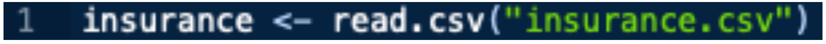
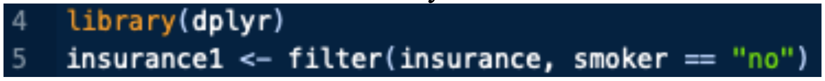
1. Download the file ‘insurance.csv’ from our class Blackboard site.

2. Read this file into your R environment. Show the step that you used to accomplish this



```
insurance <- read.csv("insurance.csv")
```

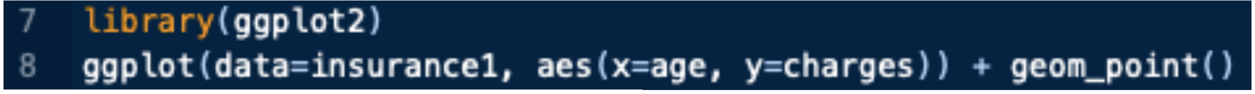
3. Filter the dataframe to create a new dataframe that only contains the records of people who are not smokers. Show the code that you used to do this.



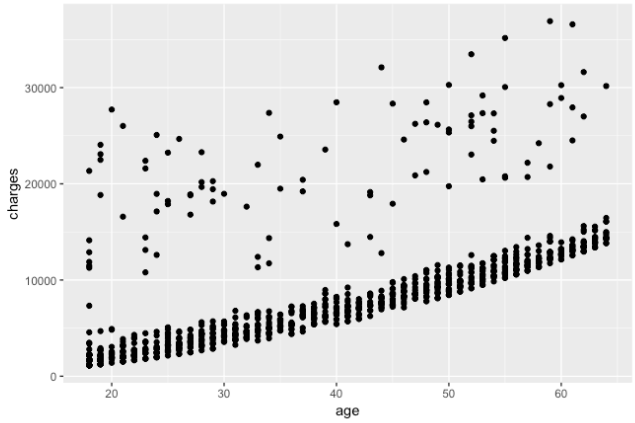
```
library(dplyr)
insurance1 <- filter(insurance, smoker == "no")
```

4. Using ggplot, create a scatterplot to depict the relationship between the input variable age and the output variable charges. Show your scatterplot, along with the code that you used to build it. What does this scatterplot suggest about the relationship between the two variables? Why (or why not) does this make intuitive sense to you?

There is a positive relationship between age and charges. As people’s age increases, their individual medical costs billed by health insurance become larger. It makes sense because when people get older, they are more likely to have body defects, compared with the young, even though they are not smokers. Therefore, the insurance company will charge more on them than the young.

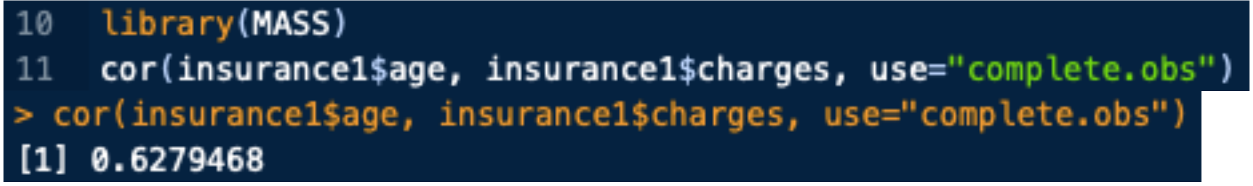


```
library(ggplot2)
ggplot(data=insurance1, aes(x=age, y=charges)) + geom_point()
```



5. Find the correlation between age and charges. Show the code that you used, and the results from your console, in a screenshot.

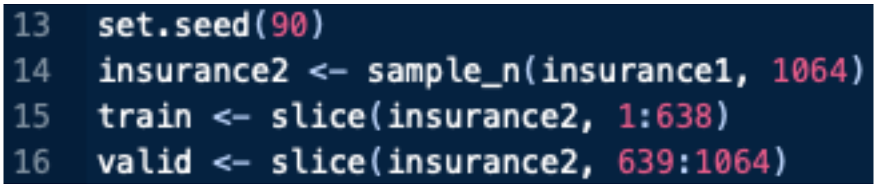
The correlation between age and charges is 0.6279468



```
library(MASS)
cor(insurance1$age, insurance1$charges, use="complete.obs")
```

```
> cor(insurance1$age, insurance1$charges, use="complete.obs")
[1] 0.6279468
```

