**Acumen Data Analysis Exercise**

Name Dhruv Singh

**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 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."

**Task**

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."

**Requirements**

1. Spend no more than two hours on this data exercise.

2. Include a written explanation of each table, chart, and figure you create.

3. Write out any assumptions about the data you have.

• We are interested in understanding how you are think about the data.

• We cannot answer any questions about the prompt, so it is important that you state your thoughts.

4. Place all output in this Excel workbook and show all work, calculations, and code, if you use it.

5. Do not send any additional files outside of the completed excel workbook. If you used other software to conduct the analysis please include all code in the workbook.

6. Save the file as ""Acumen\_Data\_Analysis\_Exercise\_[last name]\_[first name]"" and email it to recruiting@acumenllc.com."

**Questions**

**1. Understanding the Data**

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





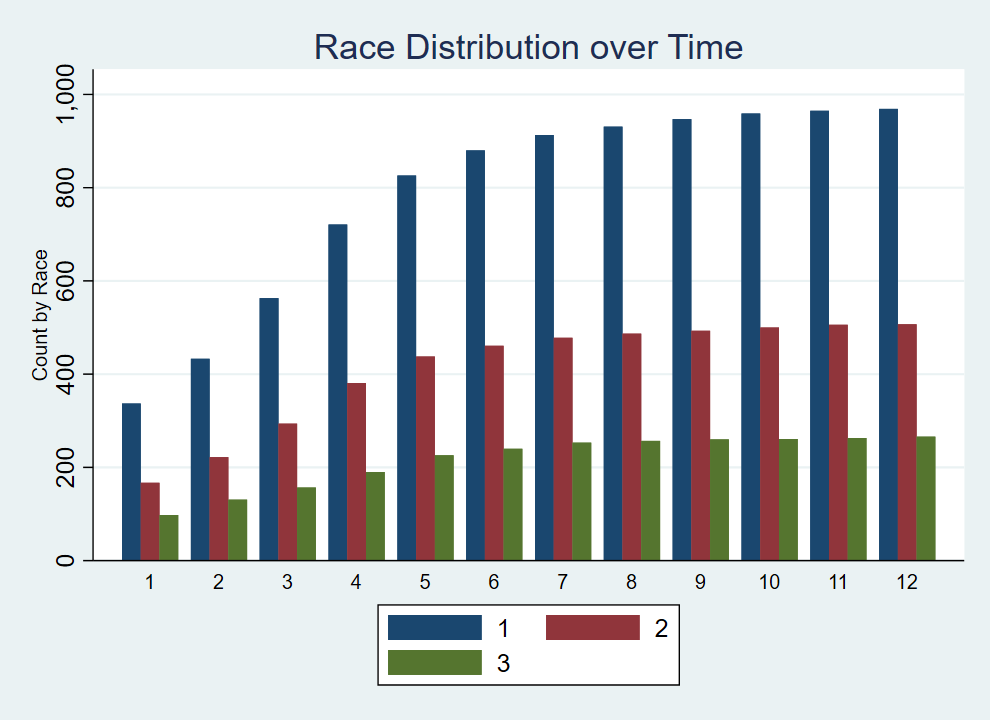
Answer: There are missing values under the race column – 2,2123 ~ or 11.11%

