Acumen LLC Data Exercise

Kunal Mishra

3/10/2018

Github with .RMD Source: https://github.com/kmishra9/AcumenDataExercise.git

Introduction & Background Information

1. A large company, Company A, provides health insurance to its employees.
2. Every four years, Company A's insurer, InsurAHealth, reviews the health status of the employees.
   - To do this, InsurAHealth calculates a health score between 0 and 6 for each employee on a quarterly basis
   - 0 denotes a very health person, and 6 denotes a very sick person
   - The health score is a proprietary tool used by InsurAHealth. The items that go into its formula are not public.
3. This past review cycle, InsurAHealth claimed that the employees at Company A have gotten sicker. The mean health score in Quarter 1 was 3.4, in Quarter 6 it was 3.5, and in Quarter 12, it was 3.9.

```
library(DataExplorer)
library(data.table)
library(ggplot2)
library(reshape2)
library(dummies)

InsurAHealthData = data.table::fread(input="InsurAHealthData.csv", header=TRUE)
InsurAHealthData
```

```
##        Observation Number Quarter Employee Id Sex (Male=1) Race Age
##     1:                  1       1           1            0    3  27
##     2:                  2       2           1            0    3  28
##     3:                  3       3           1            0    3  28
##     4:                  4       4           1            0    3  28
##     5:                  5       5           1            0    3  29
##    ---
## 19099:              19099       8        2000           NA    1  28
## 19100:              19100       9        2000           NA    1  28
## 19101:              19101      10        2000           NA    1  28
## 19102:              19102      11        2000           NA    1  28
## 19103:              19103      12        2000           NA    1  29
##        Hospital Visit This Quarter (1=Yes)  Salary Health Score
##     1:                                   0 $36,907          3.7
##     2:                                   0 $37,907          5.0
##     3:                                   0 $38,907          4.0
##     4:                                   0 $39,907          2.3
##     5:                                   0 $40,907          2.1
##    ---
## 19099:                                   0 $59,905         10.0
## 19100:                                   1 $60,905         10.0
## 19101:                                   0 $61,905         10.0
## 19102:                                   0 $62,905         10.0
## 19103:                                   0 $63,905         10.0
```

Initial Observations

The initial heads and tails of the data indicate that each row is an employee's data for a given quarter, and that the data is sorted by Employee ID, then Quarter, so that all of a unique employee's data is consecutive. We also immediately see that there is a variety of demographic information, including Race, Age, and Sex, all of which is categorical (age could also be considered continuous but it is truncated to years). The only two continuous variables we have are then health score and salary. Another thing I can immediately see is missing values (denoted NA) and health scores that don't correspond with what InsurAHealth claims, with a value of 10.0 exceeding the maximum score of 6.0.

In addition, though the salary column is meant to be a continuous numeric column, it comes in character type due to the '$' and ',' symbols within the numbers. This also makes any standard type casting a bit more difficult, so that will need to be corrected. Finally, column names with spaces in them will make my life just a bit more difficult in this analysis, so that will need to be changed as well.

