Assignment: ASSIGNMENT 8.3, 10.3, 11.3 Final Project Step 1, 2, 3

Name: Shekhar , Manish

Date Created: 2021-05-12

Date Modified: 2021-05-17, Change log 1

Date Modified: 2021-05-31, Change log 2

Change log description:

i). Adding step 2 of the final project. Data analysis and preparation.

ii). Adding step 3 of the final project. Summarization and narrative.

1. Introduction

Reasearch Topic: Medical insurance costs

Health insurance provides important financial protection in case you have a serious accident or sickness.People without health coverage are exposed to these costs. This can sometimes lead people without coverage into deep debt or even into bankruptcy.

It's easy to underestimate how much medical care can cost:

1. Fixing a broken leg can cost up to $7,500

2. The average cost of a 3-day hospital stay is around $30,000

3. Comprehensive cancer care can cost hundreds of thousands of dollars

Having health coverage can help protect you from high, unexpected costs like these.

Thus, it is important that everyone, all the time, have affordable health insurance

regardless of where they work, their income, their age, or their health status

Affordable health insurance is the key to a productive work force, small business

innovation, and the economic as well as health security of our nation's families.

It is important to reaserch and understand components of rising health care costs

and propose probable changes to keep health insuance affordable for all.

Research can help understand medical cost -

1. variation by various age groups

2. variation amongst gender

3. variation by BMI (body mass index)

4. variation by number of dependents

5. variation by smoking habits

6. variation by US region etc.

Reasearch can help -

1. National, state, and local healthcare bodies to draft appropriate healthcare policies to keep the medical insurance affordable to larger population with variety of conditions.

2. Insurance companies to market policies and provide customizations to consumers based on certain demographic behaviors.

3. Consumers to understand regional medical costs and affordability and consumer behavior.

It's data science problem because it involves -

1. Data collection

2. Data cleansing

3. Data transformation

4. Data visualization

5. EDA - Exploratory data analysis

6. Modeling and Prediction

7. Validation and generalization

8. Integration and implementation

2. Research Questions

1. What is the average cost of health insurance in US by diffrent age groups?

2. What is the average cost of health insurance in US by gender?

3. What is the effect on health insurance cost by variation in BMI?

4. What is the effect on health insurance cost by number of dependents?

5. Why is the average health insurance cost varies in different US regions?

6. Predict the health insurance cost for given gender, age, BMI, number of dependents, smoking habits, region etc.

7. What is the effect on health insurance cost by change in smoking habits?

8. What US region has most obese cases?

9. What US region has most smokers?

10. How is distribution of age groups amongst different US regions?

11. How is distribution of gender amongst different age groups and in different US regions?

12. Is there any correlation between BMI and smokers?

13. Do we see spike in any region in number of dependents?

14. What other factors can effect the insurance cost?

15. How much on an average American's pay for health insurance?

16. Which US region have most people with BMI less than average?

3. Approach

Approach involves analyzing data to discover correlations, patterns and create

machine learning model to predict cost of health insurance based of various

factors i.e. age, gender, bmi, region, smoking habit, number of dependents etc.

4. How does approach addresses the problem fully or partially?

Approach targets to give enough inputs to be able to address the problem completely.

It will help uncover various data patterns to answer multiple research questions.

It will help understand cause and effect relationship between health insurance cost

and various other factors i.e. age, gender, bmi, region, smoking habit, number of dependents etc.

It also intends to develop a model to predict health insurance cost given

various variables.

5. Data

```r
# Insurance data files for United States, one each for each region
ins_data_southwest <- read.csv("insurance_southwest.csv")
ins_data_southeast <- read.csv("insurance_southeast.csv")
ins_data_northwest <- read.csv("insurance_northwest.csv")
ins_data_northeast <- read.csv("insurance_northeast.csv")

# Combining the insurance data files into one data frame
# I manually inspected and they are all in same structure and thus can be combined using
# rbind into one data frame
ins_data <- rbind(ins_data_southwest,
                  ins_data_southeast,
                  ins_data_northwest,
                  ins_data_northeast)

# checking structure of the data
str(ins_data)
```

```
## 'data.frame':    1338 obs. of  8 variables:
##  $ X       : int  1 13 16 19 20 22 30 31 33 35 ...
##  $ age     : int  19 23 19 56 30 30 31 22 19 28 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 34.4 24.6 40.3 35.3 32.4 36.3 35.6 28.6 36.4 ...
##  $ children: int  0 0 1 0 0 1 2 0 5 1 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southwest" "southwest" "southwest" ...
##  $ charges : num  16885 1827 1837 10602 36837 ...
```

```r
# These datasets are inspired by the book Machine Learning with R by Brett Lantz. The data contains med

# Column definition
# ------------------
# age: age of primary beneficiary
# sex: insurance contractor gender, female, male
# bmi: Body mass index, providing an understanding of body, weights that are relatively high or low rel
# children: Number of children covered by health insurance / Number of dependents
# smoker: Smoking
# region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
# charges: Individual medical costs billed by health insurance

# Download link - https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv

# Information about how missing data values are recorded or how they were imputed is not provided.
# Checking for missing values in the data set
apply(ins_data, 2, function(x) any(is.na(x) | is.infinite(x)))
```

```
##        X      age      sex      bmi children   smoker   region  charges
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
```

```
# There no missing values in the data set.
```

6. Packages needed for the project

Packages for data transformation

1. dplyr

2. purrr

Packages to Regression diagnostics

1. QuantPsyc - To get standard regression coefficients

2. car - Use durbinWatsonTest() to test the assumption of independent error

3. lmtest - Use dwtest() to test the assumption of independent error

Package for interactive plotting, model fitting, and stats about data

Rcmdr

Packages for data visualization and visual evaluation

1. ggplot2 - Useful to plot various charts to evaluate assumptions of linear regression

2. qqplotr - Useful to plot various charts to evaluate assumptions of linear regression

7. Questions for future steps

1. We delve deep into linear regression this week and definitely touched a variety of topics including linear regression diagnostics, fitting a linear model, selecting parameters, generlizing the model etc and associated statistical measures. One thing is definitely needed is more practice. Taking up different datasets and getting dirty while applying these concepts.

2. We learned about linear model assumptions and how to measure them. But we did not clearly cover what to do when each of these assumptions fail i.e. what are our options. Further reading and exploration on this topic is needed.

3. Looking forward to learn logistic regression as well. I am not sure as of now if I will see a use case to in this current topic I have chosen to be able to apply logistic regression. May be along additional consumer behavioral features we can use logistic regression classification to predict whether a lead will get converted or not i.e. will someone buy an insurance or any product or not.

Final Project Step 2 - data analysis and preparation

1. How to import and clean my data?

```
# checking the structure of the data
str(ins_data)
```

```
## 'data.frame':    1338 obs. of  8 variables:
##  $ X       : int  1 13 16 19 20 22 30 31 33 35 ...
##  $ age     : int  19 23 19 56 30 30 31 22 19 28 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 34.4 24.6 40.3 35.3 32.4 36.3 35.6 28.6 36.4 ...
##  $ children: int  0 0 1 0 0 1 2 0 5 1 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southwest" "southwest" "southwest" ...
##  $ charges : num  16885 1827 1837 10602 36837 ...
```

```
# Just by looking at the data we can can have following observations -
# 1. Variable is just an index and thus can be ignored or filtered
# 2. Variable sex is character with two values male or female and thus should be
# encoded as factor with distinct numeric values
# 3. Variable smoker is character with two values yes or no and thus should be
# encoded as factor with distinct numeric values. We can use 1 for yes and 0 for no.
# 4. Variable region is character with four distinct values "southeast",
# "southwest", "northeast", "northwest" and should be encoded as factor with
# four distinct numeric values.
# Checking the summary of data set to gauge the value range of each numerical variable
summary(ins_data)
```

```
##        X               age            sex                 bmi
##  Min.   :   1.0   Min.   :18.00   Length:1338        Min.   :15.96
##  1st Qu.: 335.2   1st Qu.:27.00   Class :character   1st Qu.:26.30
##  Median : 669.5   Median :39.00   Mode  :character   Median :30.40
##  Mean   : 669.5   Mean   :39.21                      Mean   :30.66
##  3rd Qu.:1003.8   3rd Qu.:51.00                      3rd Qu.:34.69
##  Max.   :1338.0   Max.   :64.00                      Max.   :53.13
##     children        smoker             region             charges
##  Min.   :0.000   Length:1338        Length:1338        Min.   : 1122
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.: 4740
##  Median :1.000   Mode  :character   Mode  :character   Median : 9382
##  Mean   :1.095                                         Mean   :13270
##  3rd Qu.:2.000                                         3rd Qu.:16640
##  Max.   :5.000                                         Max.   :63770
```

```
# 5. Scale of numeric variables are different i.e. age varies between 18 and 64,
# while children varies between 0 to 5, and bmi between 15.96 and max 53.13.
# As we are planning to do multiple linear regression and use lm() which takes care
# of scaling variables appropriately, we do not need to worry about this data
# values variations.
# 6. We can change the data types of integer variables i.e. age and children to
# be numeric.
# Creating a copy of ins_data data frame and then applying above data transformations
# ignoring or filtering out variable x
ins_data_select <- ins_data[,c(2:8)]
# changing age and children to numeric
ins_data_select$age <- as.numeric(ins_data_select$age)
ins_data_select$children <- as.numeric(ins_data_select$children)
# changing sex to factor
ins_data_select$sex <- factor(ins_data_select$sex,
                              levels = c("female","male"),
                              labels = c(1,2))
# changing smoker to factor
ins_data_select$smoker <- factor(ins_data_select$smoker,
                              levels = c("no","yes"),
                              labels = c(0,1))
# changing region to factor
ins_data_select$region <- factor(ins_data_select$region,
                              levels = c("southwest","southeast","northwest","northeast"),
                              labels = c(1,2,3,4))
```

2. What does the final data set look like?

```
str(ins_data_select)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : num  19 23 19 56 30 30 31 22 19 28 ...
##  $ sex     : Factor w/ 2 levels "1","2": 1 2 2 2 2 1 2 2 1 2 ...
##  $ bmi     : num  27.9 34.4 24.6 40.3 35.3 32.4 36.3 35.6 28.6 36.4 ...
##  $ children: num  0 0 1 0 0 1 2 0 5 1 ...
##  $ smoker  : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 2 2 1 2 ...
##  $ region  : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ charges : num  16885 1827 1837 10602 36837 ...
```

```
head(ins_data_select)
```

```
##   age sex  bmi children smoker region   charges
## 1  19   1 27.9        0      1      1 16884.924
## 2  23   2 34.4        0      0      1  1826.843
## 3  19   2 24.6        1      0      1  1837.237
## 4  56   2 40.3        0      0      1 10602.385
## 5  30   2 35.3        0      1      1 36837.467
## 6  30   1 32.4        1      0      1  4149.736
```

3. Questions for future steps.

a). What I do not know right now is how to scale the variables for linear regression. I would love to discover more on application of variable scaling and techniques. As lm() function takes care of variable scaling I am not very much worries about this part.

b). I learned few new ways to handle missing values in the data i.e. median imputation and mean imputation methods, where we intend to replace the missing values with mean of the data. This method will not be applicable for all kinds of variables and data.

c). Need to learn how to visualize more than two variables. Reasearching various options i.e. utilizing aesthetics i.e. colour, size, shape etc.