There are missing values under the sexmale dummy column – 71 ~ or 0.37%

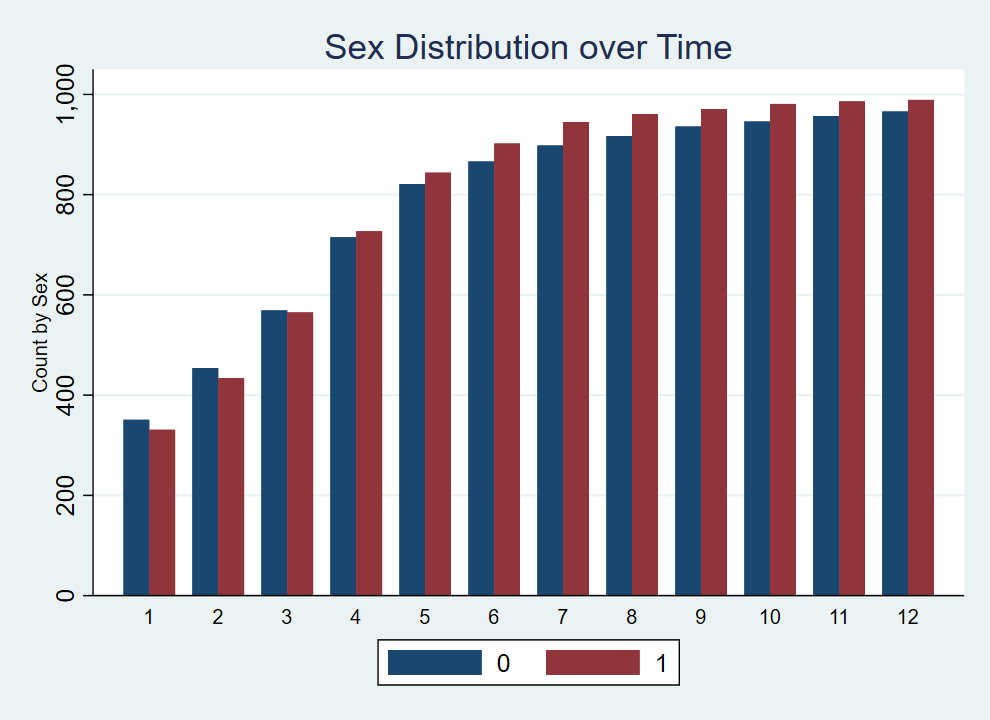
b) What are the characteristics of employees at Company A? Do these demographics change over time?

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

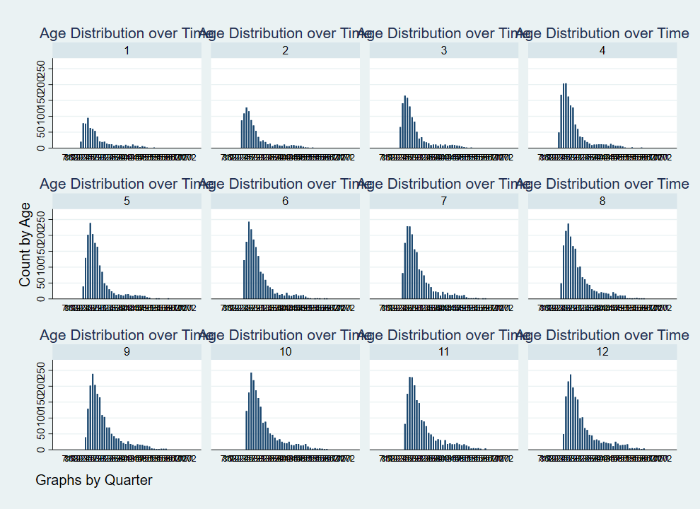
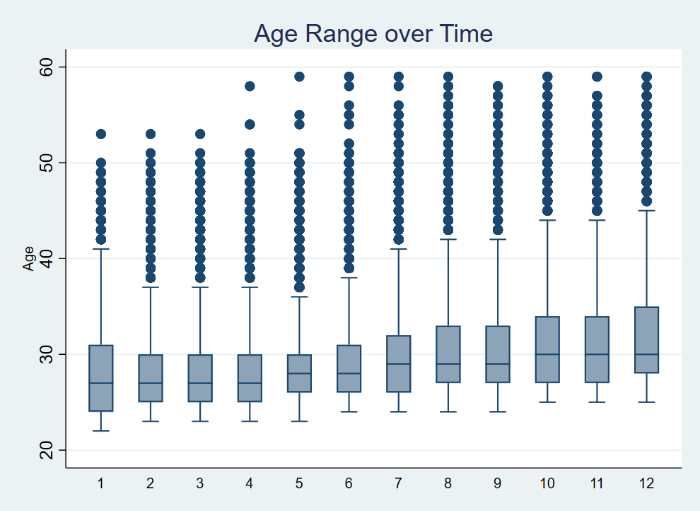
1. Answer: The distribution seems to favor a majority of race 1 over time. While numbers of all races grow, numbers of race 1 seems to grow more rapidly than those of race 2 and 3.