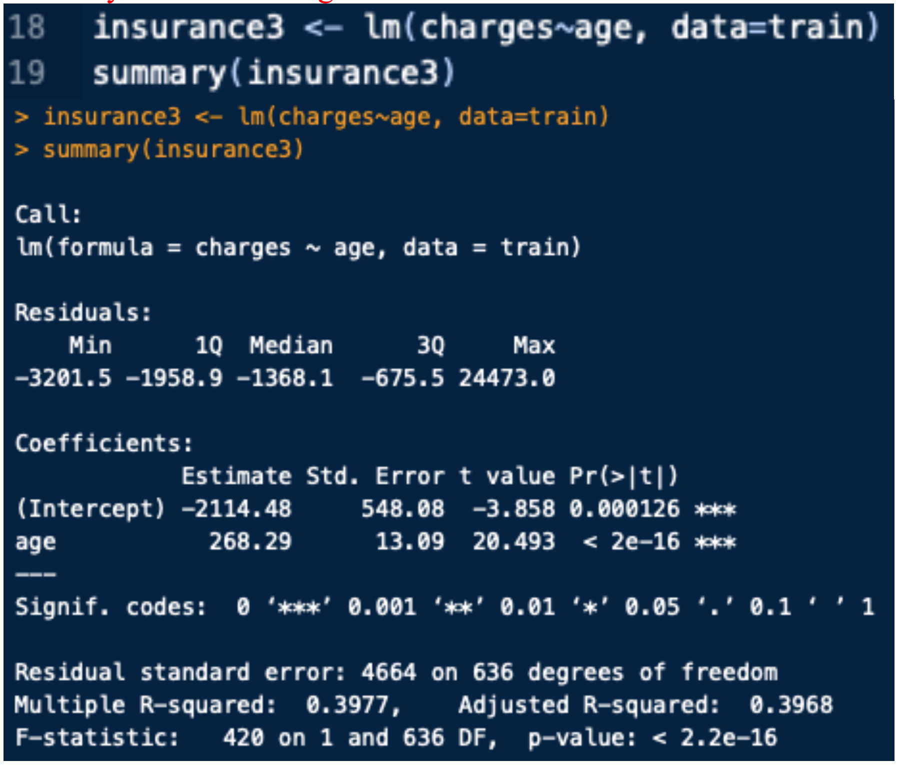
6. Using your assigned seed value, create a data partition. Assign approximately 60% of the records to your training set, and the other 40% to your validation set. Show the code that you used to do this.



```
set.seed(90)
insurance2 <- sample_n(insurance1, 1064)
train <- slice(insurance2, 1:638)
valid <- slice(insurance2, 639:1064)
```

7. Using your training set, create a simple linear regression model. Show the step(s) that you used to do this. Include a screenshot of the summary of your model, along with the code you used to generate that summary.

I first create a new variable called "insurance3" and store the value by inputting the linear regression function (lm) with my training set. Then, I use the Summary function to generate a summary of the linear regression.



```
insurance3 <- lm(charges~age, data=train)
summary(insurance3)
```

```
> insurance3 <- lm(charges~age, data=train)
> summary(insurance3)

Call:
lm(formula = charges ~ age, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3201.5 -1958.9 -1368.1  -675.5 24473.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2114.48     548.08  -3.858 0.000126 ***
age           268.29      13.09  20.493  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4664 on 636 degrees of freedom
Multiple R-squared:  0.3977,	Adjusted R-squared:  0.3968 
F-statistic:   420 on 1 and 636 DF,  p-value: < 2.2e-16
```

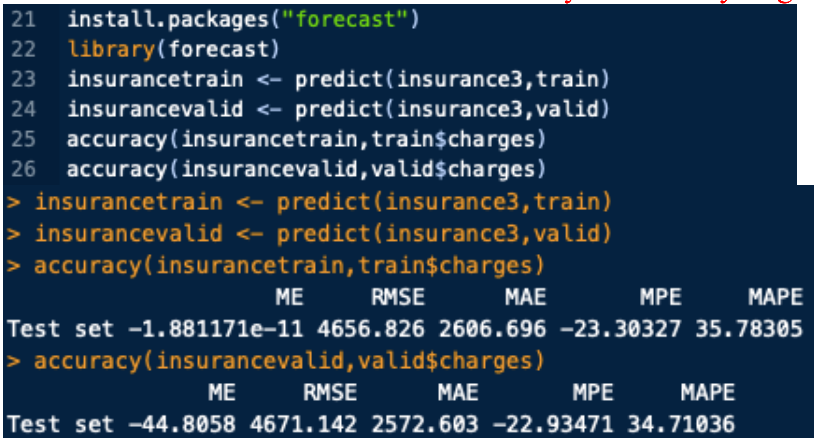
8. What is the regression equation generated by your model? Make up a hypothetical input value and explain what it would predict as an outcome. To show the predicted outcome value, you can either use a function in R, or just explain what the predicted outcome would be, based on the regression equation and some simple math.

