Final Project

Huynh Huu Phuc

2023-07-30

Table of Contents

Including Library

Activity 1

About the dataset

Dataset 2

The dataset has 11 columns and 680 observations

1. attend: classes attended out of 32
2. termGPA: GPA for term
3. priGPA: cumulative GPA prior to term
4. ACT: ACT score
5. final: final exam score
6. atndrte: percent classes attended
7. hwrte: percent homework turned in
8. frosh: 1 if freshman, otherwise is 0
9. soph: 1 if sophomore, otherwise is 0
10. skipped: number of classes skipped
11. stndfnl: standardizing the values of final, which is defined by formula (final - mean)/sd, where mean, sd are the median, standard deviation of final, respectively

These data were collected by Professors Ronald Fisher and Carl Liedholm during a term in which they both taught principles of microeconomics at Michigan State University. Professors Fisher and Liedholm kindly gave me permission to use a random subset of their data, and their research assistant at the time, Jeffrey Guilfoyle, provided helpful hints.

Goal: The dataset aims to determine the impact of several factors, including term GPA, cumulative term GPA, ACT score, percentage of classes attended, percentage of assignments turned in, and student level (freshman, sophomore, other), on the final score exam.

Cleaning datasets

Consider the type of attribute

```
'data.frame':	680 obs. of  11 variables:
 $ attend : int  27 22 30 31 32 29 30 26 24 29 ...
 $ termGPA: num  3.19 2.73 3 2.04 3.68 3.23 1.54 2 2.25 3 ...
 $ priGPA : num  2.64 3.52 2.46 2.61 3.32 2.93 1.94 2.12 2.06 2.73 ...
 $ ACT    : int  23 25 24 20 23 26 21 22 24 21 ...
 $ final  : int  28 26 30 27 34 25 10 34 26 26 ...
 $ atndrte: num  84.4 68.8 93.8 96.9 100 ...
 $ hwrte  : chr  "100" "87.5" "87.5" "100" ...
 $ frosh  : int  0 0 0 0 0 0 1 0 1 0 ...
 $ soph   : int  1 0 0 1 1 1 0 1 0 1 ...
 $ skipped: int  5 10 2 1 0 3 2 6 8 3 ...
 $ stndfnl: num  0.4727 0.0525 0.8929 0.2626 1.7332 ...
```

Check whether there is missing data.

```
0
```

However, the **hwrte** attribute should be numeric instead of characters. There might be some unusual value here.

```
"100"  "87.5" "75"   "50"   "62.5" "25"   "."    "12.5" "37.5"
```

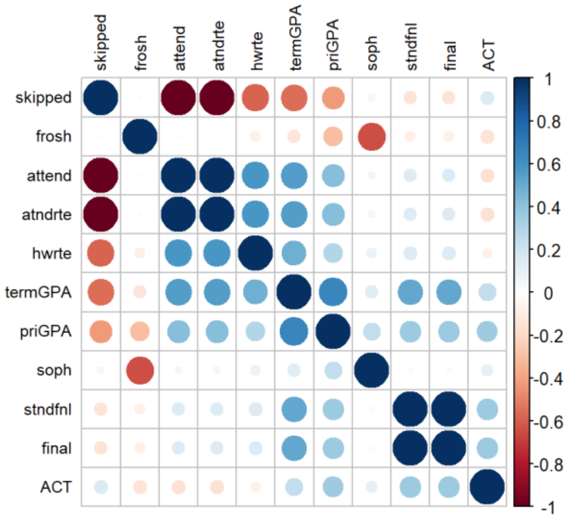
There is wrong data which is "." value.

```
6
```

There are 6 row that is not correct so we replace it by the median of hwrte.

```
100.0  87.5  75.0  50.0  62.5  25.0  12.5  37.5
```

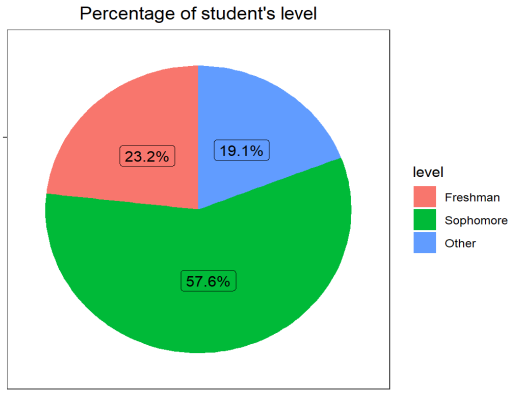
Correlation

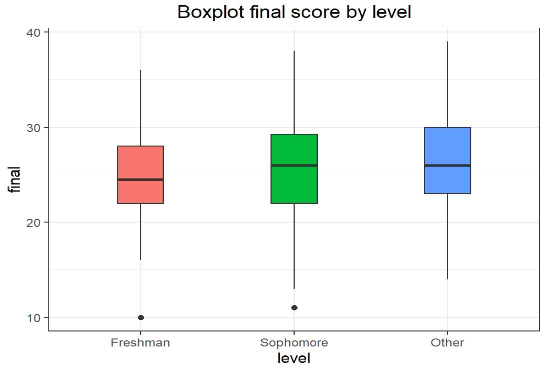


There are strong correlation between final score and termGPA, priGPA and ACT. The stndfnl attribute is a normalized value of final, as a result, it has strong correlation with final.

Descriptive statistics

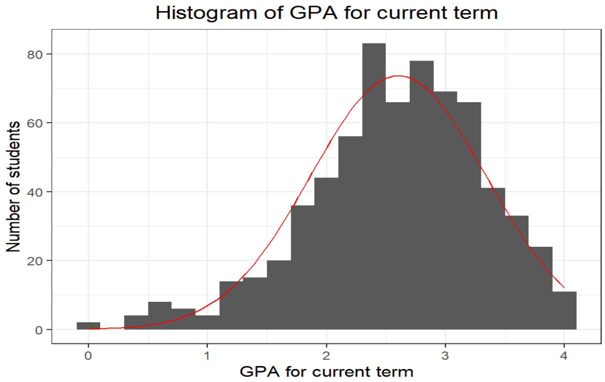
Levels

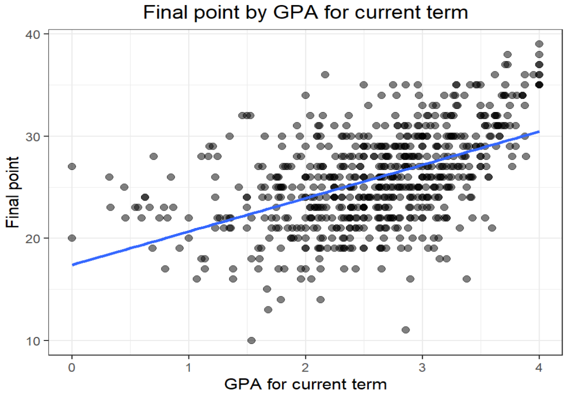




Majority of students participating in the survey are Sophomores, accounting for 56.7%. Additionally, the Sophomore category is around 2.445 times larger than the Freshman category and approximately 2.96 times larger than the Other category, which may consist of Juniors and Seniors. This indicates that Sophomores are significantly more prevalent in the dataset compared to others category. Moreover, students in the Other and Sophomore categories have a wider and higher range of final scores compared to Freshman. This indicates that students in these two categories tend to achieve higher scores on their final exams and have a greater diversity of scores. Additionally, in both the Freshman and Sophomore categories, there is an outlier with a significantly lower score, approximately 10 points.

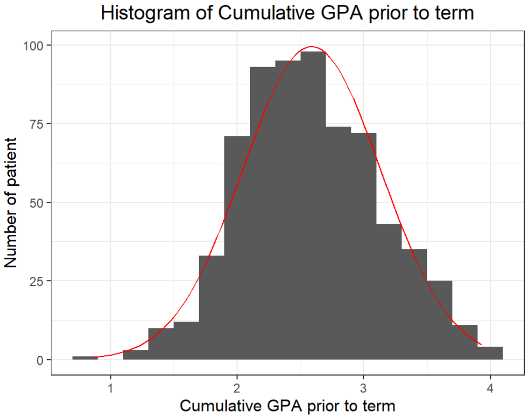
Current term GPA

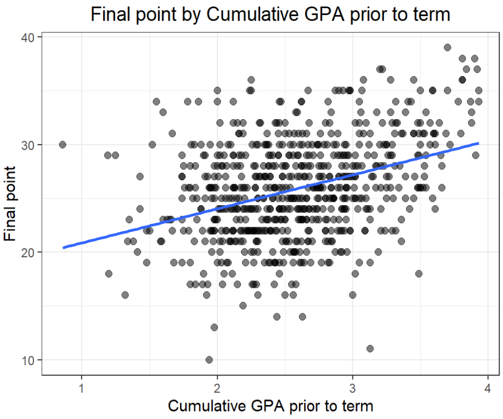




The histogram is bell-shaped, the mean and median are around 2.6 and the histogram is trending upwards between 2.4 and 3.3, corresponding to a score of C to B+. This proves that the majority of students in the survey have average to near good academic achievement. In addition, the graph has a skewed direction with a variety of grades, showing that there are a few students who are still not focused on their studies, leading to a rather low GPA in the term, specifically below 1. Besides, the correlation between the two variables final score and termGPA shows that there is a positive relationship between these two types of scores. Students who score well in the final exam will often be more likely to achieve a higher GPA.

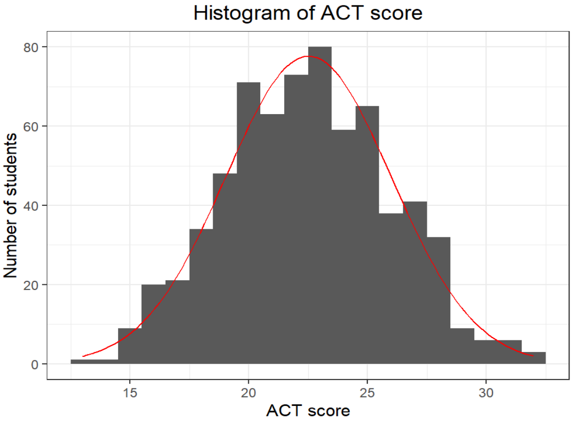
Cumulative GPA

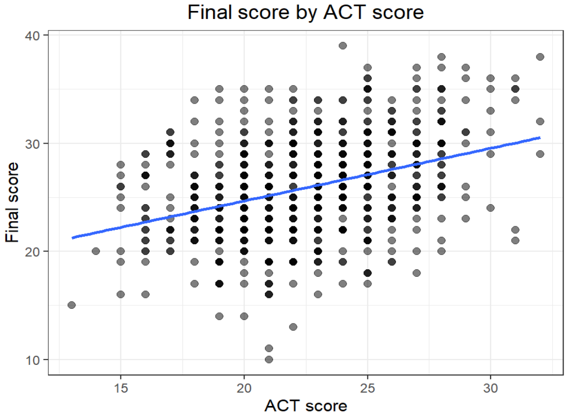




The chart with standard deviation value is about 0.545 shows a diverse range from 0.857 to 3.93 and uneven distribution in the priGPA scores of students in the data set. In which, mean and median both fall in the range of approximately 2.5, showing that students tend to get average grades, from C to C+. Moreover, it can be seen in the scatter plot that there is a positive relationship between cumulative GPA score and final exam score, which predicts that students with higher cumulative GPA scores are more likely to score well in the final exam.

