ACCUMEN DATA EXERCISE CODING

JIASHU MIAO

2019/1/20

```
library("readr")
require("readxl")
```

```
## Loading required package: readxl
```

```
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
```

```
library("ggplot2")
```

Understanding the Data

a. Are all the values in the data reasonable? Are there missing values?
b. What are the characteristics of employees at Company A? Do these demographics change over time? Use tables and charts to understand the data and demographic characteristics of employees at Company A.

```
data_original <- read_excel("/Users/MichaelMiao/Documents/career/Acumen_Data_Analysis
_Exercise.xlsx", sheet = "Data")
```

```
## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.
```

```
attach(data_original)
#View(data_original)
par(mfrow=c(2,5))
for (i in 1:length(data_original)) {
        boxplot(data_original[,i], main=names(data_original[i]), type="l")

}

# so there might be unreasonable variables in Age, Health Score
detach(data_original)
dataNEW1 <-  subset(data_original,`Health Score` >=0 & `Health Score` <=6)
range(data_original$Age)
```

```
## [1]   7 172
```

```
dataNEW2 <- subset(dataNEW1, dataNEW1$Age >= 18 & dataNEW1$Age <= 63 )
summary(dataNEW2)
```

```
##  Observation Number    Quarter        Employee Id      Sex (Male=1)
##  Min.   :    1      Min.   : 1.000   Min.   :   1.0   Min.   :0.0000
##  1st Qu.: 4784      1st Qu.: 5.000   1st Qu.: 499.0   1st Qu.:0.0000
##  Median : 9533      Median : 8.000   Median : 994.0   Median :1.0000
##  Mean   : 9527      Mean   : 7.329   Mean   : 995.4   Mean   :0.5034
##  3rd Qu.:14258      3rd Qu.:10.000   3rd Qu.:1491.0   3rd Qu.:1.0000
##  Max.   :19031      Max.   :12.000   Max.   :1993.0   Max.   :1.0000
##
##       Race            Age        Hospital Visit This Quarter (1=Yes)
##  Min.   :1.0    Min.   :18.00   Min.   :0.0000
##  1st Qu.:1.0    1st Qu.:26.31   1st Qu.:0.0000
##  Median :1.0    Median :28.53   Median :0.0000
##  Mean   :1.6    Mean   :30.45   Mean   :0.1082
##  3rd Qu.:2.0    3rd Qu.:32.34   3rd Qu.:0.0000
##  Max.   :3.0    Max.   :61.85   Max.   :1.0000
##  NA's   :1973
##      Salary       Health Score
##  Min.   :28351   Min.   :0.6266
##  1st Qu.:44538   1st Qu.:2.3047
##  Median :48178   Median :3.0200
##  Mean   :48269   Mean   :3.1436
##  3rd Qu.:51924   3rd Qu.:3.8816
##  Max.   :68826   Max.   :5.9981
##
```

```
dataNEW2 <- na.omit(dataNEW2)
colSums(is.na(dataNEW2))
```

```
##                   Observation Number                                Quarter
##                                    0                                      0
##                          Employee Id                          Sex (Male=1)
##                                    0                                      0
##                                 Race                                   Age
##                                    0                                      0
## Hospital Visit This Quarter (1=Yes)                                 Salary
##                                    0                                      0
##                         Health Score
##                                    0
```

```
# filter out the observations that is not in correct score range 0:6, some of them 10
!
# filter out the observations that the employ is beyond age 18 or over 62.
# dataNEW2 becomes the new data




# we could see there are 1973 missing values for the varible "Racec" at our new dataN
EW2.
#z we could remove them or impuate with other numbers or NULL.
quartermeanhealth <- aggregate(dataNEW2$`Health Score`,by = list(dataNEW2$Quarter), F
UN = mean)
quartermeanhealth <- as.data.frame(quartermeanhealth)
colnames(quartermeanhealth) <- c("Quarter","QuarterMeanHealthScore")
quartermeanhealth
```

```
##    Quarter QuarterMeanHealthScore
## 1        1               2.957046
## 2        2               3.064161
## 3        3               3.052074
## 4        4               3.080241
## 5        5               3.080552
## 6        6               3.120600
## 7        7               3.193676
## 8        8               3.134246
## 9        9               3.178046
## 10      10               3.148329
## 11      11               3.234421
## 12      12               3.300193
```

