**M5 CT Option 1**

Krista O’Neill

Colorado State University Global

MIS542 M5 CT Option 1

Dr. Emmanuel Tsukerman

December 19, 2021

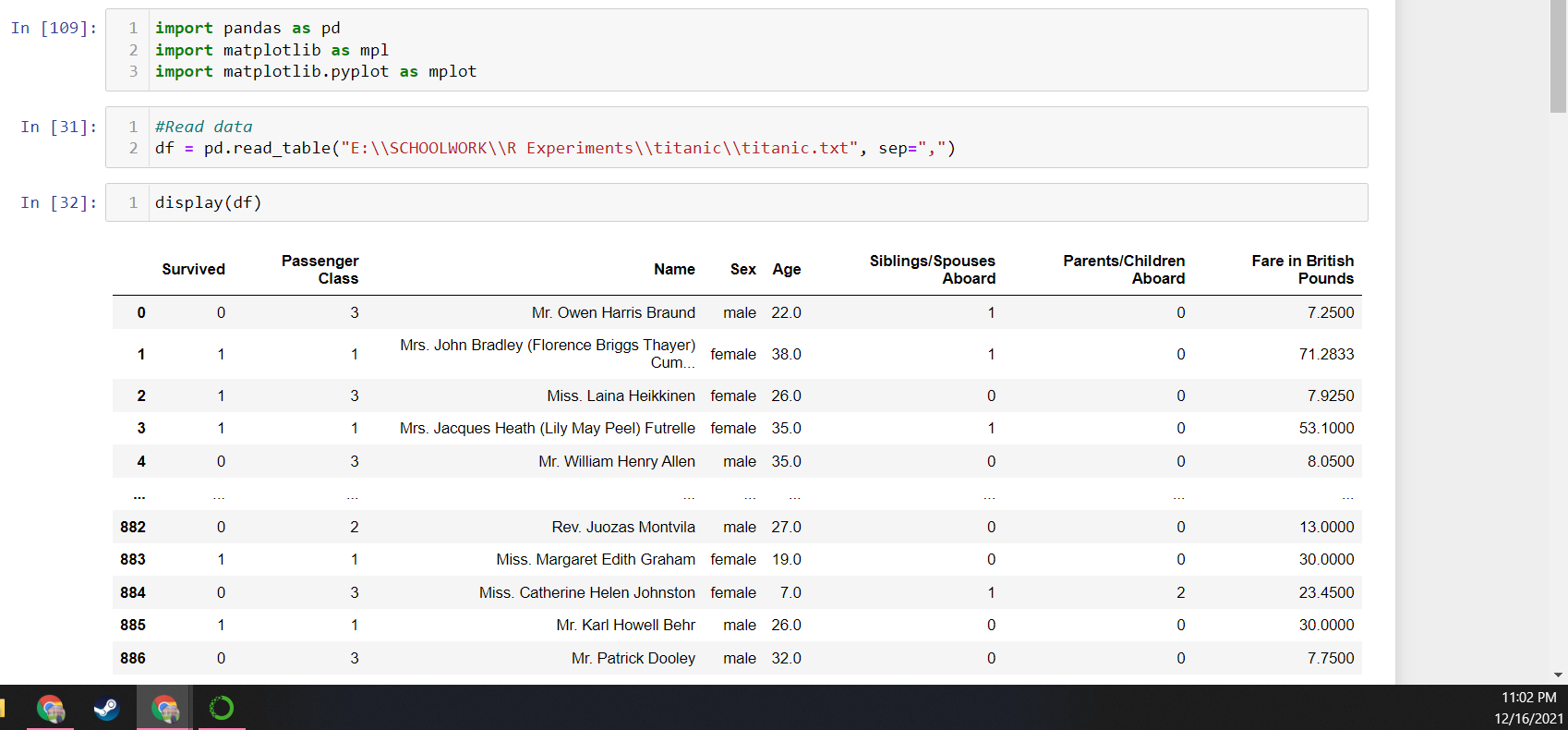
**Introduction**

Effective data analysis includes an ability to select specific variables, display them through visualization tools, and to create hypotheses and inductions based on the evidence. The following will explore a section of data regarding the passengers of the Titanic. Basic statistical functions will be performed, to demonstrate basic data understanding, and then a few visualization techniques using the Python package matplotlib.