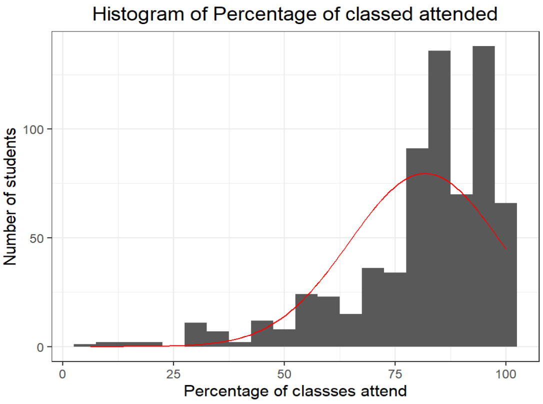
ACT score



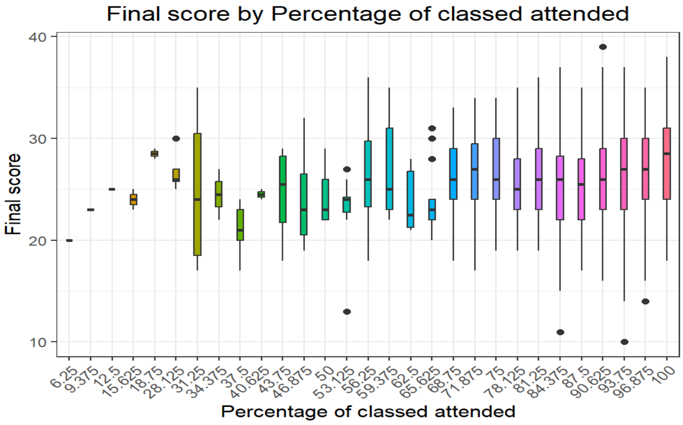


The histogram of the ACT scores displays a sawtooth pattern, indicating periodic fluctuations in the number of students achieving specific scores within the range of 13 to 32. Additionally, with an average score of approximately 22.5, a median score of around 22.0, and the narrowing shape of the histogram between 20 and 25, it suggests that students tempt to attain scores fluctuating around this range. Furthermore, there seems to be a positive correlation between the ACT scores and the final grades obtained by the students. However, the dispersion is quite large.

Percentage of classes attended

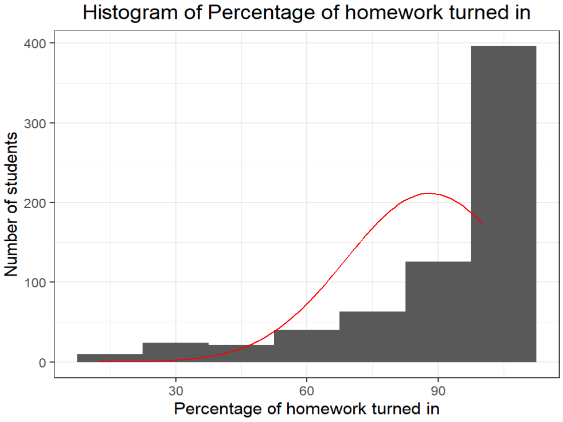


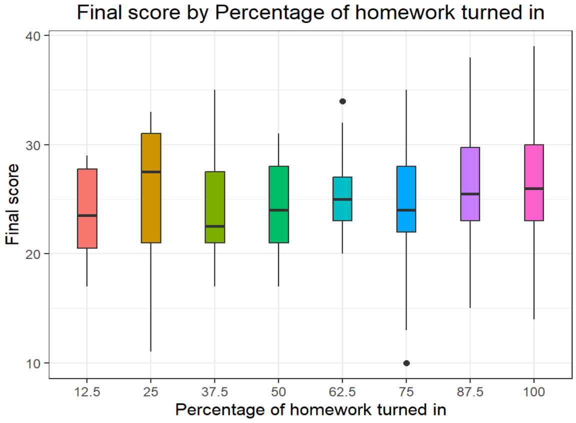
From the histogram, it can be observed that students tend to participate in classes quite regularly and actively, as the percentage of students attending classes is relatively high, usually ranging from 80% to 95%. In addition, the outliers of the variable are determined to be below 30%, causing the histogram to be skewed to the left. This indicates that there are a few students with very low attendance, lying outside the acceptable range of the overall distribution.



While the students attend nearly one-half of class has the mean of final score similar to that of the students who attend all classes. Thus, the data reveals the rate of percentage of class does not reflect the final score exam.

Percentage homework turned in

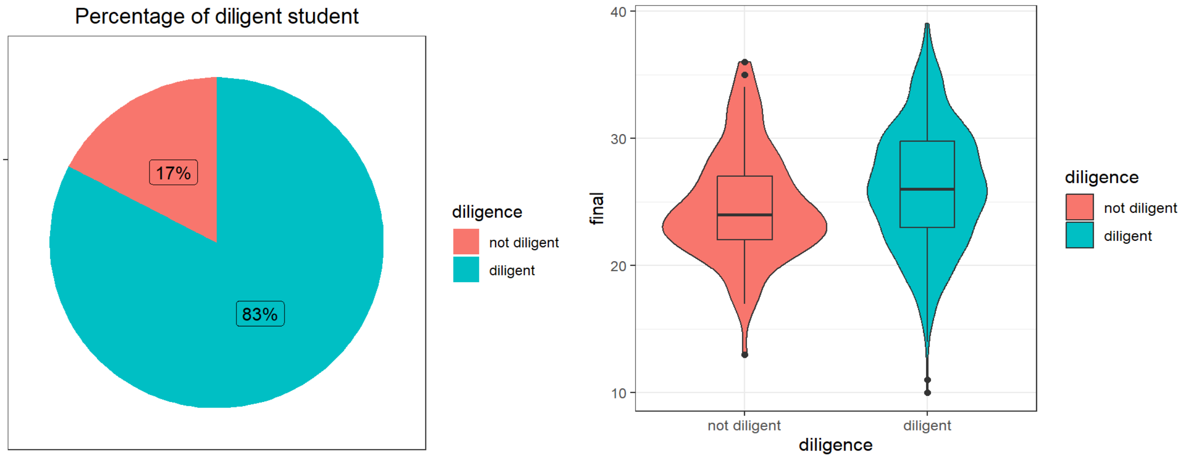




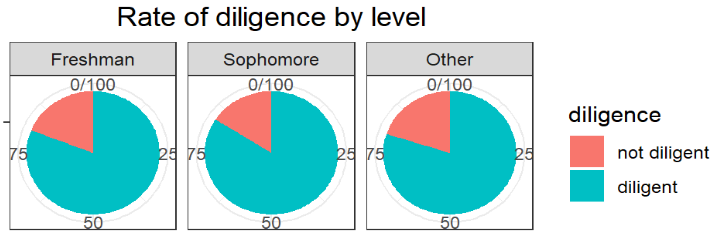
The mean and standard deviation value of cumulative GPA prior to term is about 88.015 and 19.217, respectively. In more detail, histogram shows that the data appears to be separate on one side compared to the other data. This shows that the percentage of homework turn in is not concentrated in a certain range, however, students still tend to submit their papers fully when the top of the graph is skewed towards 100% and this is also the median of the dataset. According to the boxplot, students who turn in more homework tend to receive a higher score at final exam. However, there is a small percentage of students rarely submit their work, and their final grade is greater than those who submit more than 75% of their work. As a result, the percentage of homework turned in has little bearing on the final score.

Diligence

The diligence attribute explains how often students show up for class. They are diligent if they attend more than 70% of the scheduled classes. Otherwise, they are considered as not diligence.



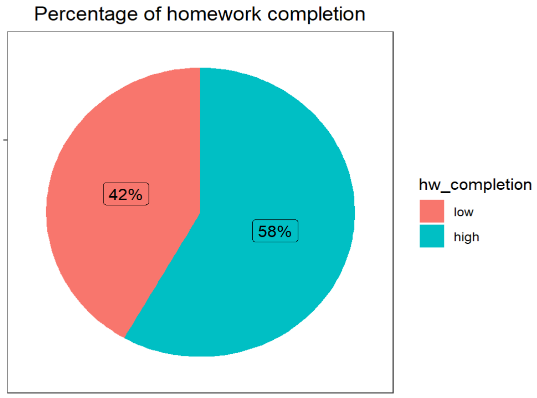
There is a little bit different on final score between student diligent and not diligent student. In more detail, the final score of diligent student is temp to larger than the group of not diligent student.

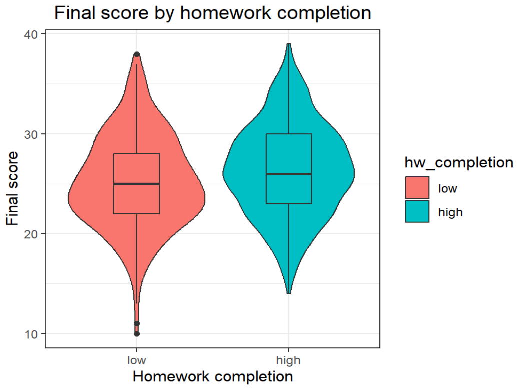


The percentage of diligent student is fluctuated between student's level. From the first pie charts, students that are sophomore, are more diligent than freshman.

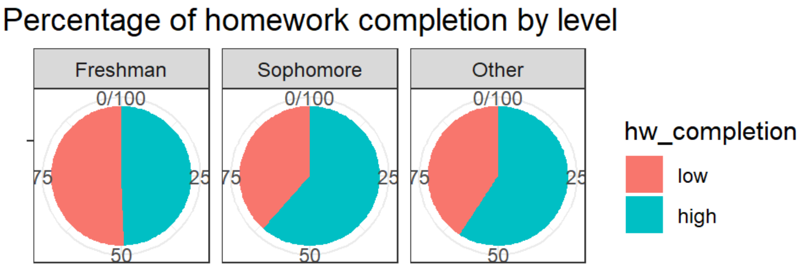
Homework Completion Level

The level of homework completion is high if they complete more than 90% of it.



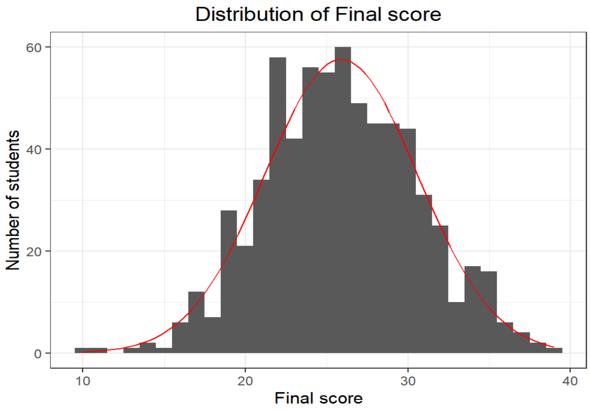


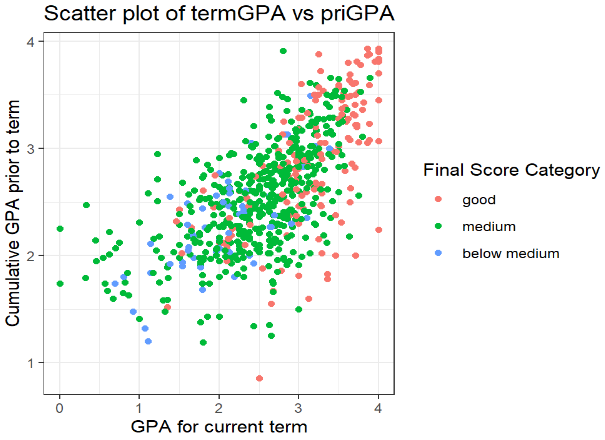
The percentage of students who are regarded as high appears to be slightly higher than that of students who are judged as low. According to the boxplot, students who turn in the majority of their homework tend to receive higher grades than those who do not.



The percentage of sophomores who do their homework is higher than that of freshmen, as seen by this figure.

