Health Care Cost Prediction with:

- ★ Linear Regression
- ★ Logistic Regression
- ★ Linear Mixed model
- ★ Generalized Mixed Models

Presenters:
Ayush Chakraborty
Gibran Erlangga
Hyunjoe Yoo
Parinitha Kompala
Shrey Khetrapal

Introduction

Medical expenses are difficult to estimate because the most costly conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

Goal

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

Dataset

For this analysis, we will use a simulated dataset containing hypothetical medical expenses for patients in the United States. This data was created for using demographic statistics from the US Census Bureau, and thus, approximately reflect real-world conditions.And got this from Kaggel.

The insurance.csv file includes 1,338 examples of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year. The features are:

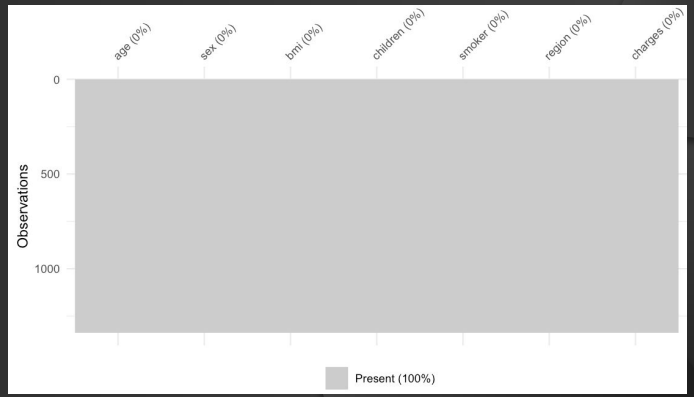
- *age:* An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
- *sex:* The policy holder's gender, either male or female.
- *bmi:* The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
- *children:* An integer indicating the number of children/dependents covered by the insurance plan.
- *smoker:* A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
- *region:* The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.

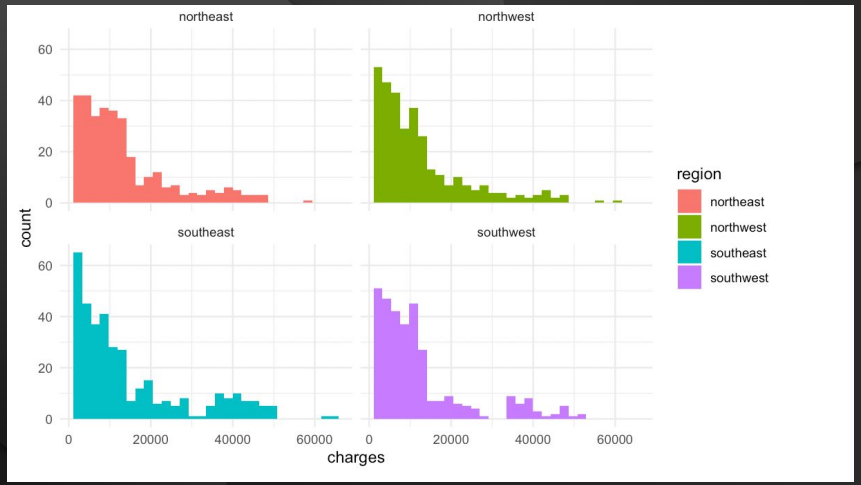
DATA EXPLORATION

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

← The structure of our dataset.

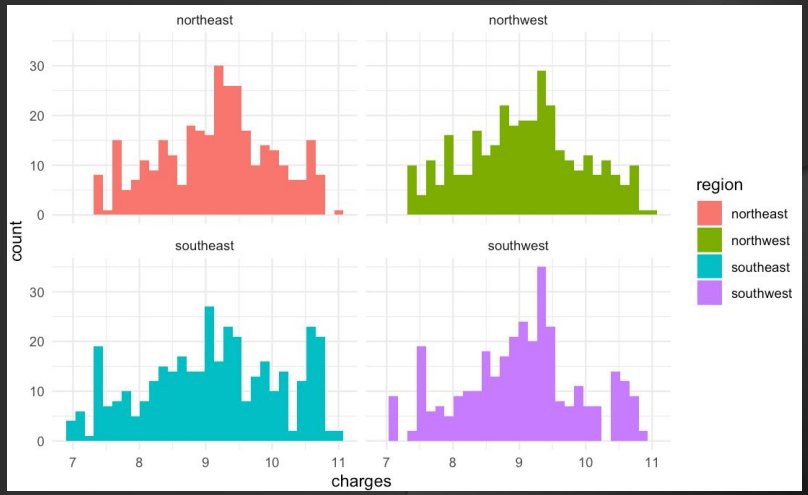
Checking for missingness in the dataset →

