Data Analysis Exercise

Name [redacted]

[redacted]

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 healthy 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 have gotten sicker.
   • Mean Health Score in Quarter 1 was 3.4, in Quarter 6 it was 3.5, and Quarter 12 it was 3.9.

Company A has hired you to evaluate InsurAHealth's claim that employees are sicker.
To facilitate your analysis, InsurAHealth has provided you with data for 12 quarters that includes 2,000 employees from Company A.
   • Each quarter is a representative sample of the employees at Company A in that quarter.
   • The demographic information included in this data is not part of InsurAHealth's health score calculation.

[redacted]

Use the data in the Data tab to answer Questions 1 - 3.
   • The tab is locked, so you will need to copy and paste the information into a new tab to manipulate it.
   • While you may use regression analysis, it is not necessary to adequately answer these questions.
The goal of this task is to demonstrate the way you think about data and outline the way you approach data-driven analyses.

[redacted]

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?
Use tables and charts to understand the data and demographic characteristics of employees at Company A.

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.

3. 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.

For this analysis, I chose not to pay too much attention to the ‘Race’ variable since the instructions stated the demographic information included in the data is not part of InsurAHealth’s score calculation.

Question 1A

Missing Values:
The variable ‘Sex’ has 71 missing entries. The variable ‘Race’ has 2123 missing entries. The are no other missing values.

Outliers:
Outliers are to be expected but there are certain variables with values that are simply unreasonable. For example, the ‘Health Score’ variable had 1,238 values with scores of 10 while the Health Score range only encompasses scores of 0-6. Since the values of 10 do not fit our scale, they are too extreme or could be typos and therefore I dropped those rows from the data frame.

When we look at the summary statistics for ‘Age’, I see the minimum is 7 while the maximum is 172 — I know immediately those two values are too extreme. I found that Age has 1,452 outliers. Of the outliers, I found the values 7, 8, 16, 70, 71, 72, 170, 171, 172 are completely unreasonable. I dropped these values from the data frame.

I found that Salary has 75 outliers which range from ($28,351, $33,505) and ($62,952
, $68,826). Of the outliers, the salaries appear to be reasonable and are not extreme enough for me to justifiably drop any rows.

‘Employee ID’ does not have any outliers, ’Race’ does not have outliers, ‘Quarter’ does not have any outliers, ‘Sex’ does not have any outliers, and ‘Hospital Visit this Quarter’ does not have any outliers.