Final score

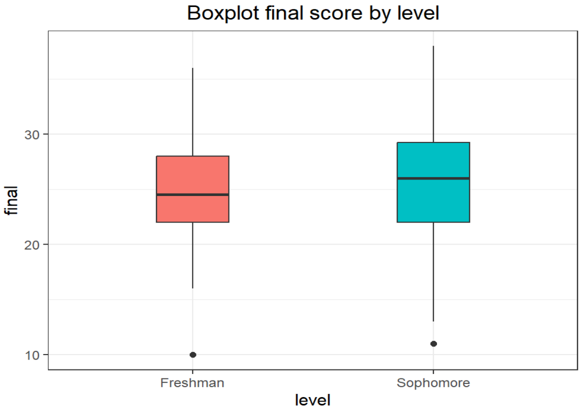




The chart above indicates that students who score “Good” in the final exam tend to have a high GPA during the term, with cumulative GPA scores ranging from 2.5 to 4. This suggests that the abilities of these students are reflected in their final scores. Additionally, students who achieve a “Medium” score at the end of the term have a broader and more diverse score distribution. Although a few students in this group have a relatively low GPA in the semester or cumulatively, a significant percentage still falls within the average and good range. The “Below medium” group consists of a small number of students but exhibits a high degree of dispersion. Interestingly, some students in this group still achieve GPAs between 3 and 3.5 in both categories. This suggests that these students may have made mistakes or encountered other negative factors that resulted in a low score on the final test. Hence, it is not possible to solely classify students’ learning abilities based on their final exam scores.

Hypothesis Testing

Level & Final score



H_0: The mean final score of freshman greater or equal to that of sophomore.

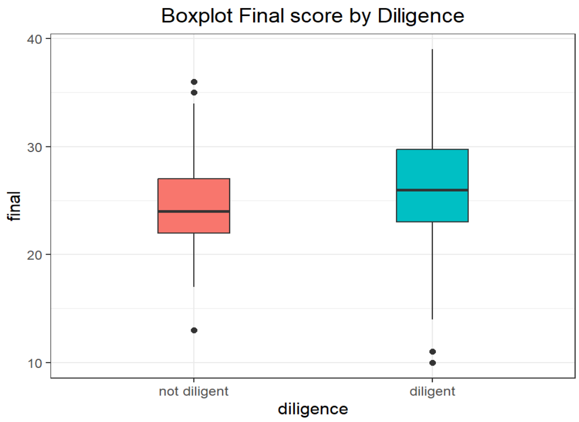
H_α: The mean final score of freshman less than that of sophomore.

Hypothesis Testing: Using t.test function to test whether ***The mean final score of freshman less than the mean final score of sophomore.***

```
    Welch Two Sample t-test

data:  final by level
t = -2.0708, df = 322.34, p-value = 0.01958
alternative hypothesis: true difference in means between group Freshman
and group Sophomore is less than 0
95 percent confidence interval:
       -Inf -0.1768979
sample estimates:
 mean in group Freshman mean in group Sophomore 
               25.13291                26.00255 
```

Since p-value = 0.01958 which is less than 0.05. Thus, we reject the null hypothesis H_0 at significant level α = 5%. Therefore, there is sufficient evidence to conclude that ***The mean final score of freshman less than the mean final score of sophomore.***

Diligence & Final score



Hypothesis Testing: Using t.test function to test whether ***The mean final score of not diligent student is less than to the mean final score of diligent student.***

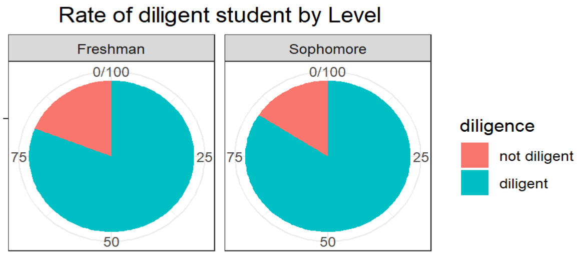
H_0: The mean final score of not diligent student less than that of diligent student.

H_α: The mean final score of not diligent student is less than that of diligent student.

```
    Welch Two Sample t-test
data:  final by diligence
t = -3.2189, df = 184.22, p-value = 0.0007603
alternative hypothesis: true difference in means between group not
diligent and group diligent is less than 0
95 percent confidence interval:
       -Inf -0.6891063
sample estimates:
mean in group not diligent     mean in group diligent
                  24.72034                   26.13701
```

Since p-value = 0.0007603 which is extremely less than 0.05. Thus, we reject the null hypothesis H_0 at significant level α = 5%. Therefore, there is sufficient evidence to conclude that ***The mean final score of not diligent student is less than to the mean final score of diligent student.***

Diligence & Level



Hypothesis Testing: Using prob.test function to test whether ***The rate of diligent student in first year different from the rate of diligent student in second year.***

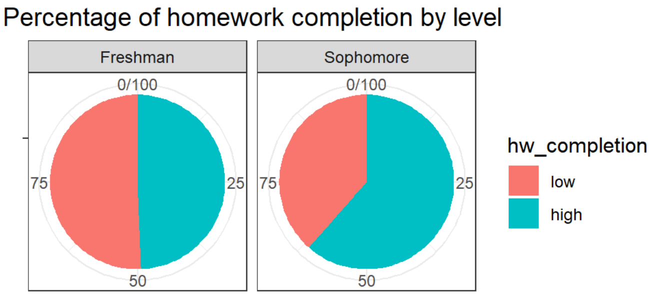
H_0: The rate of diligent student in first year equal to the rate of diligent student in second year.

H_α: The rate of diligent student in first year different from the rate of diligent student in second year.

```
            diligent not diligent
  Freshman       128           30
  Sophomore      330           62


    2-sample test for equality of proportions without continuity
correction

data:  table_level_diligence
X-squared = 0.81292, df = 1, p-value = 0.3673
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.10273596  0.03931566
sample estimates:
   prop 1    prop 2
0.8101266 0.8418367
```

Since p-value = 0.3673 which is greater than 0.05. Thus, we do not reject the null hypothesis H_0 at significant level α = 5%. Therefore, there is ***not sufficient evidence*** to conclude that ***the rate of diligent student in first year different from the rate of diligent student in second year.***

Level & Homework Completion Level



Hypothesis Testing: Use prob.test function to test whether the percentage of sophomores doing homework is greater than the percentage of freshman does.

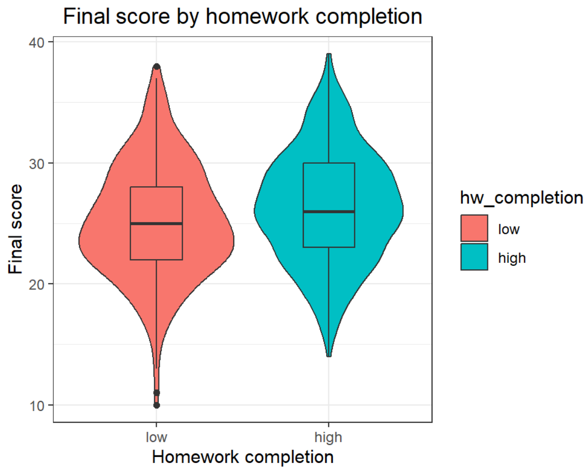
H_0: The percentage of sophomores doing homework is less than or equal to the percentage of freshman does.

H_α: The percentage of sophomores submit homework is greater than the percentage of freshman does.

```
            high low
  Sophomore  241 151
  Freshman    78  80


    2-sample test for equality of proportions without continuity
correction

data:  table_level_hwcomplete
X-squared = 6.7822, df = 1, p-value = 0.004604
alternative hypothesis: greater
95 percent confidence interval:
 0.04421757 1.00000000
sample estimates:
   prop 1    prop 2 
0.6147959 0.4936709
```

Since p-value = 0.001584 which is extremely small. We reject the null hypothesis H_0 at significant level α = 5%. Therefore, there is ***sufficient evidence*** to conclude that ***the percentage of sophomores submit their homework is greater than the percentage of freshman does.***

Homework Completion Level & Final



Hypothesis Testing: Using t.test function to test whether ***The average final grade of a student who consistently submits their homework is higher than that of a student who does not.***

H_0: The mean final score of a student who highly submits their homework less than or equal to that of a student who does not.

H_α: The mean final score of a student who highly submits their homework greater than that of a student who does not.

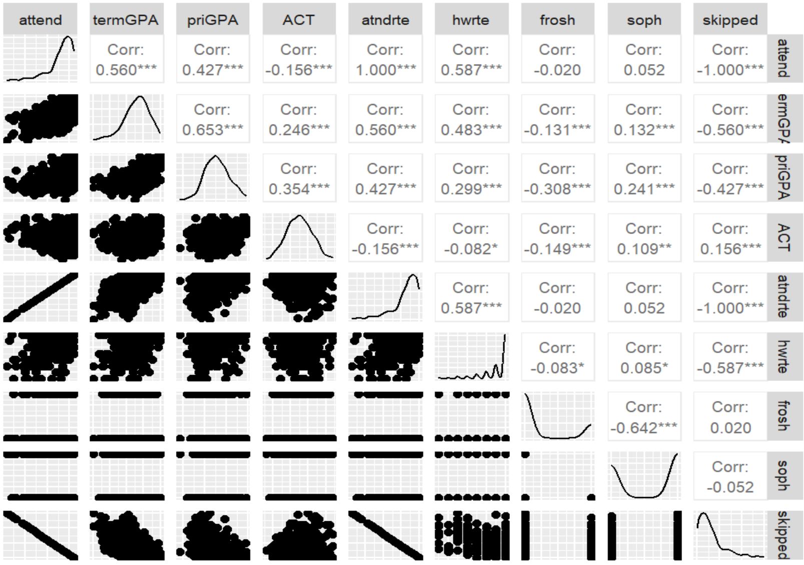
```
    Welch Two Sample t-test

data:  final by hw_completion
```

```
t = -3.3815, df = 600.48, p-value = 0.0003839
alternative hypothesis: true difference in means between group low and
group high is less than 0
95 percent confidence interval:
       -Inf -0.6328298
sample estimates:
 mean in group low mean in group high
          25.17254           26.40657
```

Since p-value = 0.000426 which is extremely small and less than 0.05. Thus, we reject the null hypothesis H_0 at significant level α = 5% or even at α = 1%. Therefore, there is ***sufficient evidence*** to conclude that ***The mean final score of a student who highly submits their homework greater than that of a student who does not.***

Regression Model

Multicolinear



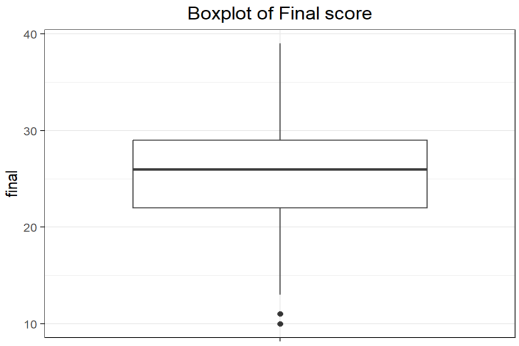
From figure, we observe that there are strong correlation between Percent classes attended (atndrte) and class skipped (skipped) and class attendance (attend).

Data preparation

The response in our linear model is final score exam which is **final** attribute.

As a result of descriptive statistics, we should not include the attribute Percentage classes attended (atndrte) and Percentage homework turned in (hwrte).

Therefore, we start to build multiple linear regression model from these predictors: termGPA, priGPA, ACT, dummy variables (frosh, soph).



There are 2 outliers but this number is quite small so we omit it.

New dimension of attend are 678 instances and 8 attributes.

Instead of normalizing the final score (**final**), we transform it into a 4-point scale by dividing it by 10.

Split data into train and validation sets with ratio 80% - 20%, respectively.

The dimension of attend train data

```
542    8
```

The dimension of attend validation data

```
136    8
```

Modeling

Model 1: we build models using variables with descriptive statistics insight.