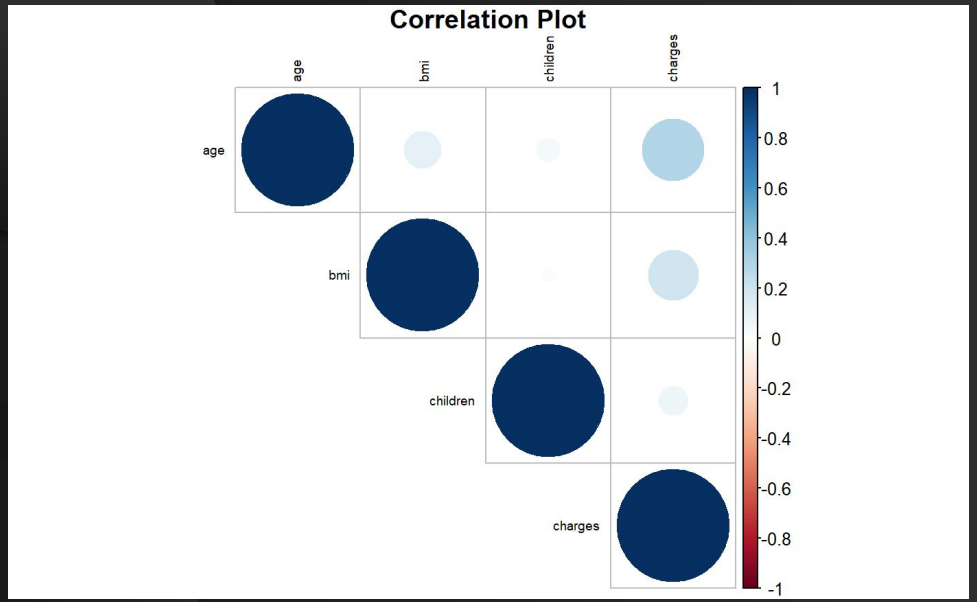




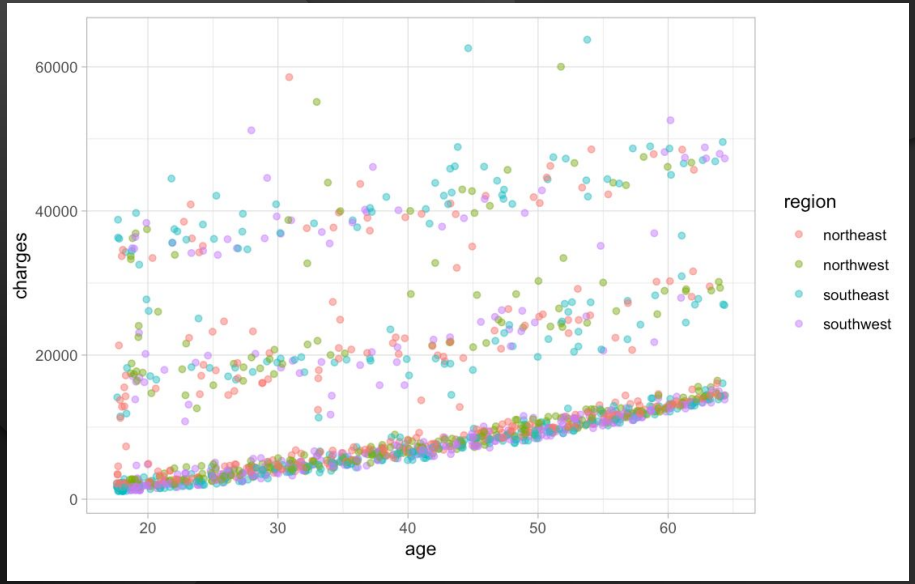
From this we can conclude that it is not normally distributed and is skewed right.

