Titanic Data Discovery

Michael Phung

Course: MIS542

Module 5 Option 1

January 17, 2021

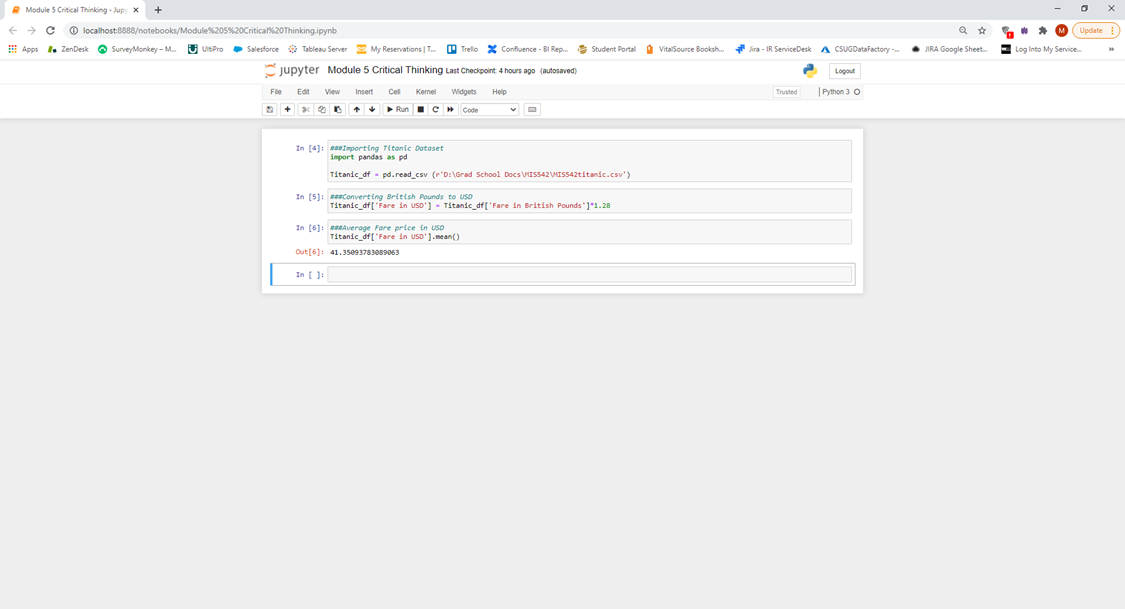
Colorado State University – Global Campus

Purpose

The purpose of this paper is to dive deeper into how Python can be a useful tool in data analytics. I got to practice how to open a CSV file to bring in a data frame. Another purpose is to dive deeper into looking more deeply into analyzing attributes and categorizing data points. I also learned how to build visualization surrounding the data and grouping by filtering the data.

