduction

The insurance industry plays an important role in ensuring individuals against unexpected risks. The relevant insurance data correspondingly continues to grow, and we'd like to explore some underlying patterns and factors that drive insurance charges for potential customers and policyholders. In particular, this analysis seeks to answer three key questions: How does health insurance charges change with age? To what degree do our variables vary with each other? And finally, what factors affect insurance charges the most?

Our dataset is a US health insurance dataset taken from Kaggle. There are 1338 observations in the dataset with 7 variables. Each of our variables are focused on characteristics of someone in the US who pays insurance charges. There are no missing data values or undefined values in our dataset.

During our analysis we decided to create a new response variable called "Tier". This is a categorical variable that we thought would suit our analysis better by representing different ranges of insurance charges. Additionally it would help us address our non-normal data which we will discuss more in the EDA section.

Table 1: Summarizes the 7 variables we chose to focus on for this analysis

Variable Name	Variable Type	Description
Age	Quantitative Discrete	Age of primary beneficiary
Sex	Categorical	Sex of contractor
BMI	Quantitative Continuous	Body mass index, weight in kg divided by the square of height in meters.
Children	Quantitative Discrete	Number of children covered by health insurance / number of dependents
Smoker	Categorical	Smoker/Non-smoker yes, no
Region	Categorical	The beneficiary's residential area in the US, northeast, southwest, southwest, northwest
Charges	Quantitative Continuous	Individual medical costs billed by health insurance
Tier	Categorical	We have 4 tiers and each tier

		represents a quartile from the Charges variable

EDA

Figure : Histogram of Distributions

Before analyzing the data, we examined the distributions of the seven variables. Upon inspection, we observed that age appeared to follow a uniform distribution, bmi seemed to follow a normal distribution, and both children and charges appeared to be right-skewed. This indicated the need to transform our response variable. Additionally, we noted that the majority of individuals in our dataset had yearly medical expenses ranging between 0 and $15,000. This implies that there will be higher variance between the people who tend to have higher insurance charges.

Regarding the boxplots, the relationship between sex and charges did not show a distinct difference due to a considerable amount of overlap between the two boxes. However, a notable disparity was observed between smokers and non-smokers. Furthermore, there was considerable overlap observed between region and the charges paid. Overall, the difference between people who smoked versus those who did not seems to be important in determining insurance charges.

Figure : corr matrix

METHODS: Linear Model

First, we use the linear model to predict our response variable, *charge*, using our six predictors – *Age, BMI, Children, Sex, Smoker,* and *Region*. The linear model we got will be in the form of

Charge = -\beta0+\beta_1**age**+\beta_2**children**+\beta_3**bmi**+\beta_4**sexmale**+ \beta_5**smokeryes**+\beta_6**regionnorthwest**+\beta_7**regionsoutheast**+\beta_8**regionsouthwest**

Resampling approach: Cross Validation

To assess the accuracy of our linear model, we choose the resampling method - cross-validation with 5 folds. In this method, we split our data into 5 equal-size folds and ran the train and test. This method allows us to see the accuracy of our predictive model by comparing the mean squared error.

Results: Linear Model

The coefficients and p-values for the predictors fitted in the linear model are shown as below,
Table 2: Coefficients and p-values of the predictors in the linear model

Variables	Estimated Coefficients	P-value
Intercept	-11938.5	<2e-16
age	256.9	<2e-16
children	475.5	0.000577
bmi	339.2	<2e-16
sexmale	-131.3	0.693348
smokeryes	23848.5	<2e-16
regionnorthwest	-353.0	0.458769
regionsoutheast	-1035.0	0.030782
regionsouthwest	-960.0	0.044765

We got our linear model should be,

Charge = -11938.5+256.9**age**+475.5**children**+339.2**bmi**-131.3**sexmale**+
23848.5**smokeryes**-353**regionnorthwest**-1035**regionsoutheast**-960r**egionsouthwest**

To test if the variable is significant in the linear model, we found the p-values of variables, *sexmale* and *regionnorthwest*, are comparatively large, 0.693348 and 0.458769 respectively. The p-value indicates that these two variables are not significant under the significance level of 0.1. The other six variables *age*, *children, bmi, smokeryes, regionsoutheast*, and *regionsouthwest* are significant under the significance level of 0.1.