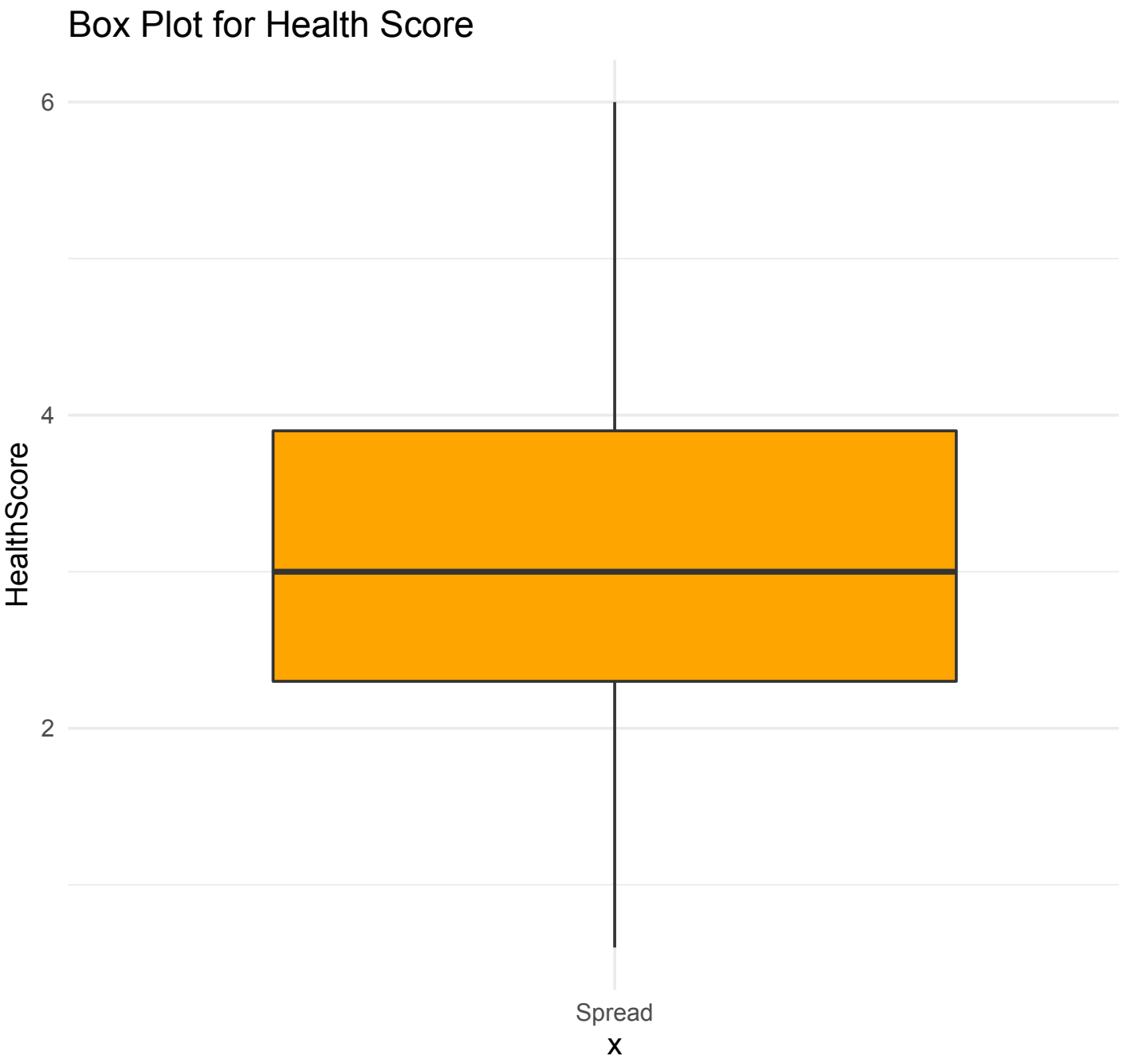
Question 1B

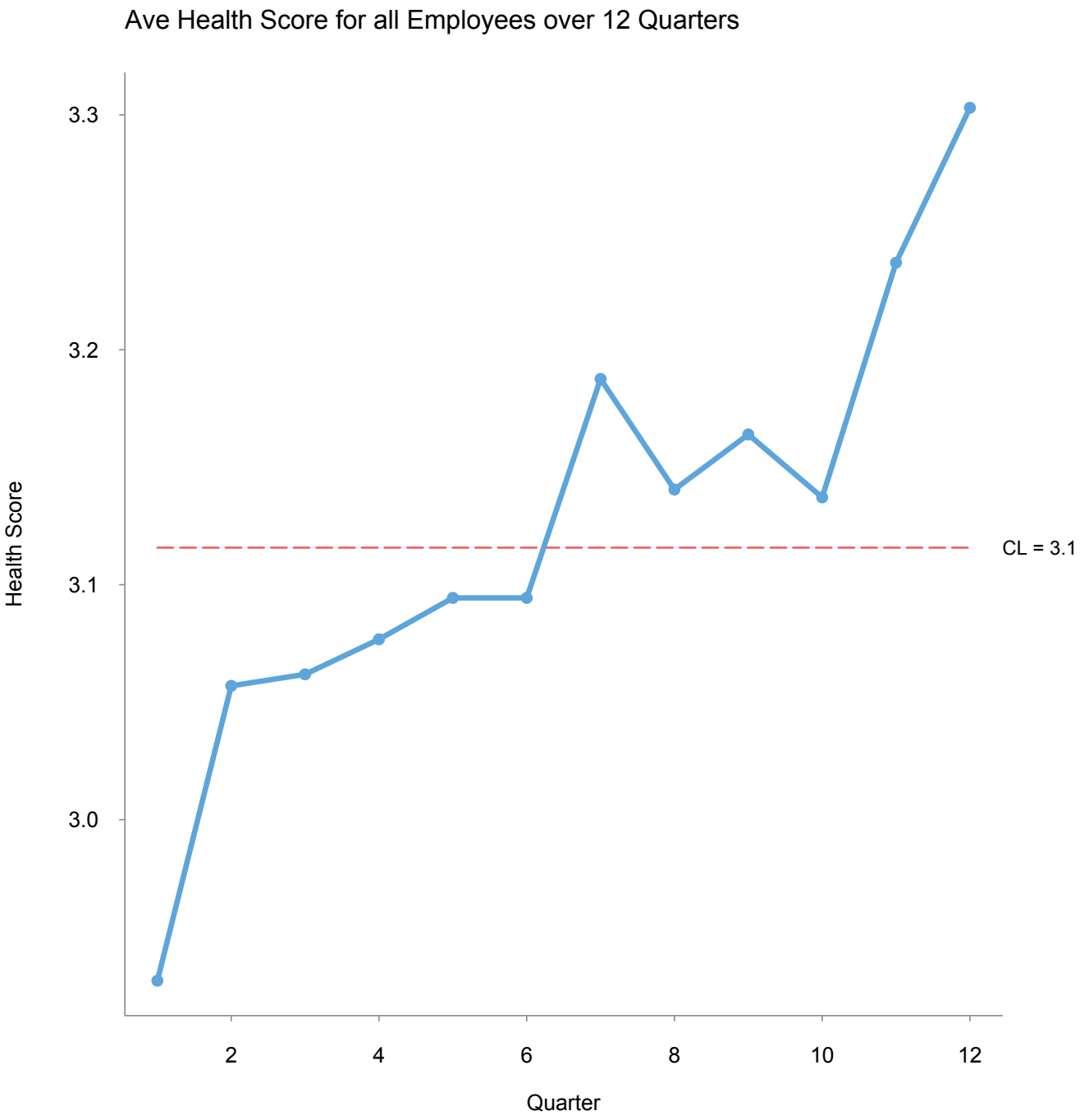
Health Score

Based on this data set, when we look at the box plot titled, “Box Plot for Health Score”, we see 50% of employees who work for Company A have health scores between 2.3-3.9.

I made a run chart to view whether the entire company’s average health score increased or decreased over 12 quarters. When we look at the run chart titled, “Ave Health Score for all Employees over 12 Quarters”, we see that the health score has increased over 12 quarters:

Quarter1=2.93 Quarter6= 3.09 Quarter12=3.3.

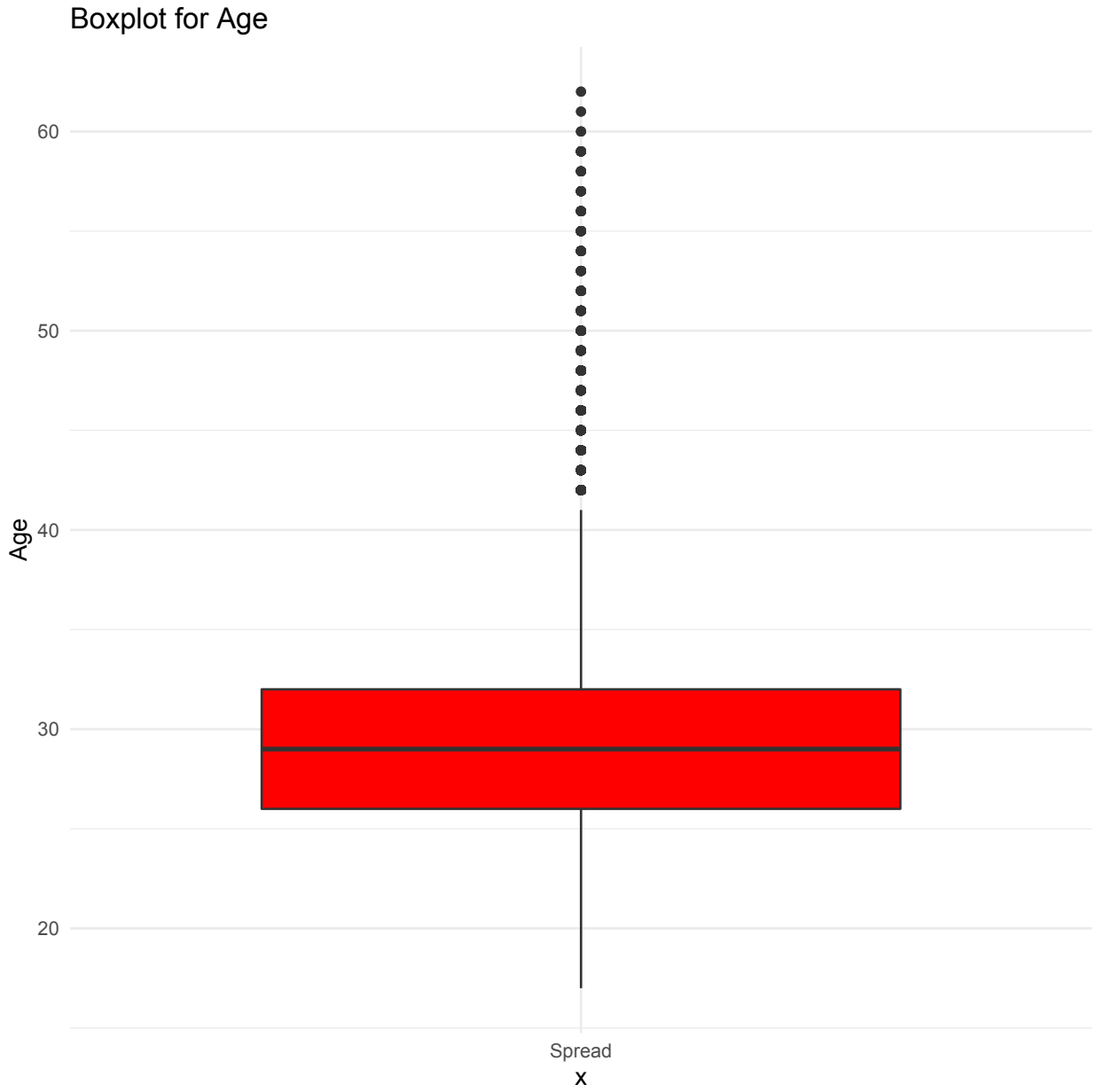



Ave Health Score for all Employees over 12 Quarters
Health Score
3.3
3.2
3.1
3.0
CL = 3.1
2
4
6
8
10
12
Quarter

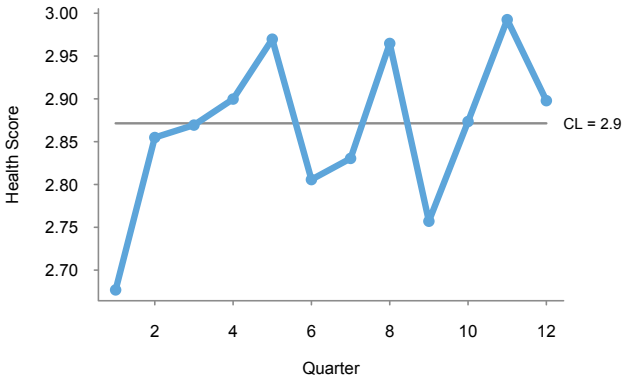
Age

Based on the data, when we look at the box plot titled, “Box Plot for Age”, we see that 50% of employees who work for Company A are between the ages of 26-32 years old while 25% of employees who work for company A are between the ages of 17-25 and the other 25% of employees who work for Company A are 41 years and older.