The predictors are: termGPA, priGPA, ACT

The summary of model 1

```
Call:
lm(formula = final ~ termGPA + priGPA + ACT, data = attend_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.15374 -0.25546  0.00626  0.25218  1.06297 



Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.147272   0.120002   9.560  < 2e-16 ***
termGPA      0.279189   0.029197   9.562  < 2e-16 ***
priGPA      -0.020191   0.041476  -0.487    0.627    
ACT          0.034447   0.005092   6.765 3.49e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3837 on 538 degrees of freedom
```

```
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2962 
F-statistic: 76.89 on 3 and 538 DF,  p-value: < 2.2e-16
```

Based on the result of summary function:

- The p-value of F-test is less than 2.2e-16 which is less than significant level α = 5%. This means there at least 1 variable can estimate the final score. In the other words, the model can be used to estimate the final score.

- The p-value of t test in termGPA and ACT are less than significant level α = 5%. So we can conclude that these coefficients different from 0 except priGPA. Hence, attributes termGPA, ACT has a significant effect on final score.

- There is a negative correlation between final score (outcome) and the priGPA. There may some effect on this due to the cumulative GPA to term (priGPA) of freshman is not significant. So we construct a different model that consider the interaction between the student's level and priGPA.

Model 2: Replace variable priGPA in Model 1 by the variables that account for interaction between student's level and priGPA.

```

Call:
lm(formula = final ~ termGPA + ACT + I(priGPA * frosh) + I(priGPA * 
    soph), data = attend_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.12906 -0.26473 -0.00057  0.25446  1.05117 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.141277   0.118937   9.596  < 2e-16 ***
termGPA            0.284547   0.023562  12.076  < 2e-16 ***
ACT                0.034657   0.004968   6.975 9.01e-12 ***
I(priGPA * frosh) -0.023614   0.020931  -1.128   0.2597    
I(priGPA * soph)  -0.033927   0.015188  -2.234   0.0259 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3823 on 537 degrees of freedom
Multiple R-squared:  0.3063,    Adjusted R-squared:  0.3011 
F-statistic: 59.27 on 4 and 537 DF,  p-value: < 2.2e-16
```

We test if coefficient of priGPA*frosh is equal to 0 using significant level α = 5%. From the output of summary function, the p-value of t.test is 0.2597 which is higher than 0.05. Thus, we fail to reject the null hypothesis H_0.

Build model 2 again without predictor priGPA*frosh.

The summary of Model 2

```

Call:
```

```
lm(formula = final ~ termGPA + ACT + I(priGPA * soph), data = 
attend_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.12831 -0.27873  0.00077  0.25589  1.05402 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.111885   0.116078   9.579  < 2e-16 ***
termGPA           0.281566   0.023420  12.023  < 2e-16 ***
ACT               0.035087   0.004955   7.081  4.5e-12 ***
I(priGPA * soph) -0.024030   0.012402  -1.938   0.0532 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3824 on 538 degrees of freedom
Multiple R-squared:  0.3046,    Adjusted R-squared:  0.3008 
F-statistic: 78.56 on 3 and 538 DF,  p-value: < 2.2e-16
```

Based on the result of summary function:

- The p-value of F-test is less than 2.2e-16 which is less than significant level α = 5%. This means there at least 1 variable can estimate the final score. In the other words, the model can be used to estimate the final score.
- The p-value of t test in termGPA and ACT are less than significant level α = 5%. So we can conclude that coefficients of termGPA and ACT different from 0.
- However, the one of priGPA*soph is 0.0532 which is higher than 0.05. Hence, the coefficient of priGPA*soph may be equal to 0.

Model 3: Using stepwise algorithm for both direction (BIC criteria) on **Model 1**.

The summary of Model 3

```
Call:
lm(formula = final ~ termGPA + ACT, data = attend_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.15610 -0.25305  0.00801  0.24967  1.05350 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.132269   0.115895   9.770  < 2e-16 ***
termGPA     0.270285   0.022742  11.885  < 2e-16 ***
ACT         0.033822   0.004924   6.868  1.8e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3834 on 539 degrees of freedom
Multiple R-squared:  0.2998,    Adjusted R-squared:  0.2972 
F-statistic: 115.4 on 2 and 539 DF,  p-value: < 2.2e-16
```

Based on the result of summary function:

- The p-value of F-test is less than 2.2e-16 which is less than significant level α = 5%. This means there at least 1 variable can estimate the final score. In the other words, the model can be used to estimate the final score.
- The p-value of t test of all coefficients are less than significant level α = 5%. So we can conclude that all coefficients different from 0.

Comparing model 1, model 2 and model 3

```
  Model      AIC      BIC NumPredictors                        Predictors
1     1 505.6465 527.1229             3           termGPA, priGPA, ACT
2     2 502.1160 523.5923             3 termGPA, ACT, I(priGPA * soph)
3     3 503.8852 521.0663             2                     termGPA, ACT
```

As we see, the model 3 has lowest number of predictors but has the smallest BIC. Howerver, the difference between BIC values is not much significant. We will test if we could use the model 3.

The hypothesis test (model 1 and model 3):

H_0: coefficient of priGPA= 0

H_α: coefficient of priGPA$\neq$ 0

Hypothesis testing: we use **anova** function to test whether the reduced model can be used

```
Analysis of Variance Table

Model 1: final ~ termGPA + priGPA + ACT
Model 2: final ~ termGPA + ACT
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    538 79.191
2    539 79.226 -1 -0.034883 0.237 0.6266
```

Since p-value = 0.6266 which is larger than significant level, 0.05. Thus, we fail to reject the null hypothesis H_0. In other words, we can conclude that removing variable priGPA does not affect much the fit of models.

The hypothesis test (model 2 and model 3):

H_0: coefficient of priGPA*soph= 0

H_α: coefficient of priGPA*soph$\neq$ 0

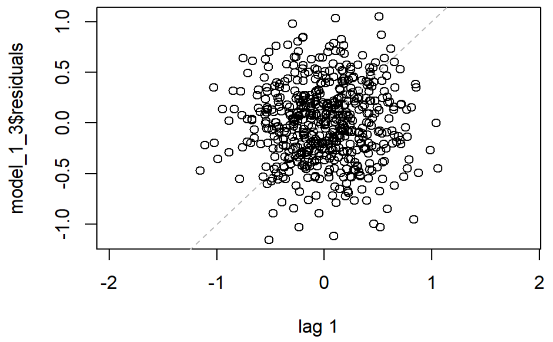
Hypothesis testing: we use **anova** function to test whether the reduced model can be used

```
Analysis of Variance Table

Model 1: final ~ termGPA + ACT + I(priGPA * soph)
Model 2: final ~ termGPA + ACT
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    538 78.677
```

```
2    539 79.226 -1  -0.54905 3.7544 0.05319 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value = 0.05319 which is larger than significant level, 0.05. Thus, we fail to reject the null hypothesis H_0. In other words, we can conclude that removing variable priGPA*soph does not affect much the fit of models.

Independence testing



H_0: There is no correlation among the residuals.

H_α: The residuals are autocorrelated.

```
 lag Autocorrelation D-W Statistic p-value
   1    -0.005041313      2.003933   0.946
 Alternative hypothesis: rho != 0
```

The D-W test statistic is 2.003933 which lies between the between 1.5 and 2.5 and p-value is 0.946. ***As a result, autocorrelation is probably not a cause for concern.***

Stability testing

H_0: Homoscedasticity is present (the residuals are distributed with equal variance).

H_α: Heteroscedasticity is present (the residuals are not distributed with equal variance).

```

	studentized Breusch-Pagan test

data:  model_1_3
BP = 1.5008, df = 2, p-value = 0.4722
```

The test statistic for the studentized Breusch-Pagan test is 1.5008, with 2 degrees of freedom. To make more sense, the p-value of the test is 0.4722, which is greater than the commonly used significance level $\alpha = 0.05$. Thus, we fail to reject the null hypothesis H_0 that homoscedasticity is present. In other words, there is **enough evidence** to conclude that ***the residuals of regression model are distributed with equal variance.***

Normality testing

H_0: The residuals of regression model is normally distributed.

H_α: The residuals of regression model is not normally distributed.

```

	Shapiro-Wilk normality test

data:  model_1_3$residuals
W = 0.99813, p-value = 0.8267
```

The test statistic for the Shapiro-Wilk normality test is 0.99813. And the p-value of the test is 0.8267, which is much greater than the commonly used significance level of 0.05. Thus, we fail to reject H_0. In other words, there is ***enough evidence*** to conclude that ***The residuals of regression model is normally distributed.***

Model expression

Based on independence, stability and normality testing, we accept model 3 for estimate the final score

The equation:

$$\widehat{final} = (0.270285 \times termGPA + 0.033822 \times ACT + 1.132269) * 10$$

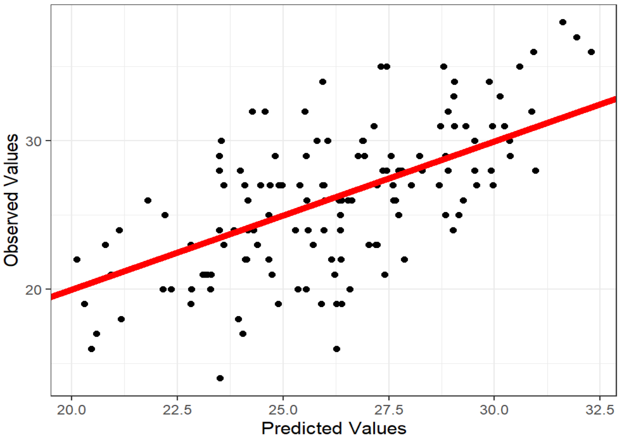
The fitted line reveals that:

- If the termGPA, ACT, are all 0, the maximum final score is 11.323.
- If the termGPA is increased by 1 unit and the other preidcot remain constanst, the final score is increased by 2.756 unit
- If the ACT is increased by 1 unit, the final score is increased by 0.338 unit

Prediction

We will do predict on the validation test and the Root mean square error is:

```
rmse = 3.783746
```



Activity 2

About the dataset

This dataset contains insightful information related to insurance claims, giving us an in-depth look into the demographic patterns of those receiving them. By analyzing these key factors across geographical areas and across different demographics such as age or gender we can gain a greater understanding of who is most likely to receive an insurance claim. This understanding gives us valuable insight that can be used to inform our decision making when considering potential customers for our services. On a broader scale it can inform public policy by allowing for more targeted support for those who are most in need and vulnerable. These kinds of insights are extremely valuable and this dataset provides us with the tools we need to uncover them.

The dataset has 11 columns and 1340 observations

1. index: surrogate key index
2. PatientID: ID of patient
3. age: age of patient
4. gender: gender of patient
5. bmi: the body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy
6. bloodpressure: percent classes attended
7. diabetic: whether the insured person is diabetic or not
8. children: number of children of the insured person
9. smoker: whether the insured person is a smoker or not.
10. region: region where patient live
11. claim: amount of the insurance claim

Goal: Identifying trends in insurance claims based on age, gender, BMI, and blood pressure. Investigating correlations between health traits with the likelihood of making a claim.

Reference link: Insurance Claim Analysis: Demographic and Health

Cleaning datasets

Consider about the type of attributes:

```
'data.frame':   1340 obs. of  11 variables:
 $ index        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ PatientID    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age          : num  39 24 NA NA NA NA NA 19 20 30 ...
 $ gender       : chr  "male" "male" "male" "male" ...
 $ bmi          : num  23.2 30.1 33.3 33.7 34.1 34.4 37.3 41.1 43 53.1 ...
 $ bloodpressure: int  91 87 82 80 100 96 86 100 86 97 ...
 $ diabetic     : chr  "Yes" "No" "Yes" "No" ...
 $ children     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ smoker       : chr  "No" "No" "No" "No" ...
 $ region       : chr  "southeast" "southeast" "southeast" "northwest" ...
 $ claim        : num  1122 1132 1136 1136 1137 ...