Cross validation results

To compute the error, each time we resampled 1071 observations and leave 268 observations for test set. The linear model, using six predictors – *Age, BMI, Children, Sex, Smoker,* and *Region,* is fitted on the 1071 observations. We got our final linear model as

Charge = -11938.5+256.9**age**+475.5**children**+339.2**bmi**-131.3**sexmale**+
23848.5**smokeryes**-353**regionnorthwest**-1035**regionsoutheast**-960.1**regionsouthwest**

which is really similar to the linear model we got before. For this model, we got the R^2 of about 0.7477 and the RMSE is about 6089.185. The RMSE we got is pretty large, which indicates that the linear model is not the appropriate model to predict the *charge.*

Residual & QQplot

Next, we did the residual analysis based on the linear model to explore the problems within that model.

Figure: QQPLOT and Residual Plot

As we observe, our qq-plot is really non-normal and the residual plot appears really strange as the points in the residual plot are clustered in three parts.These strange points shown in the plot with our huge RMSE all indicate the linear model is not fit to predict our response variable, *charge.* One possible explanation for that may be the values for the response variable, *charge* are extraordinarily larger than the values for the six predictors.

Since our linear model is not appropriate for predicting, to fit other models, we may consider doing the log transformation to the response variable *charge,* or we will add the polynomial in the model.

METHOD: Multinomial Logistic Regression Model

As we saw from our linear model, considering the qq-plot and residual plot of the regression model, running purely based off of numerical prediction is incredibly inaccurate and doesn't leave us with many insights behind our data. Predicting a certain range of the insurance charges seems doable and reasonable for our targeted customers and policymakers, since many insurance policies also classify targeted values into tiers. Therefore, we considered transforming the response variable, *charges* and created a categorical one called ***tier*** to see whether it would yield more accuracy.

To create the new ***tier*** variable, we divided *charges* based on its quartiles. For those smaller than the 25% quantile, we assigned them to tier 1, denoted as *1* in *tier*. For those between the first quartile (inclusive) and the median (exclusive), we classified them as *2* in *tier*. For values larger than or equal to the median but smaller than the 75% quantile, we denoted them as in tier *3*. And for those greater than or equal to the third quartile, we classified them as *4*.
Here in our data, the three quartiles are 4740, 9382, and 16639.913 respectively.

Based on our new response variable with four levels, we then considered using a multinomial logistic regression model since our goal is to predict whether an observation's charge belongs to tier 1, 2, 3, or 4, and we used all six predictors– *Age, BMI, Children, Sex, Smoker,* and *Region*.

The multinomial logistic regression model includes the following three equations, each representing the log of the probability of charges being in tier 2, 3, or 4, over the probability of charges being in the baseline, tier 1.

$$\ln(\text{tier2}/\text{tier1}) = \beta_{10} + \beta_{11} \cdot \text{age} + \beta_{12} (\text{sex} = \text{male}) + \beta_{13} \cdot \text{bmi} + \beta_{14} \cdot \text{children} + \beta_{15} (\text{smoker} = \text{yes}) + \beta_{16} (\text{region} = \text{northwest}) + \beta_{17} (\text{region} = \text{southeast}) + \beta_{18} (\text{region} = \text{southwest})$$

$$\ln(\text{tier3}/\text{tier1}) = \beta_{20} + \beta_{21} \cdot \text{age} + \beta_{22} (\text{sex} = \text{male}) + \beta_{23} \cdot \text{bmi} + \beta_{24} \cdot \text{children} + \beta_{25} (\text{smoker} = \text{yes}) + \beta_{26} (\text{region} = \text{northwest}) + \beta_{27} (\text{region} = \text{southeast}) + \beta_{28} (\text{region} = \text{southwest})$$

$$\ln(\text{tier4}/\text{tier1}) = \beta_{30} + \beta_{31} \cdot \text{age} + \beta_{32} (\text{sex} = \text{male}) + \beta_{33} \cdot \text{bmi} + \beta_{34} \cdot \text{children} + \beta_{35} (\text{smoker} = \text{yes}) + \beta_{36} (\text{region} = \text{northwest}) + \beta_{37} (\text{region} = \text{southeast}) + \beta_{38} (\text{region} = \text{southwest})$$

Bootstrapping

To assess model accuracy, we performed Empirical Bootstrap resampling on our multinomial model to find the standard error of our coefficients. Because of the normality issues within our data, estimating standard error through bootstrapping can give us a more accurate estimate of our model's accuracy and quality.

Results

Linear Model
Here is our fitted linear regression model.

Charge = -11938.5+256.9**age**+475.5**children**+339.2**bmi**-131.3**sexmale**+
23848.5**smokeryes**-353**regionnorthwest**-1035**regionsoutheast**-960r**egionsouthwest**

Taking predictors *age* as an example, the interpretation of its coefficients will be holding other predictors fixed, as the age increases one unit, the charge of insurance will increase 256.9 respectively.