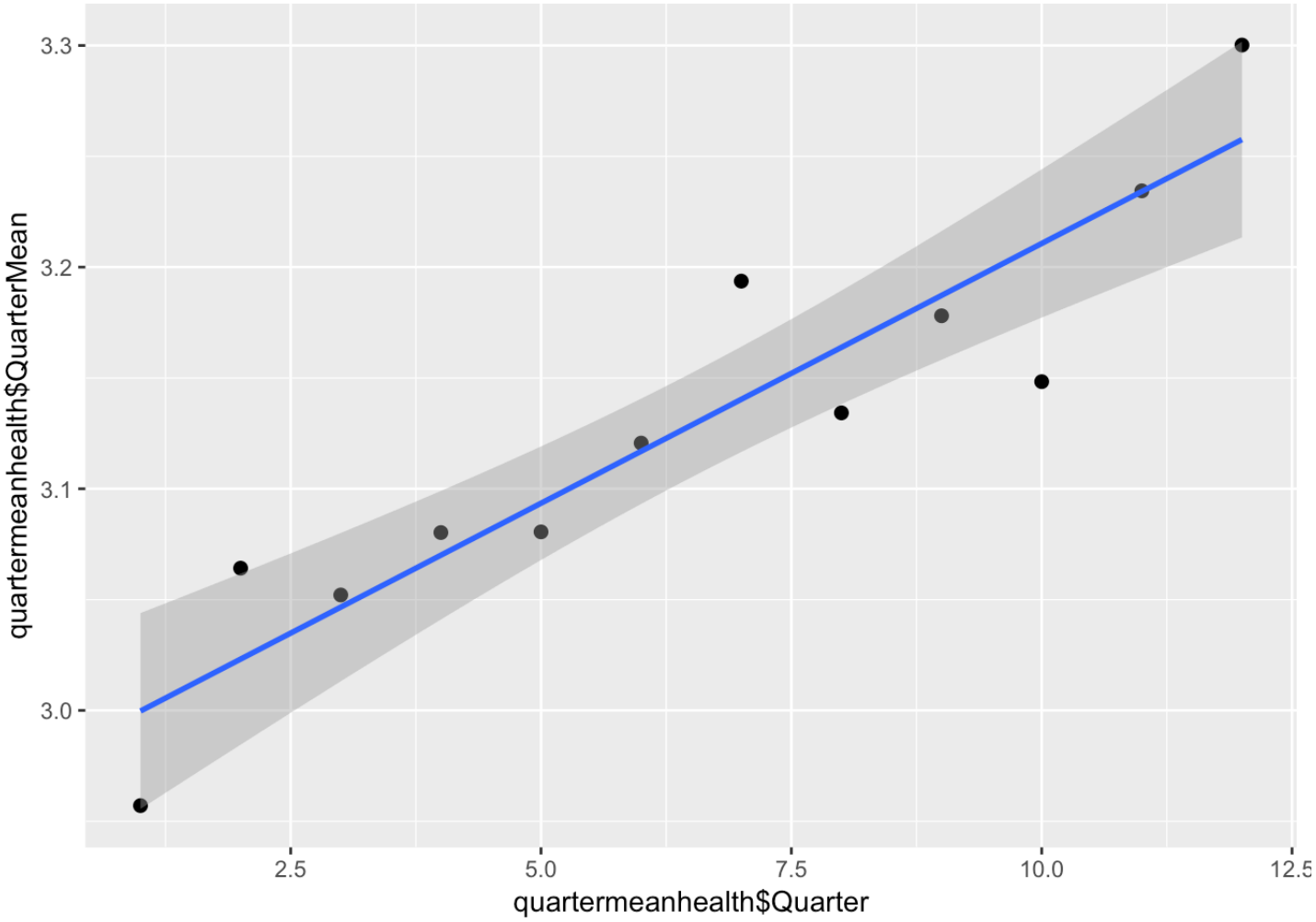
```
summary(dataNEW2)
```

```
##  Observation Number    Quarter        Employee Id      Sex (Male=1)
##  Min.   :    1      Min.   : 1.000   Min.   :   1.0   Min.   :0.0000
##  1st Qu.: 4900      1st Qu.: 5.000   1st Qu.: 510.2   1st Qu.:0.0000
##  Median : 9576      Median : 8.000   Median : 998.0   Median :1.0000
##  Mean   : 9538      Mean   : 7.331   Mean   : 996.4   Mean   :0.5062
##  3rd Qu.:14223      3rd Qu.:10.000   3rd Qu.:1487.0   3rd Qu.:1.0000
##  Max.   :19031      Max.   :12.000   Max.   :1993.0   Max.   :1.0000
##       Race           Age        Hospital Visit This Quarter (1=Yes)
##  Min.   :1.0   Min.   :18.00   Min.   :0.0000
##  1st Qu.:1.0   1st Qu.:26.29   1st Qu.:0.0000
##  Median :1.0   Median :28.53   Median :0.0000
##  Mean   :1.6   Mean   :30.41   Mean   :0.1081
##  3rd Qu.:2.0   3rd Qu.:32.33   3rd Qu.:0.0000
##  Max.   :3.0   Max.   :61.85   Max.   :1.0000
##      Salary       Health Score
##  Min.   :28351   Min.   :0.6266
##  1st Qu.:44630   1st Qu.:2.3062
##  Median :48328   Median :3.0283
##  Mean   :48399   Mean   :3.1482
##  3rd Qu.:52090   3rd Qu.:3.8908
##  Max.   :68826   Max.   :5.9981
```

```
ggplot(quartermeanhealth,aes(x=quartermeanhealth$Quarter,y=quartermeanhealth$QuarterM
ean))+geom_point(size=2)+geom_smooth(method = lm)


saqmean <- aggregate(dataNEW2$Salary, by = list(dataNEW2$Quarter), FUN = mean) %>% as
.data.frame()
colnames(saqmean) <- c("Quarter", "QuarterMeanSalaries")
saqmean
```

```
##    Quarter QuarterMeanSalaries
## 1        1            43723.67
## 2        2            44344.43
## 3        3            45117.32
## 4        4            45581.48
## 5        5            46235.16
## 6        6            47029.61
## 7        7            47934.42
## 8        8            48796.07
## 9        9            49661.03
## 10      10            50621.95
## 11      11            51547.64
## 12      12            52546.71
```

