USE PYTHON!!!!

You are going to conduct a data science analysis from conception through simple linear regression and interpretation. You may select any topic and use any dataset that you like as long as it’s publicly available and it contains two continuous variables whose association you are interested in examining. We will also provide datasets that you can use if you wish, though we encourage you to explore and ﬁnd one that’s relevant to your interests and goals.

1 Pre-step

1. Describe brieﬂy the question you would like to answer or the topic you would like to explore. Essentially, what do you hope to learn from your analysis?

2 Data

2. Find a dataset that may help you explore at least some of these questions. First, describe where you found the data set. Second, describe how you found it. Third, describe at least two variables in the dataset that are relevant to the analysis you described above. Finally, describe the unit of observation (individual, city, etc.).

3. If you could change this dataset in one way to make it better for your analysis, what would that change be and how could it improve your analysis?

4. Import the dataset into Jupyter using any method you like and show the ﬁrst ﬁve observations. If you had to do any pre-work to get the data into an uploadable format please describe it brieﬂy. (If you didn’t, please say so as well.)

3 Initial analysis

5. Conduct at least two diﬀerent manipulations of your now-ready table that help you understand something of interest about the dataset (e.g., you might explore options like sort, shape, value counts, groupby, etc.). Why did you choose these two, and what have you learned? (Hint: You may need to do a bit work to get the data into a format that is usable for you – e.g., renaming columns, changing data types, etc. If any of this was necessary, show your code and brieﬂy explain why you made these changes)

6. Generate two diﬀerent types of graphs of any kind that are useful to you to better understand what you’re interested in. They don’t need to be formatted particularly beautifully, but you do need to use two diﬀerent types of graphs (e.g., a bar chart and a scatterplot) and explain what you hoped to understand, why you chose these graphs, and whether they’re useful in improving your understanding.

4 Hypothesis formation

7. What is your dependent variable and independent variable? Brieﬂy describe how they are measured in this dataset. (Remember, they’ll both need to be continuous variables.)

8. Calculate the correlation coeﬃcient between your two variables and interpret the result.

9. Write out your regression model as an equation.

10. Write out your null and alternative hypotheses.

5 Regression analysis

11. Estimate the regression equation you speciﬁed above and show the regression output.

12. What do the results in the regression output tell you? Interpret the coeﬃcient, p-value, and conﬁdence interval for your independent variable (you don’t have to do the intercept) and the R2.

13. Which hypothesis do you reject and fail to reject, and why?

14. Generate the residual plot and comment on any heteroskedasticity. What does this imply for your inference?

6 Conclusions

15. What biases might be present in the sample itself that could be aﬀecting the outcome? Discuss at least two sources of bias.

16. Considering all the work you’ve done, including the regression output, the results of your hypothesis tests, and any biases present in the data, what conclusions, however tentative, can

you draw from your analysis about the relationship between your two variables of interest?

17. What is your analysis’s greatest weakness? In other words, what are the best reasons to be cautious about what we can learn from it?