4. Discuss how you plan to uncover new information in the data that is not self-evident.

To uncover new information in the data that is not self-evident I plan to check - #### 1. correlation among variables

2. visualize data to uncover patterns and trends

3. Check data distribution of variables

4. do case analysis, detect outliers and influencial cases

```
# To get correlation plot for factors or mixed-type, we can also use model.matrix
# to one-hot encode all non-numeric variables.
# It considers factor as separate variables, as many regression models do
# We can then use our favorite correlation-plot library. One such library is
# ggcorrplot and it has ggplot2 compatibility.
library(ggcorrplot)
```

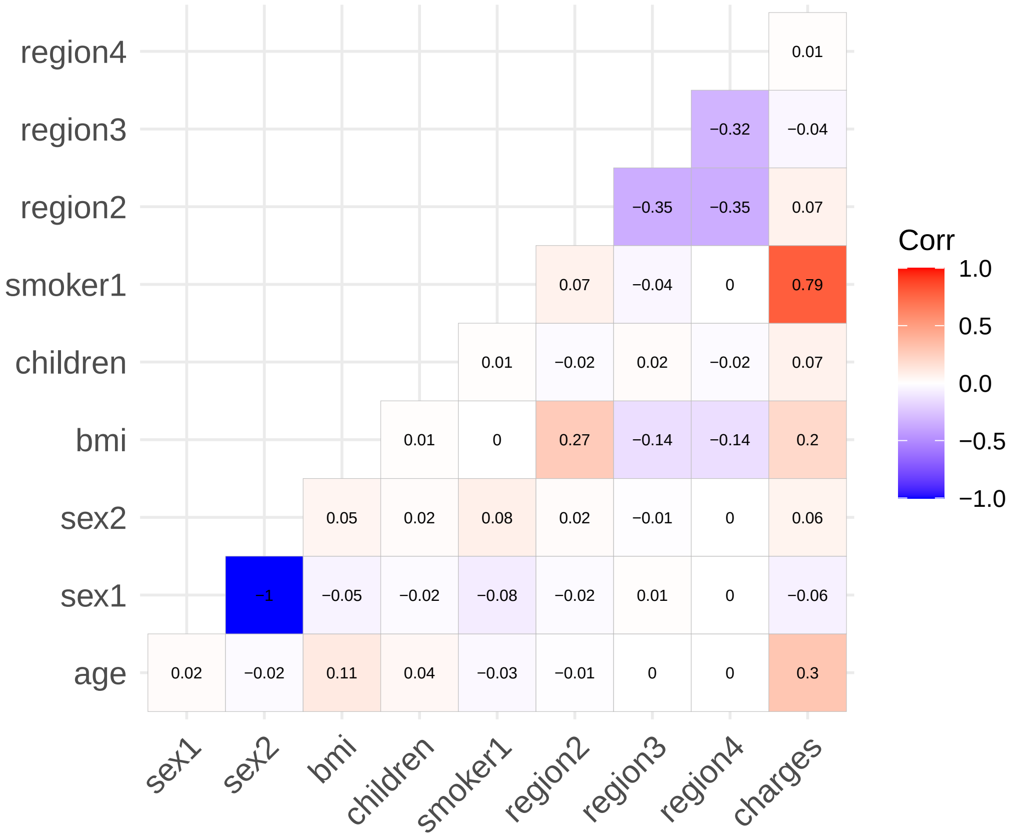
(will cover this later after creating the regression model)

```
## Loading required package: ggplot2
```

```
library(dplyr)
```

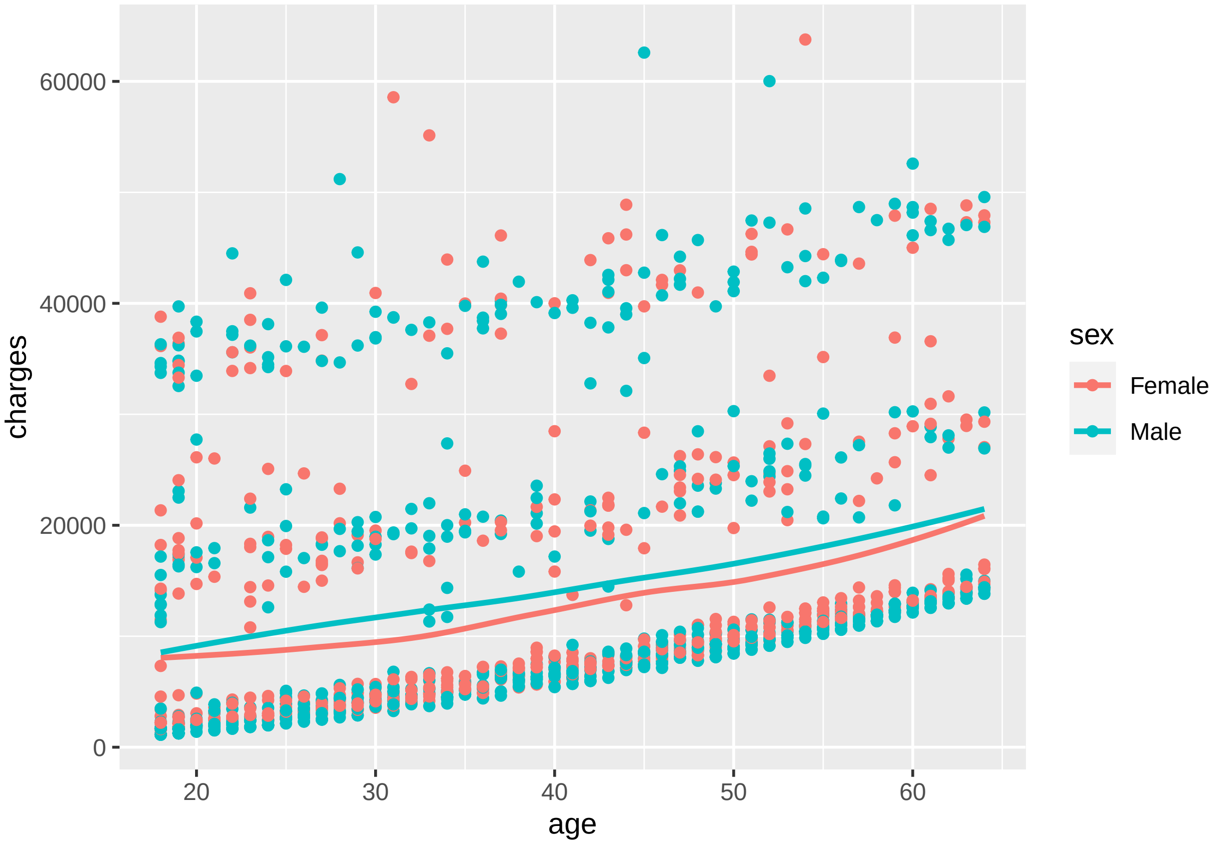
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
model.matrix(~0+., data=ins_data_select) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag = F, type="lower", lab=TRUE, lab_size=2)
```



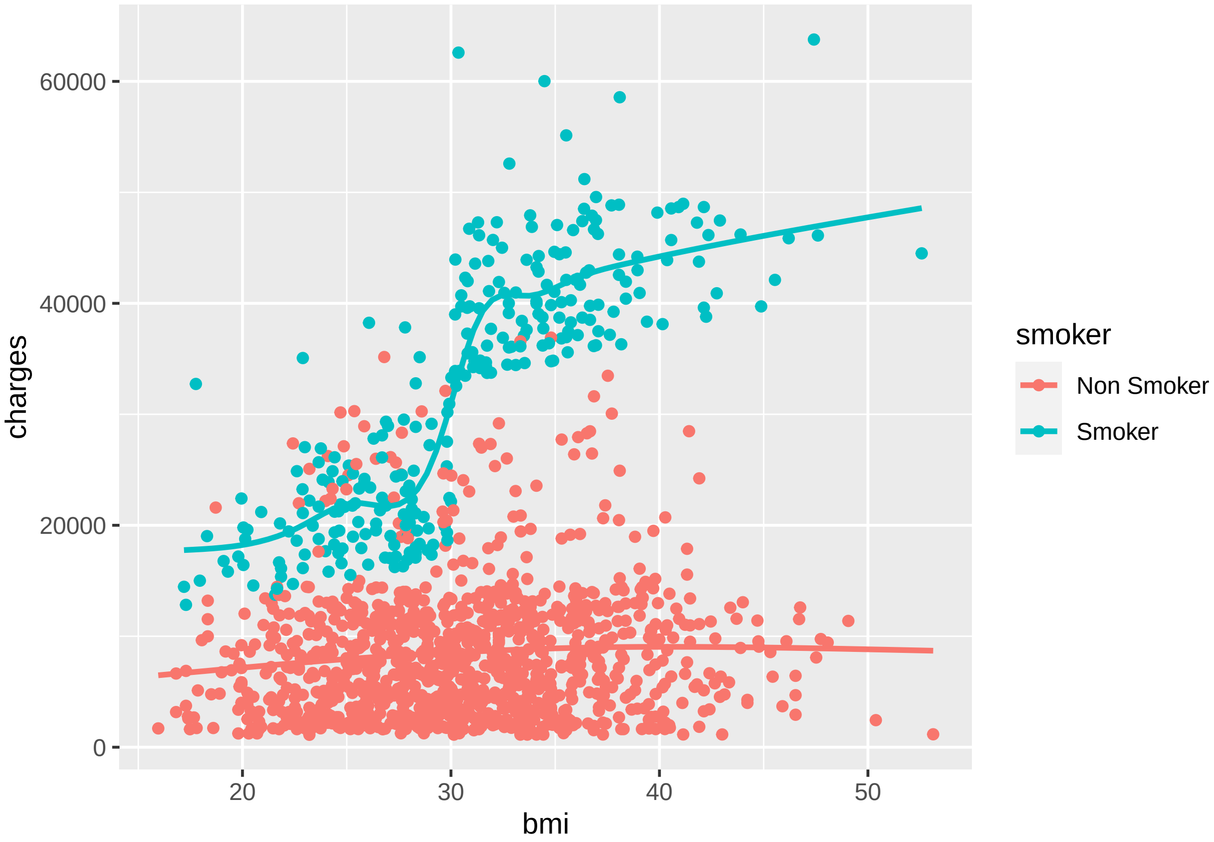
```
# Observation from correlation plot
# 1. It is obvious and we can see that being smoker is highly correlated to the
# insurance change. Being smoker shares 62% of variability (coefficient of
# determination, R^2 = 0.79*0.79 = 0.62) in insurance charges
# 2. Age shares 9% (coefficient of determination, R^2 = 0.3*0.3 = 0.09) of
# variability in the insurance charges
# 3. bmi shares 4% (coefficient of determination, R^2 = 0.2*0.2 = 0.04) of variability
# in the insurance charges
# 4. bmi in region2 (southeast) seems higher than that in region3 (northwest) and
# region4 (northeast). Bmi is negatively correlated to northwest and northeast
# region which suggests that people in north are more fit. We also know that higher
# your BMI, the higher your risk for certain diseases such as heart disease,
# high blood pressure, type 2 diabetes, gallstones, breathing problems, and
# certain cancers. So, we may consider people in southeast as higher risk and thus
# they may be considered for higher insurance premium.
# 5. Purely based on correlation sex1 (Female) seems negatively correlated to
# being a smoker while sex2 (Male) is positively correlated to being a smoker.


