Python Setup, Data Analysis

Veeral Patel

Regis University

March 8th, 2021

1. **Introduction**

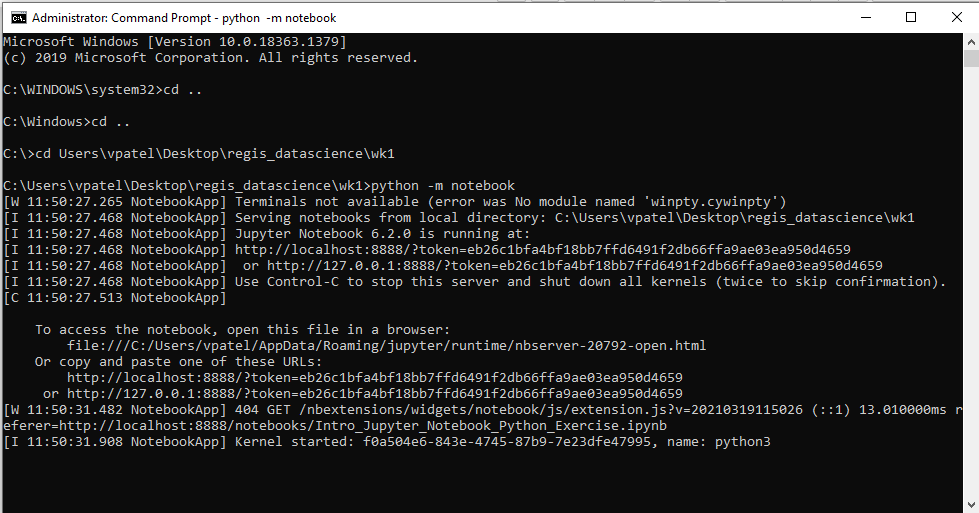
For this week’s assignment, we were to get python installed on our local computer as well as any necessary packages that would be required. I decided to install PyCharm as well as Jupyter notebook. After doing that, we were expected to understand how to assign variables, do basic arithmetic, do string concatenation, and understand the print statement. Briefly, we were asked to look at graphic with matplotlib as well as how and when to use functions in python. After that, we were asked to look at a day.csv file and figure out the mean and std, as well as plot a scatter plot with the count and dteday fields.

1. **Methods/Code/Screenshots**

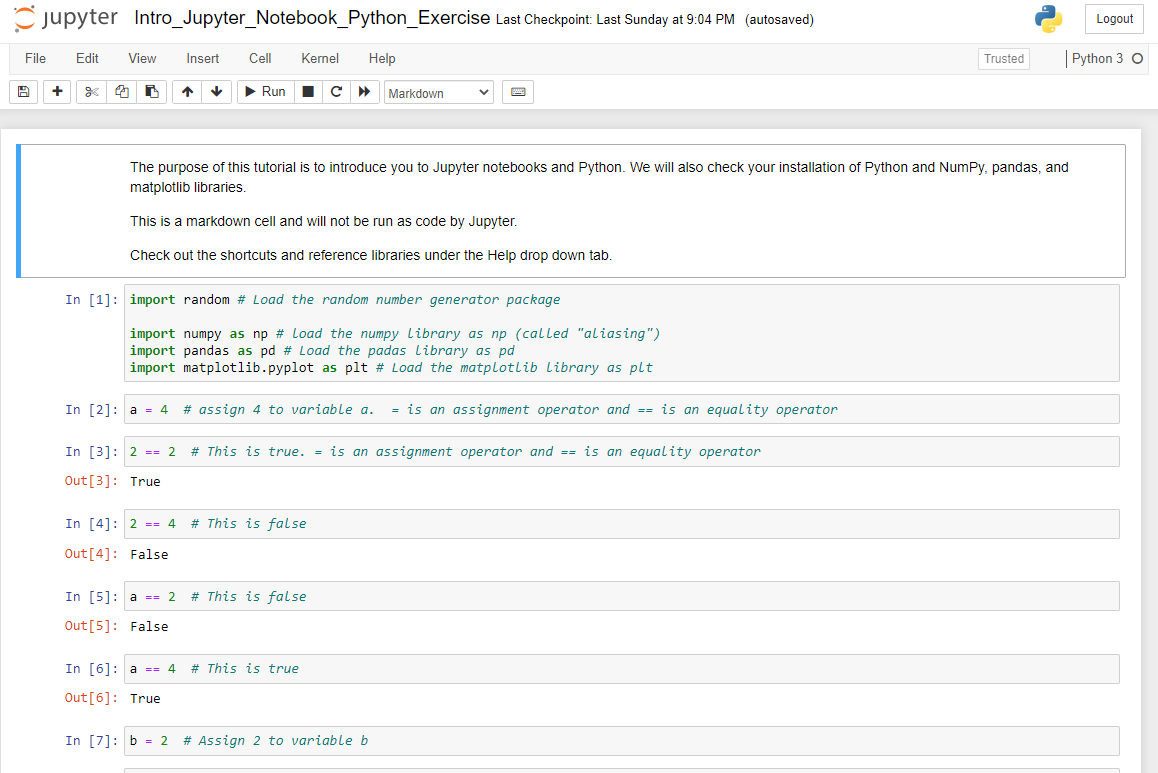
I used a windows computer and I had to install Python, Pip, Git Bash, Jupyter, and all of the dependency modules that were needed to perform the week one homework. I have worked with python before but it was for scripting. This was a new adventure for me to work with Jupyter Lab. I watched a few quick tutorials on YouTube in order to assist me with getting everything installed on my computer. Most of the code in this assignment was supposed by the professor. I also used a lot of the assignment code as a refresher into python.