1. Answer: The distribution of sexes is tilted in favor of females at the outset, in quarter 1. As overall numbers grow over time, the ratio tilts in favor of males beginning in quarter 4. The gap does not widen by a large margin.



1. Answer: There is a wide amount of variation in age range over each quarter. After trimming outliers for a clearer picture (eg, nonsensical values well over a 100), a clearer picture emerges. Median age starts somewhere around late 20s in quarter 1, and gradually increases until in quarter 12 it hovers around age 30. Long tails and right-skewed distributions in each quarter are confirmed in graph2.

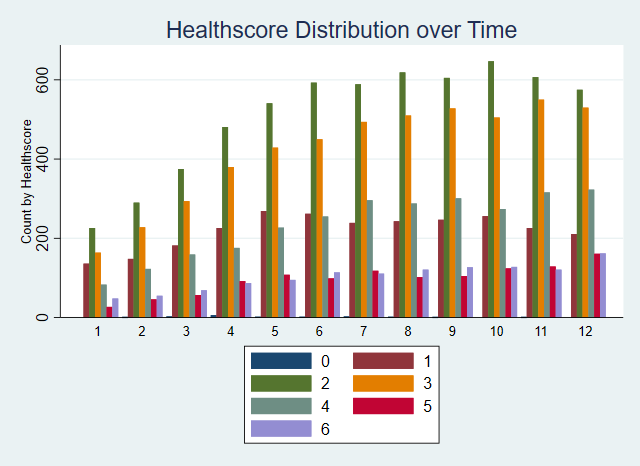


**2. Exploring Relationships**

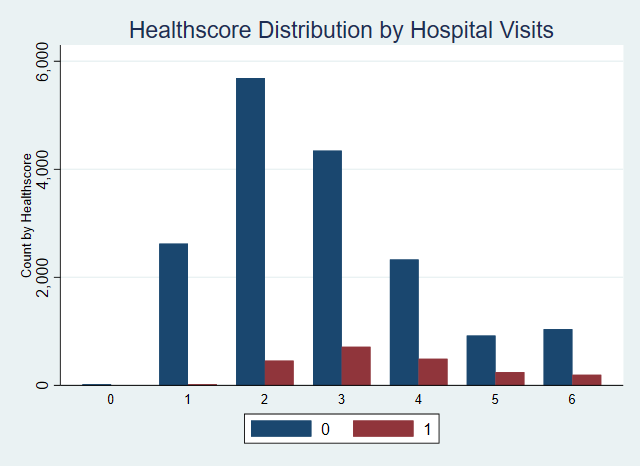
a) Which characteristics are associated with the health score?



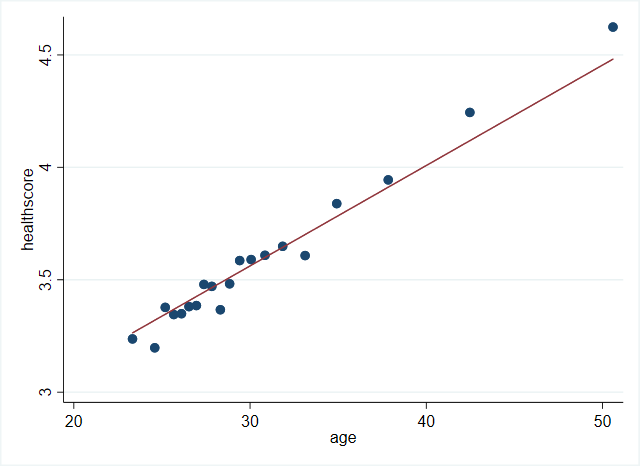
Positively related variables: quarter, sexmale1, age, hospitalvisits, and salary



