**Umuzi assignment: Predicting salaries differences**

**Start: Monday 11 June**

**Due: Tuesday 12 June 12pm**

**Submission:** Commit your notebook changes to GitHub. Once you are done, create an html report from the notebook. The last GitHub commit done before 12pm will be marked.

**To complete this assignment, you should watch the following videos:**

Crash Course Statistics: Measures of Spread

Crash Course Statistics: Plots, Outliers and Justin Timberlake

Crash Course Statistics: Correlation Doesn’t Equal Causation

In the next series of challenges, we will predict employee salaries from different employee characteristics (or features). We ask employees working in four different fields to state their salary and some background information.

Import the data salary.csv to a Jupyter Notebook and answer the following questions:

**Question 1**

1. How many responders are there? Are there any missing values in any of the variables?
2. What is the lowest salary and highest salary in the group?
3. What is the mean salary for the sample? Include the standard error of the mean.
4. What is the standard deviation for the years worked?
5. What is the median salary for the sample?
6. What is the interquartile range for salary in the sample?
7. How many men are there in the sample? How many women are there in the sample? Present this information in a table.
8. How many women are executives compared to men?
9. Create a histogram for the variable Salary.
10. Examine the histogram and describe the distribution for Salary.
11. Create a bar graph to show the different average salaries of men and women. (Bonus: Add error bars to the bars showing the 95% confidence interval). What does the graph tell you about the difference between men and women’s salaries?
12. Create a scatterplot showing the relationship between Years Worked and Salary (don’t forget to insert a trend line). What is the relationship between Years Worked and Salary?
13. Describe any patterns in the scatterplot. Do you notice any unusual/extreme values that do not fit the general trend? If you see any unusual values, briefly describe them (Who are they? In what way are they different?)
14. Using the pearsonr function from the scipy.stats package, calculate the Pearson correlation coefficient (and its corresponding *p* value) to determine the nature of the relationship between Years Worked and Salary. See help(pearsonr) for help on this function.
    1. Interpret the size and direction of the correlation statistic.
    2. Is the relationship statistically significant? Report the appropriate statistic(s) to support your answer.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Explanation |
| salary | Continuous | Annual salary in dollars |
| exprior | Continuous | Years of experience prior to working in this field |
| yearsworked | Continuous | Years worked in this field |
| yearsrank | Continuous | Years worked at current rank |
| market | Continuous | Market value (1 = salary at market value for position, <1 salary lower than market value for position, >1 salary higher than market value for position) |
| degree | Binary | Has degree (0 = no, 1 = yes) |
| otherqual | Binary | Has other post-secondary qualification (0 = no, 1 = yes) |
| position | Categorical | Position (1 = Junior Employee, 2 = Manager, 3 = Executive) |
| male | Binary | 0 = no, 1 = yes |
| Field | Categorical | Field of work (1 = Engineering, 2 = Finance, 3 = Human Resources, 4 = Marketing) |
| yearsabs | Continuous | Years absent from work (e.g. due to illness / child rearing / personal reasons) |