The regression equation: charges = 268.29*age – 2114.48, which indicates that each one-unit increases in age are associated with a 268.29 increase in charges, holding other variables fixed. For instance, if the age of a person is 25, his individual medical costs will be $4,592.77 (268.29*25 – 2114.48)

9. Using the accuracy() function from the forecast package, assess the accuracy of your model against both the training set and the validation set. What do you notice about these results? Describe your findings in a couple of sentences.

Running the Accuracy function against two sets of data, I found out that compared with the training set, the validation set has a higher value of RMSE and MPE, and a lower value of ME, MAE and MAPE.

MAE and RMSE are the common metrics used to measure accuracy for continuous variables. Since both are negatively-oriented scores, lower values are better. Besides, both MPE and MAPE measure the difference between the measured value and the targeted value. Therefore, an MPE and MAPE close to zero mean that I am very close to my targeted value, which is good.

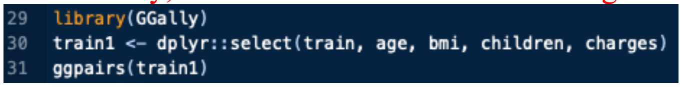


```
install.packages("forecast")
library(forecast)
insurancetrain <- predict(insurance3,train)
insurancevalid <- predict(insurance3,valid)
accuracy(insurancetrain,train$charges)
accuracy(insurancevalid,valid$charges)
```

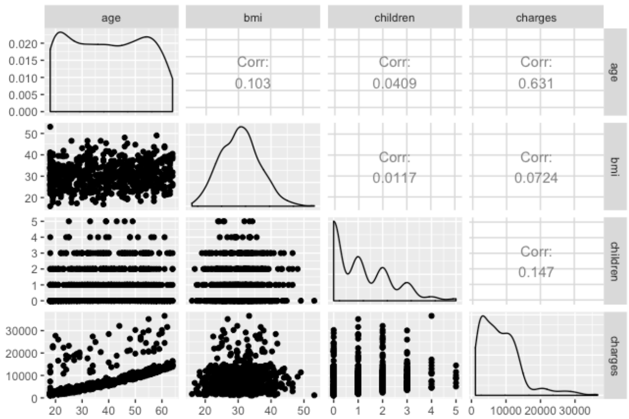
```
> insurancetrain <- predict(insurance3,train)
> insurancevalid <- predict(insurance3,valid)
> accuracy(insurancetrain,train$charges)
                    ME     RMSE      MAE       MPE     MAPE
Test set -1.881171e-11 4656.826 2606.696 -23.30327 35.78305
> accuracy(insurancevalid,valid$charges)
               ME     RMSE      MAE       MPE     MAPE
Test set -44.8058 4671.142 2572.603 -22.93471 34.71036
```

1. Create a scatterplot matrix that depicts the relationships among all of the numerical variables that you might use as predictors (use your training set to build this). Show the code you used to build the scatterplot matrix, and show your scatterplot matrix. Describe your scatterplot matrix in a few sentences. Are there any variable relationships that suggest that multicollinearity could be an issue here?

In my scatterplot matrix, age has the most impact on charges with a moderate correlation of 0.631. In addition, age has a lower correlation with bmi (0.103) and children (0.0409), and bmi also has a lower correlation with children (0.0117). Therefore, I think that multicollinearity is not an issue here because the correlation of the input variables of age, bmi, and children is low enough to not cause the problem. Moreover, the matrix indicates that bmi has the least impact on charges, and children are discrete variables where the graph displays a horizontal line. Additionally, the matrix demonstrates that as age increases, charges increase as well.



```
library(GGally)
train1 <- dplyr::select(train, age, bmi, children, charges)
ggpairs(train1)
```

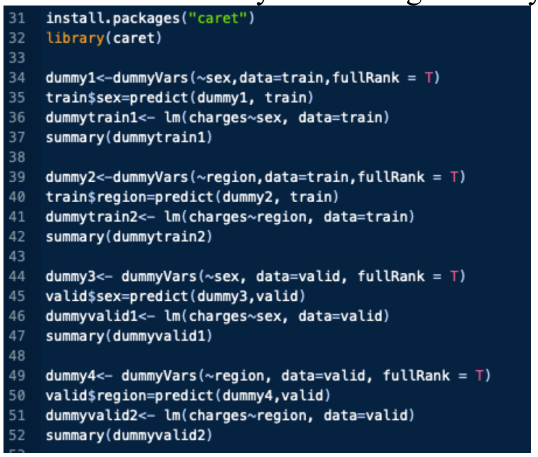


2. What are dummy variables? In a couple of sentences, describe what they are and explain their purpose.

A dummy variable is a variable that used to distinguish different treatment groups. In regression analysis, a dummy variable is displayed by either 0 or 1, in which 0 indicates the control group and 1 indicates the treated group. The purpose of the dummy variable is to allow people to use a single regression equation to represent multiple groups.