# Checking relation between age and charges using ggplot()
library(ggplot2)
ggplot(data = ins_data_select, aes(x = age, y = charges, colour = sex)) +
  geom_point() + geom_smooth(fill=NA) +
  scale_color_discrete(labels = c('Female','Male'))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

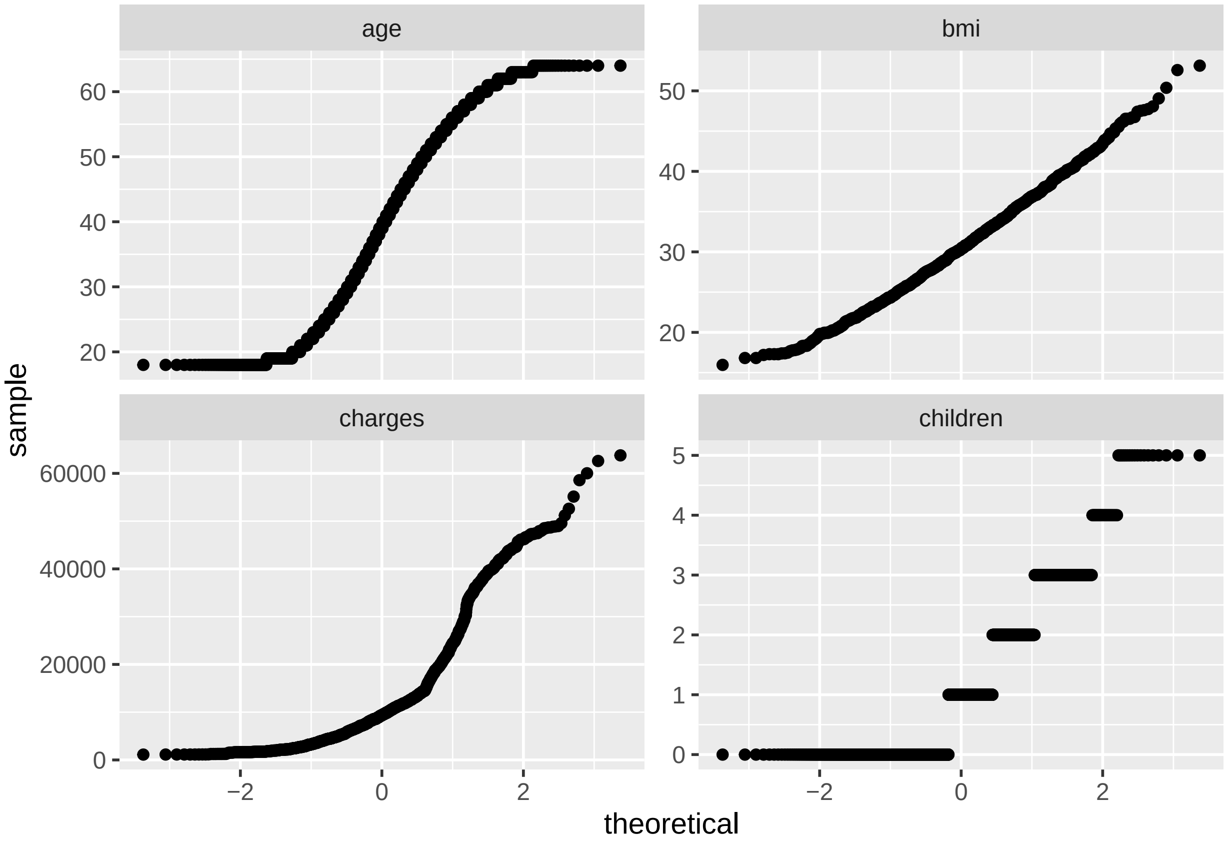


```
# We have already seen that there is positive correlation between age and charges
# i.e. positive change in age will have positive change in insurance charges.
# We also see that men are almost always paying more than women.


# Plotting bmi against insurance changes and color the data points by smoker
ggplot(data = ins_data_select, aes(x = bmi, y = charges, colour = smoker)) +
  geom_point() + geom_smooth(fill=NA) +
  scale_color_discrete(labels = c('Non Smoker','Smoker'))
```

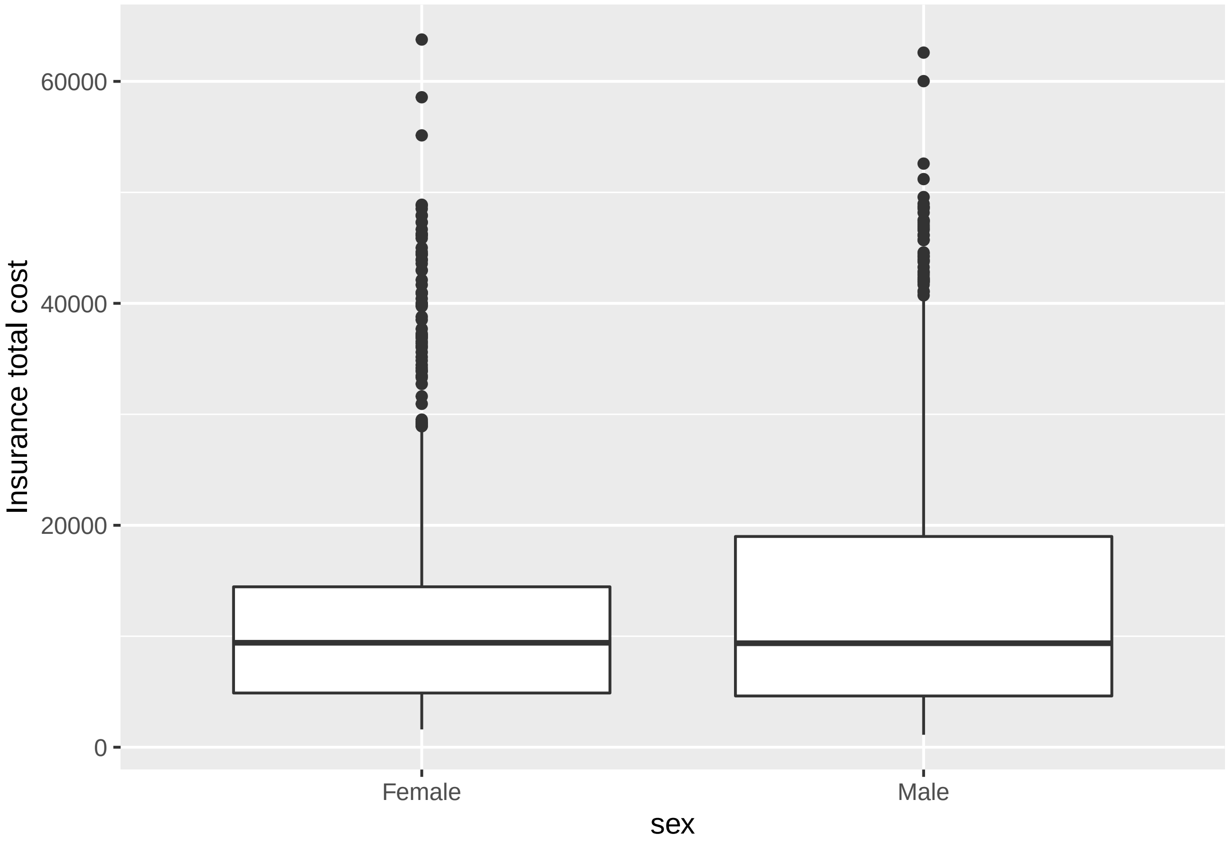
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