5-fold Cross-Validation Result:
The final linear model is

Charge = -11938.5+256.9**age**+475.5**children**+339.2**bmi**-131.3**sexmale**+
23848.5**smokeryes**-353**regionnorthwest**-1035**regionsoutheast**-960.1**regionsouthwest**

which is really similar to the linear model we got before. For this model, we got the R^2 of about 0.7477 and the RMSE is about 6089.185. The RMSE we got is pretty large, which indicates that the linear model is not the appropriate model to predict the *charge.*

Multinomial
Here are the results from the fitted multinomial logistic regression model.

The three rows in the results for coefficients correspond to the fitted values for the coefficients of the three equations.

ln(tier2/tier1) = -9.559 + 0.301 * age - 0.545(sex = male) + 0.005 * bmi + 0.78 children + 20.794(smoker = yes) -0.126 (region = northwest) - 0.755(region = southeast) - 0.555(region = southwest)

ln(tier3/tier1) = -17.752 + 0.476 * age - 0.945(sex = male) + 0.023 * bmi + 0.858 children + 30.014(smoker = yes) -0.742 (region = northwest) - 1.767(region = southeast) - 1.244(region = southwest)

ln(tier4/tier1) = -15.568 + 0.371 * age - 0.701(sex = male) + 0.073 * bmi + 0.842 children + 38.003(smoker = yes) -0.522 (region = northwest) - 1.41(region = southeast) - 1.59(region = southwest)

Some examples of the interpretation for the coefficients:

- for b_{11}, one-unit increase in the variable age is associated with the increase in the log odds of being in tier 2 vs. in tier 1 in the amount of 0.301.
- For b_{22}, The log odds of being in tier 3 vs. in tier 1 will decrease by 0.954 if moving from sex="female" to sex = "male".

To have an approximate estimate for the accuracy, we used 80% for training and made the rest 20% as our validation set to calculate the prediction accuracy, which is around 0.825.

Bootstrapping Results:

Table 3:

	(Intercept)	age	children	bmi	sexmale	smokeryes	regionnorthwest	regionsoutheast	regionsouthwest
2	-9.735505584	0.3075843969	0.8945053351	-0.000880764406 8	-0.7133852941	13.44375433	-0.1921073772	-0.6631007596	-0.6111449743
3	-17.34151242	0.4693642999	0.9433629421	0.0185265 4884	-1.091512547	26.55458556	-0.6494645944	-1.641863458	-1.311721482
4	-17.07218604	0.3994424642	0.9940535247	0.0676884 5098	-0.8844988257	30.25068953	-0.4573442674	-1.127513867	-1.583989072

se2	1.22775524	0.03366653407	0.1277803497	0.02320215471	0.2677621179	3.213739079	0.3633126585	0.4261280021	0.374392372
se3	2.025545759	0.04893808846	0.1781214637	0.03004554332	0.3471402016	3.495542728	0.4743442149	0.5610395484	0.4822906207
se4	2.467406954	0.05285706713	0.2024087158	0.03544278833	0.4206261485	3.453321891	0.6108293564	0.6481580829	0.6155112278

We can see from our bootstrapping that the estimated model coefficients are fairly close to the true model coefficients.

Standard error has increased somewhat across the board, generally with our higher tiers. This is likely because our higher tiers have much more variability in them, making it harder to estimate for. BMI, sex and regions seem to have noticeably large standard errors for them as well.

Precision and Accuracy

Table 4:

	Tier 1	Tier 2	Tier 3	Tier 4
Accuracy	0.8918919	0.8571429	0.8915663	0.9393939
Precision	0.8805970	0.8450704	0.8783784	0.9285714
Recall	0.8939394	0.9090909	0.8783784	0.8387097

Our results for accuracy, precision, and recall are all very strong. Our lowest score for accuracy was in tier 2 with 85.7% and our highest score was in tier 4 with 93.9%. With precision our lowest score was in tier 2 with 84.5% and the highest was again tier 4 with 92.8%. Lastly with recall our lowest score was in tier 4 with 83.8% and our highest was in tier 2 with 90.9%. It seems that our model is better at categorizing a charge in the 4th tier and worse when the charge is in the 2nd tier.

Overall Coefficient Analysis:

From our multinomial model we found that whether or not someone was a smoker was a primary factor in influencing the insurance charges for an individual. Age and children were also significant, although the standard error we calculated from bootstrapping indicates that children may not have the most accurate estimation of its coefficient. The standard error for gender and region indicate that they aren't as strong of coefficients.