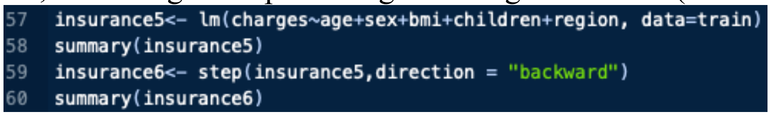
3. Create dummy variables for any categorical predictors in the data set, and show the code that you used to do this

a. To complete this step, you will need to create dummy variables for the categorical variables in your training set and your validation set.

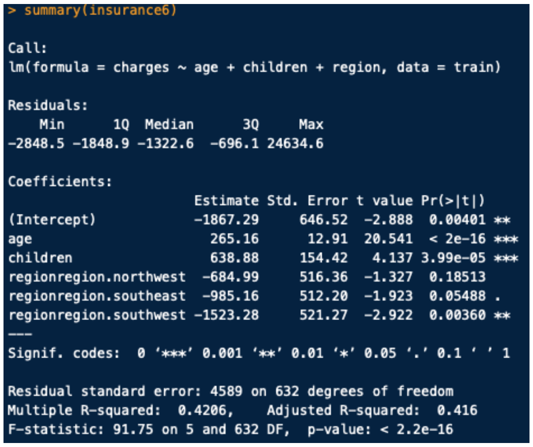


```
install.packages("caret")
library(caret)

dummy1<-dummyVars(~sex,data=train,fullRank = T)
train$sex=predict(dummy1, train)
dummytrain1<- lm(charges~sex, data=train)
summary(dummytrain1)

dummy2<-dummyVars(~region,data=train,fullRank = T)
train$region=predict(dummy2, train)
dummytrain2<- lm(charges~region, data=train)
summary(dummytrain2)

dummy3<- dummyVars(~sex, data=valid, fullRank = T)
valid$sex=predict(dummy3,valid)
dummyvalid1<- lm(charges~sex, data=valid)
summary(dummyvalid1)

dummy4<- dummyVars(~region, data=valid, fullRank = T)
valid$region=predict(dummy4,valid)
dummyvalid2<- lm(charges~region, data=valid)
summary(dummyvalid2)
```

4. Using backward elimination, build a multiple regression model with the data in your training set, with the goal of predicting the charges variable. (Start with all of the potential predictors).



```
insurance5<- lm(charges~age+sex+bmi+children+region, data=train)
summary(insurance5)
insurance6<- step(insurance5,direction = "backward")
summary(insurance6)
```

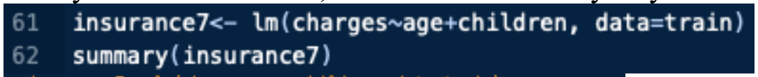


```
> summary(insurance6)

Call:
lm(formula = charges ~ age + children + region, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2848.5 -1848.9 -1322.6  -696.1 24634.6 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -1867.29     646.52  -2.888  0.00401 ** 
age                     265.16      12.91  20.541  < 2e-16 ***
children                638.88     154.42   4.137 3.99e-05 ***
regionregion.northwest  -684.99     516.36  -1.327  0.18513    
regionregion.southeast  -985.16     512.20  -1.923  0.05488 .  
regionregion.southwest -1523.28     521.27  -2.922  0.00360 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4589 on 632 degrees of freedom
Multiple R-squared:  0.4206,	Adjusted R-squared:  0.416 
F-statistic: 91.75 on 5 and 632 DF,  p-value: < 2.2e-16
```