Let's clean the data:

```
colnames(InsurAHealthData) = c("Observation_Number", "Quarter",
                               "Employee_ID", "Sex",
                               "Race", "Age",
                               "Hospital_Visit_This_Quarter",
                               "Salary", "Health_Score")

# Fixing the salary column (string --> numeric)
salaryColumn = InsurAHealthData[,Salary]
str(salaryColumn)
```

```
##  chr [1:19103] "$36,907" "$37,907" "$38,907" "$39,907" "$40,907" ...
```

```
salaryColumn = gsub("$", "", salaryColumn, fixed=TRUE)
salaryColumn = gsub(",", "", salaryColumn, fixed=TRUE)
str(salaryColumn)
```

```
##  chr [1:19103] "36907" "37907" "38907" "39907" "40907" "41907" "42907" ...
```

```
InsurAHealthData$Salary = as.numeric(salaryColumn)

# Converting categorical variables to factors
InsurAHealthCategorical = InsurAHealthData[,
                                           lapply(.SD, as.factor),
                                           .SDcols=c("Quarter", "Race", "Sex", "Hospital_Visit_This_Qua

# Converting continuous variables to numeric
InsurAHealthContinuous = InsurAHealthData[,
                                          lapply(.SD, as.numeric),
                                          .SDcols=c("Salary", "Health_Score", "Age")]

# Binding all variables back together
InsurAHealthData = cbind(InsurAHealthContinuous,
                         InsurAHealthCategorical,