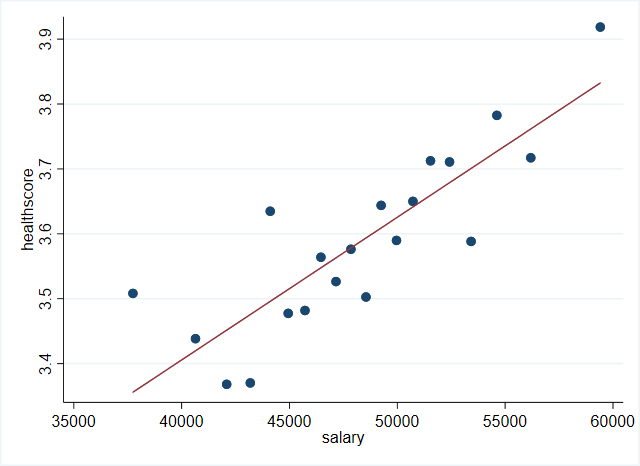
While this image doesn’t capture much of the noise, it shows a far higher frequency of employees in good health over time, and a gradually increasing number of employees in poor health.



Rather than visits to the hospital driving higher healthscores, it seems to be that people with poorer health tend to visit the hospital more.



Healthscore is positively correlated with age.



Healthscore is positively correlated with salary.

Negatively related variables: race



**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.

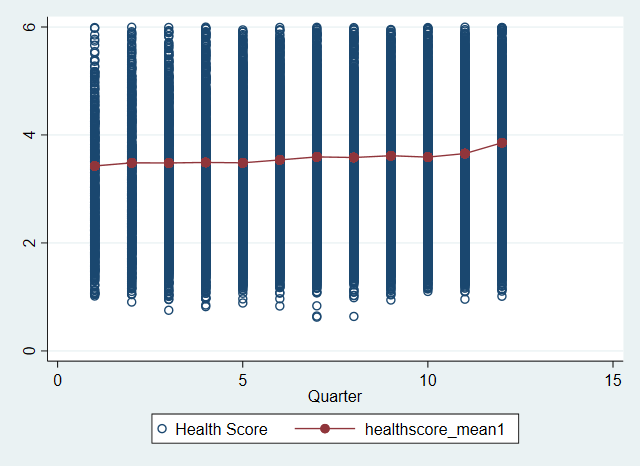
First list how you would evaluate the claim. Then, time-permitting, implement the steps you suggested."



As we know from the provided data, the panel is unbalanced, in that there are new entrants every quarter. However, we can still run a standard fixed effects model with the data we have.

In order to evaluate this claim, we would formulate a model to check whether average health scores for individuals are increasing over time. We want to account for the fact that we’re tracking individuals over time, and so, we would want to see their health outcome at the start and end of the period. To do this, we can use a fixed effects model that controls for individual characteristics, and for shocks and variation based on time period.

The claim that the mean healthscore is increasing over time is indeed valid as seen in the picture below, albeit gradually.



For the purposes of this exercise, we intuitively select a fixed effects model. Below we present 3 versions of models to see which best explain the variation in healthscores over time.

**Model 1: panel regression, time and individual fixed effects**

xtreg healthscore i.quarter sexmale1 race age hospitalvisitthisquarter1yes salary, fe



**Model 2: time and employee id fe**

xi: areg healthscore i.quarter sexmale1 race age hospitalvisitthisquarter1yes salary, absorb(employeeid)



**Model 3: Random effects**

xtreg healthscore sexmale1 race age hospitalvisitthisquarter1yes salary, re robust



All these models point to the same thing, that hospital visits are the only significantly correlated variable to healthscores.

However, judging by the fact that coefficients on all time periods are positive, we can infer that on average healthscores are increasing over time.