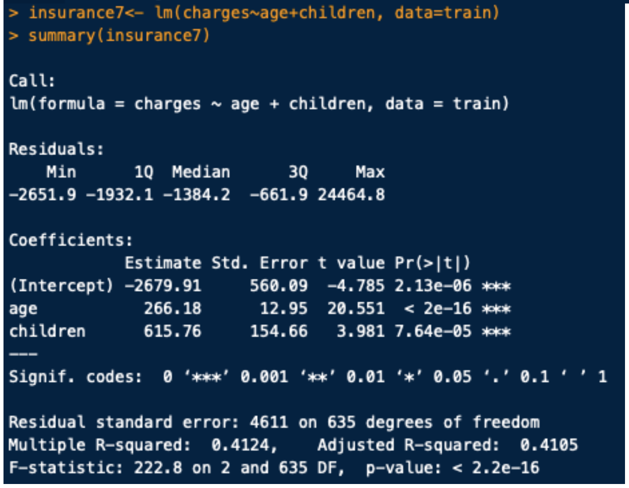
5. Based in part on what was recommended by the backward elimination process, and in part on your judgement, which variables will you keep? (Note: This is not a trick question. Part of making a multiple linear regression model involves subjective judgement).

Based on my judgment, I will keep the variable of age and children because those variables have a p-value that less than 0.05 which is statistically significant. The region has a p-value greater than 0.05, and thus, I decide to eliminate it.

6. Using the variables that you will keep, build a multiple linear regression model. Show the code you used to build it, and show a summary of your multiple regression model.



```
insurance7<- lm(charges~age+children, data=train)
summary(insurance7)
```



```
> insurance7<- lm(charges~age+children, data=train)
> summary(insurance7)

Call:
lm(formula = charges ~ age + children, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-2651.9 -1932.1 -1384.2  -661.9 24464.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2679.91     560.09  -4.785 2.13e-06 ***
age           266.18      12.95  20.551  < 2e-16 ***
children      615.76     154.66   3.981 7.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4611 on 635 degrees of freedom
Multiple R-squared:  0.4124,	Adjusted R-squared:  0.4105 
F-statistic: 222.8 on 2 and 635 DF,  p-value: < 2.2e-16
```

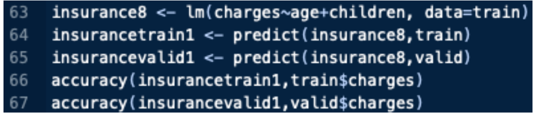
7. Make up a fictional person, and assign this person attributes for each of the predictors in your model. What does your model predict that this person's health care charges will be? To answer this, you can use a function in R or just explain it using the equation and some simple math.

The multiple regression model: charges = -2679.91 + 266.18*age + 615.76*children
For instance, if a person is 50 years old and has 2 children, his individual medical costs will be $11,860.61 (-2679.91 + 266.18*50 + 615.76*2)

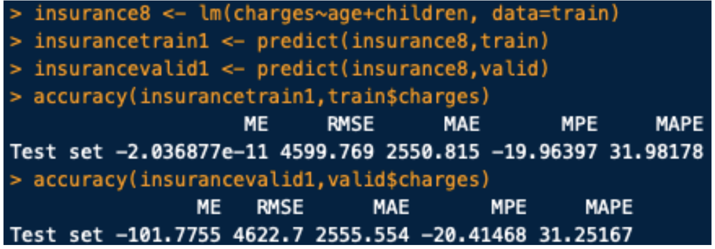
8. Using the accuracy() function from the forecast package, assess the accuracy of your model against both the training set and the validation set. What do you notice about these results? Describe your findings in a couple of sentences.

Running the Accuracy function against two sets of data, I found out that compared with the training set, the validation set has a higher value of RMSE and MAE, and a lower value of ME, MPE and MAPE.

By adding a new variable to the regression, I find out that the training set has better performance than the validation set with a lower value of RMSE and MAE, and MPE is closer to zero compared with MPE in the validation set. It is good because the training set is used to train the model, while the validation set is only used to evaluate the model's performance. Therefore, we should expect the regression to perform better in the training set than the validation set.



```
insurance8 <- lm(charges~age+children, data=train)
insurancetrain1 <- predict(insurance8,train)
insurancevalid1 <- predict(insurance8,valid)
accuracy(insurancetrain1,train$charges)
accuracy(insurancevalid1,valid$charges)
```



```
> insurance8 <- lm(charges~age+children, data=train)
> insurancetrain1 <- predict(insurance8,train)
> insurancevalid1 <- predict(insurance8,valid)
> accuracy(insurancetrain1,train$charges)
                    ME     RMSE      MAE       MPE     MAPE
Test set -2.036877e-11 4599.769 2550.815 -19.96397 31.98178
> accuracy(insurancevalid1,valid$charges)
                ME   RMSE      MAE       MPE     MAPE
Test set -101.7755 4622.7 2555.554 -20.41468 31.25167
```