Limitation

The approach of transforming the continuous response variable, *charges*, into a categorical variable with four tiers based on quantiles and using multinomial logistic regression introduces a limitation in terms of sacrificing detailed information for the sake of accuracy. By categorizing charges into broad tiers, such as the 75th quantile to the maximum, which is from 16639.913 to 63770.428, the range within each tier can be quite large, resulting in potential significant differences between charges within the same tier. Meanwhile, the span over the interval between the minimum and the 25th quantile, which is 1121.874 to 4740.287, is much smaller than the one from 75th quantile to the maximum, so instead of predicting specific charges, predicting a range leads to a loss of granularity and precision in the insurance charge predictions.

One possible alternate solution to improve this limitation a bit is to define a more reasonable set of ordinal levels using the domain knowledge in insurance and consider ordinal logistic regression instead of multinomial logistic regression. While ordinal logistic regression models capture the ordinal relationship between the tiers, it also maintains the ordering of the levels. Ordinal logistic regression allows for modeling the ordinal relationship between the tiers while preserving the ordering of the categories. The model can thus capture inherent order and potentially improve the limitation to some extent, with careful tailoring for the ordinal level definitions using insurance-related expertise.

Another alternative could be keeping the continuous response variable and using more powerful algorithms such as random forest regression, which takes the idea of bootstrapping, and gradient boosting. These algorithms may provide more accurate predictions without sacrificing the detailed information, but they would be harder to interpret, and for algorithms like gradient boosting, we have to fine-tune a lot of hyperparameters, which could be very computationally expensive.

Therefore, we should carefully consider the trade-off between accuracy and the loss of detailed information when transforming a continuous response variable into categorical tiers.

Linear model code

```
insurance_model = lm(charges ~., data = insuranceData)
summary(insurance_model)
```

5-fold cross validation code

```
library(caret)
```

```
ctrl <- trainControl(method = "cv", number = 5)

#fit a regression model and use k-fold CV to evaluate performance
linear_model <- train(charges ~ age + children + bmi + sex +
                      smoker + region, data = insuranceData,
              method = "lm", trControl = ctrl)
print(linear_model)
linear_model$finalModel
```

Multinomial model code

multinom

```{r}
library(dplyr)
library(nnet)
set.seed(403)

q = quantile(insuranceData$charges)
# q1 = 4740.287
# median = 9382.033
# q3 = 16639.913
changedata = insuranceData %>% mutate(tier= if_else(.$charges < q[2], "1", (if_else(.$charges
< q[3], "2", (if_else(.$charges < q[4], "3", "4"))))))
changedata = changedata[, -c(7)]

insurance_train_index <- sample(1:nrow(changedata), 0.8 * nrow(changedata))  # 80% for
training
insurance_train_data <- changedata[insurance_train_index, ]
insurance_test_data <- changedata[-insurance_train_index, ]

# the code is for results section

multinom_model <- multinom(tier~., data = insurance_train_data)
summary(multinom_model)

predicted_value <- predict(multinom_model, insurance_test_data)
table(predicted_value, insurance_test_data$tier)
sum(predicted_value == insurance_test_data$tier)/nrow(insurance_test_data)
```

Bootstrap code:

```{r}
B = 10000
size = nrow(changedata)
coeffs2 = matrix(NA, nrow=B, ncol=9)
coeffs3 = matrix(NA, nrow=B, ncol=9)
coeffs4 = matrix(NA, nrow=B, ncol=9)

for(i in 1:B) {
  sampleData = sample(size, size, replace=T)
  multinom_model <- multinom(tier~age + children + bmi + sex +
                  smoker + region, data = changedata, subset=sampleData, trace=F)
  coeffs2[i,] = coef(multinom_model)[1,]
  coeffs3[i,] = coef(multinom_model)[2,]
  coeffs4[i,] = coef(multinom_model)[3,]
}
coef_mat = rbind(apply(coeffs2, 2, mean), apply(coeffs3, 2, mean), apply(coeffs4, 2, mean))
colnames(coef_mat) = colnames(coef(multinom_model))
rownames(coef_mat) = c("2", "3", "4")
se_mat = rbind(apply(coeffs2, 2, sd), apply(coeffs3, 2, sd), apply(coeffs4, 2, sd))
rownames(se_mat) = c("se2", "se3", "se4")
coef_stuff = rbind(coef_mat, se_mat)
write.csv(coef_stuff, "bootstrap_coef.csv")
```