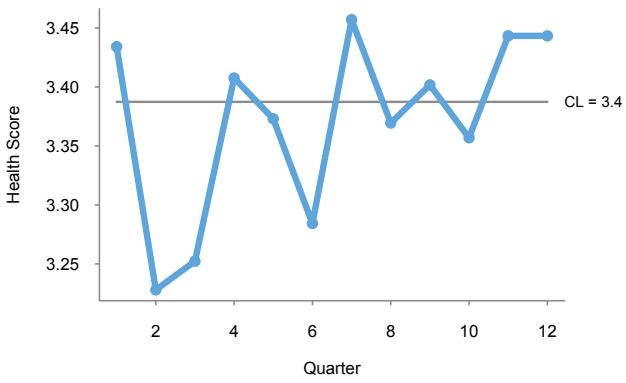
I made a side-by-side of 5 run charts for each of our age groups. The age intervals are as follows: (17-25), (26-29), (30-32), (33-41), (42 and older). Based on the run charts, while all five age groups’ health scores appear to be increasing, the 17-25 year olds have the lowest median health score of 2.9, followed by the 26-29 year olds with a median health score of 3, then the 30-32 year olds with a median health score of 3.2, followed by the 33-41 year olds with a median health score of 3.4, and lastly we have the 42 years and older group with a median health score of 3.9. In other words, older employees appear to be sicker than younger employees.



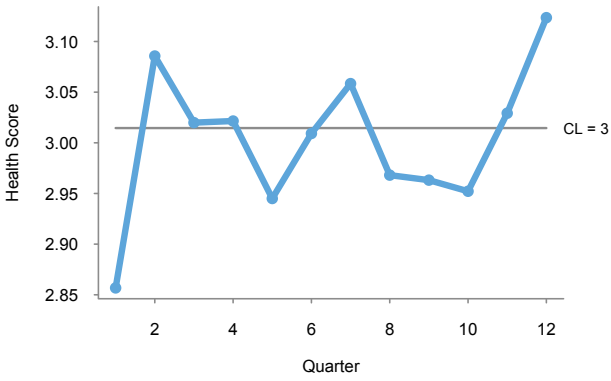
Ave Health Score of Employees 17-25 over 12 Quarters



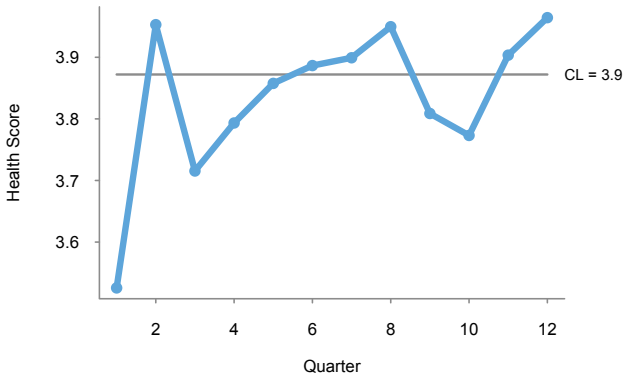
Ave Health Score of Employees 33-41 over 12 Quarters



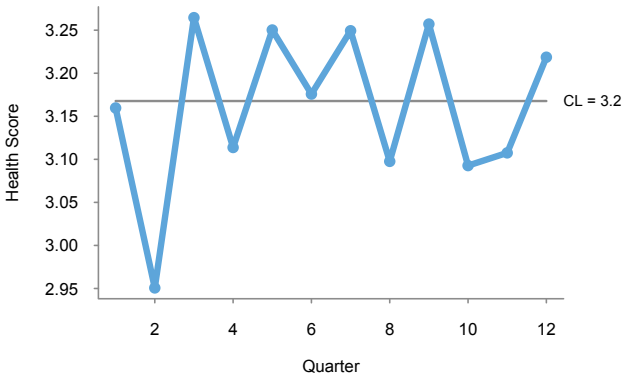
Ave Health Score of Employees 26-29 over 12 Quarters



Ave Health Score of Employees 42 and Older over 12 Quarters



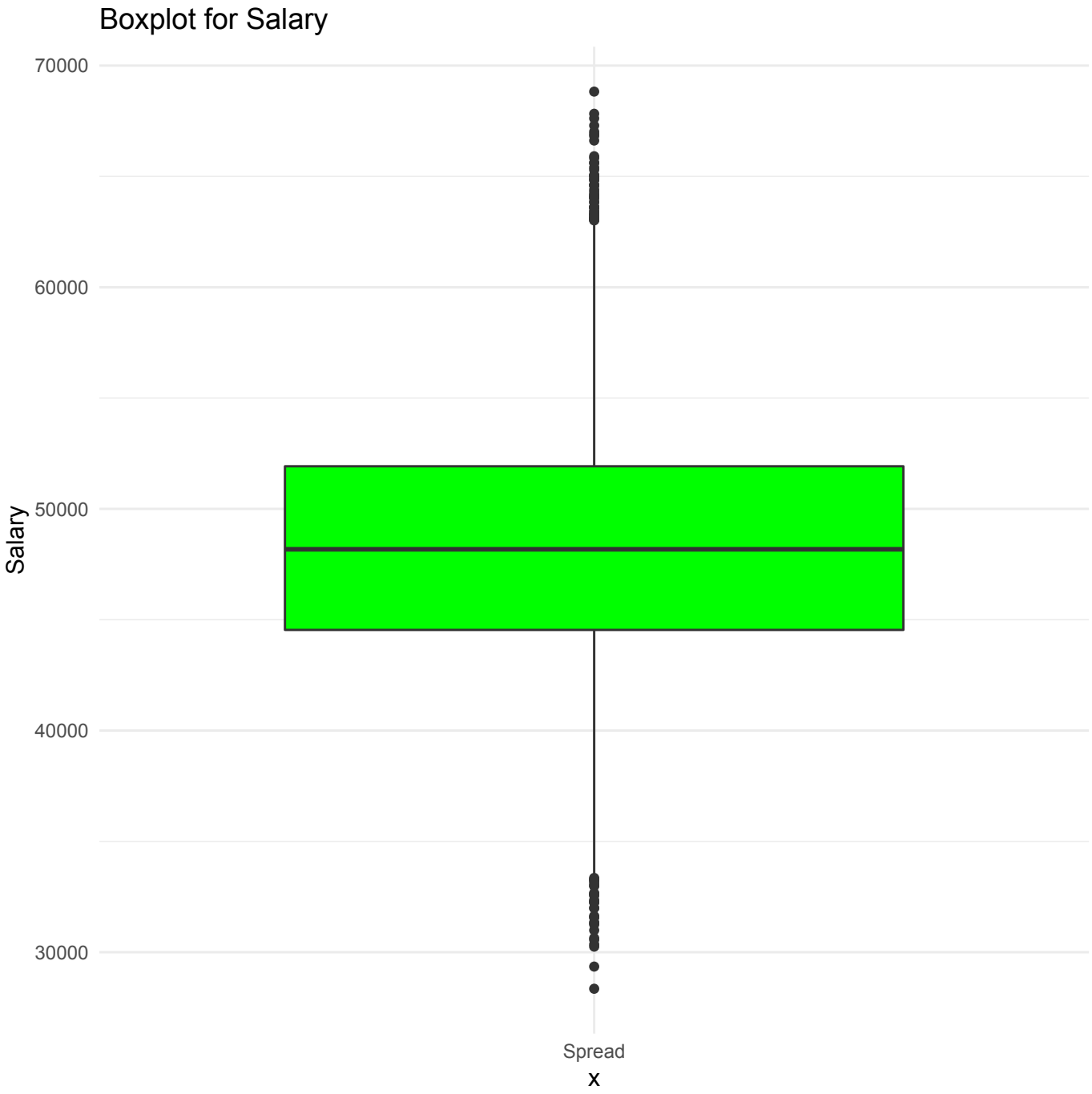
Ave Health Score of Employees 30-32 over 12 Quarters