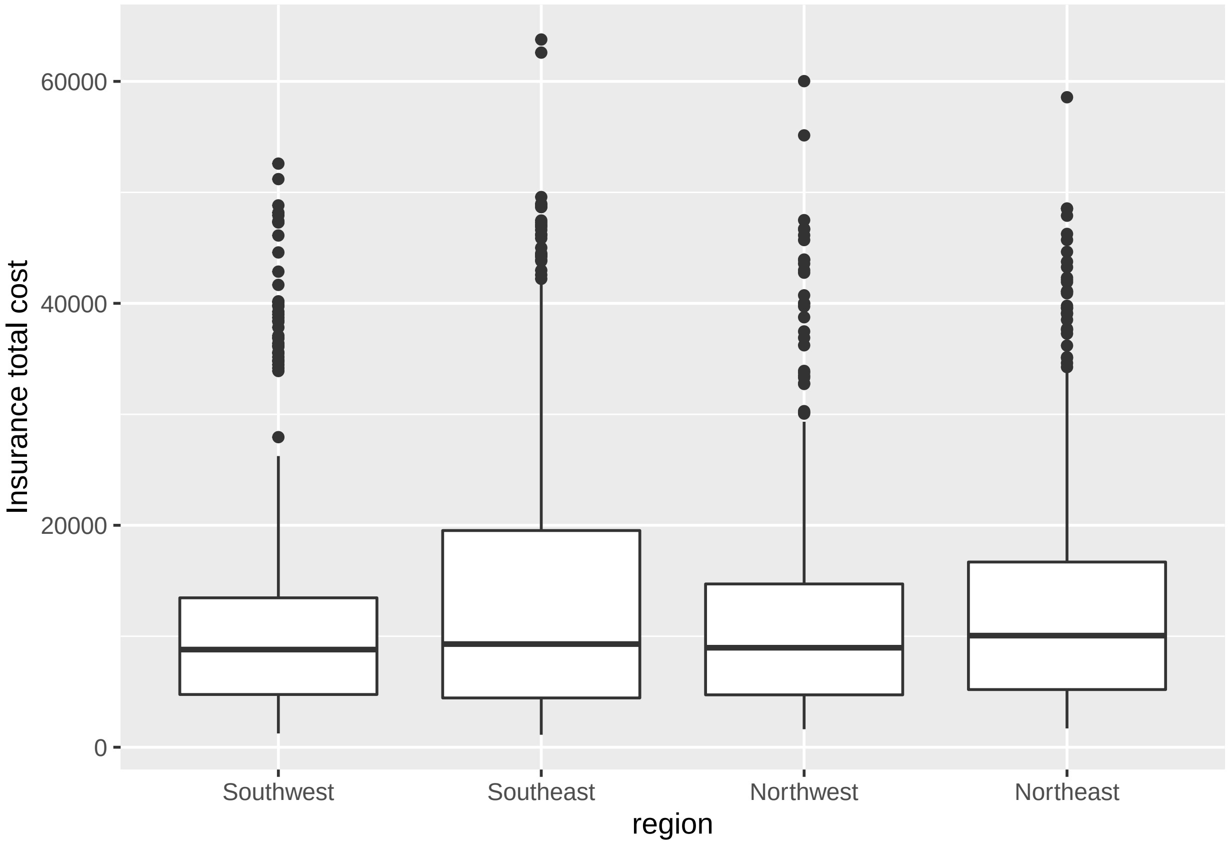
```
# we can clearly observe that smokers are almost always paying more than non smokers
# with same bmi. In fact per the data set smokers are paying a lot more than non
# smokers when their bmi is greater than 30. For non smokers insurance charges only
# is seeing very slight increase if bmi is between 25 and 45 otherwise it remains low
# and almost static.


# Checking if data distribution of numeric variables is normal
# combining pipe operator between dplyr transformation and ggplot
library(tidyr)

ins_data_select %>% select(age, bmi, children, charges) %>%
  gather() %>%
  ggplot(., aes(sample = value)) +
  stat_qq() +
  facet_wrap(vars(key), scales ='free_y')
```



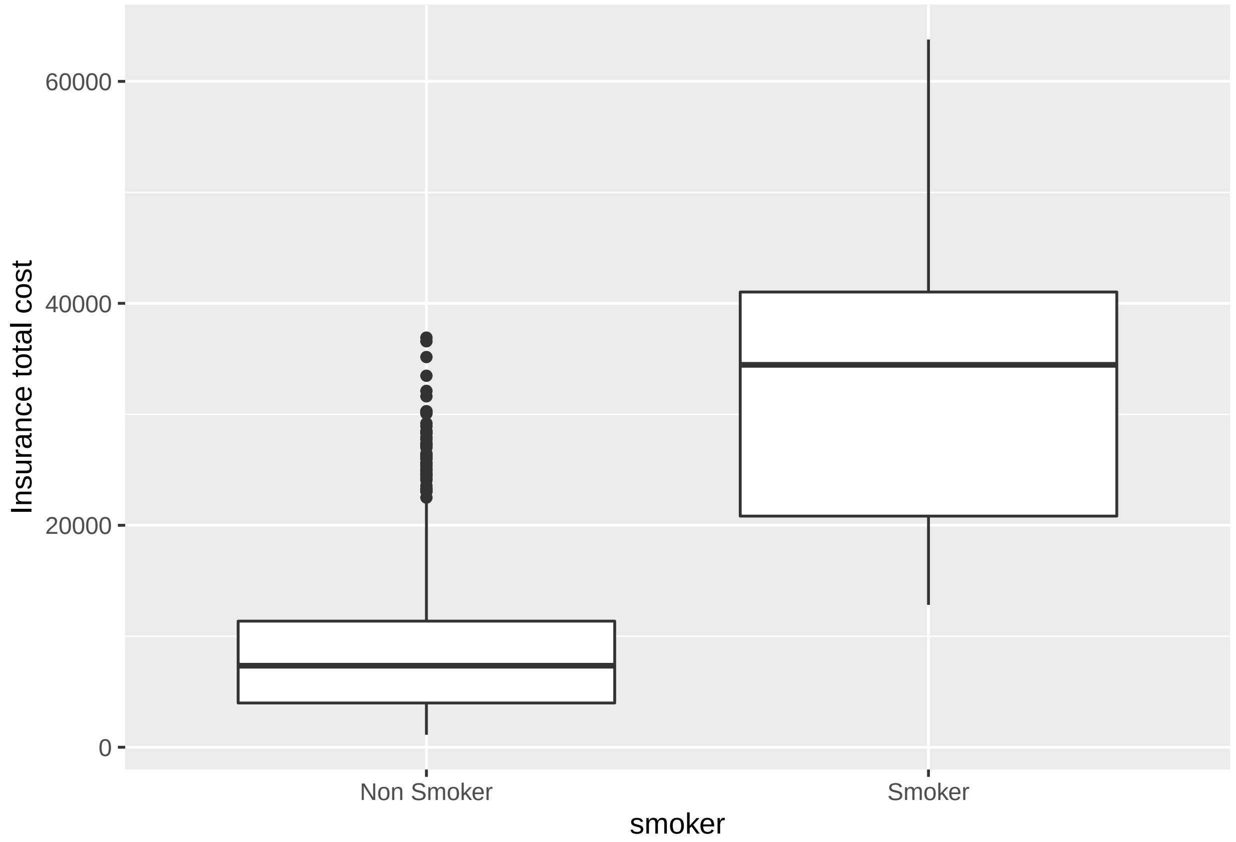
```
# Except bmi data no other attribute seems perfectly normal in distribution


ggplot(data = ins_data_select, aes(x = sex, y = charges)) +
  geom_boxplot() + ylab("Insurance total cost") +
  scale_x_discrete(labels = c('Female','Male'))
```



```
# We can observe that for both male and female we have quite a bit of outliers and
# thus data is not normally distributed


ggplot(data = ins_data_select, aes(x = region, y = charges)) +
  geom_boxplot() + ylab("Insurance total cost") +
  scale_x_discrete(labels = c('Southwest','Southeast','Northwest','Northeast'))
```



```
# We can observe that for all 4 regions we have quite a bit of outliers and thus
# it's data is not normally distributed


ggplot(data = ins_data_select, aes(x = smoker, y = charges)) +
  geom_boxplot() + ylab("Insurance total cost") +
  scale_x_discrete(labels = c('Non Smoker','Smoker'))
```



```
# If we just consider data for smokers it appears normally distributed with hardly
# any outliers while non smokers are not normally distributed with many outliers.
# Overall data in smoker column is not normally distributed
```

5. Uncovering signals from data.

5. i.What are different ways you could look at this data?

I plan to visualize data both granular and summary. I can also calculate summaries using sql like commands using dplyr.

5. ii. How do you plan to slice and dice the data?

I plan to visualize data both granular and summary. I can also calculate summaries using sql like commands using dplyr.

5. iii. How could you summarize your data to answer key questions?

Using dplyr sql like commands I can derive new variables or just explore summaries at aggregated levels.

5. iv. What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

scatterplots, boxplots, histograms, barcharts etc.