  index PatientID age gender  bmi bloodpressure diabetic children smoker
1     2         3  NA   male 33.3            82      Yes        0     No
2     3         4  NA   male 33.7            80       No        0     No
3     4         5  NA   male 34.1           100       No        0     No
4     5         6  NA   male 34.4            96      Yes        0     No
5     6         7  NA   male 37.3            86      Yes        0     No
6    13        14  32   male 27.6           100       No        0     No
7    14        15  40   male 28.7            81      Yes        0     No
8    15        16  32   male 30.4            86      Yes        0     No
     region   claim
1 southeast 1135.94
2 northwest 1136.40
3 northwest 1137.01
4 northwest 1137.47
5 northwest 1141.45
6      <NA> 1252.41
7      <NA> 1253.94
8      <NA> 1256.30
```

Omit the row has NA value in **region** attribute

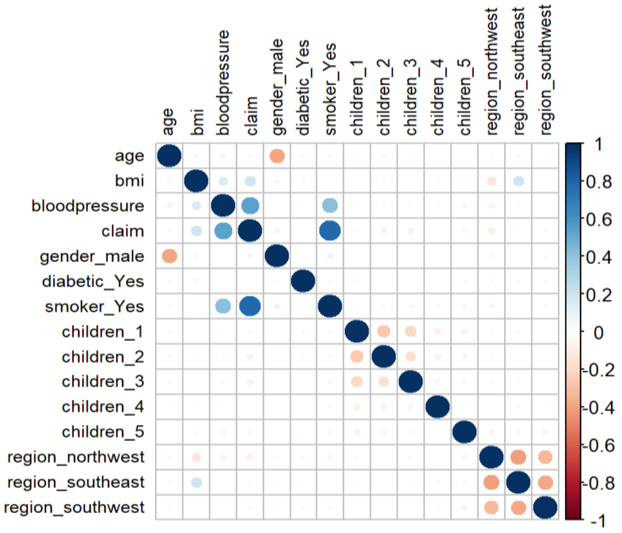
Replace the NA value in age attribute by mean

Verify the data

Remove the index and patientID from data

Encode categorical variable by factor function

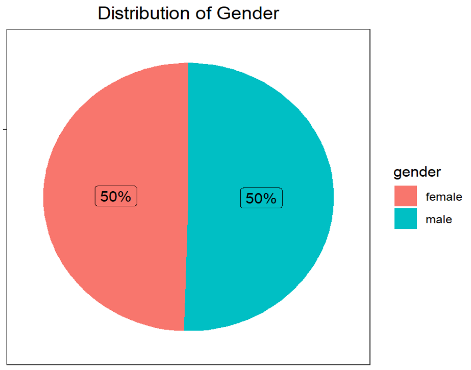
Correlation

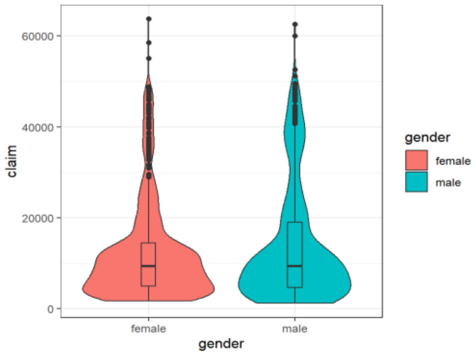


According to figure, the **claim** is strongly affected by the **smoke status**, **bloodpressure** and **bmi**.

Descriptive statistics

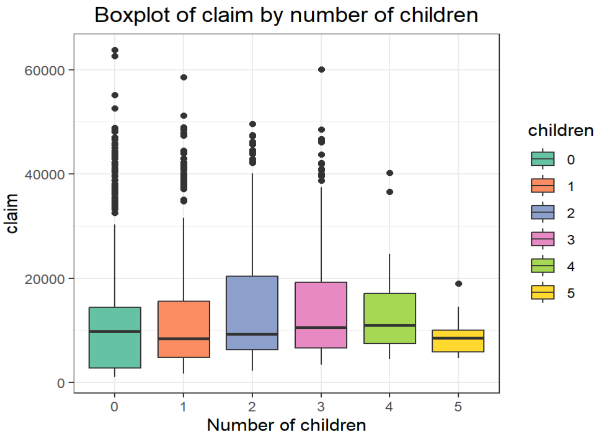
Gender





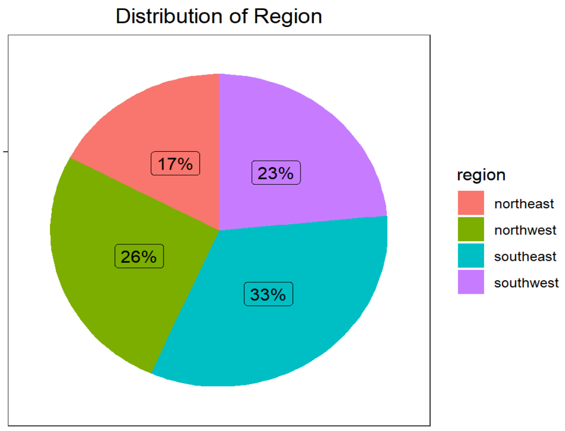
According to the boxplot, both genders have the same average claim amount. Additionally, the boxplot for both genders has the same distribution with two peaks, one at around $10,000 and the other at nearly $40,000. However, the female gender group has a higher number of outliers, indicating that there are more cases of unusually high indemnity insurance in this group.

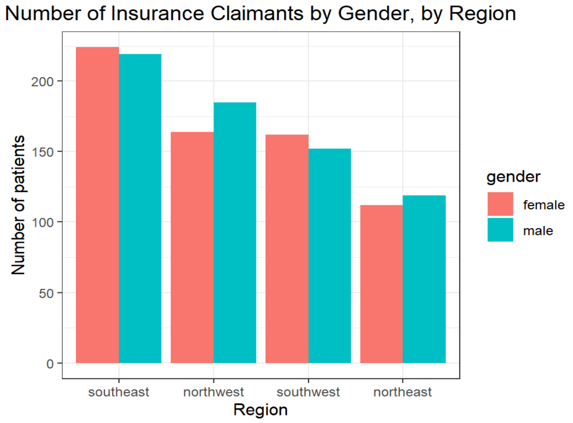
Number of children



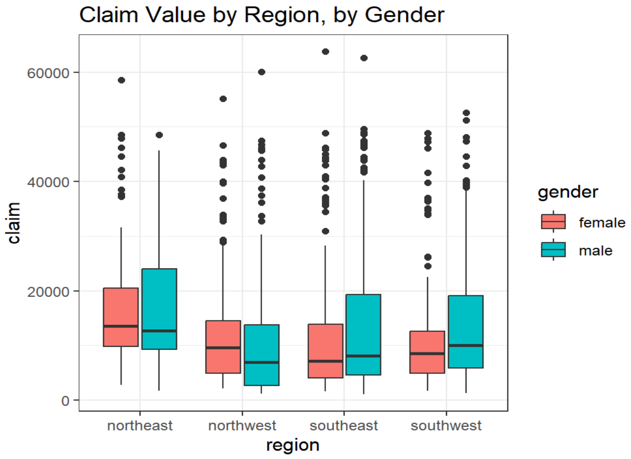
The chart shows that families with 2 to 3 children have the widest distribution, ranging from about 8000 to nearly 20000, showing that the amount of insurance claims received is higher than that of other families. On the other hand, families with 4 to 5 children have a narrower distribution, showing that these two groups have the lowest frequency of receiving insurance claims. In particular, family groups have a very low concentration level, showing that the number of insurance claims in these groups is not high, most of them are concentrated at less than 20000. However, the boxplot chart of groups with 0 and 1 children have more outliers than other groups. This shows that families with no children or only one child have relatively large fluctuations in the amount of insurance claims, with some cases having unusually high claims.

Region



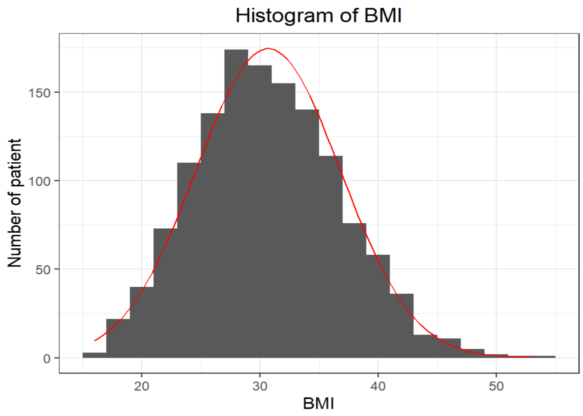


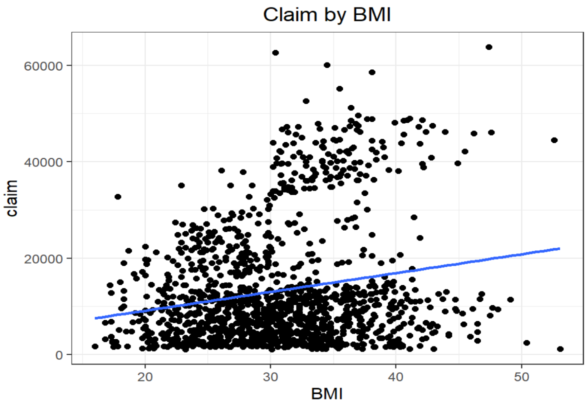
Patients in the Southeast have significantly more claims than any other region.



The plot reveals that claim median value lies in the rang of around 10,000-15,000 for all the regions, for both the genders. Moreover, the claim value outliers are rampant for all the regions, for both the genders

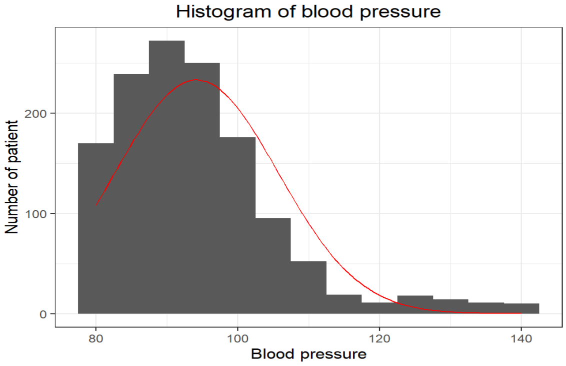
BMI

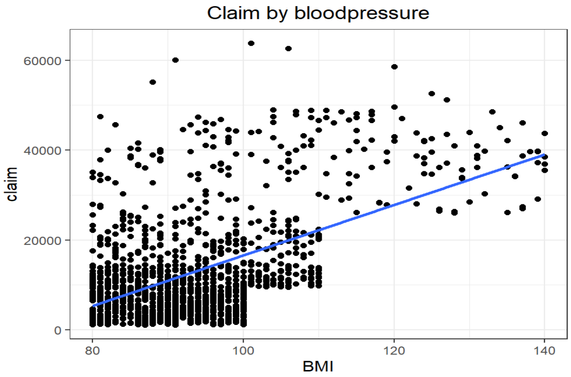




The bmi attribute has a typical normal distribution, with a tendency to concentrate more from 25 to 35. The claim seems to be not linear depending on Bmi. In fact, the graph depicts 2 cluster.

Blood pressure



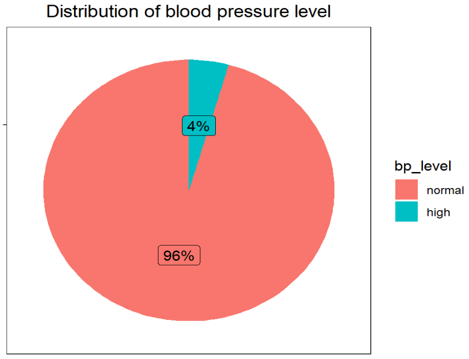


Blood pressure chart has right skewed, showing that the majority of patient have a normal blood pressure value, between 80 and 100. However, there are a few people in the pre-hypertension group when this index falls between 120 and 140. The scatter plot shows that blood pressure readings seem to have a weak and positive correlation with the amount of insurance claims. In the normal blood pressure group, this value usually falls between 0 and 20000 and there are some outliers with a much higher number of claims than the majority. Although the number of patients in the pre-hypertension group is not much, all claims are above 20000.