In order to analyze the day.csv file, I used matplotlib and the pandas library to get the data into a workable dataframe and then to plot the data onto a scatterplot.

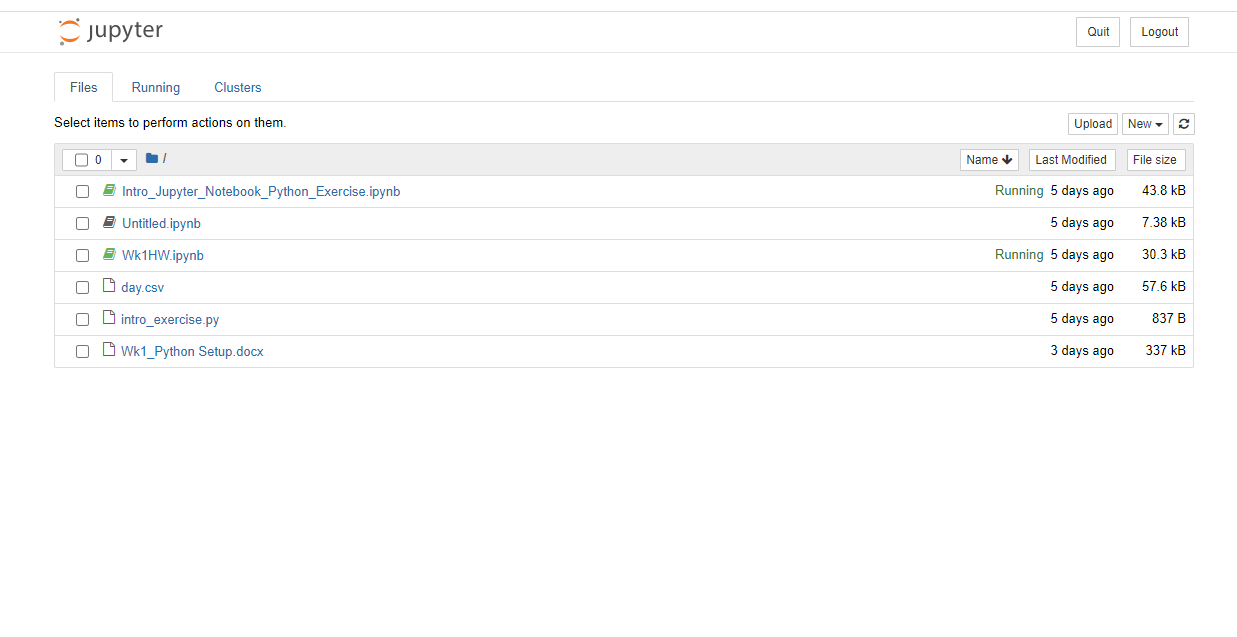
Jupyter Notebook running through command prompt:



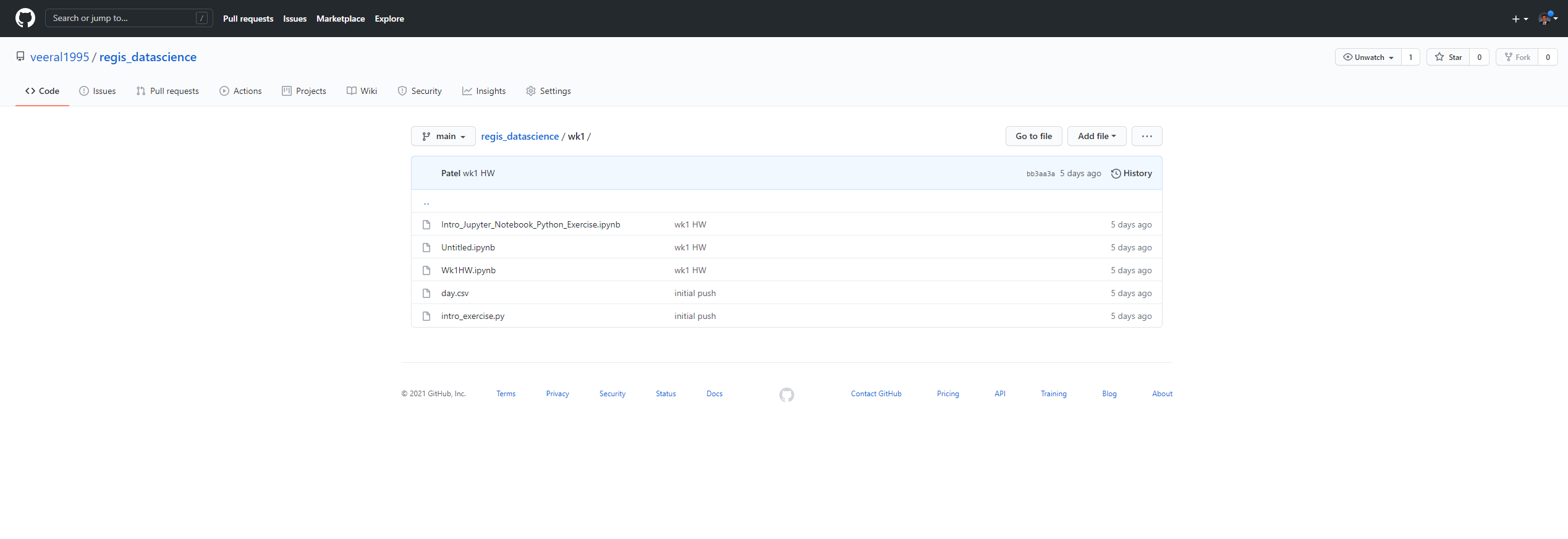
Jupyter Notebook running on my computer:



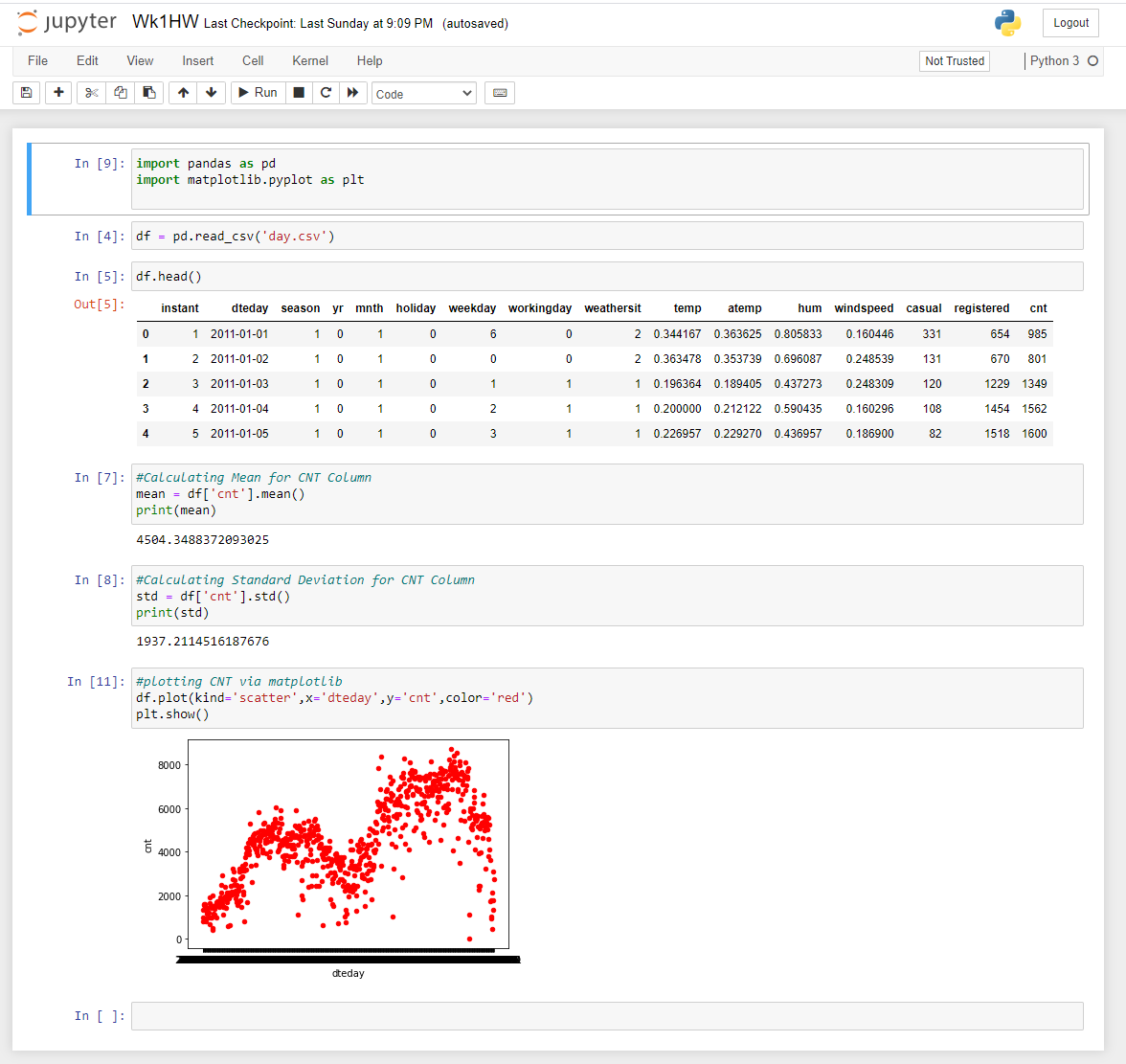
Directory of files on JupyterLab:



GitHub Repository:



1. **Results/Output**



1. **Analysis of Results:**

I felt that this week was a really good week and really served as a refresher for myself. In terms of what the results mean in the context of the lab, there were three main questions that we can breakdown:

Mean: The mean here means that the average count per day was 4504. In the larger context, this means that per day, we can expect the cnt to be around 4504

STD: The standard deviation is a measure of how spread out the numbers are. For the count column, the standard deviation was 1937. In a normal distribution, the high standard deviation signals that the numbers are far apart from each other.

Scatter plot: The scatter plot shows that there is a rising trend between count and date. We see that there are two large dips in count around the summer and winter time. All in all, a scatter plot is used to show the relationship between cnt and dteday.

1. **Conclusion:**

I thought that this exercise was a really good exercise to start to learn how to use our environment. A positive that I took from this exercise was that I was able to really understand how to start Jupyter lab and get started with a .csv file. Also, I learned the syntax of python and how to use Jupyter lab as well as getting it setup in PyCharm. This would apply to my field of software engineering because there are many times where python could be used for data validation but I have not used it in the past because I was not all that familiar with it. Also, Jupyter lab is a very good resource to use in terms of getting data analysis done in very fast and clean manner. I also love how easy it is to share code and allows non engineers to look at data with charts and graphs that show right in the browser.

I think that this would useful in really any field that is because now, there is data in every organization. Marketing, accounting, engineering, etc. can all gain an advantage by being able to use and analyze data. In this dataset, I think that the useful fields would be month, and season. I think that these things would allow us to see seasonality patterns more easily in the dataset.

This was a good experience to me, I really am looking forward to learning more in the future. I really like working with data and I want to be good at it so I can start to learn more about datasets.

1. **References:**

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & et al. (2016). *Jupyter Notebooks-a publishing format for reproducible computational workflows* (Vol. 2016, pp. 87-90).

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, *563*(7732), 145-147.

Weir, C. J., Butcher, I., Assi, V., Lewis, S. C., Murray, G. D., Langhorne, P., & Brady, M. C. (2018). Dealing with missing standard deviation and mean values in meta-analysis of continuous outcomes: a systematic review. *BMC medical research methodology*, *18*(1), 1-14.