**Screenshots + Code**

**Figure 1**

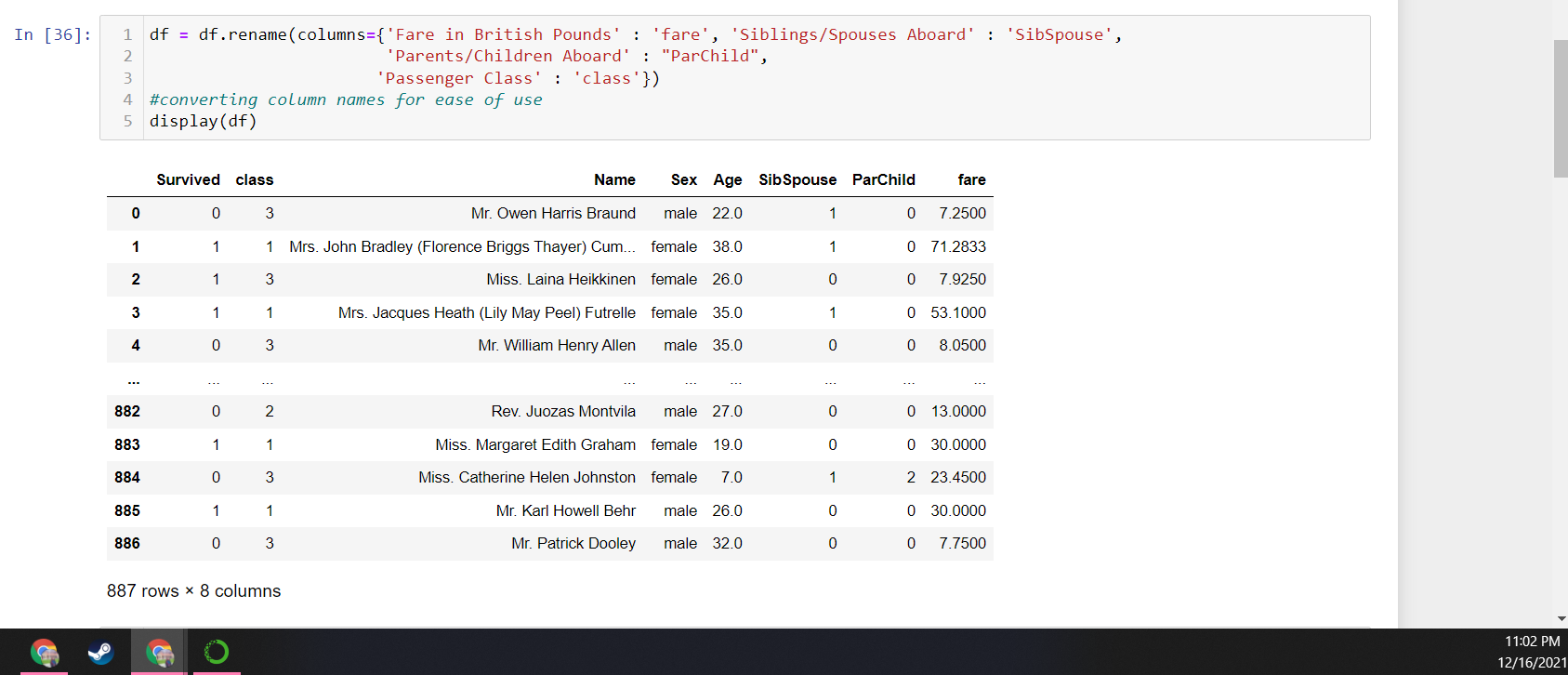
*Showing the successful import of .txt file*



First the file had to be saved from the CSU Global website as a .txt file, as it did not download appropriately when the link was clicked. After importing, it was important to view the head of the dataset just to see what variables were being used, and if there was any way to improve ease of use before the rest of the assignment.