According American Heart Association, the blood pressure can be categorized as follow:

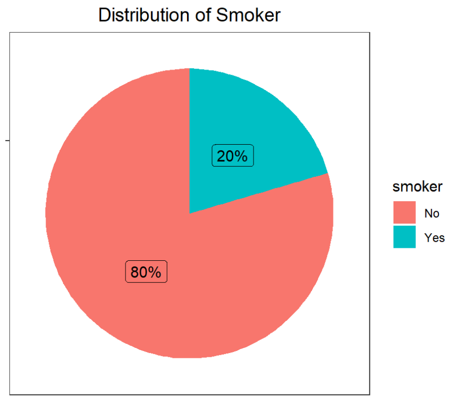
- normal: blood pressure less than 120
- high: blood pressure lies greater or equal to 120

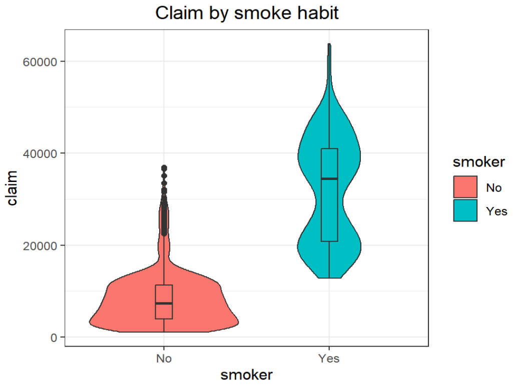
Let’s construct a attribute for blood pressure level called **bp_level**:



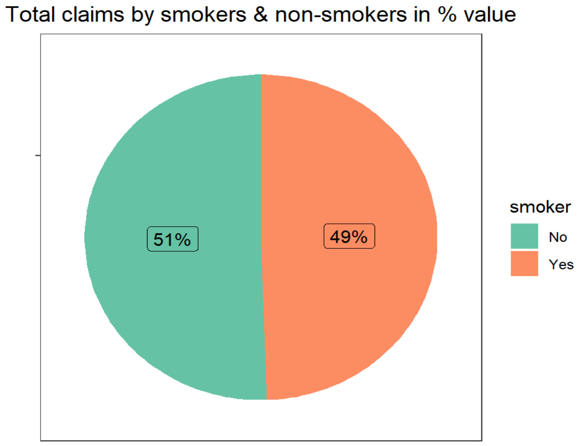
The number of patients who have high blood pressure is quite small respect to the total patients.

Smoker



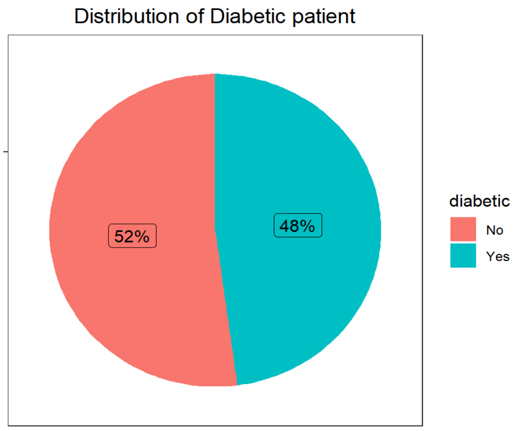


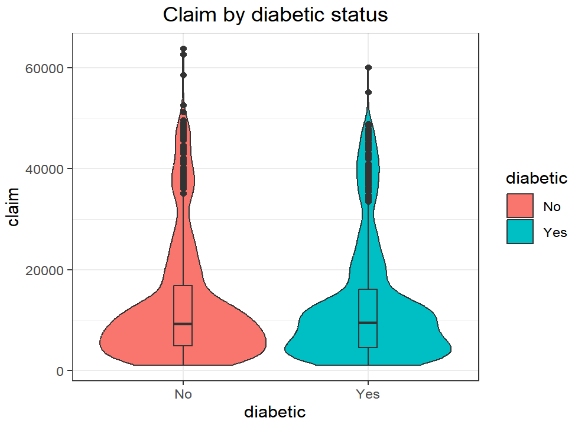
It can be seen that although the number of smokers in the data set is not high, the claim concentration of this group is much higher than that of the non-smoker group, most of which are in the range from 20000 to 40000. The non-smoker group has lower concentration and shorter allocation path length. In addition, the appearance of many outliers with a high number of claims indicates, showing that there is an instability in the number of claim receipts in the non-smoker group. This shows that a smoker tend to receive more claims than a non-smokers.



It can be confirmed that although the number in the data set is only 20%, the smoker group tends to receive more claims since the total claims received account for nearly 1/2 of the total. The reason might be the smoker group temp to face health problems more than the other.

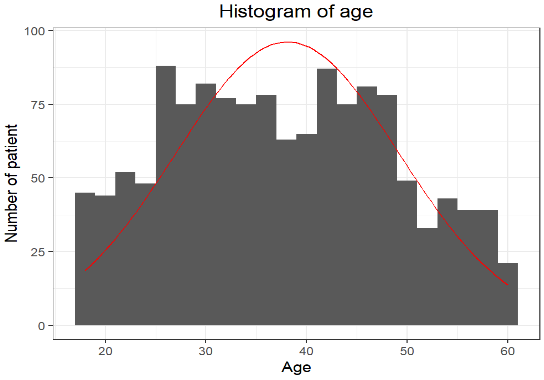
Diabetic

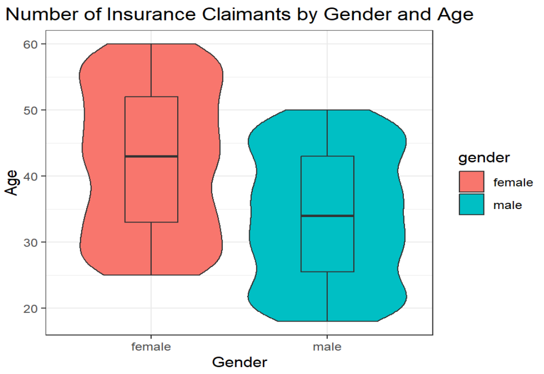




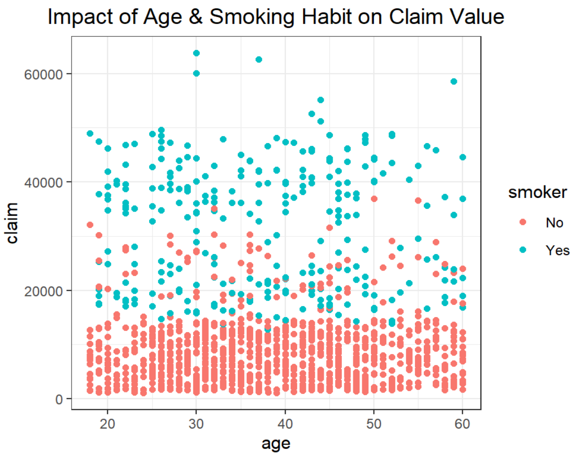
The rate of people having insurance with diabetic and non diabectic status is the same. From the boxplots, there are several cases that has an unsually claim which roundly 30000 to 50000. Moreover, the average claim of both group share the same number with nearly 10000. In advance, considering the distribution between the two groups in the boxplot, it seems that there is not much correlation between the amount of claim and diabetic status, or in other words, being diabetic does not affect much the amount of claims.

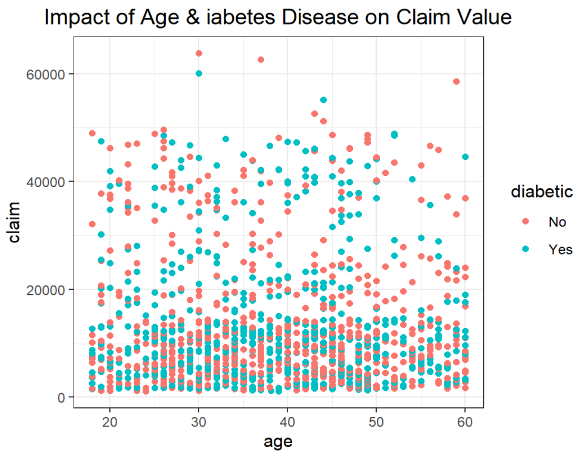
Age





The graph shows that there is a diverse distribution of ages in the data set. The boxplot shows age of female insurance claimants is higher, has a higher median than males.

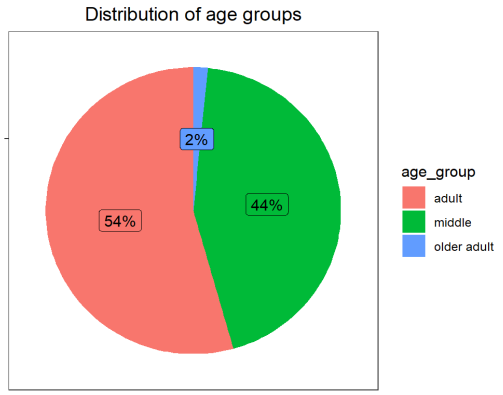


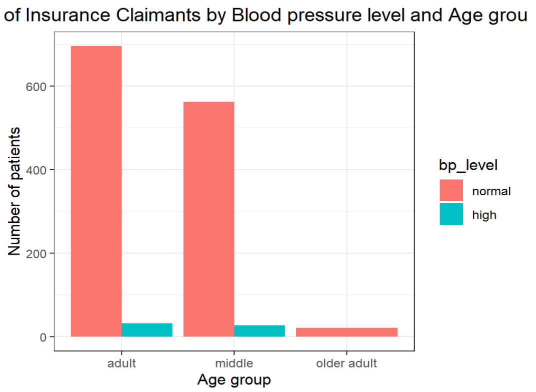


The plot reveals that claim value is typiclly high for people with smoking habit. Moreover, there is no significant correlation between claim value and prevalence of diabetes.

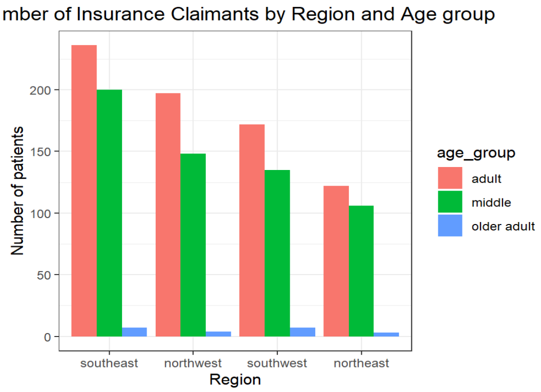
We create a attribute for age group called age_group. Since the patients in the data has age greater or equal to 18 years old, we split into 3 groups:

- adult: age lies between 18 and 39
- middle: age lies between 40 and 59
- older adults: age greater than 60



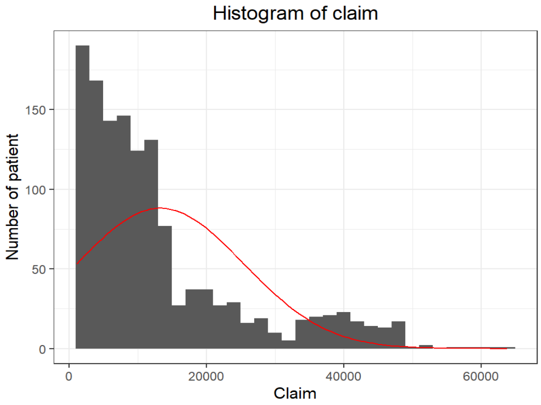


The rate of patients who are older adult is much smaller compare to the other two groups. Most of patients are adult. In the adult and middle age groups, the number of patients with high blood pressure is almost equal. However, there are no patients with high blood pressure among the older adult population.



According to the plot, most of patients or insurance claimants are in adult group. While the opposite pattern is true for older adult.

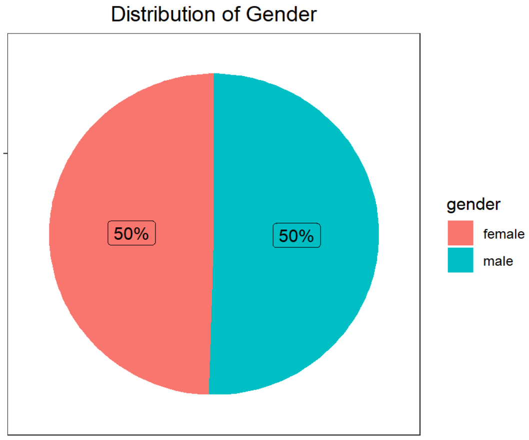
Claim



The histogram of claim has right skewed shows that the majority of patients have a claim number below 20000. However, there are still many cases where this number is relatively high due to the characteristics of special patient groups, ranging from 20000 to 50000.

Hypothesis testing

Gender



H_0: The rate of male and female patients are equal.

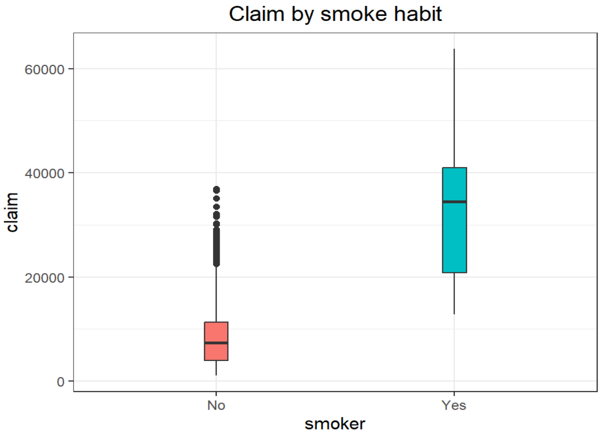
H_{α}: The rate of male and female patients are different.

Hypothesis Testing: Using prop.test function to test whether ***The rate of male and female patients are equal.***