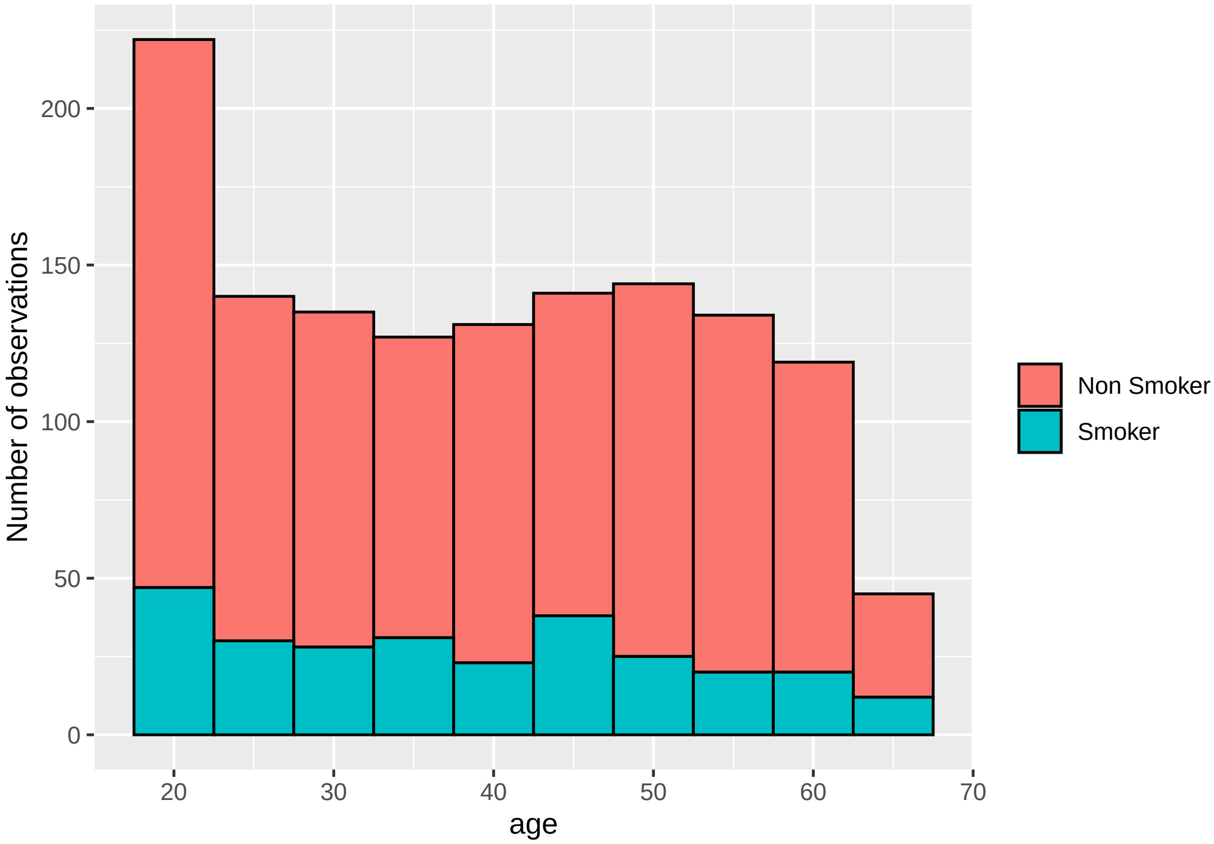
Trying to answer research questions collected earlier

```
# check the average cost of health insurance in US by different age groups
library(dplyr)
ins_data_select %>%
  group_by(age) %>% summarise(avg_insurance_cost = mean(charges)) %>%
  arrange(desc(avg_insurance_cost))
```

a). What is the average cost of health insurance in US by diffrent age groups?

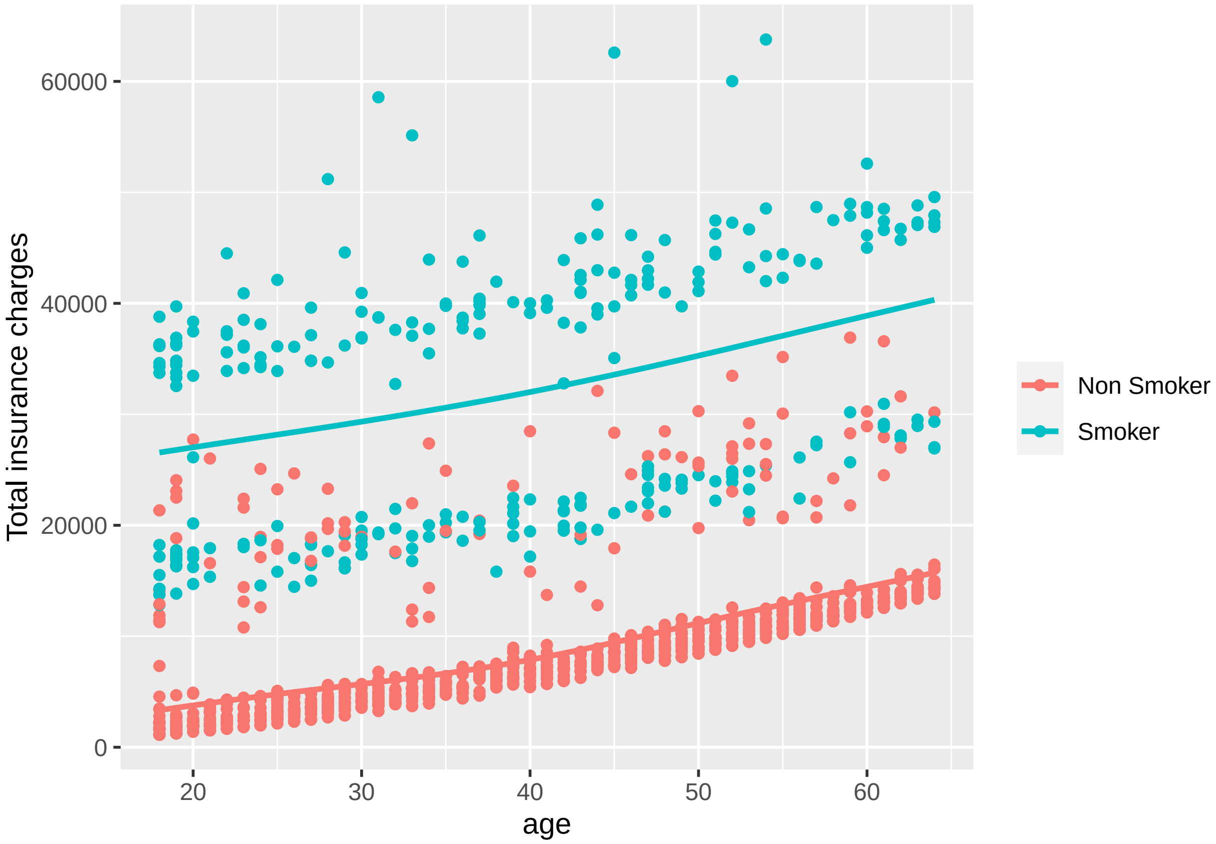
```
## # A tibble: 47 x 2
##      age avg_insurance_cost
##    <dbl>              <dbl>
##  1    64             23276.
##  2    61             22024.
##  3    60             21979.
##  4    63             19885.
##  5    43             19267.
##  6    62             19164.
##  7    59             18896.
##  8    54             18759.
##  9    52             18256.
## 10    37             18020.
## # ... with 37 more rows
```

```
# Insurance premium increases with age and vice versa

# Within our data set at what age group we have most and least smokers or what's
# the overall distribution of smokers by age
ggplot(data = ins_data_select, aes(x = age)) +
  geom_histogram(binwidth = 5, aes(fill = smoker), colour = "Black") +
  scale_fill_discrete(name = "", labels = c("Non Smoker", "Smoker")) +
  ylab("Number of observations")
```



```r
# We can observe that most smokers are between the age of 18 and 23 closely followed
# by those between 42 and 47. Least smokers are people who are above 62+ years.
# We also have highest number of young people in the data set (aged between 18 and 23).
ggplot(data = ins_data_select, aes(x = age, y = charges, colour = smoker)) +
  geom_point() + geom_smooth(fill=NA) +
  scale_color_discrete(name = "", labels = c("Non Smoker", "Smoker")) +
  ylab("Total insurance charges")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
# Irrespective of age, smokers are almost always paying higher insurance cost.
```

```
# check the average cost of health insurance in US by gender
library(dplyr)
ins_data_select %>%
  group_by(sex) %>% summarise(avg_insurance_cost = mean(charges)) %>%
  arrange(desc(avg_insurance_cost))
```

b). What is the average cost of health insurance in US by gender?

```
## # A tibble: 2 x 2
##   sex   avg_insurance_cost
##   <fct>              <dbl>
## 1 2                 13957.
## 2 1                 12570.
```

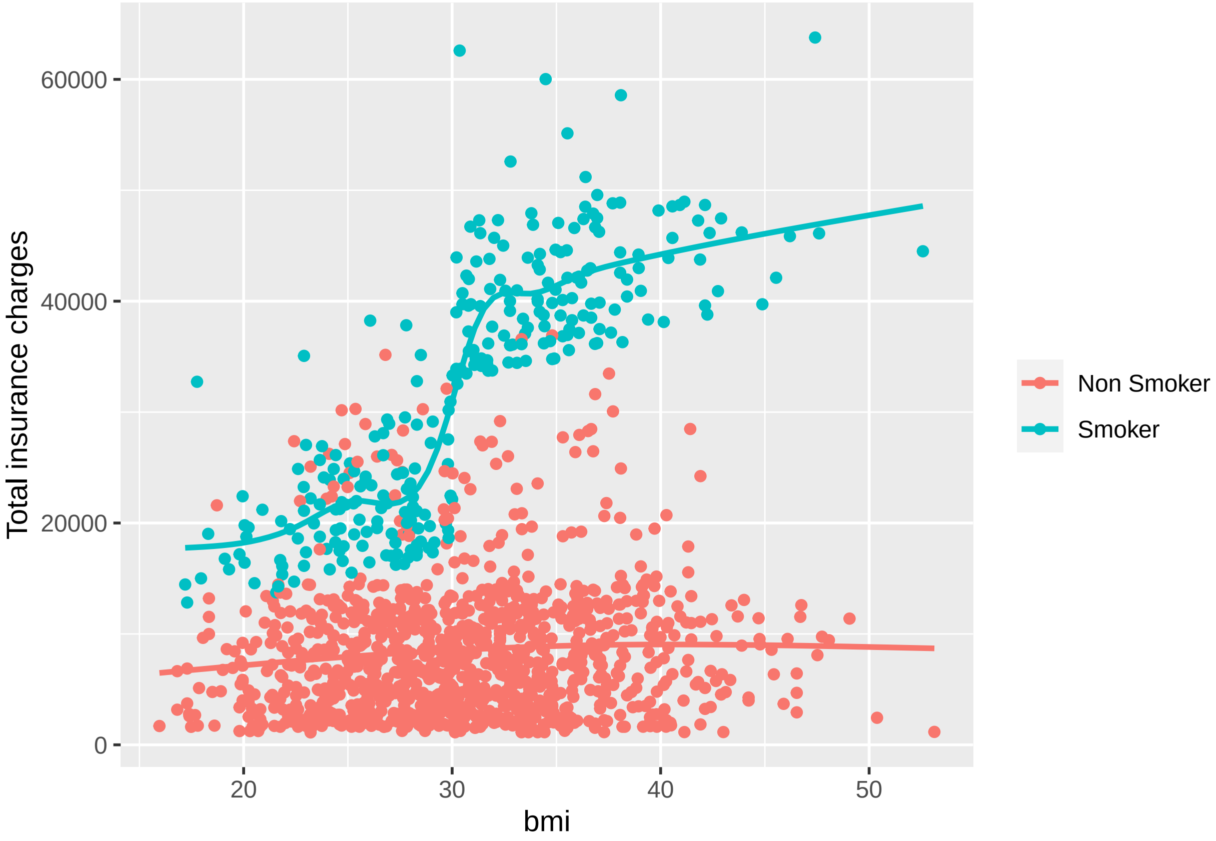
```
# According to the data set men seem to pay little more than women on an average.
```

c). What is the effect on health insurance cost by variation in BMI?

```
ggplot(data = ins_data_select, aes(x = bmi, y = charges, colour = smoker)) +
  geom_point() + geom_smooth(fill = NA) +
  scale_color_discrete(name = "", labels = c("Non Smoker", "Smoker")) +
  ylab("Total insurance charges")
```