After adding log function to the variable



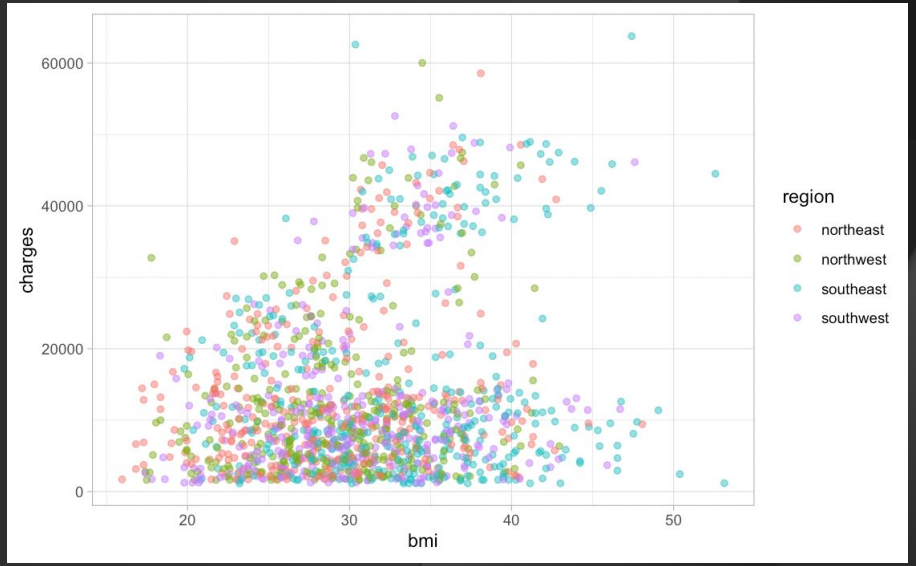


Correlation between the variables.

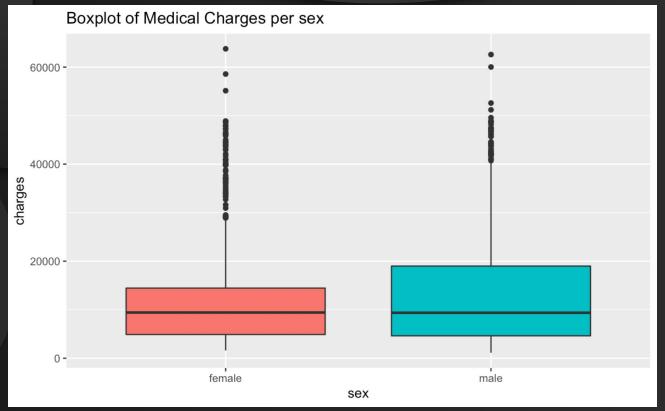


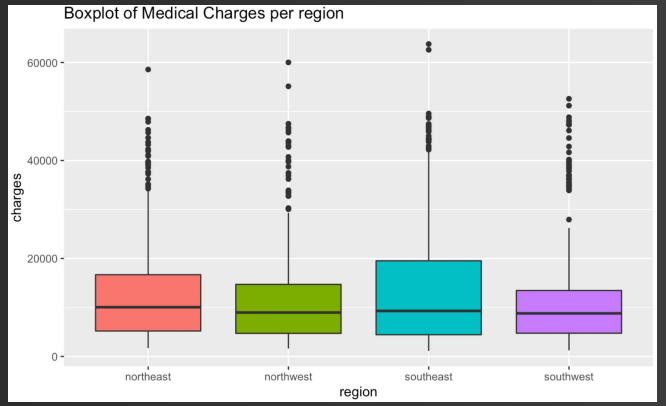
As Age go up Charges for health insurance also trends up.