                         InsurAHealthData[,c("Employee_ID", "Observation_Number"), with=FALSE])
str(InsurAHealthData)
```

```
## Classes 'data.table' and 'data.frame':   19103 obs. of  9 variables:
##  $ Salary                     : num  36907 37907 38907 39907 40907 ...
##  $ Health_Score               : num  3.7 5 4 2.3 2.1 1.5 4.7 2.3 2.8 2.8 ...
##  $ Age                        : num  27 28 28 28 29 29 29 29 30 30 ...
##  $ Quarter                    : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ Race                       : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Sex                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hospital_Visit_This_Quarter: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Employee_ID                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Observation_Number         : int  1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

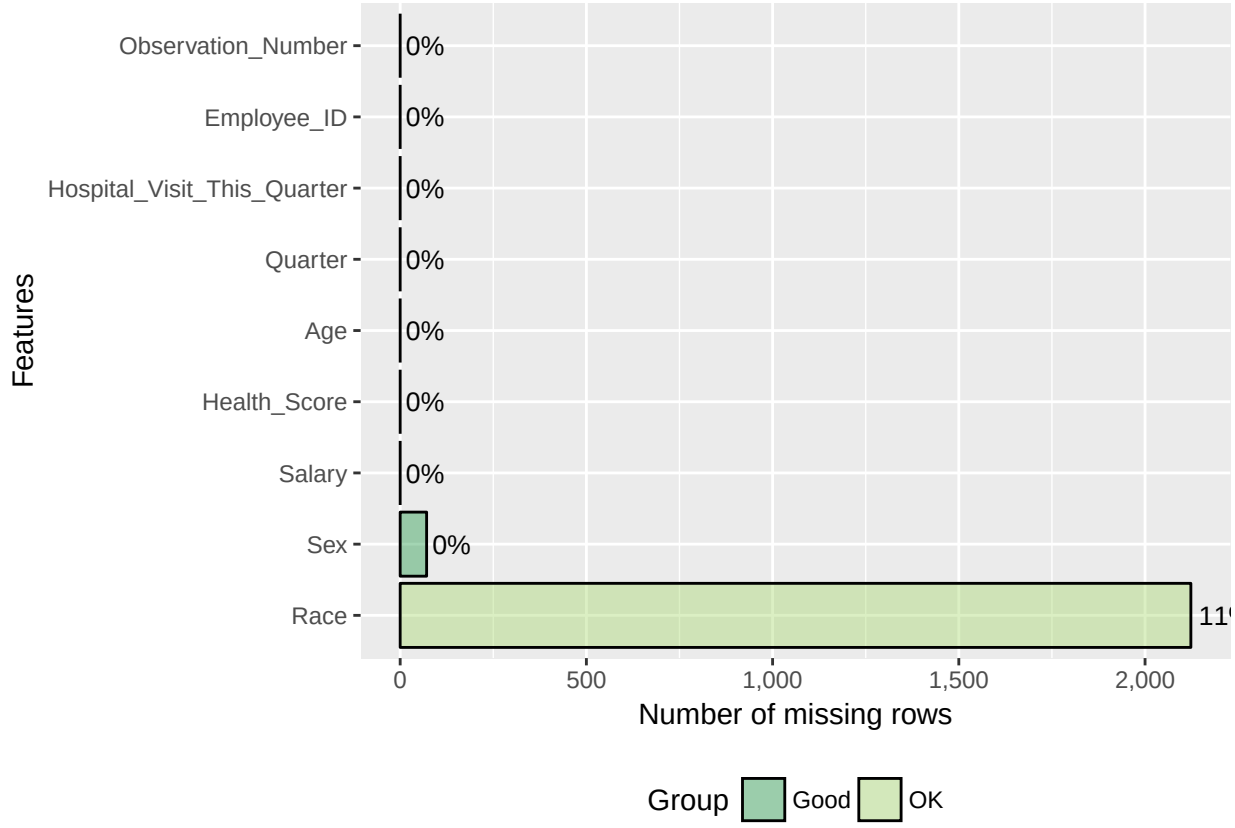
Question 1 - Understanding the Data

a) Are all the values in the data reasonable? Are there missing values?

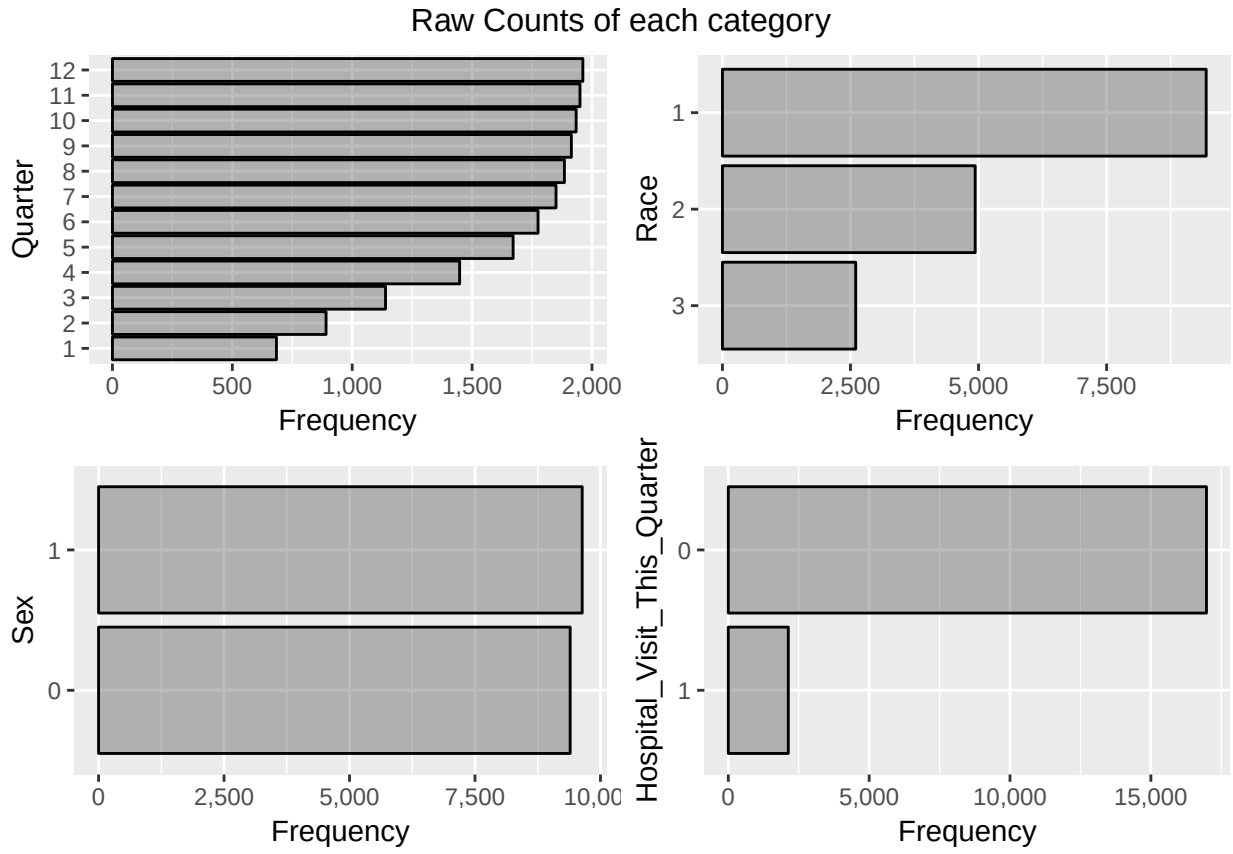
b) What are the characteristics of employees at Company A? Do these demographics change over time?

Let's start by answering 1A. We noticed that there were missing and unreasonable values in our first look at the data, but let's first try to figure out how big these issues are.

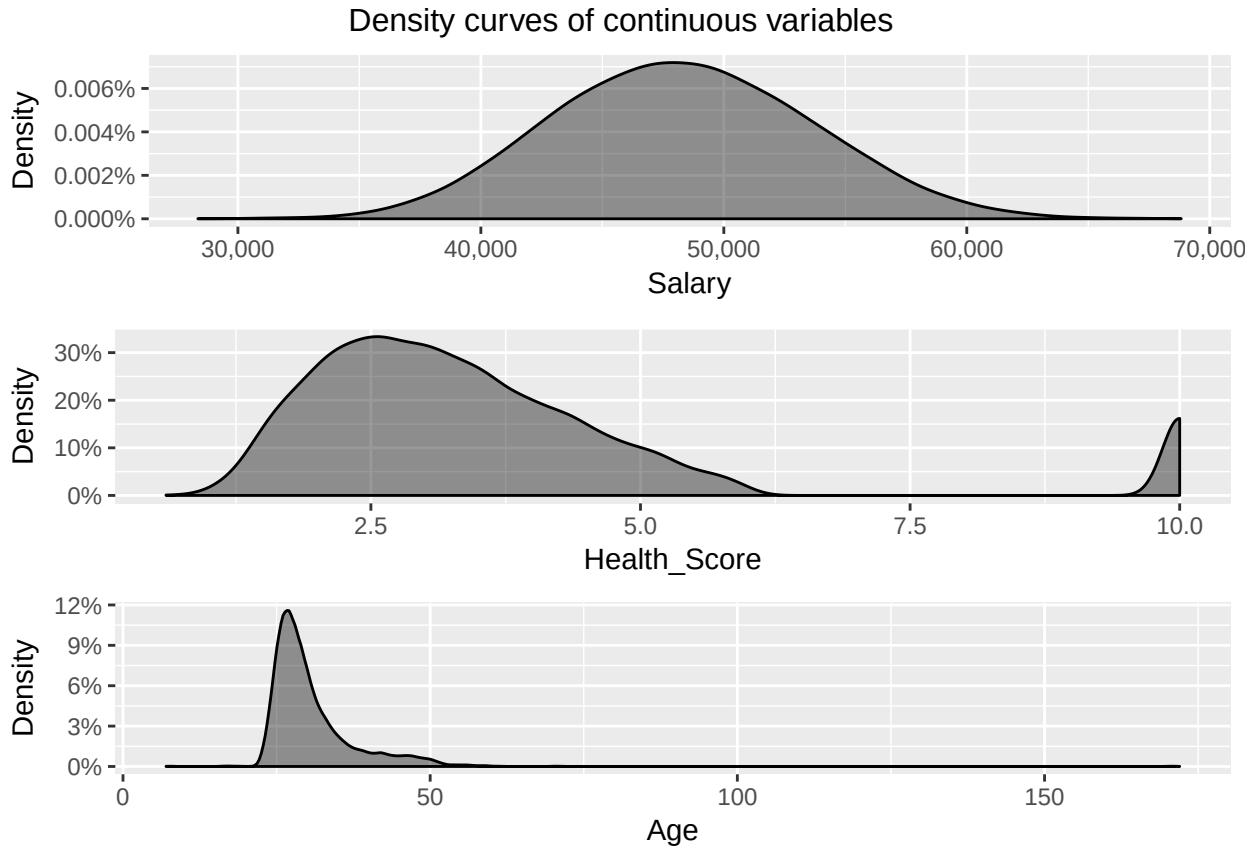
```
DataExplorer::plot_missing(InsurAHealthData)
```