We observed that bmi effects charge but not very much until being a smoker comes into mix. High bmi with smoking habbit definitely increases insurance cost. Bmi independently shares only 4% (coefficient of determination, R^2 = 0.2*0.2 = 0.04) of variability in the insurance charges.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



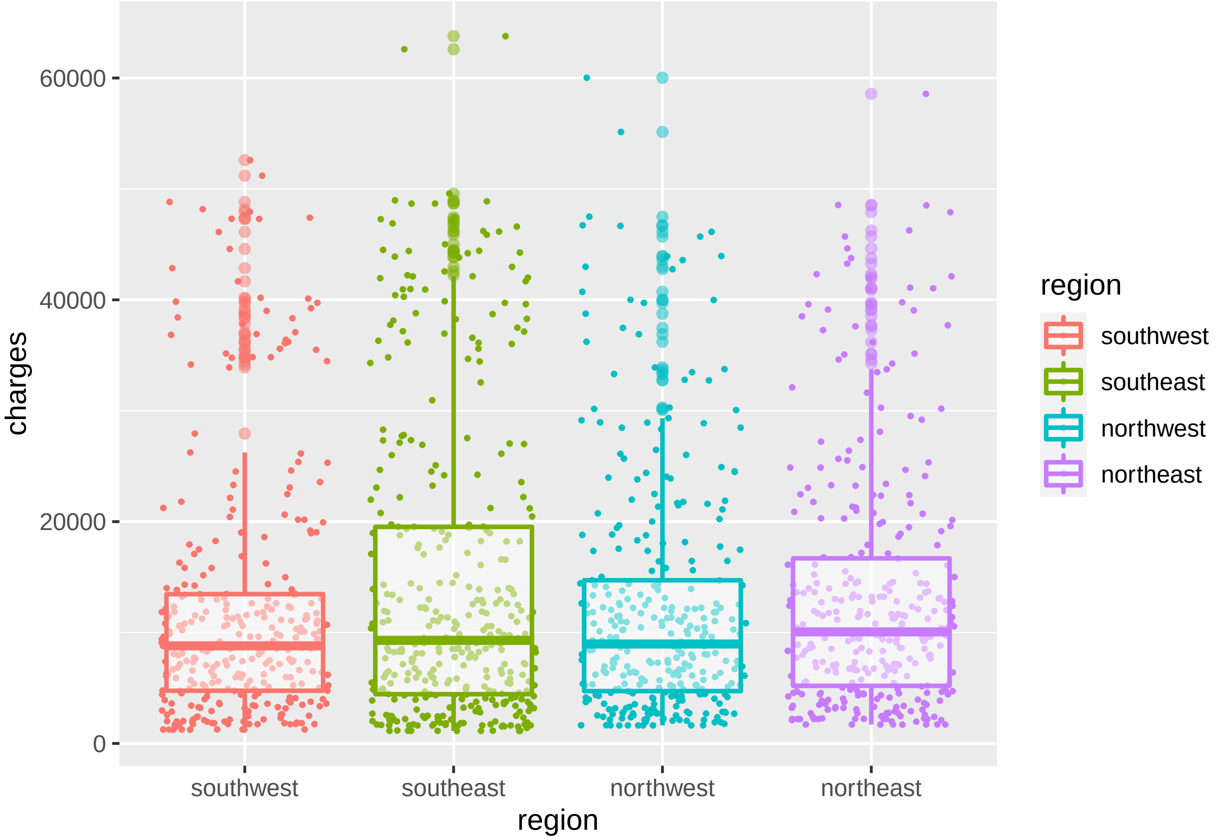
```
# check the average cost of health insurance in US by number of dependents
library(dplyr)
ins_data_select %>%
  group_by(children) %>% summarise(avg_insurance_cost = mean(charges)) %>%
  arrange(desc(avg_insurance_cost))
```

d). What is the effect on health insurance cost by number of dependents?

```
## # A tibble: 6 x 2
##   children avg_insurance_cost
##      <dbl>              <dbl>
## 1        3             15355.
## 2        2             15074.
## 3        4             13851.
## 4        1             12731.
## 5        0             12366.
## 6        5              8786.
```

```
# There appears to be a rise in insurance cost till 3 dependents and after that there
# is a drop. Needs more analysis i.e. counts of observations we have, other
# factors influencing etc.
# health insurance cost variation by region
ggplot(data = ins_data_select, aes(x= region, y = charges, colour = region)) +
  geom_jitter(size = 0.5) +
  geom_boxplot(size = 0.8, alpha = 0.5) +
  scale_color_discrete(name = "region",
                       labels = c("southwest", "southeast", "northwest", "northeast")) +
  scale_x_discrete(labels = c("southwest", "southeast", "northwest", "northeast"))
```

e). Why is the average health insurance cost varies in different US regions?



```
# variance in charges for region 2 population (southeast) and region 4 (northeast)
# population are higher. They have higher medians too in comparison to other
# regions which probability of person in region 2 and region 4 of paying higher
# for insurance is comparatively higher than other regions.


# check the average cost of health insurance in US by regions
library(dplyr)
ins_data_select %>%
  group_by(region) %>% summarise(avg_insurance_cost = mean(charges)) %>%
  arrange(desc(avg_insurance_cost))
```

```
## # A tibble: 4 x 2
##   region avg_insurance_cost
##   <fct>               <dbl>
## 1 2                  14735.
## 2 4                  13406.
## 3 3                  12418.
## 4 1                  12347.
```

```
# From correlation analysis we southeast (region 2) have more cases of bmi and
# smokers and thus we see higher average insurance cost.
# Lets check total number of smokers by region and also total number of people with
# bmi > average(bmi)
# derive is_smoker
ins_data_select$is_smoker <- ifelse(ins_data_select$smoker == 1, 1, 0)
ins_data_select %>%
  group_by(region) %>% summarise(number_of_smokers = sum(is_smoker)) %>%
  arrange(desc(number_of_smokers))
```

```
## # A tibble: 4 x 2
##   region number_of_smokers
##   <fct>              <dbl>
## 1 2                     91
## 2 4                     67
## 3 1                     58
## 4 3                     58
```

```
# we can see that southeast (region 2) and northeast (region 4) have most smokers
# per data set
# derive is_more_avg_bmi
ins_data_select$is_more_avg_bmi <- ifelse(ins_data_select$bmi > mean(ins_data_select$bmi), 1, 0)
ins_data_select %>%
  group_by(region) %>% summarise(number_of_people_more_than_avg_bmi = sum(is_more_avg_bmi)) %>%
  arrange(desc(number_of_people_more_than_avg_bmi))
```

```
## # A tibble: 4 x 2
##   region number_of_people_more_than_avg_bmi
##   <fct>                               <dbl>
## 1 2                                     235
## 2 1                                     155
## 3 4                                     131
## 4 3                                     125
```

```
# We see that in data we have most people having high bmi > average bmi are in
# southeast (region 2)
# With smokers and high bmi it thus proves the observation that insurance charges
# are high in southeast (region 2)
```

f).1. Predict the health insurance cost for given gender, age, BMI,

```
# performing multiple linear regression
# splitting the data into training and test set
library(caTools)
set.seed(123)
split = sample.split(ins_data_select$charges, SplitRatio = 0.8)
training_set = subset(ins_data_select, split == TRUE)
test_set = subset(ins_data_select, split == FALSE)

# feature scaling is not required as we will lm() which takes care of it
```

```r
# creating linear regression model object on whole data to better gauge the
# best predictors
regressor = lm(formula = charges ~ .,
               data = ins_data_select)

# check the summary of linear regression
summary(regressor)
```