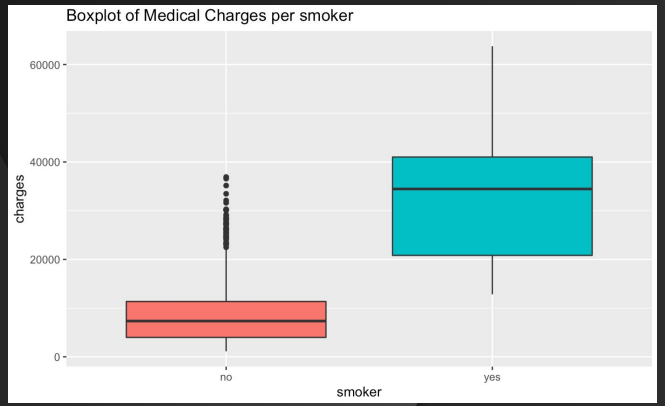
Looking into the correlation between charges and age, charges and bmi based on the region

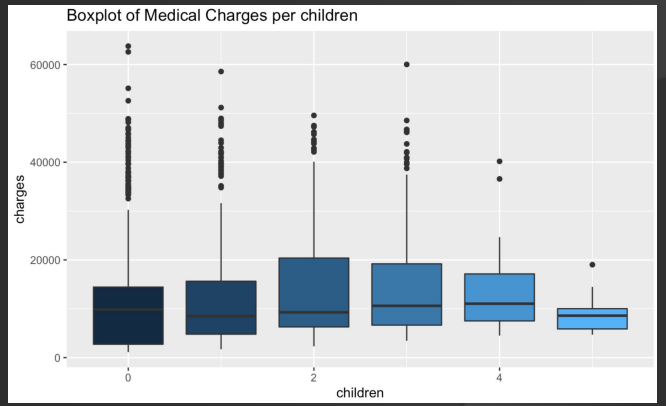


Dependent vs. Independent Variables









Linear Regression

```
Call:
lm(formula = charges ~ age + bmi + region + children + sex +
    smoker, 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 ***
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
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
---
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
```