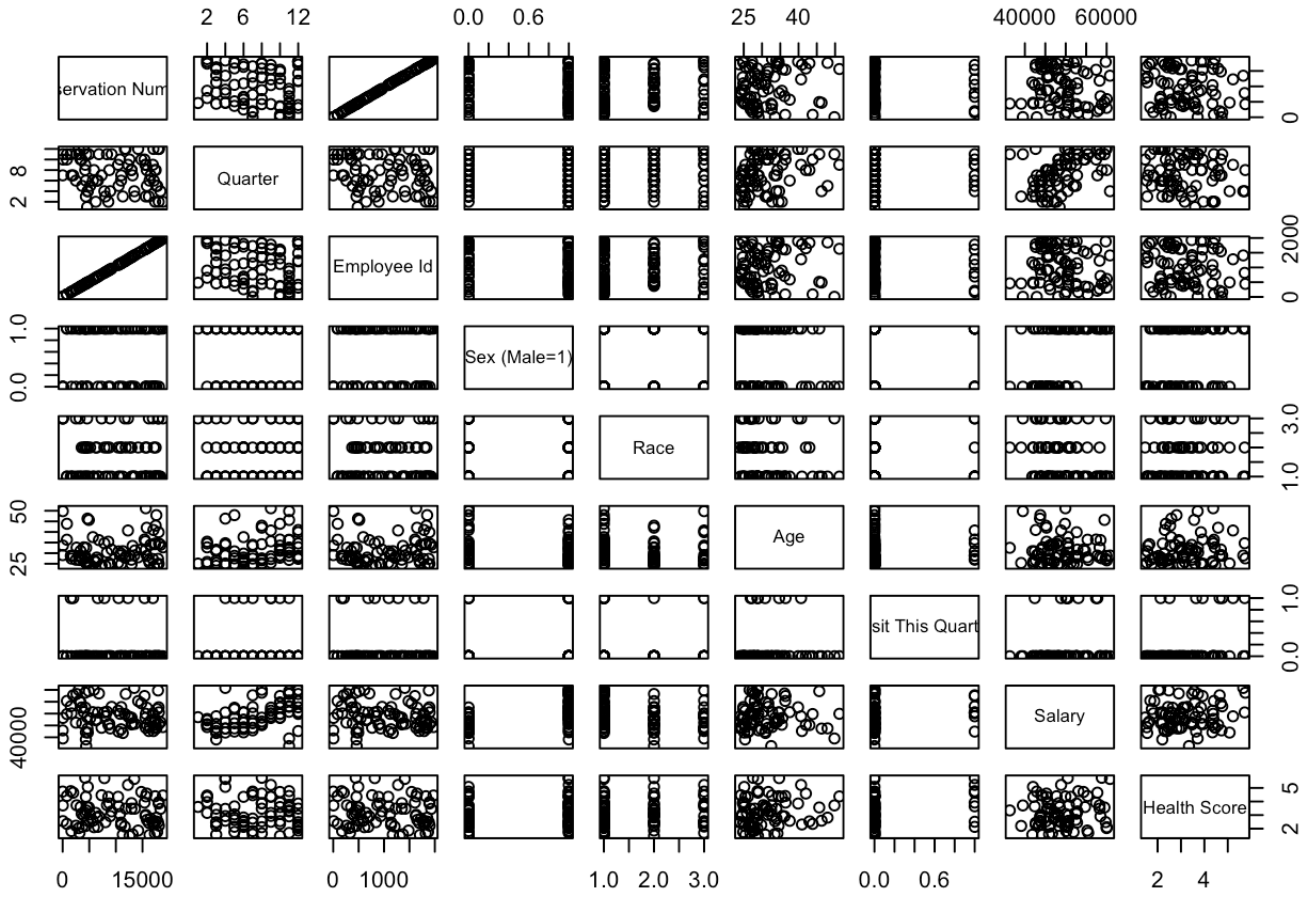
```
# We can see as the time goes by, which is when quartr goes from 1 to 12, the salaries increases and the health score increases based on the means from each quarters as we calculated above.
```