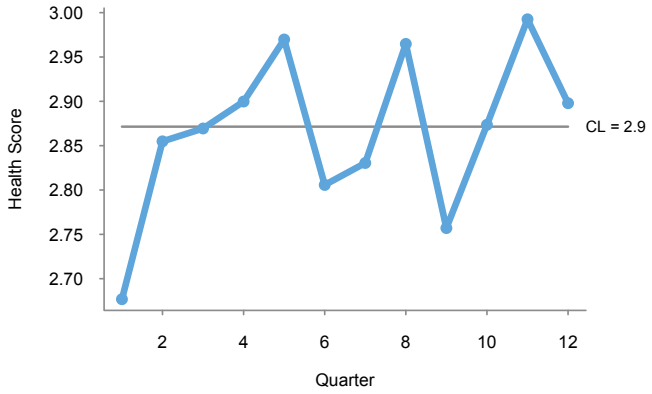
Salary

50% of employees who work for Company A have salaries between $44,540-$51,923.

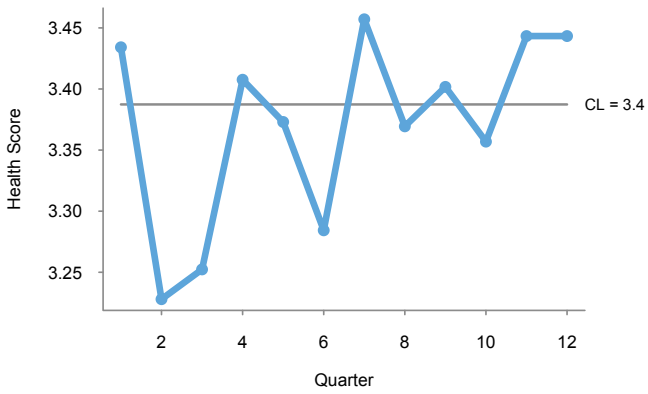
I made a side-by-side of 4 run charts for each salary group. The salary intervals are as follows: (< $44,540), ($44,540-$48,177), ($48,178-$51,923), (< $51,923). Based on the run charts, health scores appear to vary a bit between groups. The lowest salary group has an increasing health score but has one of the lower median health score's at 3.1. The second lowest salary group has an increasing health score and the same median health score as the first group at 3.1. The two higher salary groups have the same median health score but the group with salaries between $48,178-%51,923 has a slight decrease in health score.



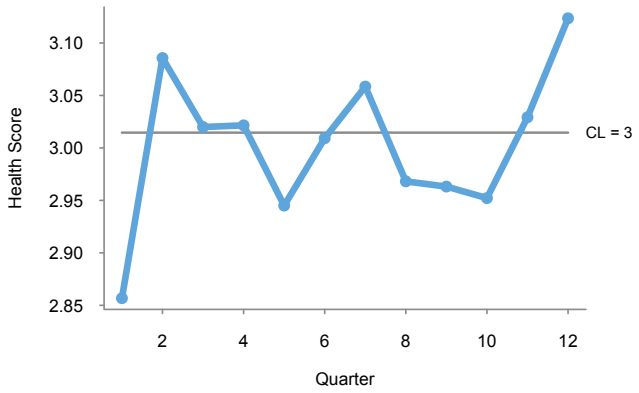
Ave Health Score of Employees 17-25 over 12 Quarters



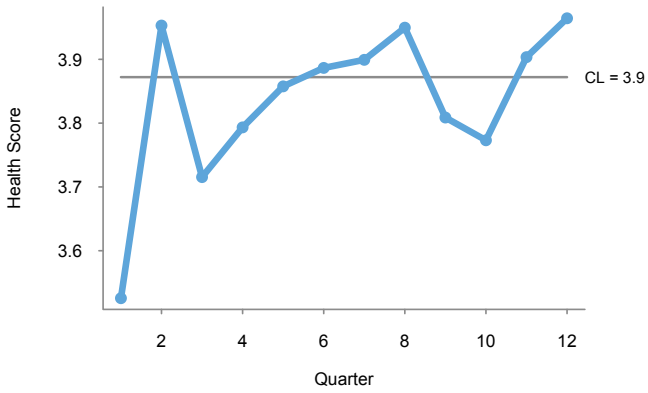
Ave Health Score of Employees 33-41 over 12 Quarters



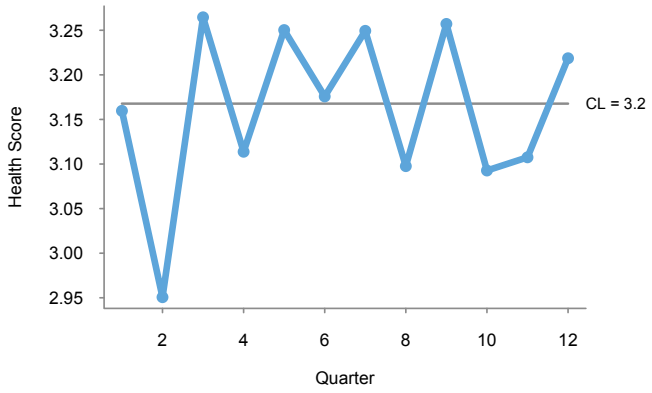
Ave Health Score of Employees 26-29 over 12 Quarters