Dependent variable: Insurance charges

Independent variable: Age, BMI, Region, Children, Sex, Smoker

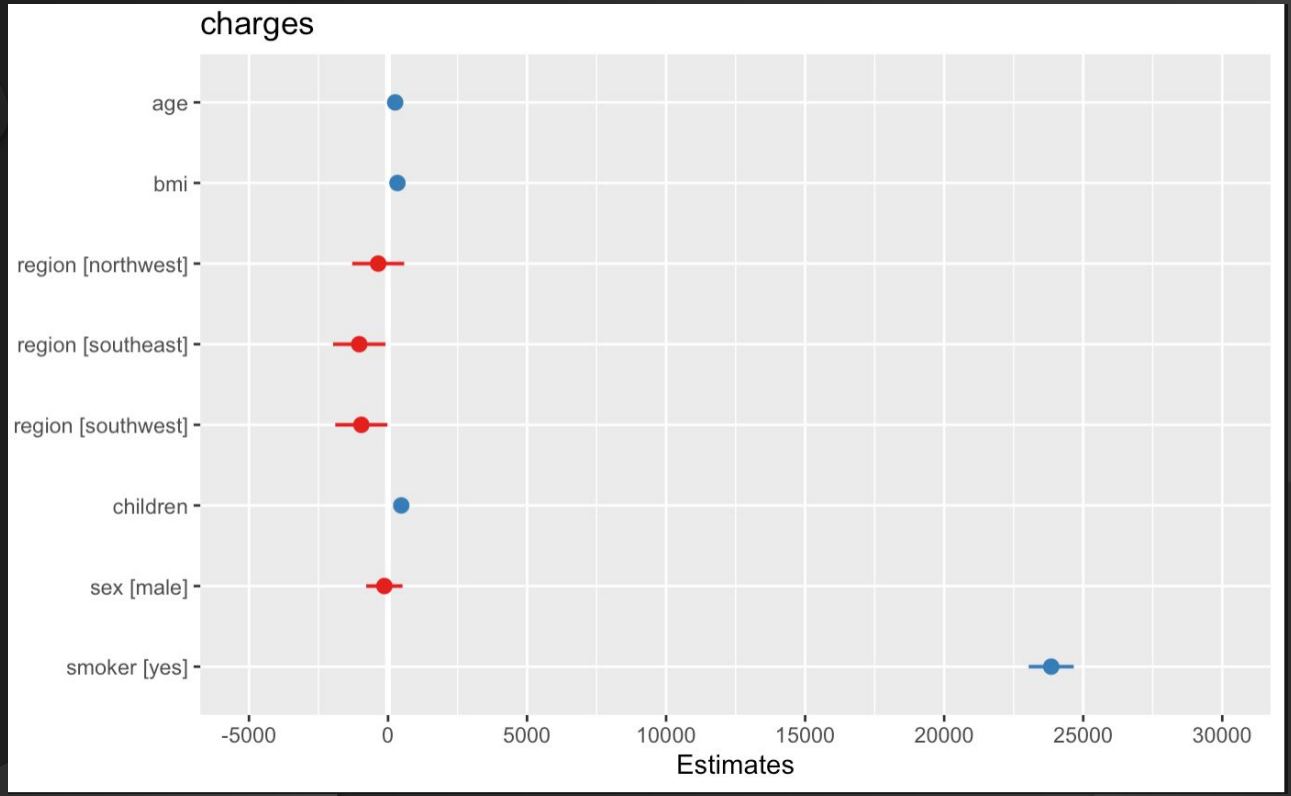
P value < 0.05 ⟶ significant

Top significant variables:

Age, BMI, Children, Smoker

R-squared: 0.75 ⟶ 75% variance explained by the model

Linear Regression



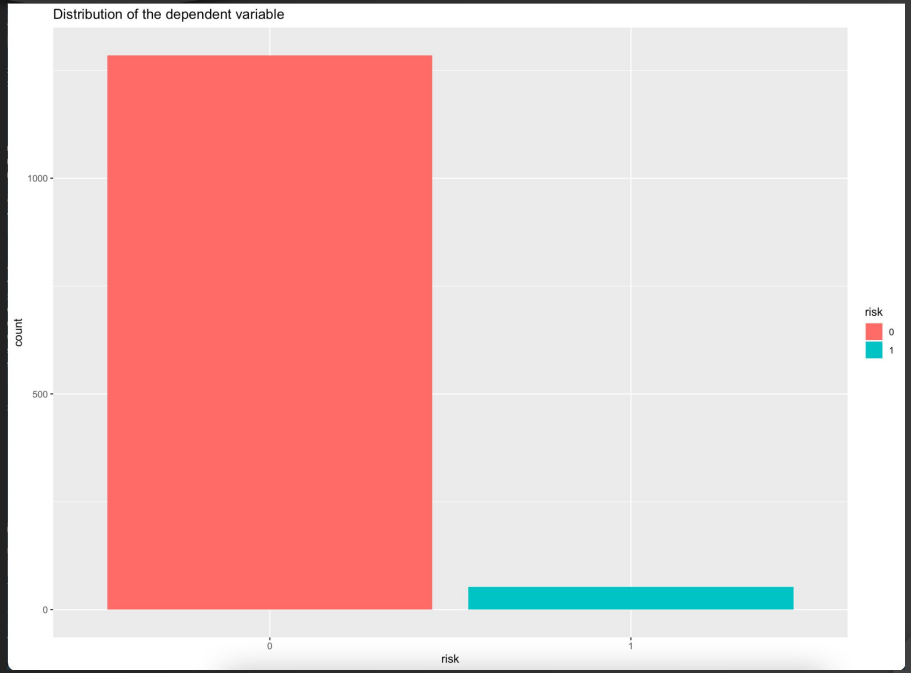
Logistic Regression

Intuition : Health care companies and High Risk Applicants

For the purpose of Logistic Regression we categorised each individual as a High Risk Applicant if he has attributes :