```
# Looking for unexpected values - discrete data
DataExplorer::plot_bar(
    data = InsurAHealthCategorical,
    title="Raw Counts of each category"
)
```




```
#Looking for unexpected values - continuous data
DataExplorer::plot_density(
    data = InsurAHealthContinuous,
    title = "Density curves of continuous variables"
)
```




```
#Ensuring Employee_IDs have no more than 12 associated quarters (would be a signal that Employee_IDs ar
max(InsurAHealthData[, .N, by=Employee_ID][,N])
```

```
## [1] 12
```

Further Observations

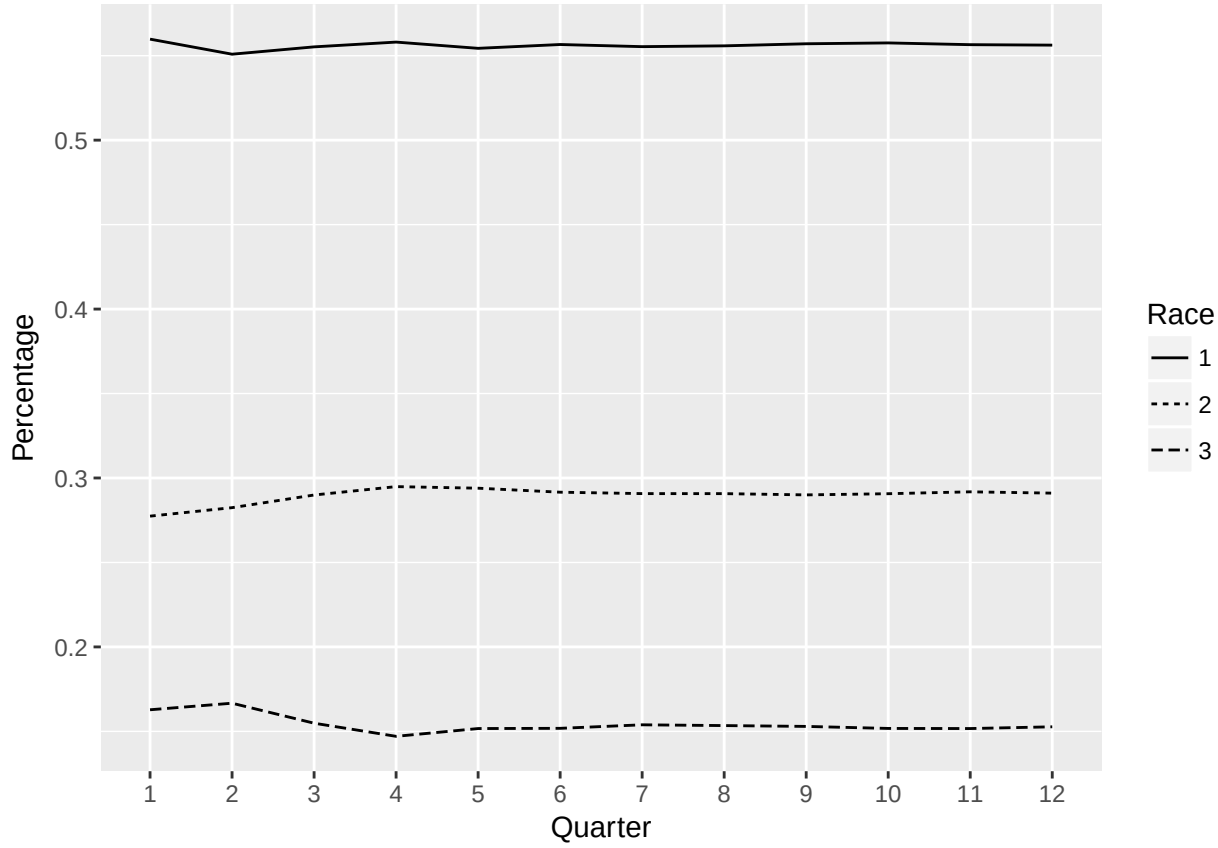
- Missing Data
  - Most of the missing data is in the Race column with 11% of rows missing a value. When aggregated by Employee_ID, we see that 0.1121305% of employees are missing a race value.
  - The Sex column is also missing a negligible amount (<1%) of data. When aggregated by Employee_ID, we see that [PLACEHOLDER]% of employees are missing a race value.