2. Exploring Relationships

a. Which characteristics are associated with the health score? Use tables and charts (suggestion: scatter plots for continuous variables) to determine which characteristics are associated with the health score.

```
datapair <- dataNEW2[sample(nrow(dataNEW2),80),] %>% as.data.frame()
pairs(datapair)
```




```
colSums(is.na(dataNEW2))
```

```
##                    Observation Number                               Quarter
##                                     0                                     0
##                           Employee Id                          Sex (Male=1)
##                                     0                                     0
##                                  Race                                   Age
##                                     0                                     0
## Hospital Visit This Quarter (1=Yes)                                Salary
##                                     0                                     0
##                          Health Score
##                                     0
```

```
dim(dataNEW2)
```

```
## [1] 15858     9
```

```
dataNEW2$`Sex (Male=1)` <- as.factor(dataNEW2$`Sex (Male=1)`)
dataNEW2$Race <- as.factor(dataNEW2$Race)
dataNEW2$`Hospital Visit This Quarter (1=Yes)` <- as.factor(dataNEW2$`Hospital Visit
This Quarter (1=Yes)`)
str(dataNEW2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    15858 obs. of  9 variables:
##  $ Observation Number                 : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Quarter                            : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Employee Id                        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex (Male=1)                       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 1 ...
##  $ Race                               : Factor w/ 3 levels "1","2","3": 3 3 3 3 3
3 3 3 3 3 ...
##  $ Age                                : num  27.3 27.8 28.1 28.3 28.6 ...
##  $ Hospital Visit This Quarter (1=Yes): Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
1 1 1 ...
##  $ Salary                             : num  36907 37907 38907 39907 40907 ...
##  $ Health Score                       : num  3.7 4.98 4.01 2.34 2.11 ...
##  - attr(*, "na.action")= 'omit' Named int  66 67 68 69 70 71 72 73 74 75 ...
##   ..- attr(*, "names")= chr  "66" "67" "68" "69" ...
```

```
model1 <- lm(data = dataNEW2,formula = dataNEW2$`Health Score` ~ dataNEW2$Quarter+dat
aNEW2$Age+dataNEW2$Salary+dataNEW2$`Sex (Male=1)`+dataNEW2$Race+dataNEW2$`Hospital Vi
sit This Quarter (1=Yes)`)
summary(model1)
```

```
##
## Call:
## lm(formula = dataNEW2$`Health Score` ~ dataNEW2$Quarter + dataNEW2$Age +
##     dataNEW2$Salary + dataNEW2$`Sex (Male=1)` + dataNEW2$Race +
##     dataNEW2$`Hospital Visit This Quarter (1=Yes)`, data = dataNEW2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1034 -0.8046 -0.1559  0.6660  3.1997
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                   1.745e+00  1.065e-01
## dataNEW2$Quarter                              4.296e-03  3.228e-03
## dataNEW2$Age                                  4.362e-02  1.340e-03
## dataNEW2$Salary                              -3.348e-06  2.248e-06
## dataNEW2$`Sex (Male=1)`1                      3.172e-01  1.991e-02
## dataNEW2$Race2                               -4.501e-02  1.976e-02
## dataNEW2$Race3                               -7.654e-02  2.423e-02
## dataNEW2$`Hospital Visit This Quarter (1=Yes)`1  6.607e-01  2.595e-02
##                                                t value Pr(>|t|)
## (Intercept)                                     16.386  < 2e-16 ***
## dataNEW2$Quarter                                 1.331  0.18328
## dataNEW2$Age                                    32.554  < 2e-16 ***
## dataNEW2$Salary                                 -1.489  0.13649
## dataNEW2$`Sex (Male=1)`1                        15.935  < 2e-16 ***
## dataNEW2$Race2                                  -2.278  0.02276 *
## dataNEW2$Race3                                  -3.159  0.00158 **
## dataNEW2$`Hospital Visit This Quarter (1=Yes)`1  25.459  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 15850 degrees of freedom
## Multiple R-squared:  0.1203, Adjusted R-squared:  0.1199
## F-statistic: 309.7 on 7 and 15850 DF,  p-value: < 2.2e-16
```