- BMI > 25
- Smoker Status : Yes
- Age > 50

Logistic Regression



0 : 1258

1 : 53

Logistic Regression

Imbalance in classification problem

- Oversample (eg SMOTE)
- Downsample

Divide into Training and Test set (80-20)

Logistic Regression

```
```{r}
model_glm <- glm(risk~age+sex+bmi+children+smoker+region,
                 family=binomial,
                 data=training_data)
```
```

Logistic Regression

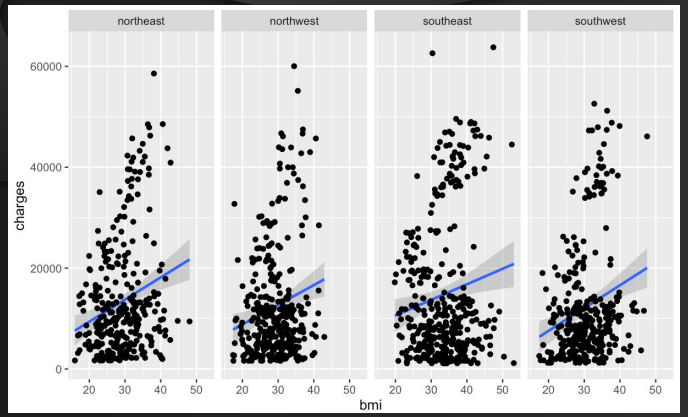
Confusion Matrix

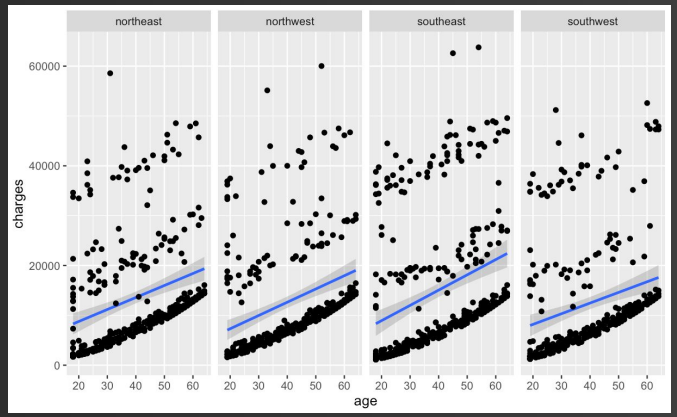
```
> confusion_matrix
     risk
pred  0  1
   0 17  2
   1  1 14
```

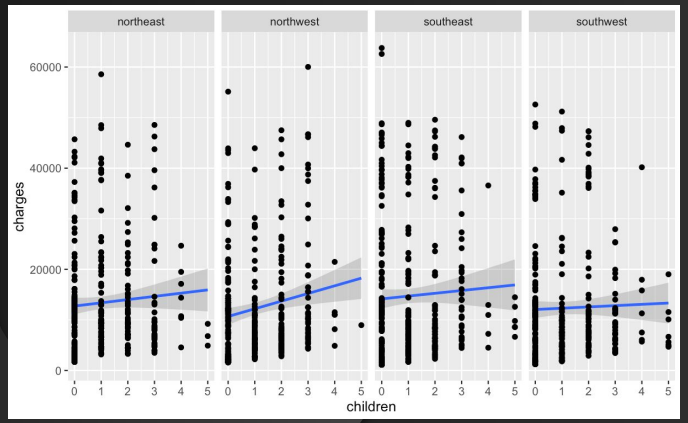
Accuracy : 91.17 %

Recall : 87.5%

Linear Mixed-Effects Model (LMM)







Different intercept and slopes on each region for different variables

Not an ideal case for LMM because these trends have same direction (positive slope, upwards)

Linear Mixed-Effects Model (LMM)