```
female   male
   662    675


    1-sample proportions test without continuity correction

data:  table_gender, null probability 0.5
X-squared = 0.1264, df = 1, p-value = 0.7222
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4683909 0.5219137
sample estimates:
        p
0.4951384
```

Since p-value = 0.7222 which is greater than 0.05. Thus, we fail to reject the null hypothesis H_0 at significant level α = 5%. Therefore, there is not sufficient evidence to conclude that The rate of male and female patients are different. In other words, we accept ***the rate of male and female patients share the same number.***

Smoker & claim



H_0: The average claim of patient who is non smoker higher than or equal to that of patient who is smoker

H_{α}: The average claim of patient who is non smoker less than that of patient who is smoker.

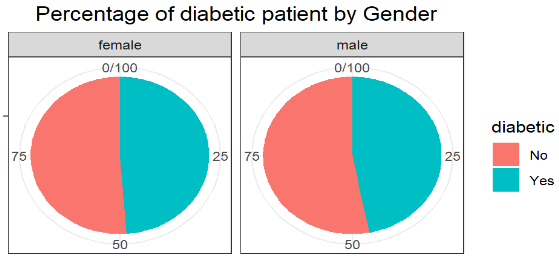
Hypothesis Testing: Using t.test function to test whether ***The average claim of patient who is non smoker less than that of patient who is smoker***

```
    Welch Two Sample t-test

data:  claim by smoker
```

```
t = -32.742, df = 311.87, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and
group Yes is less than 0
95 percent confidence interval:
     -Inf -22419.3
sample estimates:
 mean in group No mean in group Yes
         8441.348         32050.232
```

Since p-value < 2.2e-16 which is extremely small and less than 0.05. Thus we reject the null hypothesis H_0 at significant level $\alpha = 5\%$. Therefore, there is sufficient evidence to conclude that ***The average claim of patient who is non smoker less than that of patient who is smoker.***

Diabetic & gender



H_0: The rate of diabetic in female patient equal to the rate of diabetic in male patient

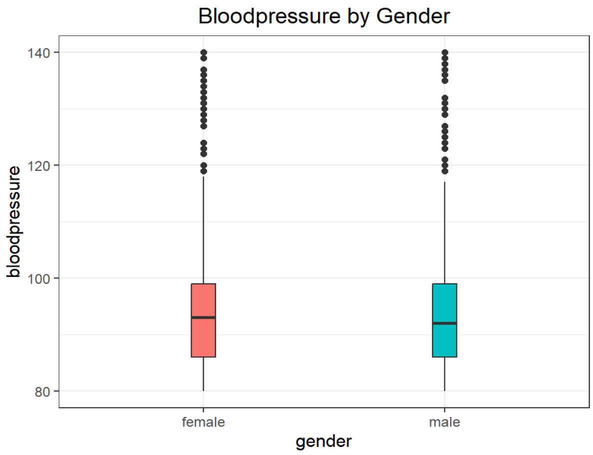
H_α: The rate of diabetic in female patient different from the rate of diabetic in male patient

Hypothesis Testing: Using prop.test function to test whether ***The rate of diabetic in female patient greater the rate of diabetic in male patient***

```
          No Yes
  female 339 323
  male   358 317


    2-sample test for equality of proportions without continuity
correction

data:  table_gender_diabetic
X-squared = 0.44782, df = 1, p-value = 0.5034
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07183359  0.03526203
sample estimates:
   prop 1    prop 2
0.5120846 0.5303704
```

We observes the p-value = 0.5034 which is larger than 0.05. Thus, we fail to reject the null hypothesis H_0. In the other words, we accept that there is no difference between the rate of diabetic patient in two gender.

Bloodpressure & gender



H_0: The mean of blood pressure in female patient equal to that in male patient.

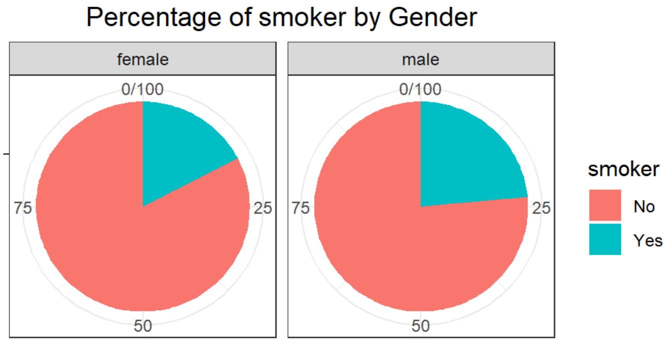
H_α:The mean of blood pressure in female patient not equal to that in male patient.

Hypothesis Testing: Using t.test function to test whether ***The mean of blood pressure in female patient not equal to that in male patient.***

```
    Welch Two Sample t-test

data:  bloodpressure by gender
t = -0.49213, df = 1334.4, p-value = 0.6227
alternative hypothesis: true difference in means between group female and
group male is not equal to 0
95 percent confidence interval:
 -1.5351837  0.9194111
sample estimates:
mean in group female   mean in group male 
            94.01360             94.32148 
```

From the test, we observe that p-value = 0.6227 > 0.05. Thus, we fail to reject null hypothesis H_0 at significant level α = 5%. In other words, we accept that there is no difference of the average bloodpressure between two gender.

Smoker & gender



H_0: The rate of smoker in female patient equal to that in male patient

H_α: The rate of smoker in female patient different from that in male patient

Hypothesis Testing: Using prop.test function to test whether ***The rate of smoker in female patient different from that in male patient***

```
         Yes  No
  female 115 547
  male   159 516


    2-sample test for equality of proportions without continuity
correction

data:  table_gender_diabetic
X-squared = 0.44782, df = 1, p-value = 0.5034
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07183359  0.03526203
sample estimates:
   prop 1    prop 2 
0.5120846 0.5303704
```

Since p-value=0.5034 which is greater than common significant level α = 5%. Thus, we fail to reject the null hypothesis H_0 at significant level α = 5%. In other words, there is no difference in percentage of smoker in two gender.

We will test whether there is a relationship between gender and smoking status.

H_0: The attirubtes smoker and gender are independent.

H_α: The attirubtes smoker and gender are not independent.

Hypothesis Testing: Using chisq.test function to test whether ***The attirubtes smoker and gender are not independent.***

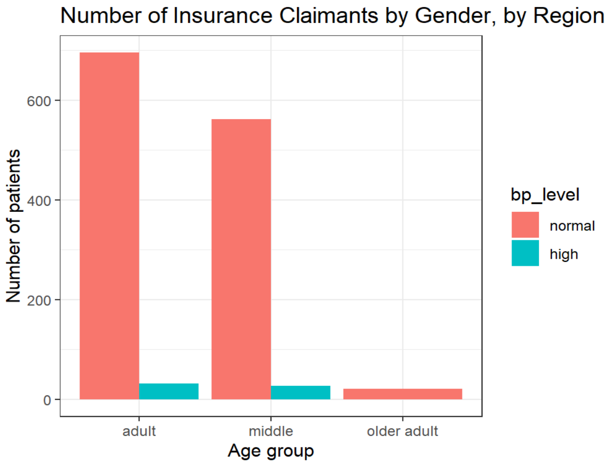
```
         Yes  No
```

```
  female 115 547
  male   159 516


    Pearson's Chi-squared test with Yates' continuity correction

data:  table_gender_smoker
X-squared = 7.4691, df = 1, p-value = 0.006277
```

Due to the fact that the p-value of 0.006277 is below the significant level of 5%. The null hypothesis H_0 is rejected. As a result, the test reveals a relationship between gender and smoking status.

Blood pressure level & age group

We will investigate whether age and blood pressure are related.



H_0: The attirubtes age and blood pressure are independent.

H_α: The attirubtes age and blood pressure are not independent.

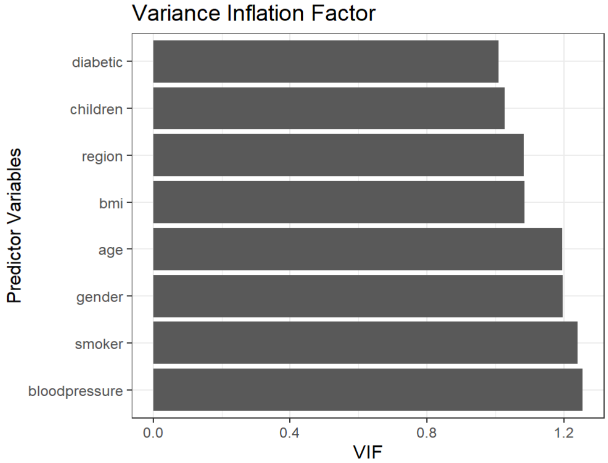
Hypothesis Testing: Using chisq.test function to test whether ***The attirubtes smoker and gender are not independent.***

```

              normal high
  adult          695   32
  middle         562   27
  older adult     21    0


    Pearson's Chi-squared test

data:  table_age_bp
X-squared = 1.0106, df = 2, p-value = 0.6033
```

Due to the fact that the p-value of 0.6033 is higher than the significant level of 5%. We fail to reject the null hypothesis H_0. As a result, the test shows that age and blood pressure are not related.

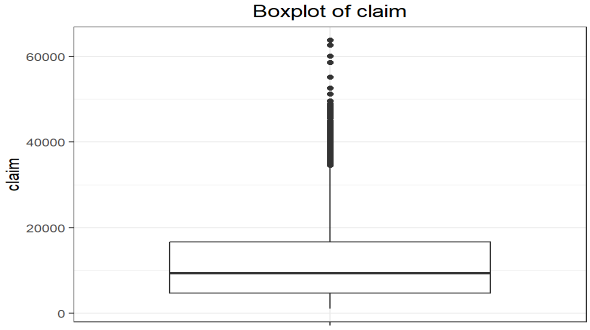
Regression model

Multicolinear



Except for the region_southeast, all predictors have a VIF that is less than 2. The VIF of region_southeast, however, slightly exceeds 2. As a consequence, we can conclude that there is no strong correlation occur between predictors.

Data preparation



The dimension of data before remove outliers:

```
1337      9
```

The dimension of data after remove outliers:

```
1198      9
```

Split data to train and validation sets with ratio 80% - 20%

The dimension of train insurance data

```
958    9
```

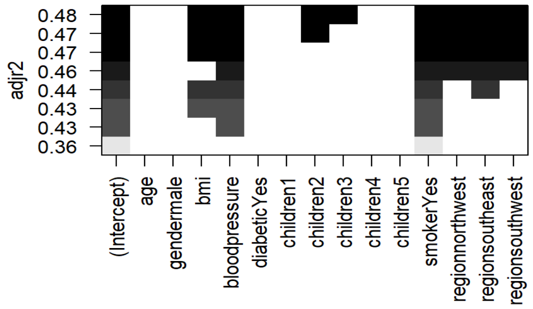
The dimension of validation insurance data

```
240    9
```

Variables selection

Best subsets regression

Using the scale of Adjust R-squared



From the figure, we observe that there are 5 models share the same value of Adjust R-squared which is 0.7. Regarding the same Adjust R-squared, adding more variable does not increase value of Adjust R-squared. Thus, we might favor the one with fewer variables.

As a result, we select the following attributes for our regression model: bmi, bloodpressure, smoker, region.

Stepwise regression

Based on AIC score to add or remove variables.