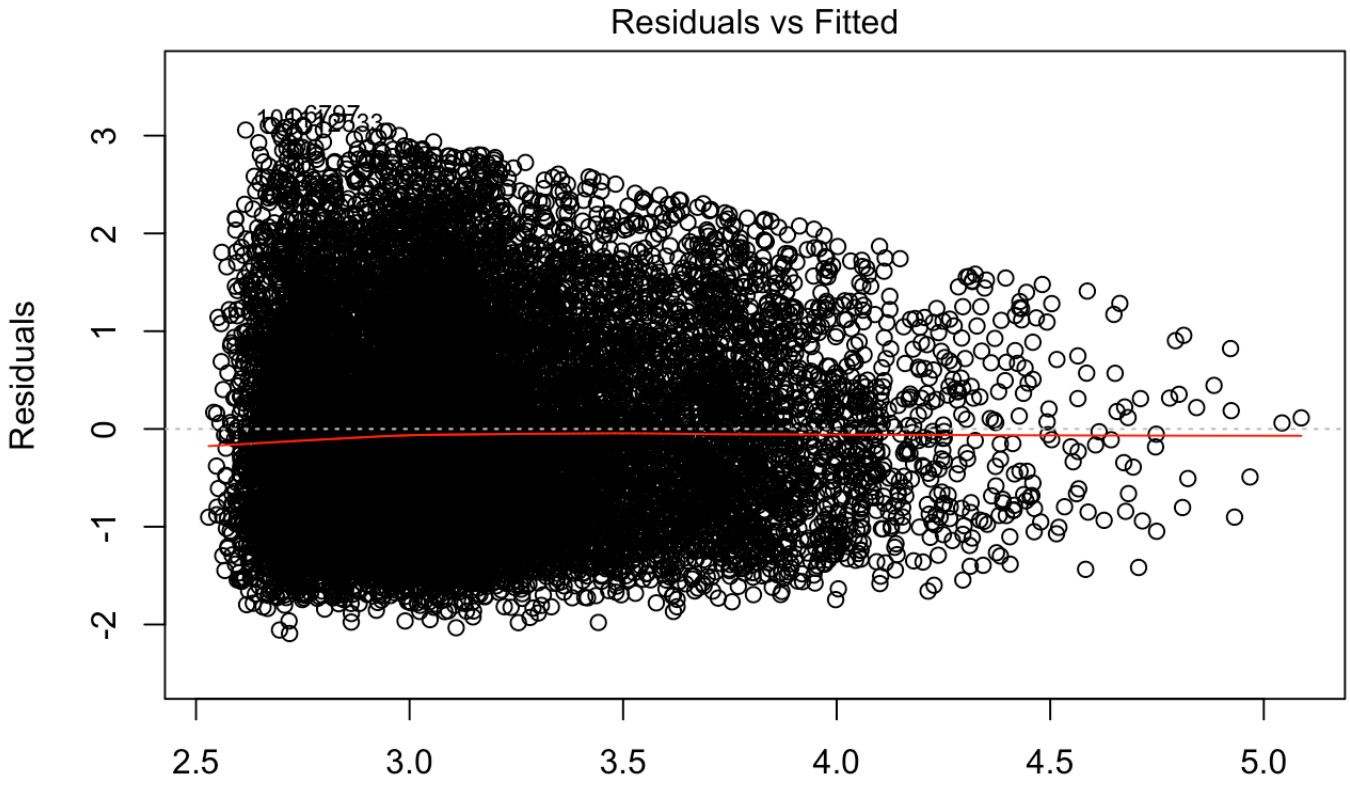
```
# Based on the plot Age, SEX, RACE, HOSPITAL VISITS matters with HEALTH SCORE.

model2 <- lm(dataNEW2$`Health Score`~ dataNEW2$Age+dataNEW2$`Sex (Male=1)`+dataNEW2$R
ace+dataNEW2$`Hospital Visit This Quarter (1=Yes)`)
summary(model2)
```