1

```
Linear mixed model fit by REML ['lmerMod']
Formula: charges ~ sex + age + bmi + (1 | region)
   Data: insurance

REML criterion at convergence: 28737.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3301 -0.6210 -0.4400  0.5972  4.1444

Random effects:
 Groups   Name        Variance  Std.Dev.
 region   (Intercept)    232625   482.3
 Residual             129135439 11363.8
Number of obs: 1338, groups:  region, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept) -6873.71    1794.59  -3.830
sexmale      1341.20     622.25   2.155
age           243.57      22.27  10.938
bmi           323.17      52.17   6.194

Correlation of Fixed Effects:
        (Intr) sexmal age
sexmale -0.145
age     -0.390  0.026
bmi     -0.827 -0.048 -0.113
```

2

```
Data: insurance
Models:
model_lr: charges ~ sex + age + bmi
model_lmer: charges ~ sex + age + bmi + (1 | region)
           npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
model_lr      5 28794 28820 -14392    28784
model_lmer    6 28796 28827 -14392    28784 0.0541  1      0.816
```

3

```
$region
          (Intercept)
northeast    225.7956
northwest   -131.7809
southeast    256.8350
southwest   -350.8496

with conditional variances for "region"
```

Conclusion:

There is not much difference in variance happened between different regions, hence LMM is **NOT** needed and LM is better fit for this case

- (1 | region) as random effect to handle non-independence by introducing different intercepts for each region
- After running ANOVA between LR and LMM, the AIC BIC values ***increase*** between LMM compared to LR -> the model unnecessarily increase the complexity without improving the fit to the data

Generalized Linear Mixed Model (GLMM) on logistic regression

- Why are we using the method?
  - Provides information about whether individual variation amongst regions is significant.
  - To see if there is a different trends compared to overall in the regions.
- What variables used in the model
  - Fixed - effect variables: Sex, Age, Children, Smoker
  - Random - effect variables: Region

Generalized Linear Mixed Model (GLMM) on logistic regression

RStudio: Notebook Output

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: charges_cat ~ sex + age + children + smoker + (1 | region)
   Data: insurance

     AIC      BIC   logLik deviance df.resid
   797.8    829.0   -392.9    785.8     1332

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4752 -0.2996 -0.0371  0.2940 12.3914

Random effects:
 Groups Name        Variance Std.Dev.
 region (Intercept) 0.05713  0.239
Number of obs: 1338, groups:  region, 4

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.707e+00  5.033e-01 -15.314   <2e-16 ***
sexmale     -3.338e-01  1.804e-01  -1.850   0.0643 .
age          1.671e-01  9.954e-03  16.788   <2e-16 ***
children     1.440e-01  7.435e-02   1.937   0.0528 .
smokeryes    2.299e+05  4.054e+06   0.057   0.9548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
          (Intr) sexmal age    chldrn
sexmale   -0.104
age       -0.920 -0.075
children  -0.365 -0.012  0.198
smokeryes  0.000  0.000  0.000  0.000
optimizer (Nelder_Mead) convergence code: 0 (OK)
unable to evaluate scaled gradient
 Hessian is numerically singular: parameters are not uniquely determined
```

Generalized Linear Mixed Model (GLMM) on logistic regression

```
Random effects:
 Groups Name        Variance Std.Dev.
 region (Intercept) 0.05713  0.239
Number of obs: 1338, groups:  region, 4
```

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.707e+00  5.033e-01 -15.314   <2e-16 ***
sexmale     -3.338e-01  1.804e-01  -1.850   0.0643 .
age          1.671e-01  9.954e-03  16.788   <2e-16 ***
children     1.440e-01  7.435e-02   1.937   0.0528 .
smokeryes    2.299e+05  4.054e+06   0.057   0.9548
```

Generalized Linear Mixed Model (GLMM) on logistic regression

RStudio: Notebook Output

```
$region
           (Intercept)
northeast  0.29605410
northwest  0.04043428
southeast -0.16775093
southwest -0.16823272

with conditional variances for "region"
```

Conclusion:
There is not much difference in variance happened between different regions, hence GLMM is **NOT** needed and GLM is better fit for this case

Conclusion

- We have figured out the features or variables which play vital role in the amount of health insurance as well as the prediction model to help the companies and the customers.
- For further analysis, dataset with more features will provide comparatively more accurate outputs.

Thank You
Any questions :) ?