```

Call:
lm(formula = claim ~ bmi + bloodpressure + children + smoker + 
    region + bp_level, data = insurance_data)

Residuals:
   Min     1Q Median     3Q    Max 
-15857  -4270  -1034   3298  32837 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -17598.34    2195.52  -8.016 2.39e-15 ***
bmi                342.60      30.49  11.238  < 2e-16 ***
bloodpressure      182.28      22.26   8.189 6.14e-16 ***
children1          582.85     456.11   1.278 0.201510    
children2         1955.85     504.61   3.876 0.000111 ***
children3         2173.13     591.77   3.672 0.000250 ***
children4         3077.69    1341.41   2.294 0.021925 *  
children5         1112.67    1577.20   0.705 0.480642    
smokerYes        20481.43     494.80  41.393  < 2e-16 ***
regionnorthwest  -2037.68     559.89  -3.639 0.000284 ***
regionsoutheast  -2863.21     542.44  -5.278 1.52e-07 ***
regionsouthwest  -2172.64     573.12  -3.791 0.000157 ***
```

```
bp_levelhigh      4032.32    1190.03   3.388 0.000724 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6555 on 1324 degrees of freedom
Multiple R-squared:  0.7096,    Adjusted R-squared:  0.707 
F-statistic: 269.6 on 12 and 1324 DF,  p-value: < 2.2e-16
```

Thus, we select following variables for our regression model: bmi, bloodpressure, children, smoker, region.

Modeling

Model 1: we construct model from variables which are selected from Best subsets regression method

Thus, the preditors are: bmi, bloodpressure, smoker, region.

```

Call:
lm(formula = claim ~ bmi + bloodpressure + smoker + region, data = 
insurance_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11637.0  -3755.2   -878.7   2379.9  23948.2 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11178.35    2011.93  -5.556 3.58e-08 ***
bmi                139.74      31.22   4.475 8.55e-06 ***
bloodpressure      199.48      19.51  10.224  < 2e-16 ***
smokerYes        12786.98     580.72  22.019  < 2e-16 ***
regionnorthwest  -2790.94     549.05  -5.083 4.47e-07 ***
regionsoutheast  -4020.10     540.07  -7.444 2.19e-13 ***
regionsouthwest  -3527.55     568.49  -6.205 8.15e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5476 on 951 degrees of freedom
Multiple R-squared:  0.4693,    Adjusted R-squared:  0.466 
F-statistic: 140.2 on 6 and 951 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\widehat{claim} = {} & 139.74 \times bmi + 199.48 \times bloodpressure + 12786.98 \times smoker \\ & - 2790 \times regionnorthwest - 4020.1 \times regionsoutheast - 3527 \times regionsouthwest\end{aligned}$$

Based on the result:

- The p-value of F-test is less than 2.2e-16 which is less than significant level α = 5%. This means there at least 1 variable can estimate the final score. In the other words, the model can use to estimate the claim.

- The p-value of t test in all coefficients is less than significant level α = 5%. So we can conclude that all coefficients different from 0. Hence, attributes bmi, bloodpressure, smoker and region has a statically significant effect on claim.

Model 2: we construct model from variables which are selected from stepwise regression method

The predictors are: bmi, bloodpressure, smoker, region and children

```
Call:
lm(formula = claim ~ bmi + bloodpressure + smoker + region + 
    children, data = insurance_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11023.3  -3607.4   -955.2   2247.9  22921.5 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12226.61    2008.91  -6.086 1.68e-09 ***
bmi                143.89      30.99   4.643 3.92e-06 ***
bloodpressure      201.90      19.36  10.430  < 2e-16 ***
smokerYes        12799.08     575.74  22.231  < 2e-16 ***
regionnorthwest  -2811.26     544.03  -5.167 2.89e-07 ***
regionsoutheast  -3959.79     536.61  -7.379 3.48e-13 ***
regionsouthwest  -3484.67     564.20  -6.176 9.73e-10 ***
children1          316.51     438.11   0.722 0.470201    
children2         1786.75     498.00   3.588 0.000351 ***
children3         2116.09     593.35   3.566 0.000380 ***
children4         2627.80    1386.76   1.895 0.058408 .  
children5          534.10    1593.82   0.335 0.737620    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5422 on 946 degrees of freedom
Multiple R-squared:  0.4824,	Adjusted R-squared:  0.4764 
F-statistic: 80.16 on 11 and 946 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\widehat{claim} = &\quad 143.89 \times bmi + 201.90 \times bloodpressure + 12799.08 \times smoker \\ - &\quad 2811.26 \times regionnorthwest - 3959.79 \times regionsoutheast \\ - &\quad 3484.67 \times regionsouthwest + 316.51 \times children1 + 1786.75 \times children2 \\ + &\quad 2116.09 \times children3 + 2627.8 \times children4 + 534.1 \times children5 - 12226.61\end{aligned}$$

Based on the result:

- The p-value of F-test is less than 2.2e-16 which is less than significant level α = 5%. This means there at least 1 variable can estimate the final score. In the other words, the model can use to estimate the claim.

- The p-value of t test of bmi, bloodpressure, smoker and region coefficients are less than significant level α = 5% except. So we can conclude that all coefficients different from 0.

- Regarding children coefficients, the p-value of coefficients of children1, children4 and children5 are ,respectively. There figures higher than significant level α = 5%. Thus these coefficients are equal to 0. However, the p-value of coefficients of children2 and children3 are much smaller than 0.05. Then, we can state that these coefficient is not equal to 0. As a result, we need a test to check if we can omit the children variable.

The hypothesis test (model 1 and model 2):

H_0: coefficient of children = 0

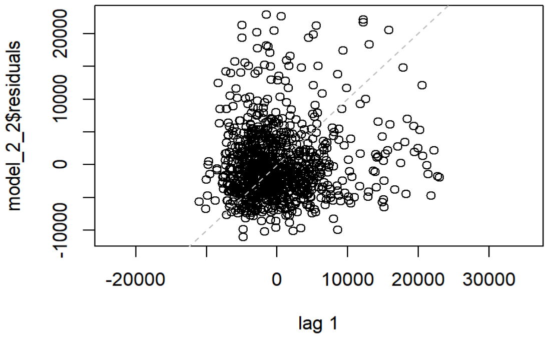
H_α: coefficient of children ≠ 0

Hypothesis testing: we use **anova** function to test whether the reduced model can be used

```
Analysis of Variance Table

Model 1: claim ~ bmi + bloodpressure + smoker + region
Model 2: claim ~ bmi + bloodpressure + smoker + region + children
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)    
1    951 2.8517e+10                                  
2    946 2.7813e+10  5 703347049 4.7845 0.0002521 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the test output, the p-value = `0.0002521` which is less than 0.05. So the null hypothesis H_0 is resoundingly rejected. In other words, we will use the 'full' model (model 2)

Independence testing



H_0: There is no correlation among the residuals.

H_α: The residuals are autocorrelated.

```
 lag Autocorrelation D-W Statistic p-value
   1      0.07969448        1.8403   0.014
 Alternative hypothesis: rho != 0
```

In practice, the D-W test statistic is `1.8403` which lies between the between 1.5 and 2.5. ***As a result, autocorrelation is probably not a cause for concern.***

Stability testing

H_0: Homoscedasticity is present (the residuals are distributed with equal variance).

H_α: Heteroscedasticity is present (the residuals are not distributed with equal variance).

```
	studentized Breusch-Pagan test

data:  model_2_2
BP = 11.226, df = 11, p-value = 0.4246
```

The test statistic for the studentized Breusch-Pagan test is `11.226`, with 11 degrees of freedom. To make more sense, the p-value of the test is `0.4246`, which is greater than the commonly used significance level $\alpha = 0.05$. Thus, we fail to reject the null hypothesis H_0 that homoscedasticity is present. In other words, there is **enough evidence** to conclude that ***the residuals of regression model are distributed with equal variance.***

Normality testing

H_0: The residuals of regression model is normally distributed.

H_α: The residuals of regression model is not normally distributed.

```
	Shapiro-Wilk normality test

data:  model_2_2$residuals
W = 0.90133, p-value < 2.2e-16
```

The test statistic for the Shapiro-Wilk normality test is `0.90133`. And the p-value of the test less than the significance level 0.05. Thus, we fail to reject H_0. In other words, there is ***not enough evidence*** to conclude ***The residuals of regression model is normally distributed.***

Model expression

Based on independence, stability and normality testing, we accept model 1 for estimate the claim.

The equation:

$$\begin{aligned}\widehat{claim} = {} & 143.89 \times bmi + 201.90 \times bloodpressure + 12799.08 \times smoker \\ - {} & 2811.26 \times regionnorthwest - 3959.79 \times regionsoutheast \\ - {} & 3484.67 \times regionsouthwest + 316.51 \times children1 + 1786.75 \times children2 \\ + {} & 2116.09 \times children3 + 2627.8 \times children4 + 534.1 \times children5 - 12226.61\end{aligned}$$

The fitted line reveals that:

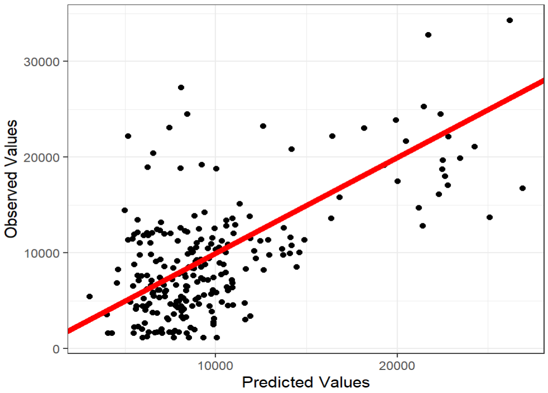
- If all the coefficients equal to 0, then the claim is -12226.61 unit.
- If the bmi is increased by 1 unit and the other predictors remain constant, the claim is increased by 143.89 unit.

- If the bloodpressure is increased by 1 unit and the other predictors remain constant, the claim is increased by 201.90 unit.
- If the the patient is smoker (smoker = 1) and the other predictors remain constant, the claim is increased by 12799.08 unit.
- If the the patient live in region northwest, southeast and southwest and the other predictors remain constant, the claim is decreased by 2811.26, 3959.79 and 3484.67 unit respectively.
- If the patient has 1, 2, 3, 4 and 5 childrens, the other predictors remain constant, the claim is increased by 316.51, 1786.75, 2116.09, 2627.8 and 534.1 unit respectively.

Prediction

We will do predict on the validation test and the Root mean square error is:

```
4912.2
```


Observed Values
Predicted Values