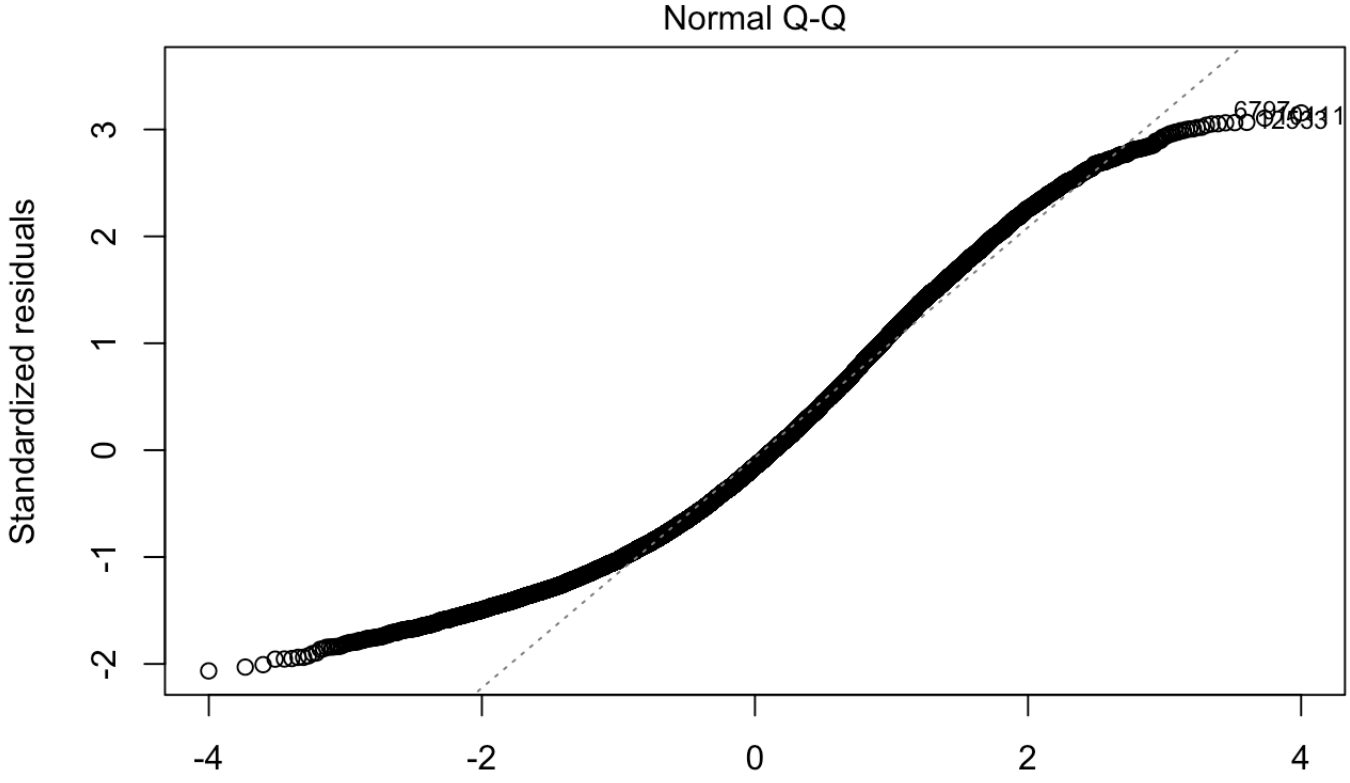
```
##
## Call:
## lm(formula = dataNEW2$`Health Score` ~ dataNEW2$Age + dataNEW2$`Sex (Male=1)` +
##     dataNEW2$Race + dataNEW2$`Hospital Visit This Quarter (1=Yes)`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0924 -0.8065 -0.1534  0.6656  3.1963
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                     1.605948   0.041929
## dataNEW2$Age                                    0.044031   0.001301
## dataNEW2$`Sex (Male=1)`1                        0.299862   0.016098
## dataNEW2$Race2                                 -0.034269   0.018401
## dataNEW2$Race3                                 -0.066131   0.023201
## dataNEW2$`Hospital Visit This Quarter (1=Yes)`1  0.661374   0.025911
##                                                 t value Pr(>|t|)
## (Intercept)                                      38.301  < 2e-16 ***
## dataNEW2$Age                                     33.852  < 2e-16 ***
## dataNEW2$`Sex (Male=1)`1                         18.627  < 2e-16 ***
## dataNEW2$Race2                                   -1.862  0.06257 .
## dataNEW2$Race3                                   -2.850  0.00437 **
## dataNEW2$`Hospital Visit This Quarter (1=Yes)`1  25.525  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 15852 degrees of freedom
## Multiple R-squared:  0.1202, Adjusted R-squared:  0.1199
## F-statistic:   433 on 5 and 15852 DF,  p-value: < 2.2e-16
```

```
par(mrow=c(2,1))
```

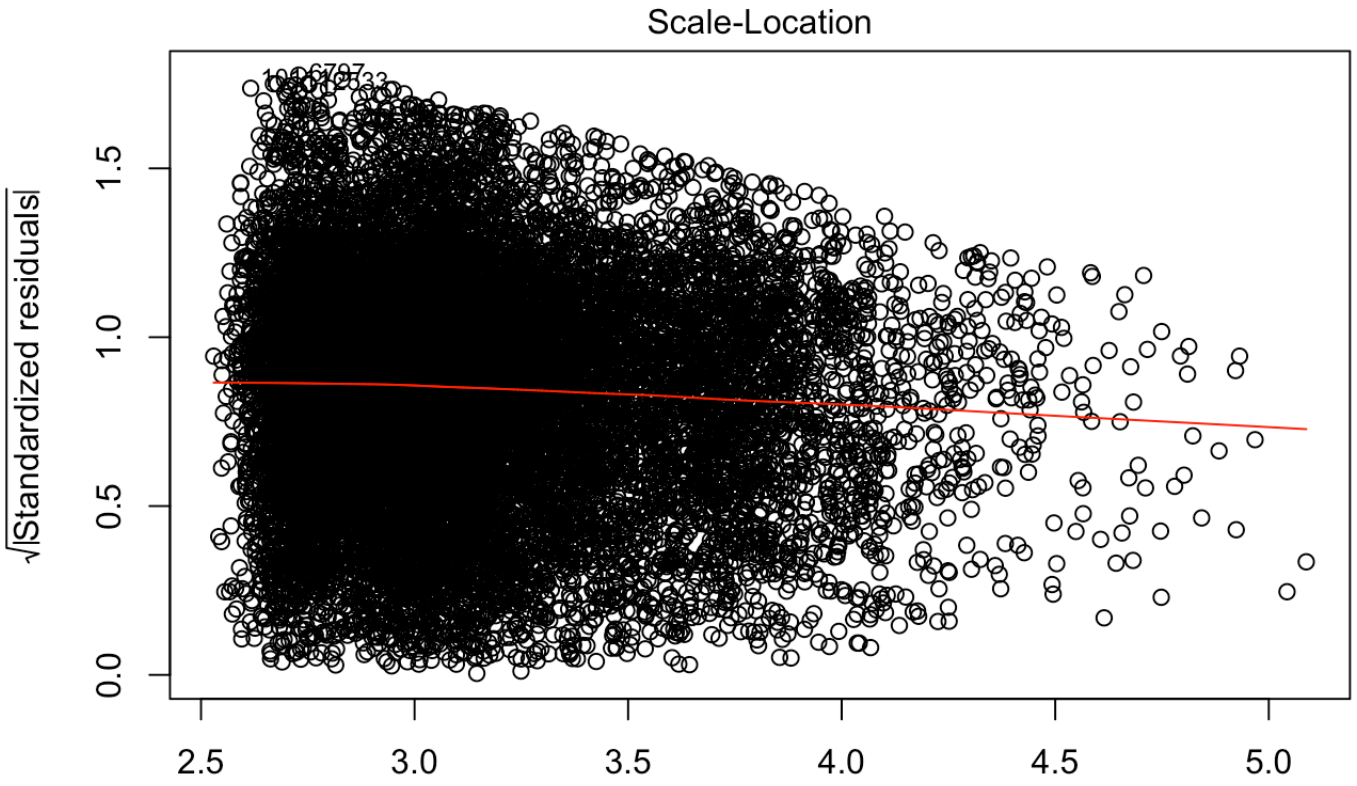
```
## Warning in par(mrow = c(2, 1)): "mrow"不是图形参数
```

```
plot(model2)
```