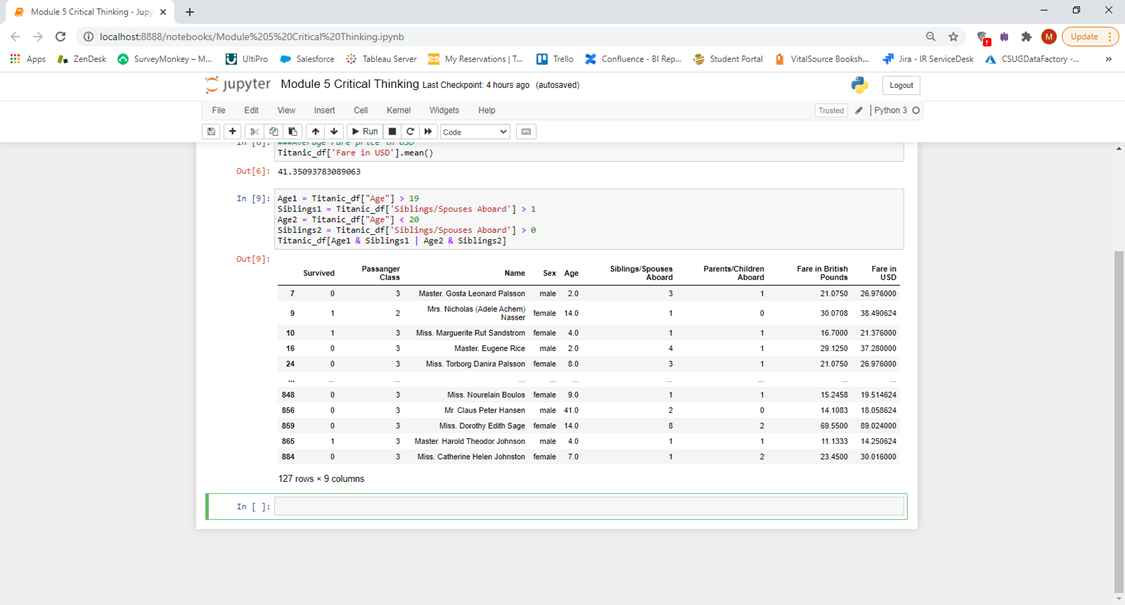
**Matrix Creation**

In this first part of this exercise, I imported the Titanic dataset and then converted the UK currency into US dollars easily by just multiplying the UK amount by 1.28, as demonstrated in figure one:



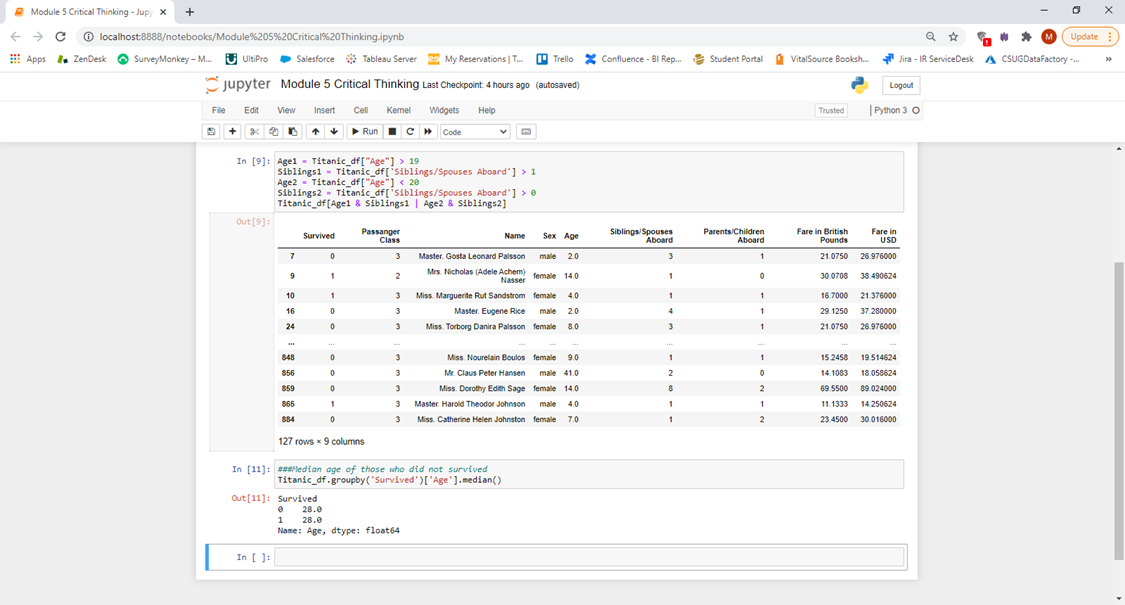
*Figure 1: Importing Titanic dataset and converting British pounds to USD*

In figure two, I am isolating passengers with siblings. How I did this is using an and if syntax by identifying each scenario based on module notes where I classified anyone over nineteen with more than one ‘sibling/spouse’ and anyone under twenty with one ore more ‘sibling/spouse’ and a person that had a sibling on board:



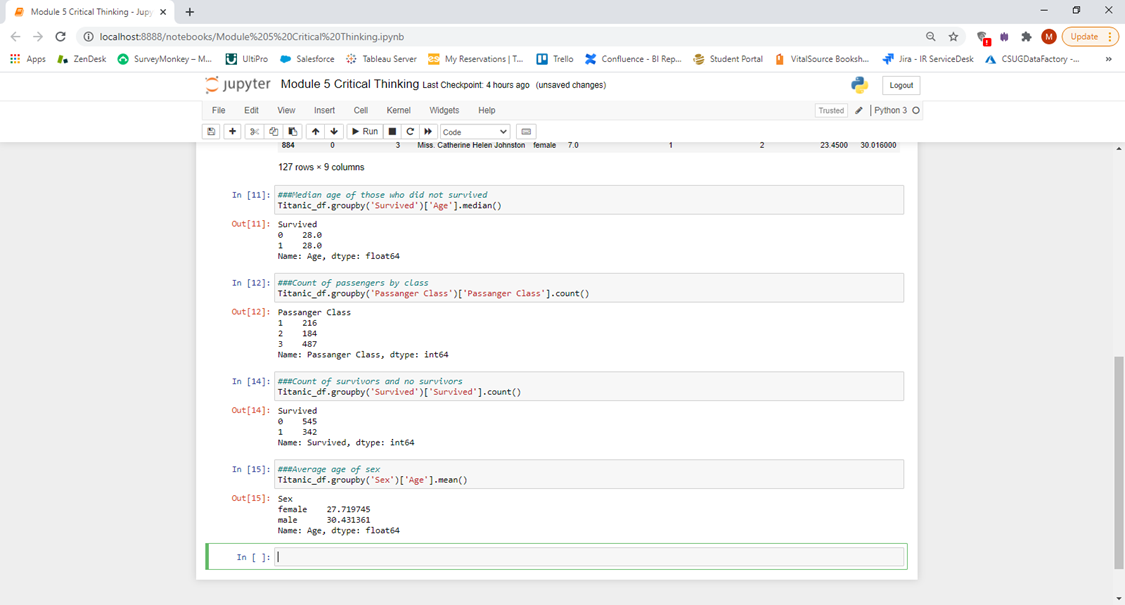
*Figure 2: Count of passengers with siblings*

In figure three, I show the meidan average age of those who didn’t survive by using the groupby and median function:



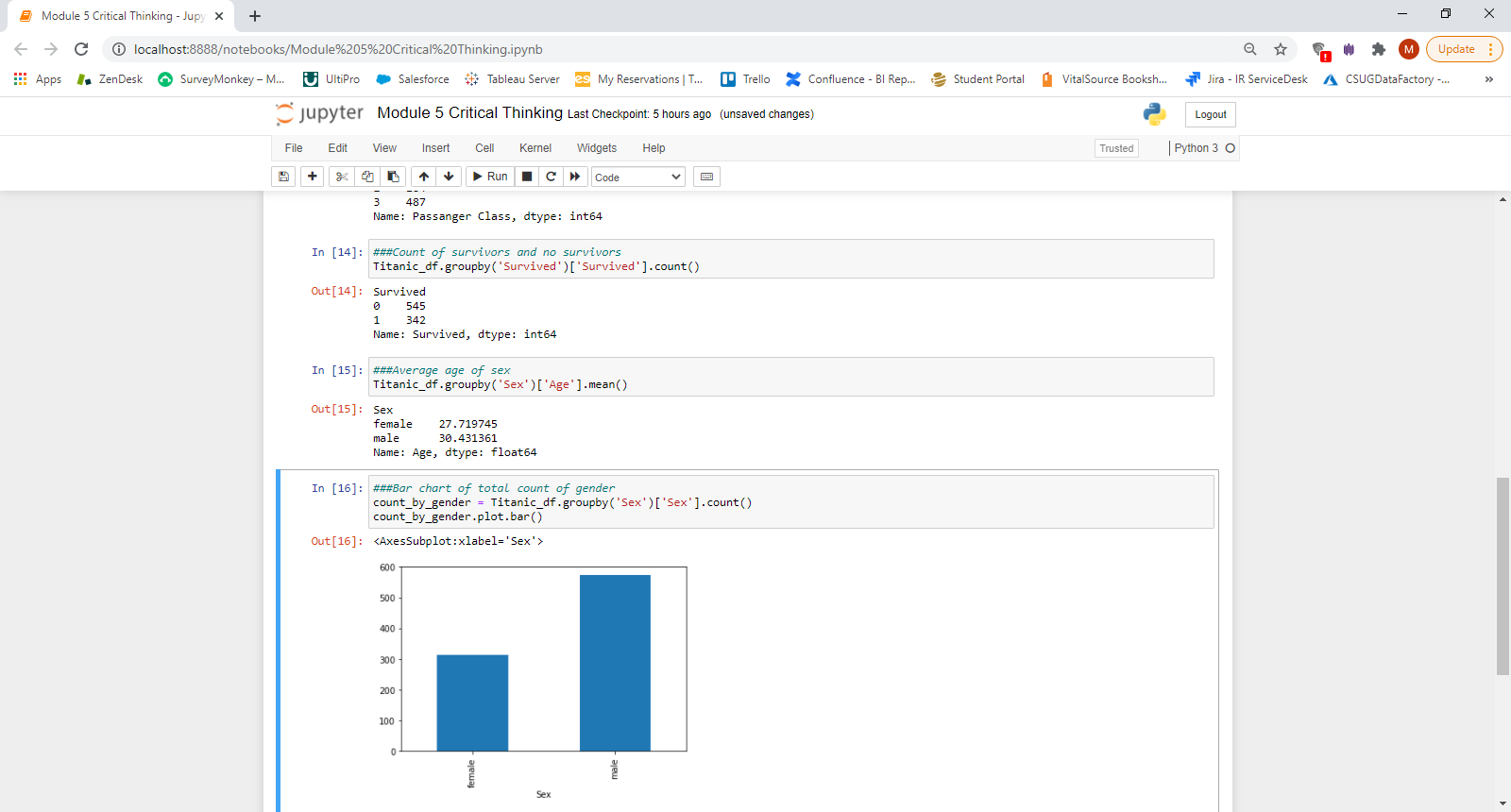
*Figure 3: Median age of those who did not survive*