Ave Health Score of Employees 42 and Older over 12 Quarters



Ave Health Score of Employees 30-32 over 12 Quarters

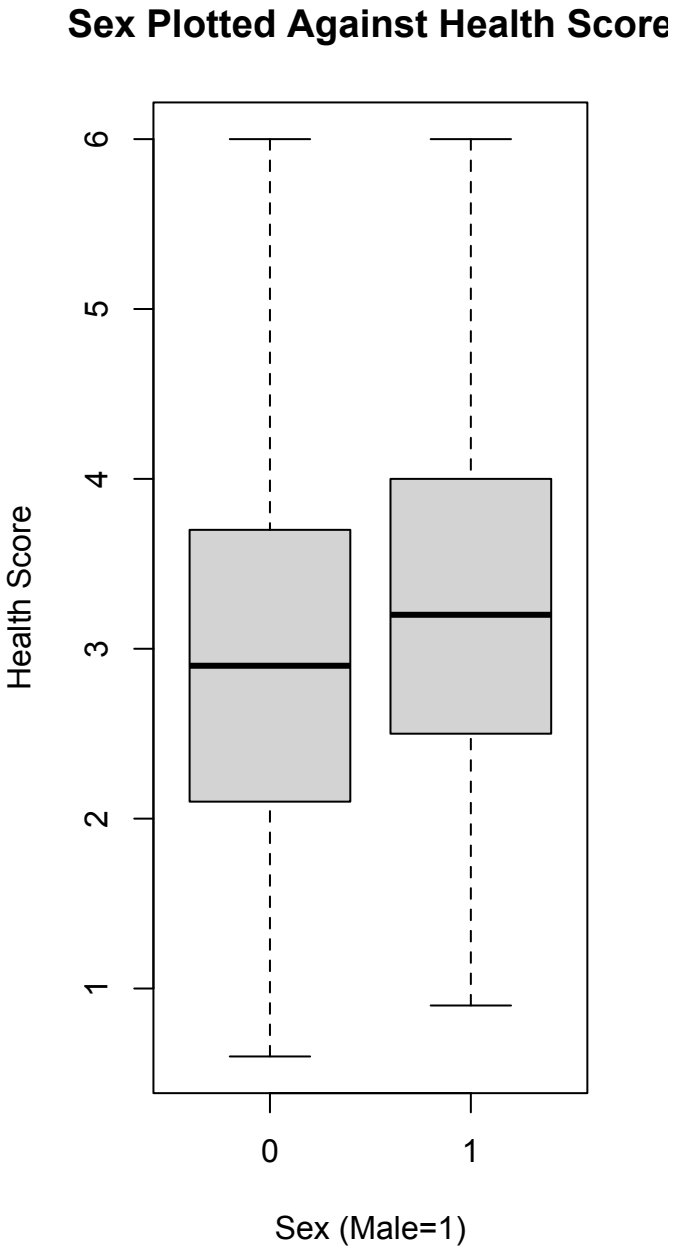


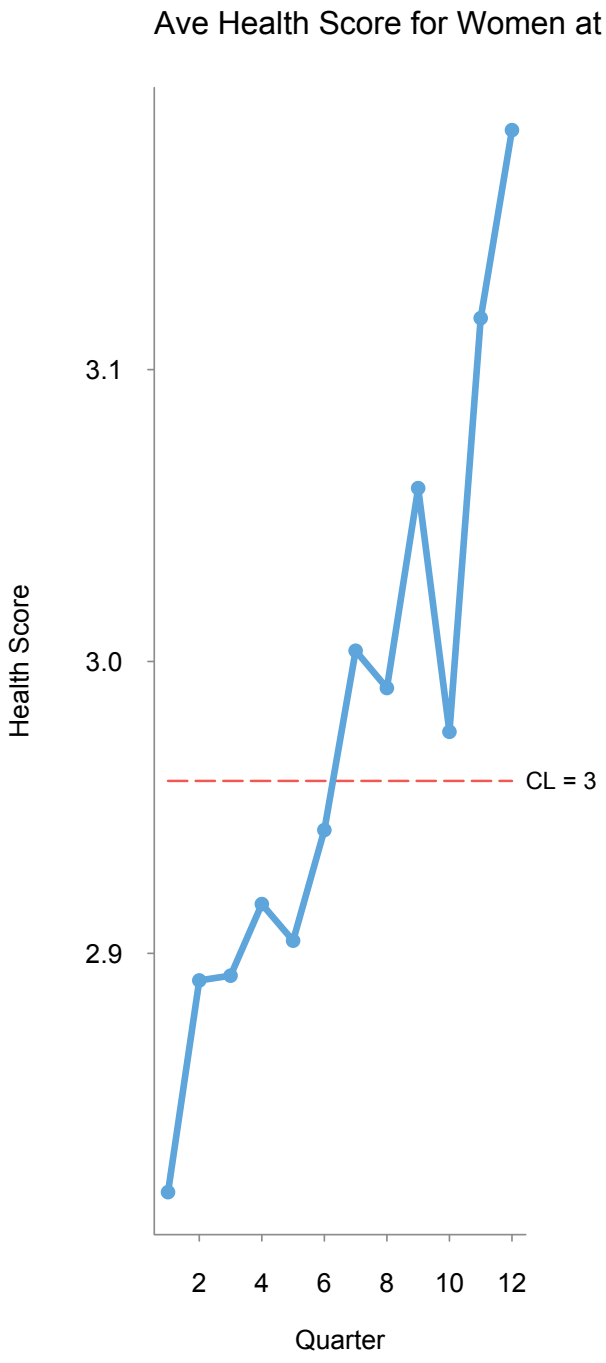
Sex

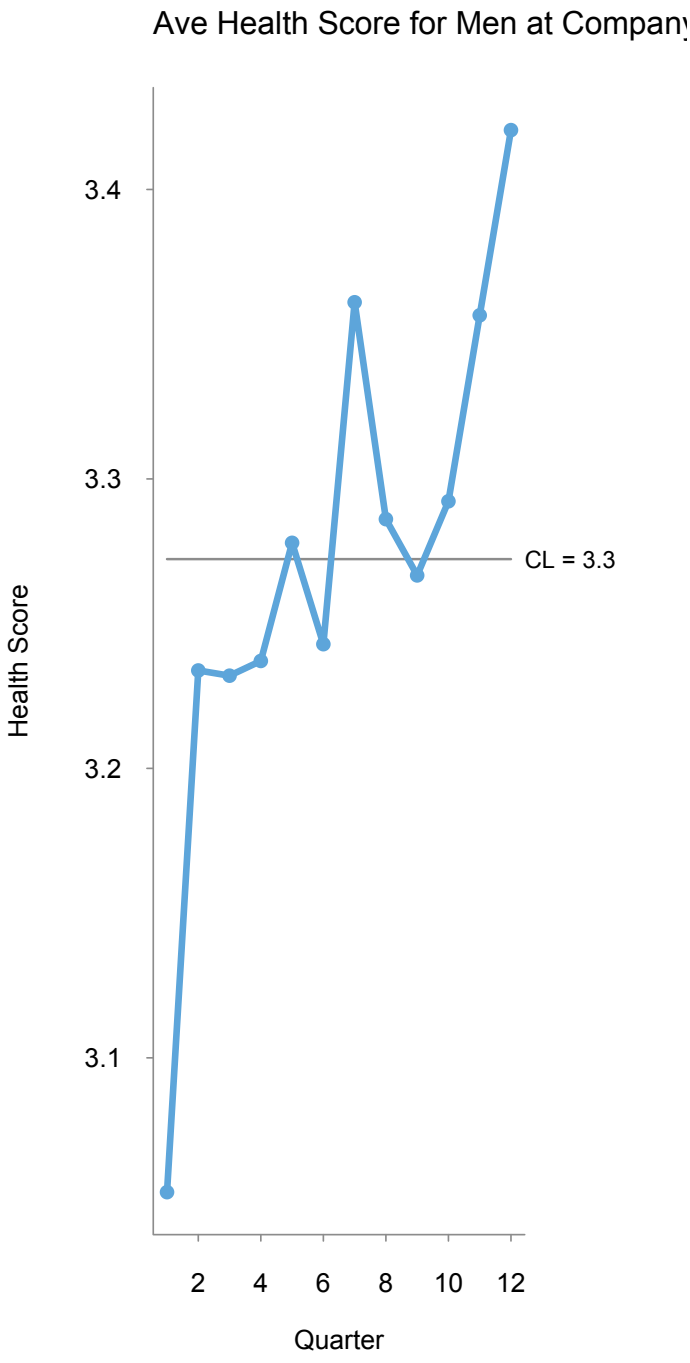
Based on this data set, I created a simple 2x2 table to show that 49.7% of employees who work for Company A are women while 50.3% of employees who work for Company A are men.