  - Though it may not be "missing", its worth noting that many employees do not have all 12 quarters worth of data, and for whatever reason employees tend to have more data towards the end of the year. This also tends to be the time of year at which disease transmission (such as Influenza, Norovirus, Rotovirus, Common Cold viruses and more) in the US is highest, so this may be a source of bias in the data.
- Unreasonable values
  - For whatever reason, we see a number of "invalid" health scores of 10.0. This warrants further investigation but could also be attributable to a data entry error (someone added an extra 0 to 1.0) or data corruption. Further analysis should exclude this data in the meantime.

Now, let's look at 1B.

- Demographics
  - The population is fairly evenly split, and the race split is 0.136209, 0.258284, 0.4943726 for race 3, 2, and 1 respectively. Age is displaying a nice bell curve with a mean age of 30.592263 years old.

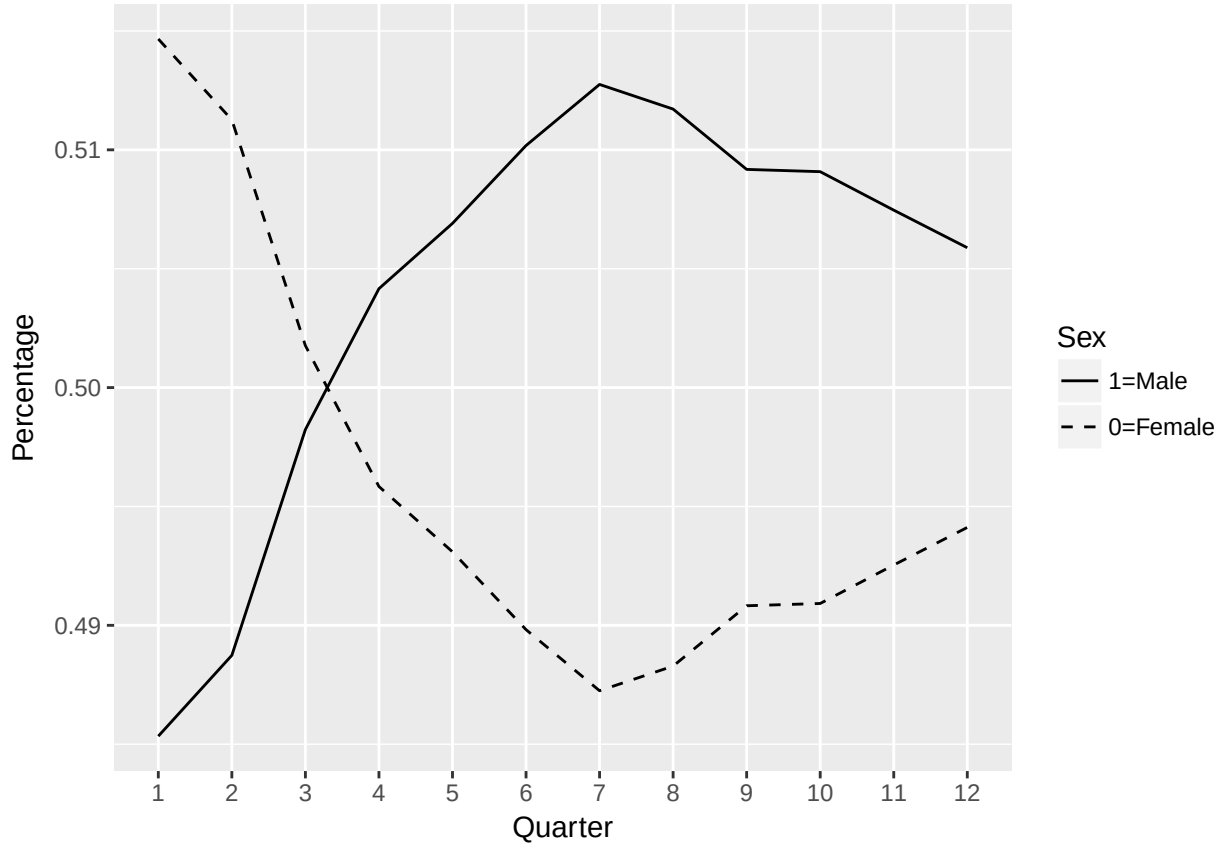
Now, let's graph plot the characteristics of the population over time:

```
# Count the number of employees per race, per quarter, then find what percentage that constitutes
InsurAHealthDataTemporalRacePercentages =
    merge(x = InsurAHealthData[!is.na(Race), .(.N), keyby=.(Quarter, Race)],
          y = InsurAHealthData[!is.na(Race), .(.N), keyby=.(Quarter)],
          keyby = Quarter
          )[, .(Quarter, Race, Percentage=N.x/N.y)]

# Graph it temporally
ggplot(data = InsurAHealthDataTemporalRacePercentages,
       aes(x=Quarter, y=Percentage, group=Race)) +
    geom_line(aes(linetype=Race))
```

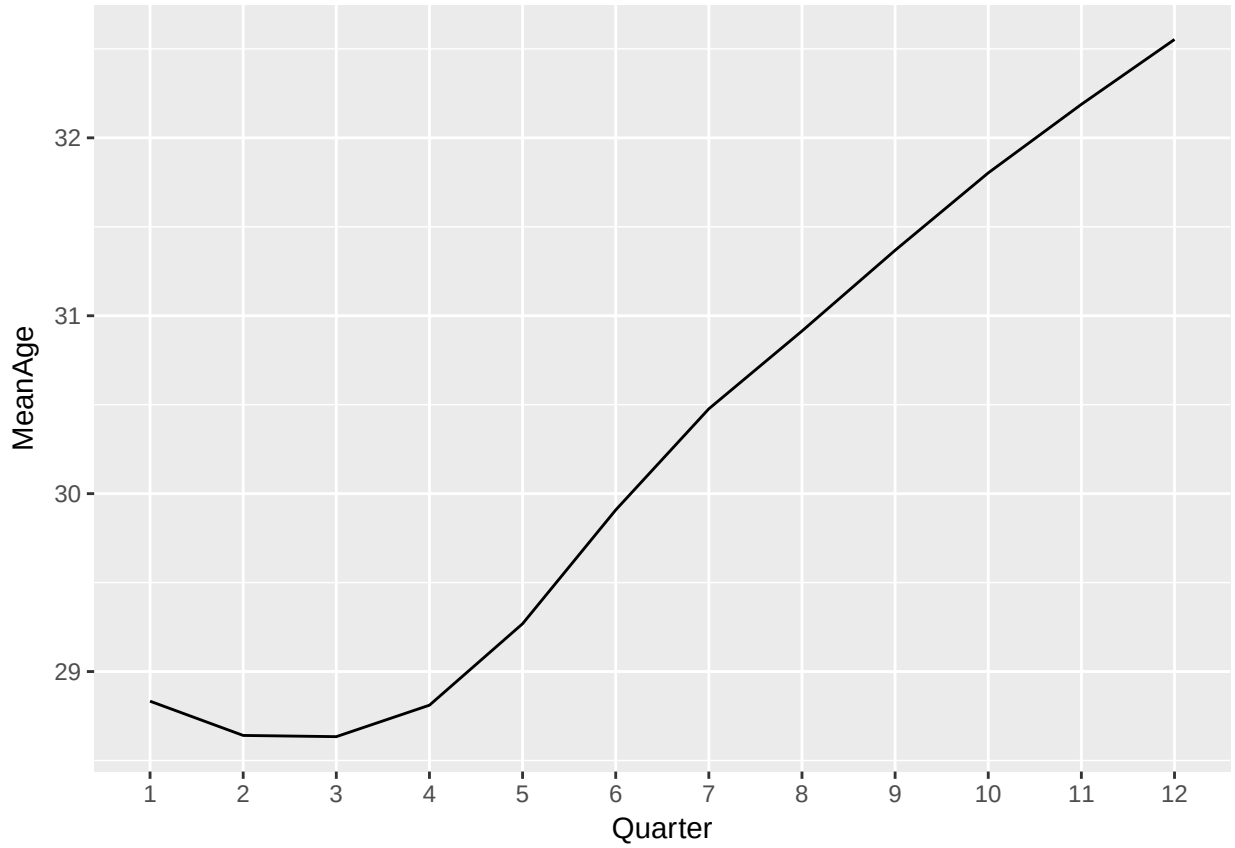



```
# Count the number of employees per sex, per quarter, then find what percentage that constitutes
InsurAHealthDataTemporalSexPercentages =
    merge(x = InsurAHealthData[!is.na(Sex), .(.N), keyby=.(Quarter, Sex)],
          y = InsurAHealthData[!is.na(Sex), .(.N), keyby=.(Quarter)],
          keyby = Quarter
          )[, .(Quarter, Sex, Percentage=N.x/N.y)]

# Graph it temporally
ggplot(data = InsurAHealthDataTemporalSexPercentages,
       aes(x=Quarter, y=Percentage, group=Sex)) +
    geom_line(aes(linetype=Sex)) +
    scale_linetype_manual(values = c("dashed","solid"),
                          labels = c("0=Female", "1=Male")) +
```

```
    guides(linetype=guide_legend(reverse = TRUE))
```




```
# Find the average age on a quarterly basis
InsurAHealthDataTemporalMeanAge = InsurAHealthData[, .(MeanAge=mean(Age)), keyby=Quarter]

# Graph it temporally
ggplot(data = InsurAHealthDataTemporalMeanAge,
       aes(x=Quarter, y=MeanAge, group=1)) +
    geom_line()
```



Based on these graphs, it is apparent that race has remained relatively stable while the population of employees has swung slightly to become a majority Male after starting out as a majority Female. This swing was fairly small, however. Finally, mean age has increased a little over 3 years, as we would expect over the course of 12 quarters (3 years).

Question 2 - Exploring Relationships

a) Which characteristics are associated with the health score?

I was initially planning on jumping into this with a quick Principal Components Analysis but my data is composed of several categorical variables and several sources, including this Stack Overflow post warned against doing so (even using dummy variables). I am **aware** that other Factor Analysis methods exist for analysis of mixed data, but these currently lie outside my comfort zone and I don't think scrounging around documentation or external packages (such as FactoMineR) for a few minutes without understanding the why and how of the math, would be very useful.

Thus, I think using Linear Regression with dummy variables is the way to **quickly** proceed here (though I would probably do a deeper dive into alternative statistical methods in a real analysis), to find what, if any, statistically significant predictors of the health score there are.