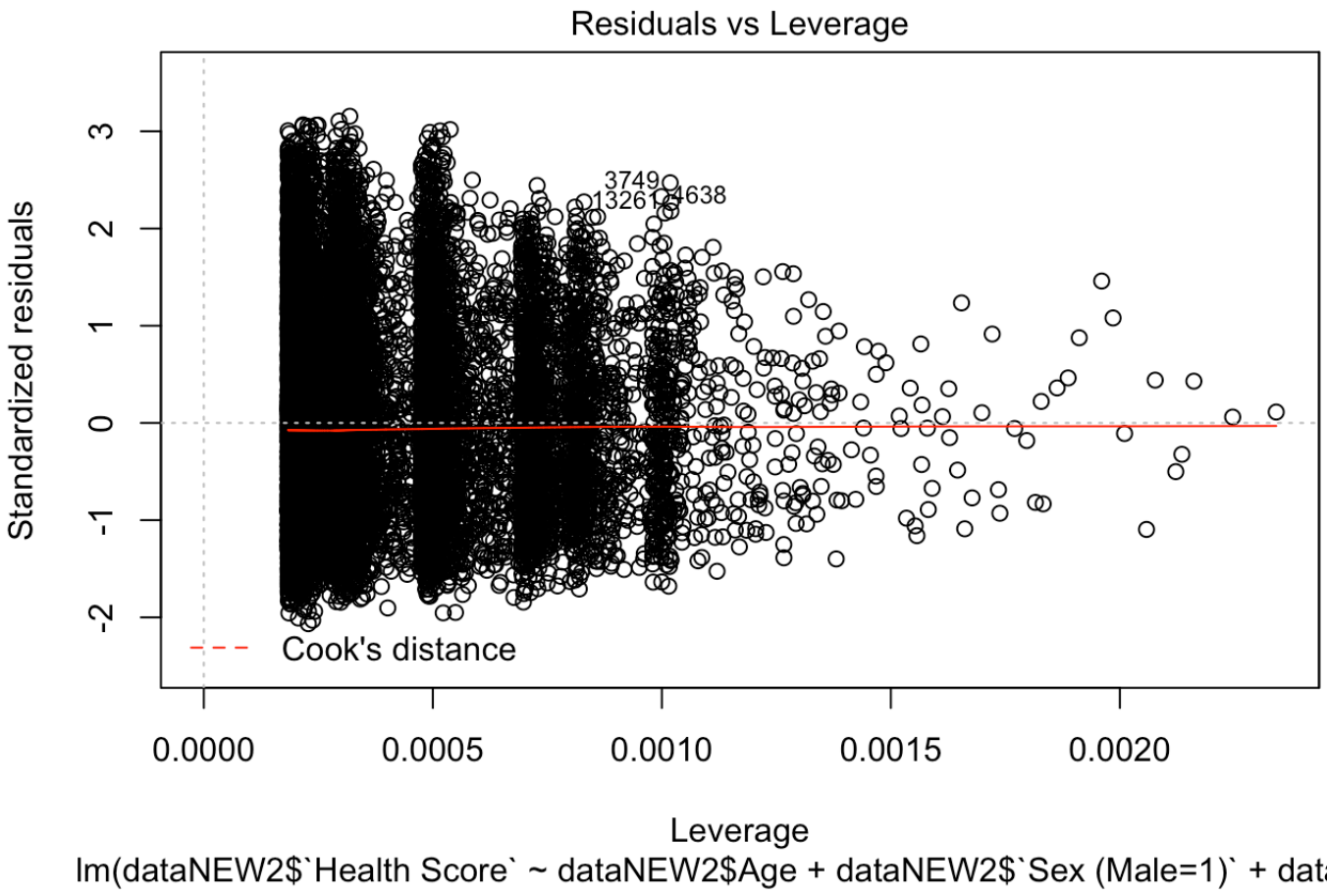

Residuals vs Fitted
Residuals
Fitted values
lm(dataNEW2$`Health Score` ~ dataNEW2$Age + dataNEW2$`Sex (Male=1)` + dataN ...