When we look at the box plot titled, “Sex Plotted Against Health Score”, we notice women have slightly lower health scores, which means they appear to be in slightly better health than compared to the men.

I made a side-by-side run chart to show the difference in health scores between the genders. Based on the run charts, while both genders’ health scores are increasing over 12 quarters, women have an overall lower median health score than men do over the course of 12 quarters. Women’s median health score = 3 while the men’s median health score = 3.3.




Ave Health Score for Women at Comp
Health Score
3.1
3.0
2.9
CL = 3
2
4
6
8
10
12
Quarter

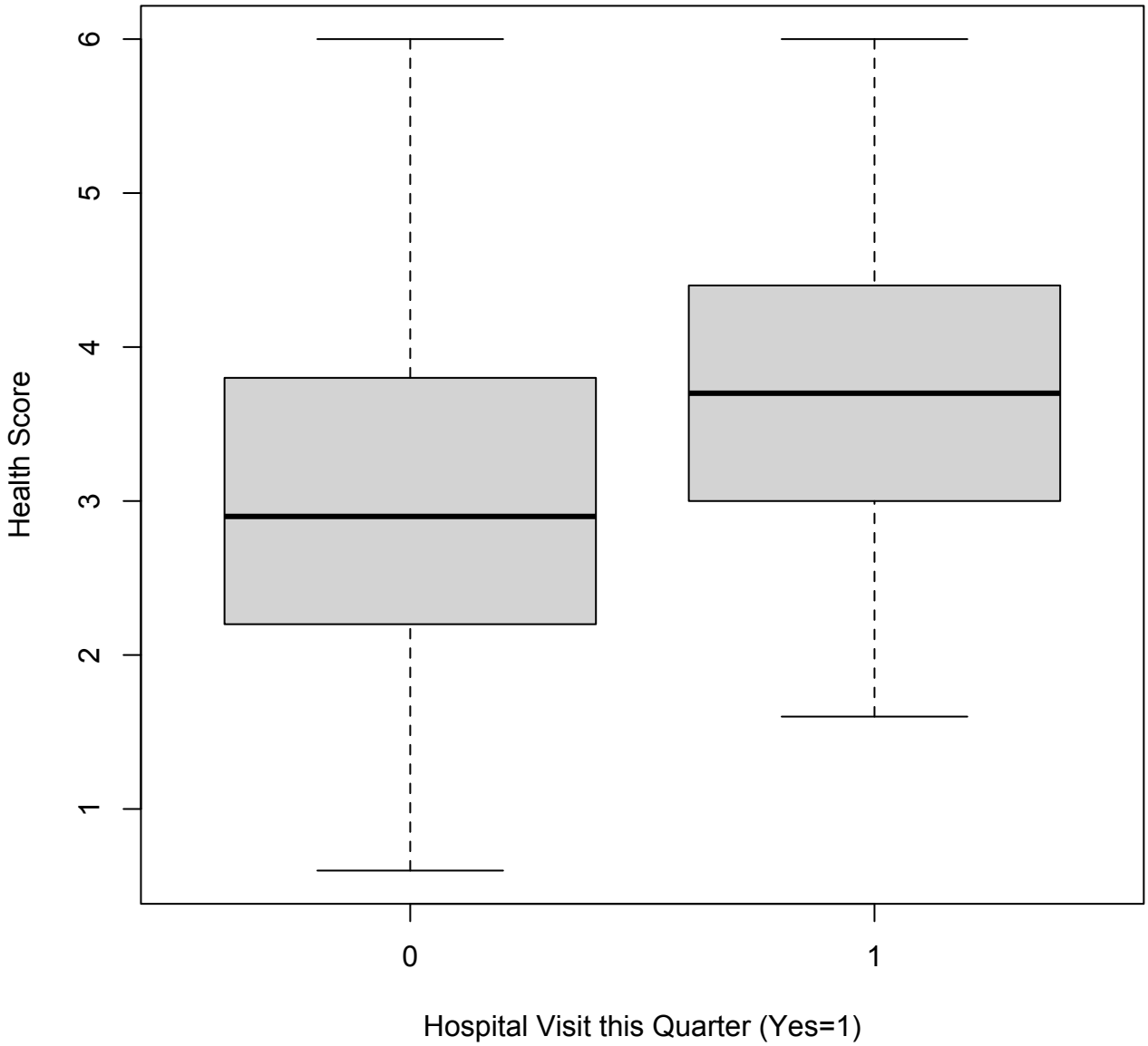

Ave Health Score for Men at Company
Health Score
3.4
3.3
3.2
3.1
CL = 3.3
2
4
6
8
10
12
Quarter