Figure four demonstrates three additional questions outside of the assignment request. The first is count of passengers of by class. I repeated the groupby syntax to not only separate by class, but able to use the count function to count it by class. I then counted the numbers of those who survived and who did not survive based on the same method where the value ‘0’ was classified as not survived, and ‘1’ as survived. The next exploration was average age by gender where the method was the same but rather than using count, I used the mean function:

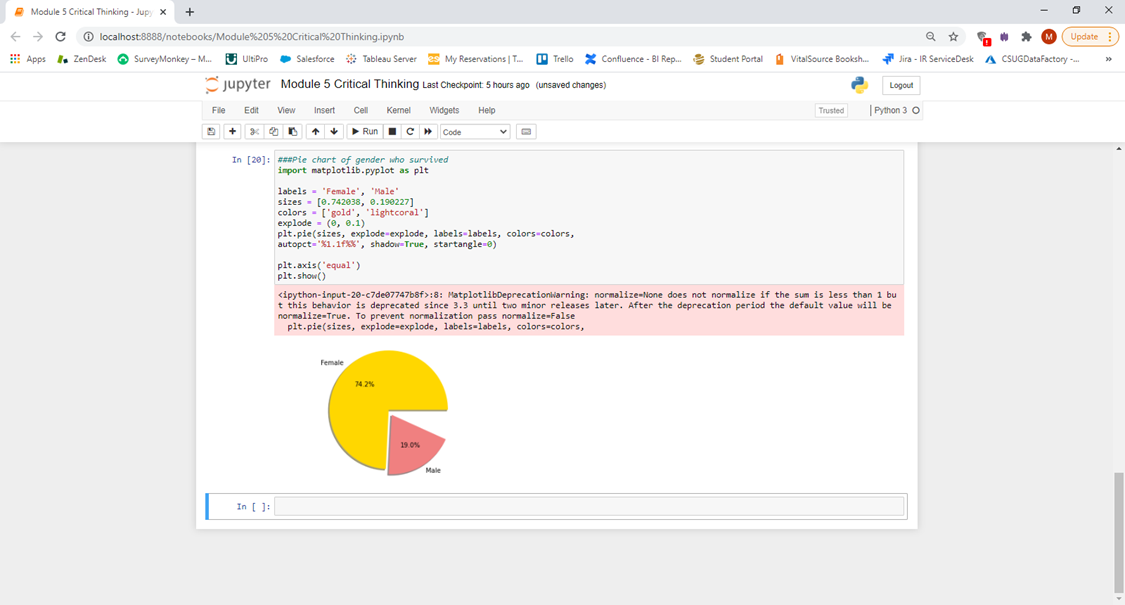


*Figure 4: Count of passengers by class, count of survivors and no survivors by class, average age of sex*