```
str(InsurAHealthData)
```

```
## Classes 'data.table' and 'data.frame':   19103 obs. of  9 variables:
##  $ Salary                    : num  36907 37907 38907 39907 40907 ...
##  $ Health_Score              : num  3.7 5 4 2.3 2.1 1.5 4.7 2.3 2.8 2.8 ...
##  $ Age                       : num  27 28 28 28 29 29 29 29 30 30 ...
##  $ Quarter                   : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Race                      : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
```

```
##  $ Sex                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hospital_Visit_This_Quarter: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Employee_ID                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Observation_Number         : int  1 2 3 4 5 6 7 8 9 10 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
fitData = InsurAHealthData[!is.na(Race) && !is.na(Sex),-c("Observation_Number", "Employee_ID"), with=FA
str(fitData)
```

```
## Classes 'data.table' and 'data.frame':   19103 obs. of  7 variables:
##  $ Salary                     : num  36907 37907 38907 39907 40907 ...
##  $ Health_Score               : num  3.7 5 4 2.3 2.1 1.5 4.7 2.3 2.8 2.8 ...
##  $ Age                        : num  27 28 28 28 29 29 29 29 30 30 ...
##  $ Quarter                    : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Race                       : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Sex                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hospital_Visit_This_Quarter: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# Configuring dummy variables for Race
fitDummyData = as.data.table(
    dummies::dummy.data.frame(data = fitData, names = c("Race"))
    )[,-c("RaceNA")]    # Useless Column - We've excluded data with missing race/sex values, but the fa
str(fitDummyData)
```

```
## Classes 'data.table' and 'data.frame':   19103 obs. of  9 variables:
##  $ Salary                     : num  36907 37907 38907 39907 40907 ...
##  $ Health_Score               : num  3.7 5 4 2.3 2.1 1.5 4.7 2.3 2.8 2.8 ...
##  $ Age                        : num  27 28 28 28 29 29 29 29 30 30 ...
##  $ Quarter                    : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Race1                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Race2                      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Race3                      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Hospital_Visit_This_Quarter: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

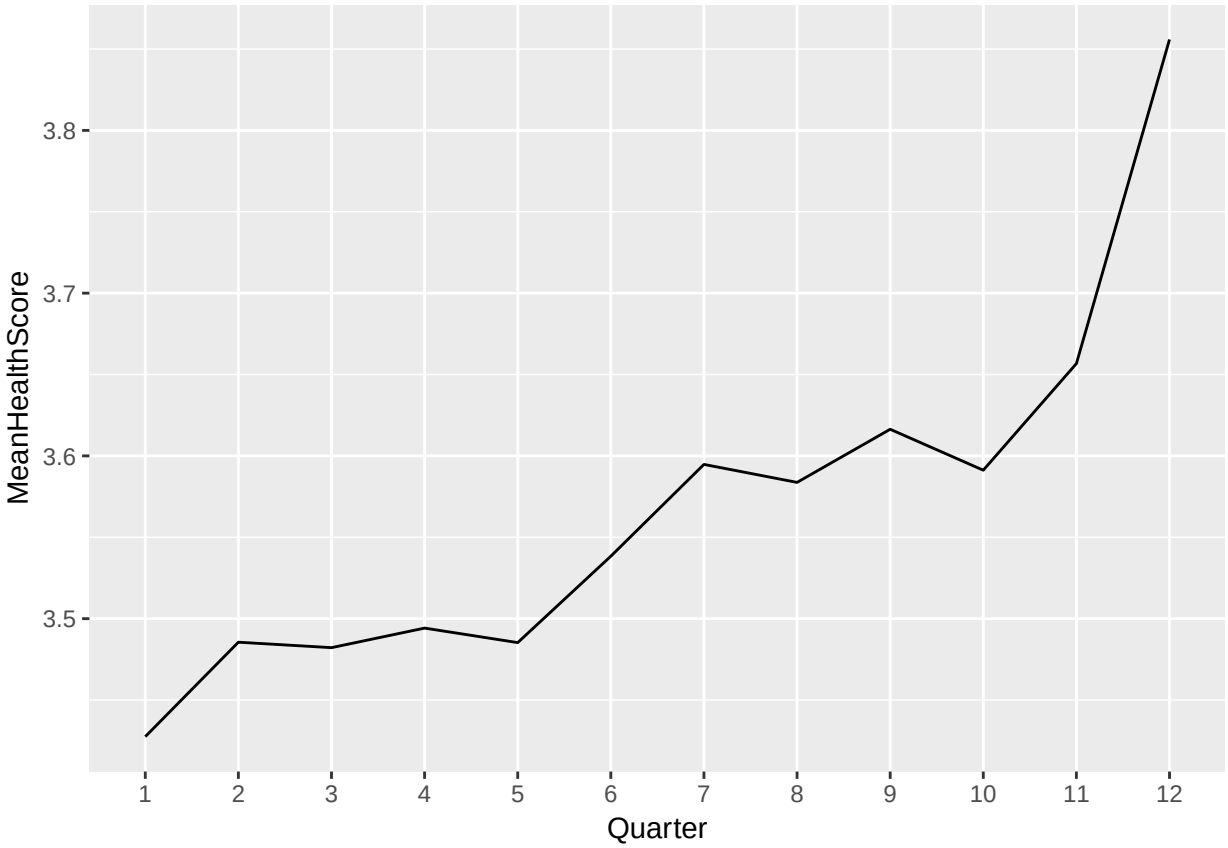
```
# Fitting a linear regression model
linearRegressionFit = lm(formula = Health_Score~.,
                         data = fitDummyData)
summary(linearRegressionFit)
```

```
##
## Call:
## lm(formula = Health_Score ~ ., data = fitDummyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8824 -1.1676 -0.4773  0.4790  7.1490
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.979e+00  1.906e-01  10.381   <2e-16 ***
## Salary                      -3.430e-06  3.824e-06  -0.897   0.3698
## Age                          4.502e-02  2.010e-03  22.394   <2e-16 ***
## Quarter2                     5.612e-02  9.657e-02   0.581   0.5612
```

```
## Quarter3                      3.699e-02  9.203e-02   0.402   0.6877
## Quarter4                      3.545e-02  8.840e-02   0.401   0.6884
## Quarter5                      1.107e-02  8.672e-02   0.128   0.8984
## Quarter6                      3.593e-02  8.642e-02   0.416   0.6776
## Quarter7                      6.907e-02  8.652e-02   0.798   0.4247
## Quarter8                      4.237e-02  8.706e-02   0.487   0.6265
## Quarter9                      4.391e-02  8.782e-02   0.500   0.6171
## Quarter10                     2.081e-02  8.878e-02   0.234   0.8146
## Quarter11                     7.357e-02  8.993e-02   0.818   0.4133
## Quarter12                     1.741e-01  9.142e-02   1.905   0.0568 .
## Race1                         9.071e-02  4.668e-02   1.943   0.0520 .
## Race2                        -5.471e-03  4.948e-02  -0.111   0.9120
## Race3                        -6.574e-02  5.572e-02  -1.180   0.2381
## Sex1                          3.818e-01  3.361e-02  11.361   <2e-16 ***
## Hospital_Visit_This_Quarter1  8.183e-01  4.390e-02  18.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.896 on 19013 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.05533,    Adjusted R-squared:  0.05443
## F-statistic: 61.86 on 18 and 19013 DF,  p-value: < 2.2e-16
```