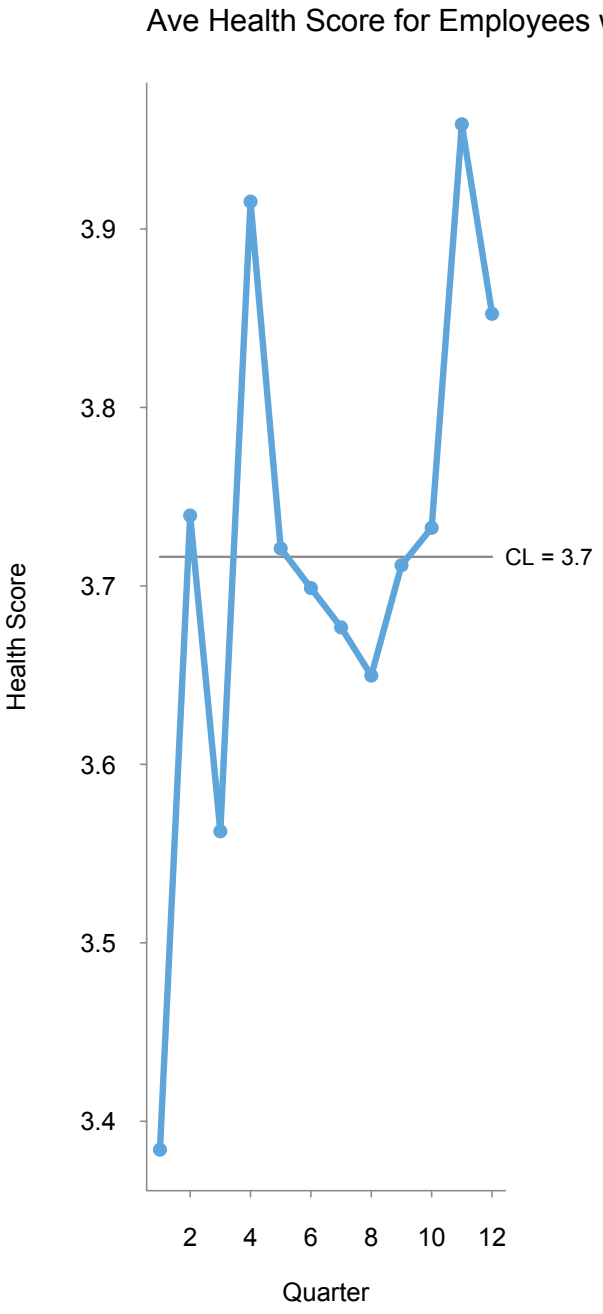
Hospital Visit this Quarter

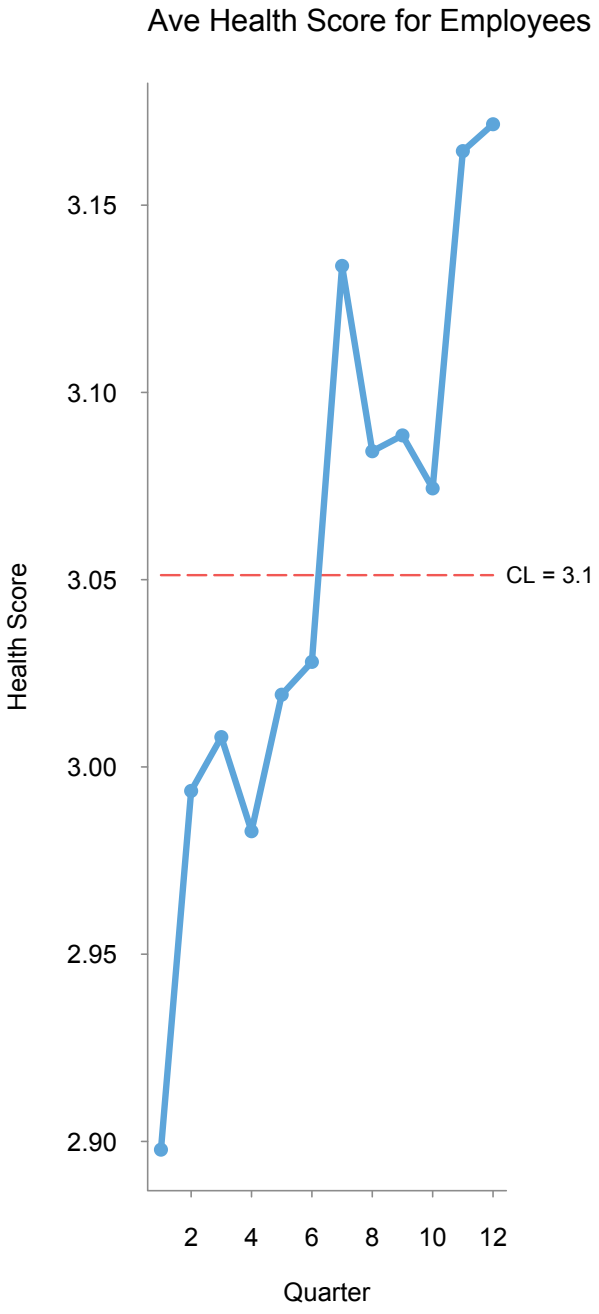
Based on the data set, I created a simple 2x2 table to show that 89% of employees who work for Company A did not visit the hospital this quarter while 11% of employees who work for Company A did visit the hospital this quarter.

When we look at the box plot titled, “Hospital Visit Plotted Against Health Score”, we see that those employees who have not visited the hospital has an overall lower median health score that compared to those employees who did visit the hospital this quarter.

I then made a side-by-side of 2 run charts to show the difference between the health scores of employees who visited the hospital vs the health score of employees who did not visit the hospital. Based on the run charts, while both groups’ health scores appear to be increasing over 12 quarters, employees who have not visited the hospital this quarter have a lower median health score compared to those employees who have visited the hospital this quarter. Those who visited the hospital this quarter: Quarter1= 3.38 Quarter6= 3.70 Quarter12 = 3.85 while those who did not visit the hospital this quarter: Quarter1= 2.90 Quarter6= 3.03 Quarter12= 3.17. In other words, those employees who visited the hospital this quarter are sicker than those employees who did not visit the hospital this quarter.