number of dependents, smoking habits, region etc.

```
## 
## Call:
## lm(formula = charges ~ ., data = ins_data_select)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -11855  -3354   -373   1389  30949 
## 
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -9605.89    1343.00  -7.153 1.40e-12 ***
## age                255.47      11.85  21.565  < 2e-16 ***
## sex2              -138.14     331.33  -0.417 0.676803    
## bmi                201.23      46.52   4.326 1.64e-05 ***
## children           472.56     137.13   3.446 0.000587 ***
## smoker1          23827.79     411.18  57.950  < 2e-16 ***
## region2            -45.14     468.41  -0.096 0.923246    
## region3            606.33     474.88   1.277 0.201889    
## region4            915.05     475.75   1.923 0.054647 .  
## is_smoker              NA         NA      NA       NA    
## is_more_avg_bmi   2083.68     555.81   3.749 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6033 on 1328 degrees of freedom
## Multiple R-squared:  0.7535, Adjusted R-squared:  0.7519 
## F-statistic: 451.1 on 9 and 1328 DF,  p-value: < 2.2e-16
```

```r
# Looking at the coefficients and applying backward elimination of predictors that
# are not significant
# sex has t-value of -0.394 with P-value of 69% way more than threshold of 5%
# region seems overall insignificant as well with P-values of 0.87, 0.20,
# and 0.044 (significant but overall insignificant as this is just dummy
# variable representing partial data)
# recreating the model after removing sex and region with training_set
regressor_2 = lm(formula = charges ~ age + bmi + children + smoker,
               data = training_set)

# check the summary of linear regression
summary(regressor_2)
```

```
## 
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = training_set)
## 
```

```
## Residuals:
##    Min     1Q Median     3Q    Max 
## -11995  -2901  -1072   1286  29539 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -11487.71    1057.75 -10.861  < 2e-16 ***
## age            249.61      13.37  18.669  < 2e-16 ***
## bmi            315.61      30.90  10.215  < 2e-16 ***
## children       471.44     157.56   2.992  0.00283 ** 
## smoker1      23727.27     467.45  50.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6091 on 1065 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7421 
## F-statistic: 770.2 on 4 and 1065 DF,  p-value: < 2.2e-16
```

```
# We can see R-squared has not changed much but all predictor variables are
# now significant with P-values well under threshold of 0.05 and model's F-statistic
# has almost doubled. Clearly this one is a better model.

# predicting on test set using regressor_2 (tuned model)
# creating test set with only age, bmi, children, and smoker
test_set$pred_charges <- predict(regressor_2, newdata = test_set)

# comparing predicted values with actual values
head(test_set, 20)
```

```
##    age sex  bmi children smoker region   charges is_smoker is_more_avg_bmi
## 4   56   2 40.3        0      0      1 10602.385         0               1
## 5   30   2 35.3        0      1      1 36837.467         1               1
## 8   22   2 35.6        0      1      1 35585.576         1               1
## 11  26   2 20.8        0      0      1  2302.300         0               0
## 16  61   1 39.1        2      0      1 14235.072         0               1
## 20  56   1 27.2        0      0      1 11073.176         0               0
## 21  64   1 31.3        2      1      1 47291.055         1               1
## 24  41   1 31.6        0      0      1  6186.127         0               1
## 31  52   1 37.4        0      0      1  9634.538         0               1
## 32  38   2 34.7        2      0      1  6082.405         0               1
## 34  19   2 34.1        0      0      1  1261.442         0               1
## 50  63   1 31.8        0      0      1 13880.949         0               1
## 53  41   1 37.1        2      0      1  7371.772         0               1
## 59  55   1 26.8        1      0      1 35160.135         0               0
## 65  45   2 30.2        1      0      1  7441.053         0               0
## 67  22   1 24.3        0      0      1  2150.469         0               0
## 68  52   1 31.2        0      0      1  9625.920         0               1
## 69  28   1 33.4        0      0      1  3172.018         0               1
## 87  19   1 21.7        0      1      1 13844.506         1               0
## 88  21   1 26.4        1      0      1  2597.779         0               0
##    pred_charges
## 4     15209.372
## 5     30868.786
## 8     28966.602
## 11     1566.761
## 16    17021.573
## 20    11074.903
## 21    39035.924
## 24     8719.455
## 31    13295.674
## 32     9891.906
## 34     4017.094
## 50    14273.959
## 53    11398.190
## 59    11170.496
## 65     9747.482
## 67     1672.957
## 68    11338.903
## 69     6042.642
## 87    23830.822
## 88     2557.571
```

```
# adding case level residual/outlier and influential stats for multi regression model
ins_data_select$residuals <- resid(regressor)
ins_data_select$standardized.residuals <- rstandard(regressor)
ins_data_select$studentized.residuals <- rstudent(regressor)
ins_data_select$cooks.distance <- cooks.distance(regressor)
ins_data_select$dfbeta <- dfbeta(regressor)
ins_data_select$dffit <- dffits(regressor)
ins_data_select$leverage <- hatvalues(regressor)
ins_data_select$covariance.ratios <- covratio(regressor)

# writing the saved stats for each case into a table
write.table(ins_data_select, "Insurance cost diagnostic with Diagnostics.dat", sep = "\t", row.names =
# nrow(ins_data_select) -- 1338

# check if about 5% of cases (<= 67 cases) have standardized residual within +-2.
sum(ins_data_select$standardized.residuals > 2 | ins_data_select$standardized.residuals < -2)
```

f).2. Do case analysis, detect outliers and influencial cases

```
## [1] 67
```

```
# There are 70 cases that are outside range or have large residuals, we are about the range of 6% outli

# To exactly identify outliers we can add a variable called large.residual in the data frame to save th
ins_data_select$large.residual <- ins_data_select$standardized.residuals > 2 | ins_data_select$standardi
# we can now select the outlier cases by select rows with large.residual = TRUE
# ins_data_select[order(-ins_data_select$large.residual),]
# As we only have 3 cases outside 5% range model is fairly accurate

# check how many cases have standard residuals > 3 which may be we can investigate further
sum(ins_data_select$standardized.residuals > 3 | ins_data_select$standardized.residuals < -3)
```

```
## [1] 29
```

```
# create a variable to flag cases with very large residual
ins_data_select$very.large.residual <- ins_data_select$standardized.residuals > 3 | ins_data_select$sta
# 28 cases out of 1338 -- about 2%
# Let's look at the leverage (hat value), cook's distance, and covariance ratio
# for cases with large.residual = TRUE
# ins_data_select[ins_data_select$large.residual, c("cooks.distance","leverage","covariance.ratios")]

# check if any outlier cases have cook's distance > 1
sum(ins_data_select[ins_data_select$large.residual, c("cooks.distance")] > 1)
```

```
## [1] 0
```

```
# None of the cases have cooks distance > 1, so none of the cases have undue
# influence on the model

# calculate average leverage using formula = (k+1/n)
avg_leverage <- (6+1)/1338
# three times average leverage
times_3_leverage <- avg_leverage*3
# check if there are outlier cases with levarage > 3 times the average leverage
sum(ins_data_select[ins_data_select$large.residual, c("levarage")] > times_3_leverage)
```

```
## [1] 0
```

```
# There are none
```

g). What is the effect on health insurance cost by change in smoking habits?

There is good positive correlation between being a smoker and paying higher charges for insurance as shown in the correlation matrix. By changing smoking habbits, definitely insurance cost may reduce.

h). What US region has most obese cases?

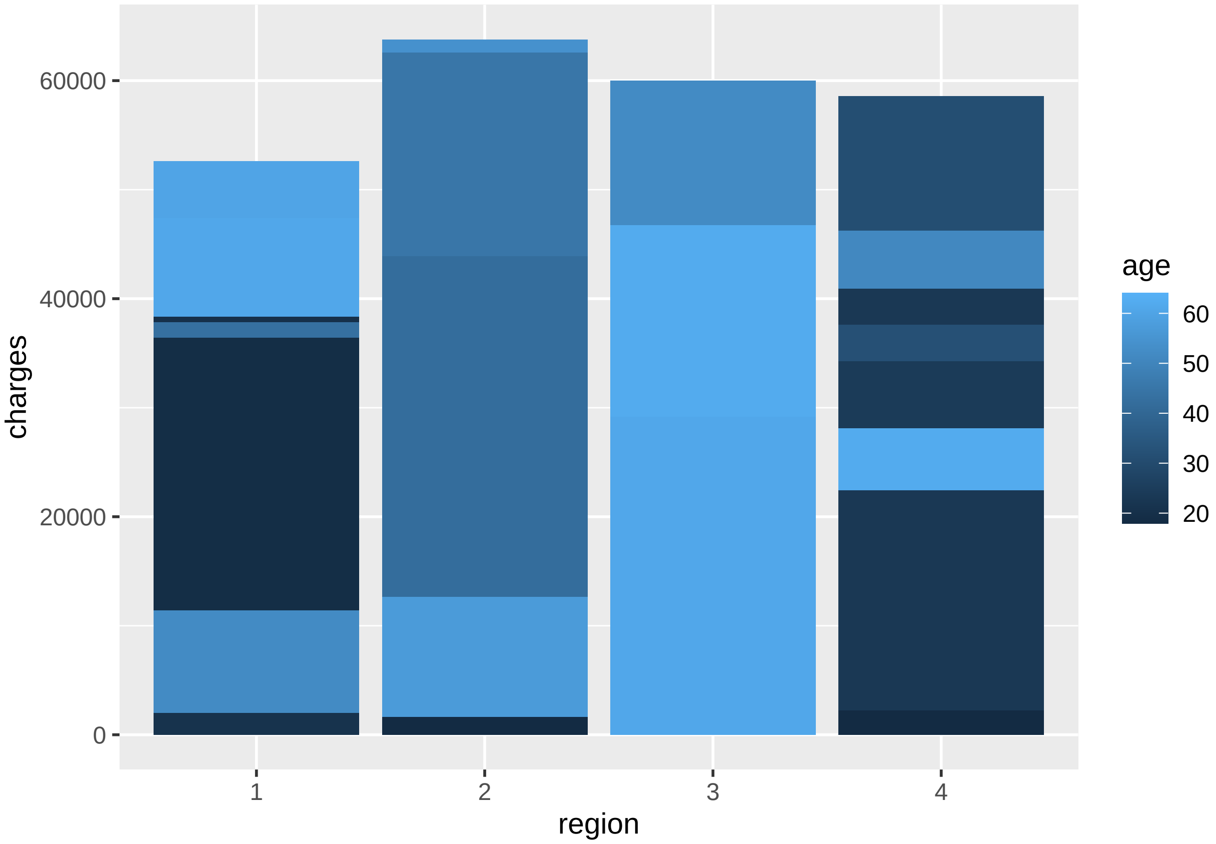
We saw that southeast (region 2) has most cases of obese or people with bmi > average bmi across US. This observation is just based on the data at hand.

i). What US region has most smokers?