Normal Q-Q
Standardized residuals

Theoretical Quantiles
lm(dataNEW2$`Health Score` ~ dataNEW2$Age + dataNEW2$`Sex (Male=1)` + dataN ...



lm(dataNEW2$`Health Score` ~ dataNEW2$Age + dataNEW2$`Sex (Male=1)` + dataN ...

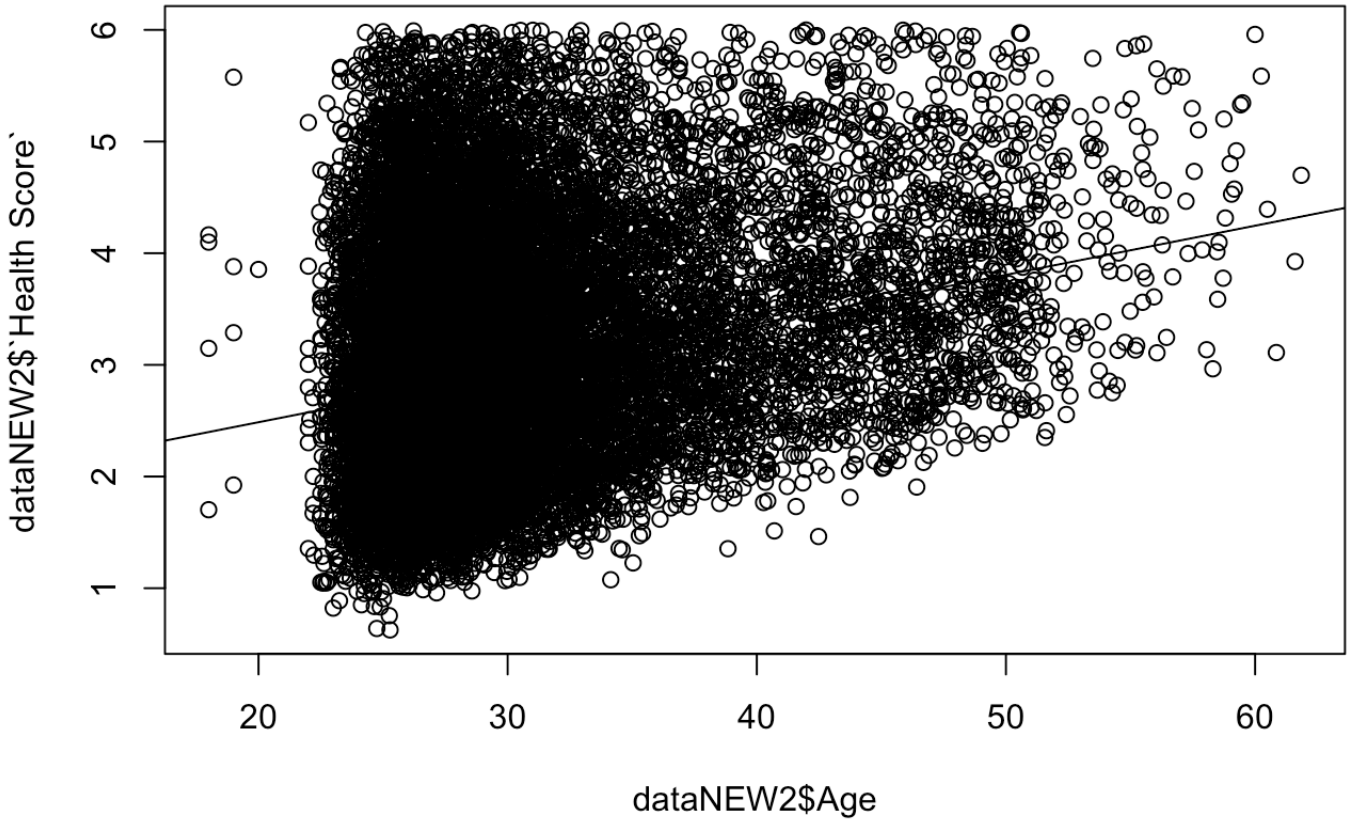



```
# follow the model assumption.

plotage <- plot(dataNEW2$Age,dataNEW2$`Health Score`)
plotage
```

```
## NULL
```

```
abline(model2)
```

```
## Warning in abline(model2): only using the first two of 6 regression
## coefficients
```



Evaluating the Claim a) Using the information from Questions 1 and 2, describe how you would evaluate InsurAHealth's claim that employees are getting sicker. First list how you would evaluate the claim. Then, time-permitting, implement the steps you suggested.