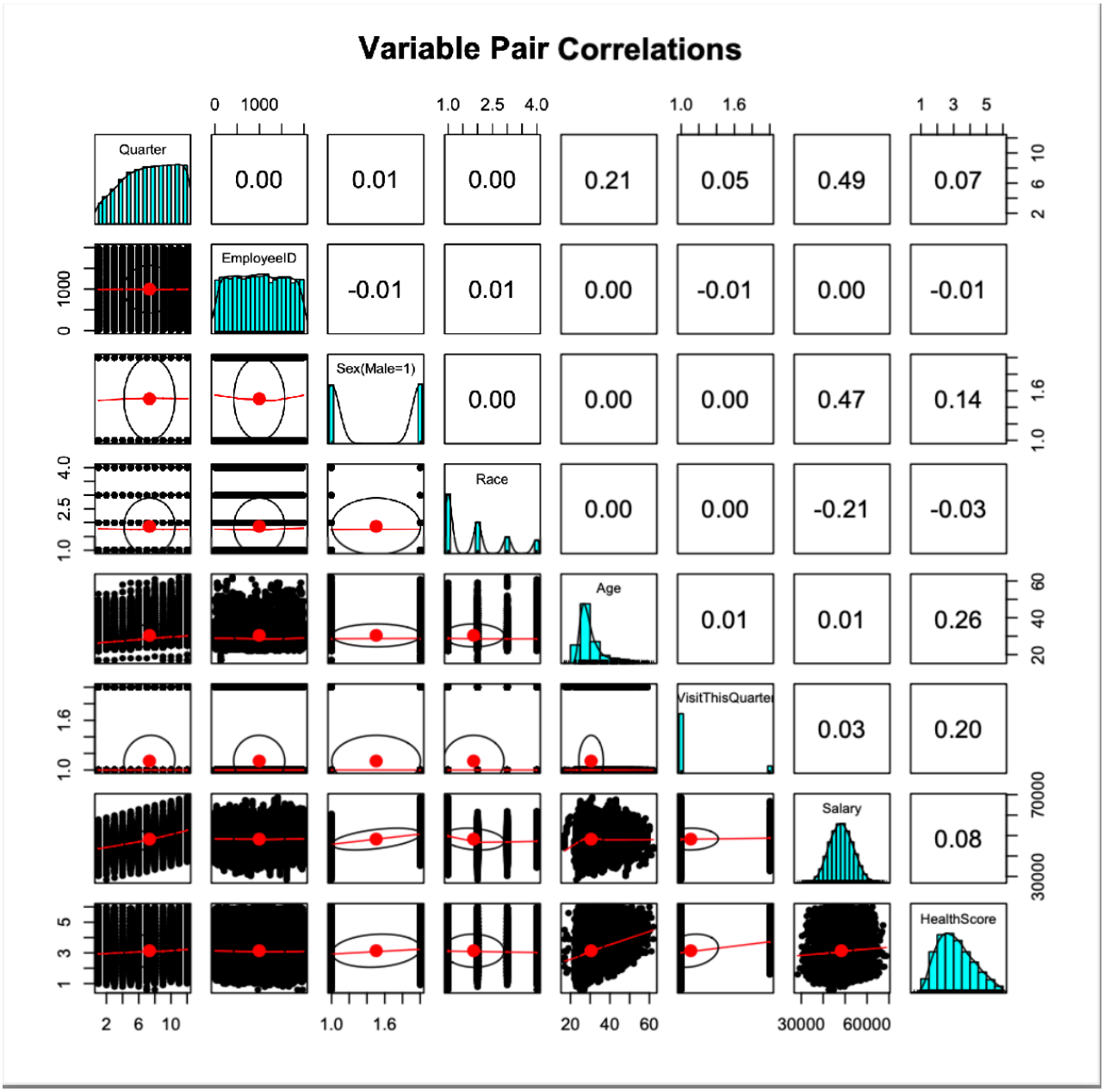

Hospital Visit Plotted Against Health Score
Health Score
1
2
3
4
5
6
0
1
Hospital Visit this Quarter (Yes=1)


Ave Health Score for Employees w/ R
Health Score
3.9
3.8
3.7
3.6
3.5
3.4
CL = 3.7
2
4
6
8
10
12
Quarter

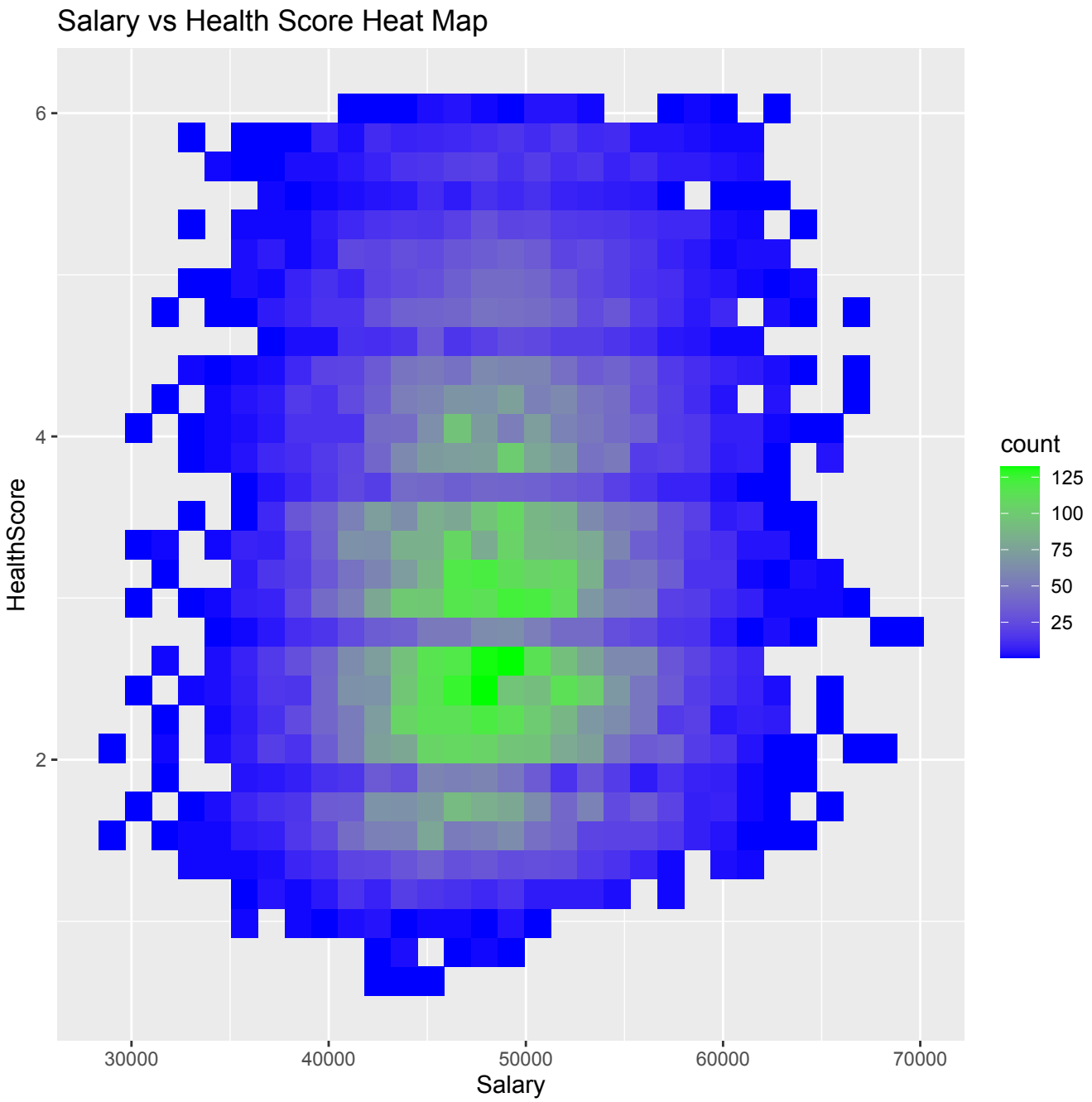

Ave Health Score for Employees w/ou
Health Score
3.15
3.10
3.05
3.00
2.95
2.90
CL = 3.1
2
4
6
8
10
12
Quarter

Question 2

Based on this data set, when we look at the correlation matrix, it appears that 'Age' is the highest correlated variable to 'Health Score' with a correlation coefficient, R = 0.26. 'Age' is followed by 'Hospital Visit this Quarter' with a correlation coefficient, R = 0.2. The third highest correlated variable with 'Health Score' is 'Sex', with a correlation coefficient, R = 0.14. The last two are 'Salary' with a correlation coefficient, R= 0.08, and then 'Quarter', with that correlation coefficient, R = 0.07.



When I plot Salary vs Health Score, it does not appear that there is a linear relationship and if there is one, it looks weak. There could be a little correlation which I would like to figure out in part 2 of this project.



Question 3

In order to evaluate the claim that employees are getting sicker, I would first start with the number of people being tested — which varies each quarter. Note that each quarter is dependent on the last. In the first quarter, there are 636 employees, in the sixth quarter there are 1,658 employees, and in the 12th quarter the sample increased to 1,798 employees. So this raises the issue of possible missing values because we consider and rely on the repeated data measures to indicate whether the mean has variation over the period of time (12 quarters). Furthermore, recall the 1,278 entries of health score with a value of 10. If those were not typos, then the health score range of 0-6 does not fully represent the data since there are employees with health scores of 10.

The first step in the process is a to apply a normality test. I went ahead and performed a test of normality on the residuals using the Shapiro-Wilks test function in R. I found that the residuals from all 12 quarters violate the normality assumption. For the sake of time, I will describe the next step and then complete this step in part 2 of this project. As the second step, I would perform a test to see if the increase in Health Score mean is significant. I am curious as to the significance of the mean change because the company could be hiring new employees that are driving up the company's average health score or there is a possibility that the employees' health is getting worse. If we find the mean change is significant, then we can conclude that the change can be attributed to the variables and not by hiring more people.