As we would expect, increasing Age, being a male, and having a hospital visit are all strong predictors of higher health score. Being Race 1 is also a moderate predictor of higher health scores, and it also seems like people were more likely to have higher health scores during Quarter 12 based on this data. We can corroborate this last interpretation by looking at mean health scores over time, as in Question 1B:

```
# Find the average health score on a quarterly basis
InsurAHealthDataTemporalMeanHealthScore = InsurAHealthData[, .(MeanHealthScore=mean(Health_Score)), keyb

# Graph it temporally
ggplot(data = InsurAHealthDataTemporalMeanHealthScore,
       aes(x=Quarter, y=MeanHealthScore, group=1)) +
    geom_line()
```



Yup. That looks about right, and this kind of data would also be useful in the next part.

Question 3 - Evaluating the Claim

a) Using the information from Questions 1 & 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.

At this point, this is a pretty in-depth analysis through the first 2 questions, so I'll try to outline my initial gut feeling, as well as some next steps I would take before giving a concrete evaluation of InsurAHealth's claim to Company A.

From my initial gut feeling, it does indeed appear that Company A's health is declining. If the graph above charting average Health Scores had less uniform of an upward trend, that would be a far more difficult conclusion to come to, but the increases from quarter-to-quarter far outnumber and outweigh the quarter-to-quarter decreases in health score. In 2A we also saw that being Male and being older were two significant predictors for an increased health score, and our temporal demographic information shows the population increasing in Male percentage and average age.

There are caveats to all of this of course, and I want to outline what is necessary to evaluate the claim.

1. First, analyzing how much standard error and random variation play into all of this. To that end, I would essentially want to graph the above health score graph, but with confidence intervals at each data point (doable because we have a large sample and the sample std. deviation). Then, if the confidence intervals of Quarter 1 and Quarter 3 overlap, we don't have a statistically significant increase in Health Score.
2. In a similar vein, it's possible that the average health scores could be rising due to increased age

polarization. I didn't specifically look for this in the dataset, but **its possible** (though not necessarily likely) that people joining the company were simply more likely to be older or younger (i.e. instead of 3 people joining at ages 29, 30, 31, a situation where 3 people join at 20, 30, 40). In this age polarization scenario, health score may exponentially increase based on age, so ensuring that this age polarization phenomena isn't occurring is a good next step. In addition, we might consider stratifying on Age to analyze whether a confounder could be affecting the company's health scores.

3. Speaking of confounders, I would want to look at other data around this time to understand whether other factors could be causing these trends. For example, in an improving economy, people may become more likely to **utilize** healthcare as they are now able to afford the necessary deductibles and copays. Another example could be looking at whether people with preexisting conditions within the company have worsening conditions, while the rest of the employees remain in stellar or even improving health. Even if a small number of them were to begin to visiting the hospital much more frequently, they may be pulling up the health scores of the entire company for the same reason outlined earlier — healthy people with good health scores have less weight to pull the average down because there's less room to actually go down (see: right-tailed Health Score distribution for the entire dataset).

4. One final note I want to make is on the ambiguity of what a Hospital visit actually entails. If it necessarily means that the employee was **admitted** to a hospital during that quarter, then these next points are moot, but otherwise elective procedures, family member visits, screening tests, preventative and primary care, etc. are all cause to "visit" the hospital. Clarifying this variable would be useful.