We observed that southeast (region 2) has most smokers followed closely by northeast (region 4) per the data.

```

ggplot(ins_data_select, aes(fill=age, y=charges, x=region)) +
  geom_bar(position="dodge", stat="identity")
```

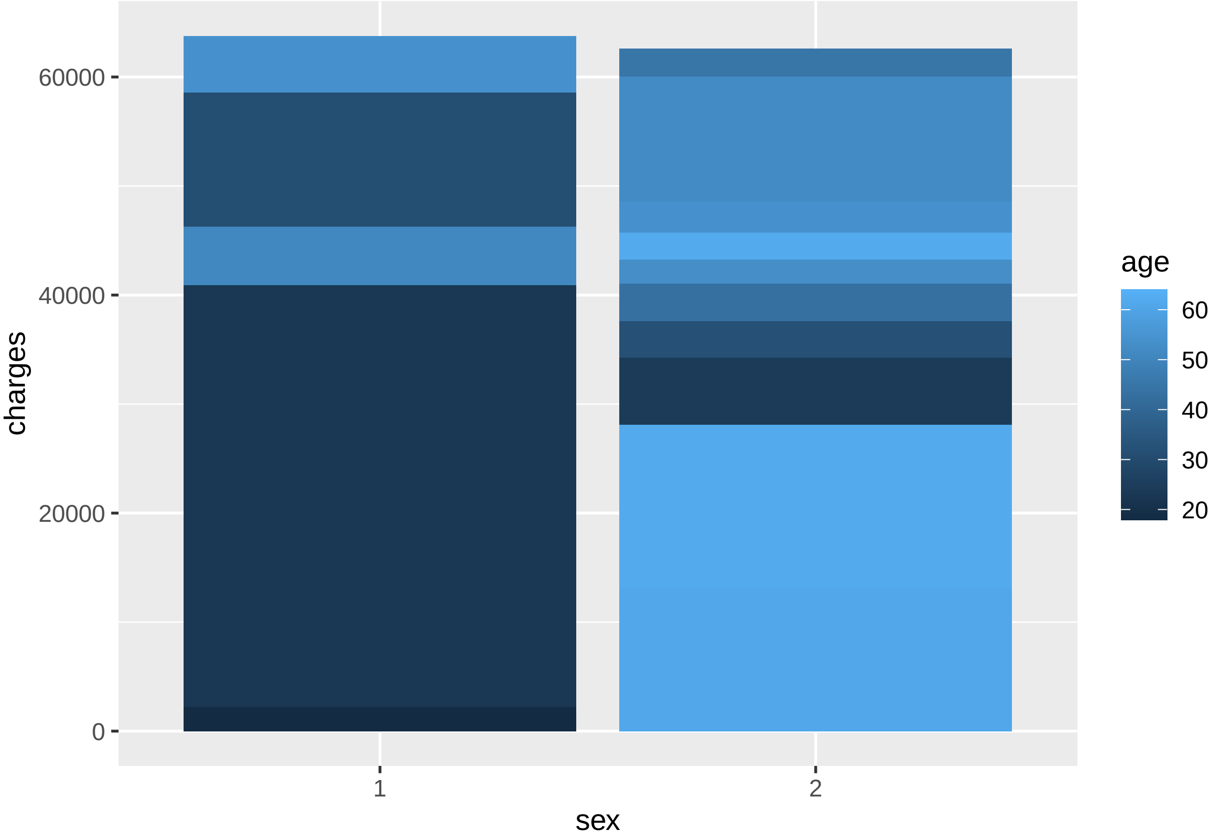
j). How is distribution of age groups amongst different US regions?



```
# We can see that most young population per data set at hand is in region 1
# (southwest) followed by region 4 (northeast). Sounds like CA and NY :).
# Region 3 (northwest) seems to have most elderly population i.e. Washington,
# Oregon, Montana, Wyoming, Idaho etc.
```

```
ggplot(ins_data_select, aes(fill=age, y=charges, x=sex)) +
  geom_bar(position="dodge", stat="identity")
```

k). How is distribution of gender amongst different age groups and in different US regions?



```
# In the data set we seem to have lot of young women while more elderly men.
```

l). Is there any correlation between BMI and smokers?

Based on correlation matrix table that I created earlier it seems there is absolutely no correlation between bmi and being a smoker.

```
# derive has_dependent
ins_data_select$has_dependent <- ifelse(ins_data_select$children > 0,1,0)
ins_data_select %>%
  group_by(region) %>% summarise(number_of_families_with_dependents = sum(has_dependent)) %>%
  arrange(desc(number_of_families_with_dependents))
```

m). Do we see spike in any region in number of dependents?

```
## # A tibble: 4 x 2
##   region number_of_families_with_dependents
##   <fct>                               <dbl>
## 1 2                                     207
## 2 3                                     193
## 3 1                                     187
## 4 4                                     177
```

```
# We can observe that most number of families with dependents are also in
# region 2 (southeast) per data.
```

n). What other factors can effect the insurance cost?

Other factors that can effect the insurance cost is absolutely the plan and coverage/benefits a person is signing up for. Most insurance companies run multi policy discounts which could be a factor. Any pre-condition and family history of known disease can also effect insurance cost.

```
ins_data_select %>% summarise(mean(charges))
```

o). How much on an average American's pay for health insurance?

```
##   mean(charges)
## 1      13270.42
```

```
# On an average an American family is paying 13,270.42$ against insurance cost
```

p). Which US region have most people with BMI less than average?

It is observed above that region 3 (northwest) has least number of people with bmi greater than average BMI and thus reflect lowest obese population.

Questions for future steps?

1. I would love to see if the given data can be combined with some other data to get more predictors i.e. patient's pre-condition or medical history data.

2. I also want to learn about other regression techniques i.e. polynomial regression, support vector rgression, random forest regression etc.

3. I can also look into creation of logistic regression model to predict risky customer i.e. probability of customers to make more claims in the future.

Final Project Step 3 - summarization and narative

Introduction:

This is an attemp to understand and uncover patterns in buying behaviour of medical insurance amongst population sample considering their attributes i.e. age, gender, bmi, number of dependents, smoking habit etc across US. Research also suggests utilizing predictive analytics and particular machine learning technique to predict the cost of medical insurance when several attributes of a person is known.

Problem Statement:

To anlyze and identify major factors effecting the cost of medical insurance in United States and predict individual cost of medical insurance given several known factors at hand.

Steps taken to address problem statement:

1. Data research and collection - Reasearched for appropriate data on various websites. Obtained medical insurance cost data in four US regions i.e. southeast, southwest, northeast, and northwest from Kaggle website. Each row of the data represents a person and his/her attributes and how much he/she pays for the medical insurance. There is one file each for each US region.

2. Data preparation and cleansing - Identified that there is no missing data or NAs present in the dataset. Filtered out irrelavant attributes to research i.e. index variable x. Converted misinterpreted data types i.e. character for categorical variables to factor. This is done for age, sex, smoker, and region. Final data set ready for reserach only had numerical or factor variables.

3. EDA (Exploratory data analysis) - Checked correlation between variables and noted the strength and weeknesses of relationships. Found that being a smoker is strong indicator of paying higher medical insurance cost. Age and BMI also showed decent positive correlation with cost of insurance. We also made some discoveries which seems specific to data at hand and may not be generalized i.e. Men appeared almost always paying more than women, smokers paid almost double when their bmi was $>$ 30 when compared to smokers with bmi $<$ 30 etc.

4. EDA to answer research questions - Summarized data and visualized it to explore the hidden relations and behaviors. I observed following -

a. Cost of insurance tops after age of 60. Cost increases in general with age.

b. Per dataset most smoker's are 18 to 23 years old followed by those in mid 40s. Smoking habit decreases with age.

c. Irrespective of age, smoker's always pay higher insurance cost.

d. Per dataset men seems to pay little higher on an average than women.

e. BMI does have a small effect on charge i.e. as BMI increases insurance charges also increase (only about 4%). When considered together with smoking habit it has greater effect. People having BMI $>$ 30 and smoking habit seems to pay almost double than people with similar BMI but are non-smoking.

f. Number of dependents does not have a significant effect on insurance cost.

g. Based on dataset, on an average insurance cost in southeast and northeast are higher.

h. Again based on data at hand, southeast region has most people with higher BMI ($>$ average BMI of population) and smoking habits (highest among all regions) which drove the avreage insurance cost highest in the region. Northeast is next in the row.

i. Most young population is in southwest followed by northeast. Northwest seems to have most elderly population.

j. Dataset contains larger population of young women and larger population of older men.

5. Linear regression to predict insurance cost - To predict cost of insurance given age, gender, bmi, smoking habit, number of dependents, region etc we can use linear regression as we want to predict continuous numeric variable - "cost". I created model with all independent variables and then applied backward elimination to take out irrelevant variables and fine tune the model to make pretty accurate prediction of insurance cost. See page 20-22.

Also did case or residual analysis to check if model is biased. Found no cases with out of range leverage and cook's distance and thus model is not biased and pretty accurate.

Implications to target audience:

1. Insurance companies usually targets younger people as they have lesser chances of making a claim / getting sick. Analysis heps them choose a group of people who are young, have less than average BMI and do not smoke. Also targetting younger men may be more benefitial than women as women may claim for pregnancy related expenses etc. Per data set most young population is in southwest and in northeast (sounds like CA and NY, NJ).

2. Consumers can clearly understand the factors affecting the insurance rates and find ways to avoid higher medical insurance by - giving up smoking, keeping BMI in check, exercising etc. Analysis also helps them understand the overall trend of insurance rates i.e. it increases with age, increases drastically for smokers, not largely affected by number of dependents etc. Regional insurance expense trend can also be understood i.e. southeast and northeast seems to have highest insurance premiums because of higher obesity and concentration of smokers. People may consider this as a factor when making decision to settle down in particular US region.

3. Analysis shows that as age increases, BMI increases and specially for smokers insurance cost can spike high. National, state, and local healthcare bodies can learn from findings and trend to draft regulations around insurance cost to keep it affordable for larger audience. Also, then can use this data to target certain behaviors to change i.e. making insurance rates higher for smokers may temp them to quit smoking habit in order to save money and ofcourse better health.

Limitations of the analysis:

One limitation of the analysis is feature set in the data that helps predict the cost of insurance. Other missing important factors in data are pre-existing medical conditions of an individual, historical pattern of medical claims, regular medical expenses, drinking habit, and type/level of coverage opted for etc. These additional factors when added should have significant impact on medical insurance cost.

Conclusion/Remarks/References:

Data science and statistics understanding gained from the course has helped to undertake this analysis and data modeling exercise. Project has helped in building overall understanding of how to write and present data science solution to the end users.

References -

1. Discussion forums with class mates have helped strenthen the undertsanding of various concepts learned through the course.

2. Course books, online material research, programing practice etc gave learning backbone as well.