Jupyter Notebook

starter-code(autosaved)Current Kernel LogoLogout

Python 3

Not Trusted

* [File](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Edit](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [View](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Insert](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Cell](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Kernel](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Widgets](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)
* [Help](http://localhost:8888/notebooks/Documents/DIS-Project-1/project-1/code/starter-code.ipynb#)

Run

CodeMarkdownRaw NBConvertHeading-

# **Project 1**

## **Step 1: Load the data and perform basic operations.**

##### ***1. Load the data in using pandas.***

In [ ]:

​

##### ***2. Print the first ten rows of each dataframe.***

In [ ]:

​

##### ***3. Describe in words what each variable (column) is.***

In [ ]:

​

##### ***4. Does the data look complete? Are there any obvious issues with the observations?***

In [ ]:

​

##### ***5. Print the types of each column.***

In [ ]:

​

##### ***6. Do any types need to be reassigned? If so, go ahead and do it.***

In [ ]:

​

##### ***7. Create a dictionary for each column mapping the State to its respective value for that column. (For example, you should have three SAT dictionaries.)***

In [ ]:

​

##### ***8. Create one dictionary where each key is the column name, and each value is an iterable (a list or a Pandas Series) of all the values in that column.***

In [ ]:

​

##### ***9. Merge the dataframes on the state column.***

In [ ]:

​

##### ***10. Change the names of the columns so you can distinguish between the SAT columns and the ACT columns.***

In [ ]:

​

##### ***11. Print the minimum and maximum of each numeric column in the data frame.***

In [ ]:

​

##### ***12. Write a function using only list comprehensions, no loops, to compute standard deviation. Using this function, calculate the standard deviation of each numeric column in both data sets. Add these to a list called sd.***

σ=1n∑i=1n(xi−μ)2⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯σ=1n∑i=1n(xi−μ)2

In [ ]:

​

## **Step 2: Manipulate the dataframe**

##### ***13. Turn the list sd into a new observation in your dataset.***

In [ ]:

​

##### ***14. Sort the dataframe by the values in a numeric column (e.g. observations descending by SAT participation rate)***

In [ ]:

​

##### ***15. Use a boolean filter to display only observations with a score above a certain threshold (e.g. only states with a participation rate above 50%)***

In [ ]:

​

## **Step 3: Visualize the data**

##### ***16. Using MatPlotLib and PyPlot, plot the distribution of the Rate columns for both SAT and ACT using histograms. (You should have two histograms. You might find*** [***this link***](https://matplotlib.org/users/pyplot_tutorial.html#working-with-multiple-figures-and-axes) ***helpful in organizing one plot above the other.)***

In [ ]:

​

##### ***17. Plot the Math(s) distributions from both data sets.***

In [ ]:

​

##### ***18. Plot the Verbal distributions from both data sets.***

In [ ]:

​

##### ***19. When we make assumptions about how data are distributed, what is the most common assumption?***

In [ ]:

​

##### ***20. Does this assumption hold true for any of our columns? Which?***

In [ ]:

​

##### ***21. Plot some scatterplots examining relationships between all variables.***

In [ ]:

​

##### ***22. Are there any interesting relationships to note?***

In [ ]:

​

##### ***23. Create box plots for each variable.***

In [ ]:

​

##### ***BONUS: Using Tableau, create a heat map for each variable using a map of the US.***

In [ ]:

​

## **Step 4: Descriptive and Inferential Statistics**

##### ***24. Summarize each distribution. As data scientists, be sure to back up these summaries with statistics. (Hint: What are the three things we care about when describing distributions?)***

In [ ]:

​

##### ***25. Summarize each relationship. Be sure to back up these summaries with statistics.***

In [ ]:

​

##### ***26. Execute a hypothesis test that evaluates whether or not the mean SAT and ACT participation rates are the same. Use*** α=0.05α=0.05***. Be sure to interpret your results.***

In [ ]:

​

##### ***27. Generate and interpret 95% confidence intervals for SAT and ACT participation rates.***

In [ ]:

​

##### ***28. Given your answer to 26, was your answer to 27 surprising? Why?***

In [ ]:

​

##### ***29. Is it appropriate to generate correlation between SAT and ACT math scores? Why?***

In [ ]:

​

##### ***30. Suppose we only seek to understand the relationship between SAT and ACT data in 2017. Does it make sense to conduct statistical inference given the data we have? Why?***

In [ ]:

​