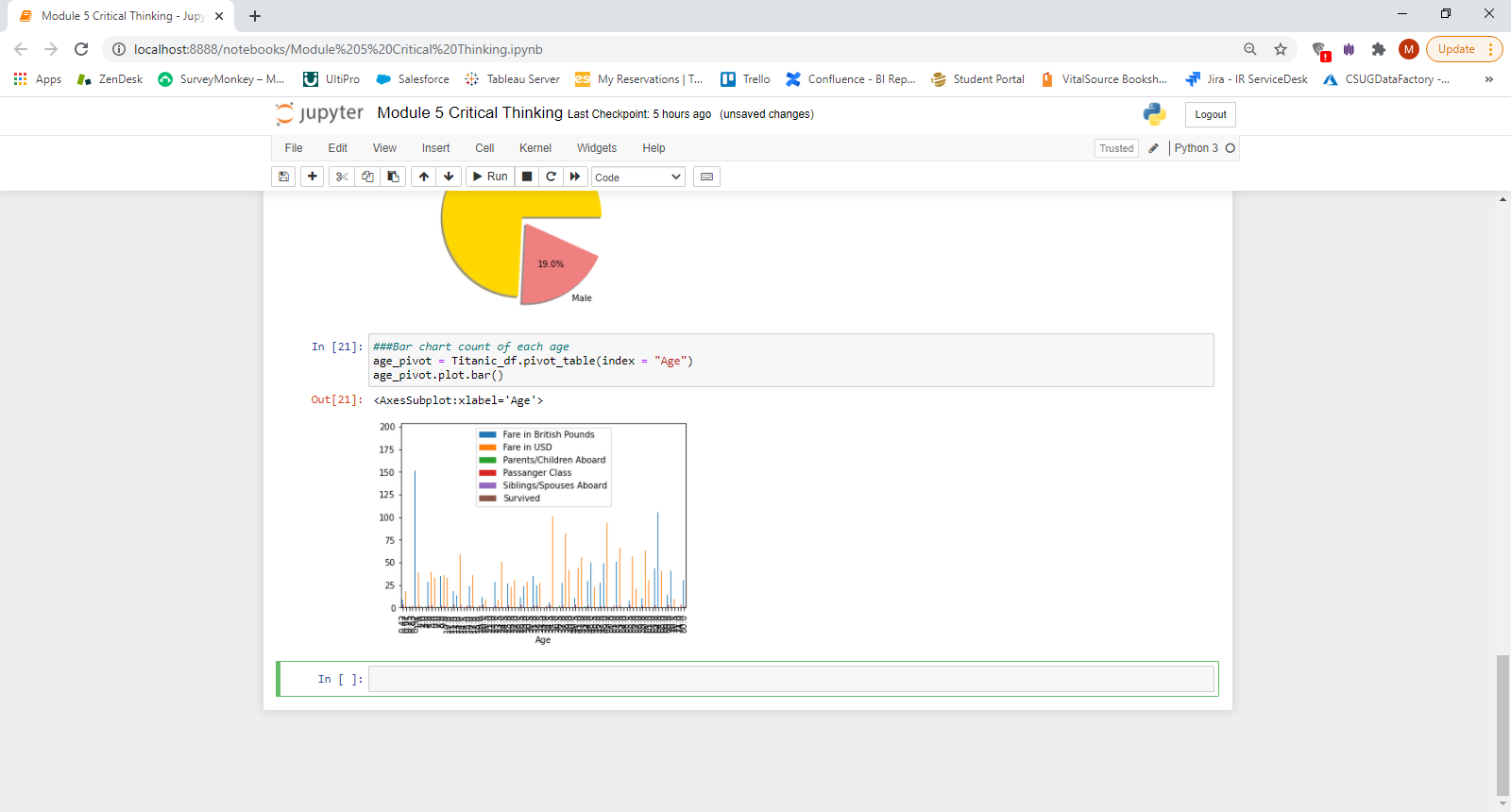
After some basic exploratory data by the numbers, I learned how to plot and visualization using the matplotlib package. I rearranged some table of the requirements to make it match the graphs better than was requested. First, I changed the pie chart for total counts of gender to a bar chart. I also changed the gender survival to a pie chart into a percentage because that visualization was more fitting. I left the counts of age into a bar chart the same. Figures five through seven demonstrates the code and visualization:



*Figure 5: Bar plot of total counts of gender*

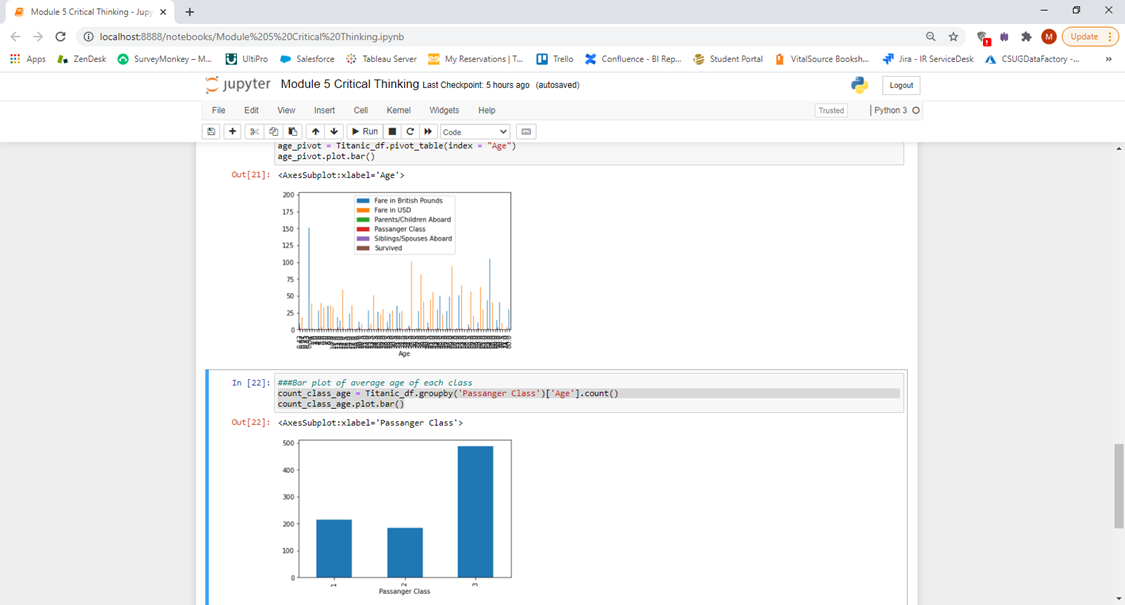


*Figure 6: Pie chart of percentage of gender who survived*

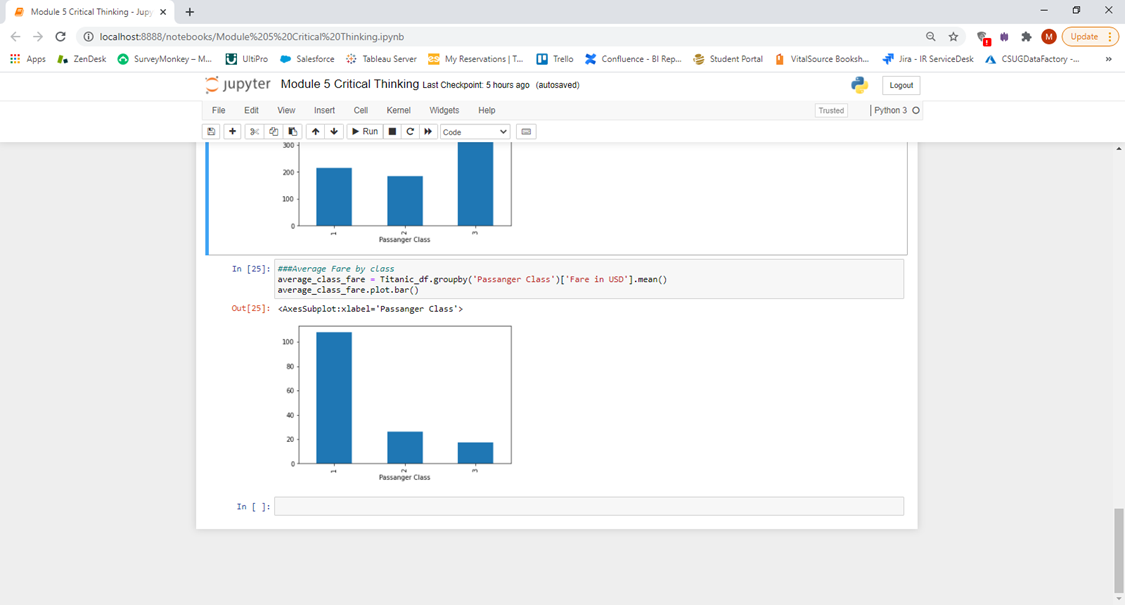


*Figure 7: Bar chart of count of each age*

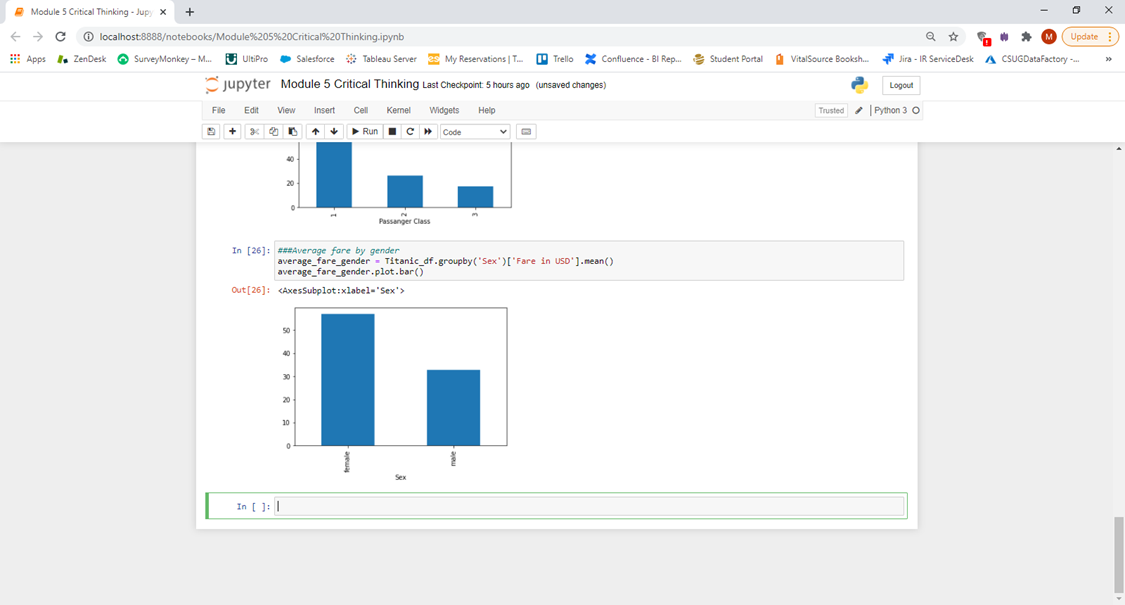
The next three figures demonstrates additional exploratory analysis of the data. I used the same bar plot method as figure five to show average age of each class in figure eight, average fare of each class in figure nine, and average fare by gender in figure ten:



*Figure 8: Average age of each class*



*Figure 9: Average fare of each class*



*Figure 10: Average fare by gender*

**Conclusion**

In this assignment, I learned to explore more deeply into a data set by learning to look at some basic individual summary statistics, but also learning a little more deeply based on categories of each data point. Also, I learned to visualize the data set by creating bar plots and pie plots. Granted I still need to learn more to enhance the visualizations with colors, data points and other enhancements.

References:

Yavuz, S. (2019, April 10). Getting started with Data Analysis with Python Pandas. Retrieved from: <https://towardsdatascience.com/getting-started-to-data-analysis-with-python-pandas-with-titanic-dataset-a195ab043c77>

Friedman, D. (2018, November 25). Value\_counts() Method: Count Unique Occurrences of Values in a Column. Retrieved from: <https://dfrieds.com/data-analysis/value-counts-python-pandas.html>

Devlin, J. (2017, October 25). Kaggle Fundamentals: The Titanic Competition. Retrieved from: https://www.dataquest.io/blog/kaggle-fundamentals/