**Figure 2**

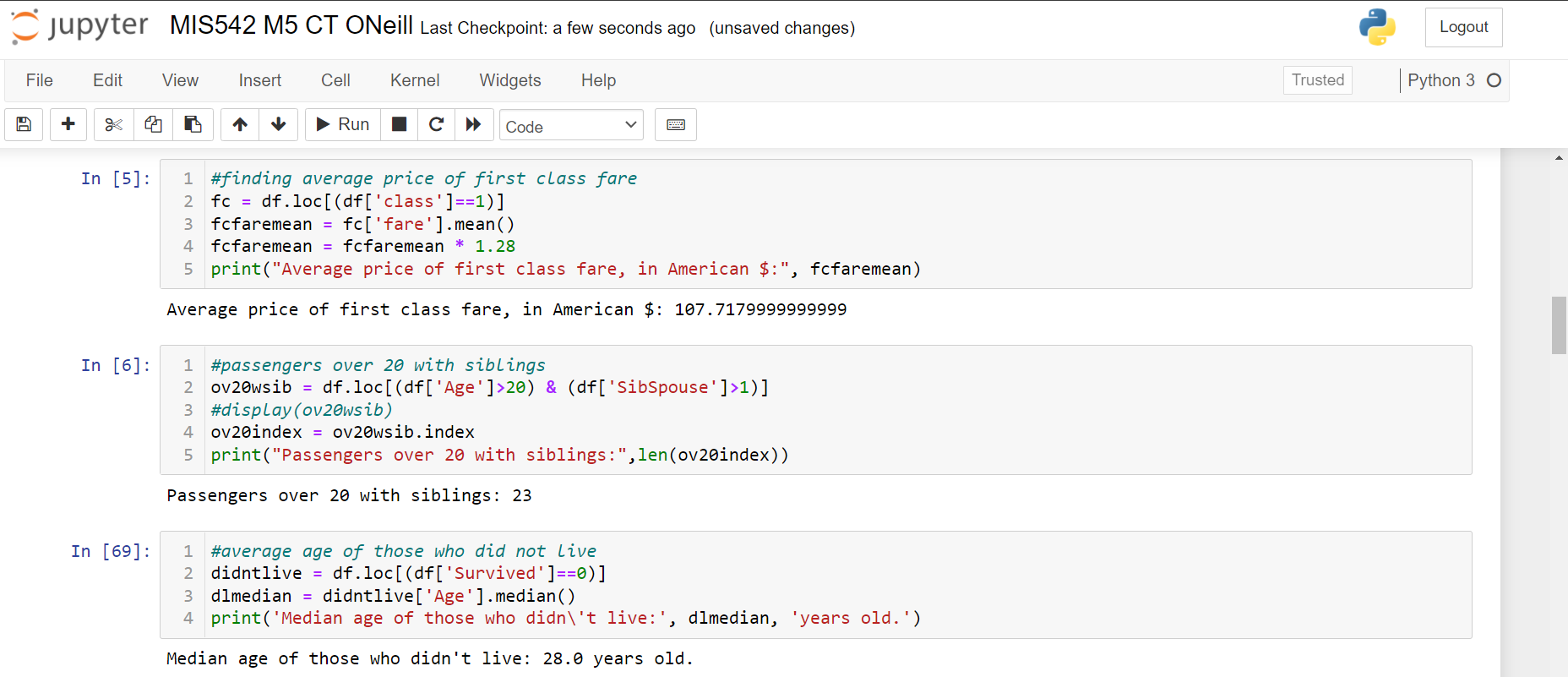
*Renaming columns for ease of use.*

**

It was useful to rename some of the columns with more complicated names before beginning with any analysis or visualization. This reduces errors in spacing, capitalization, etc.

**Figure 3**

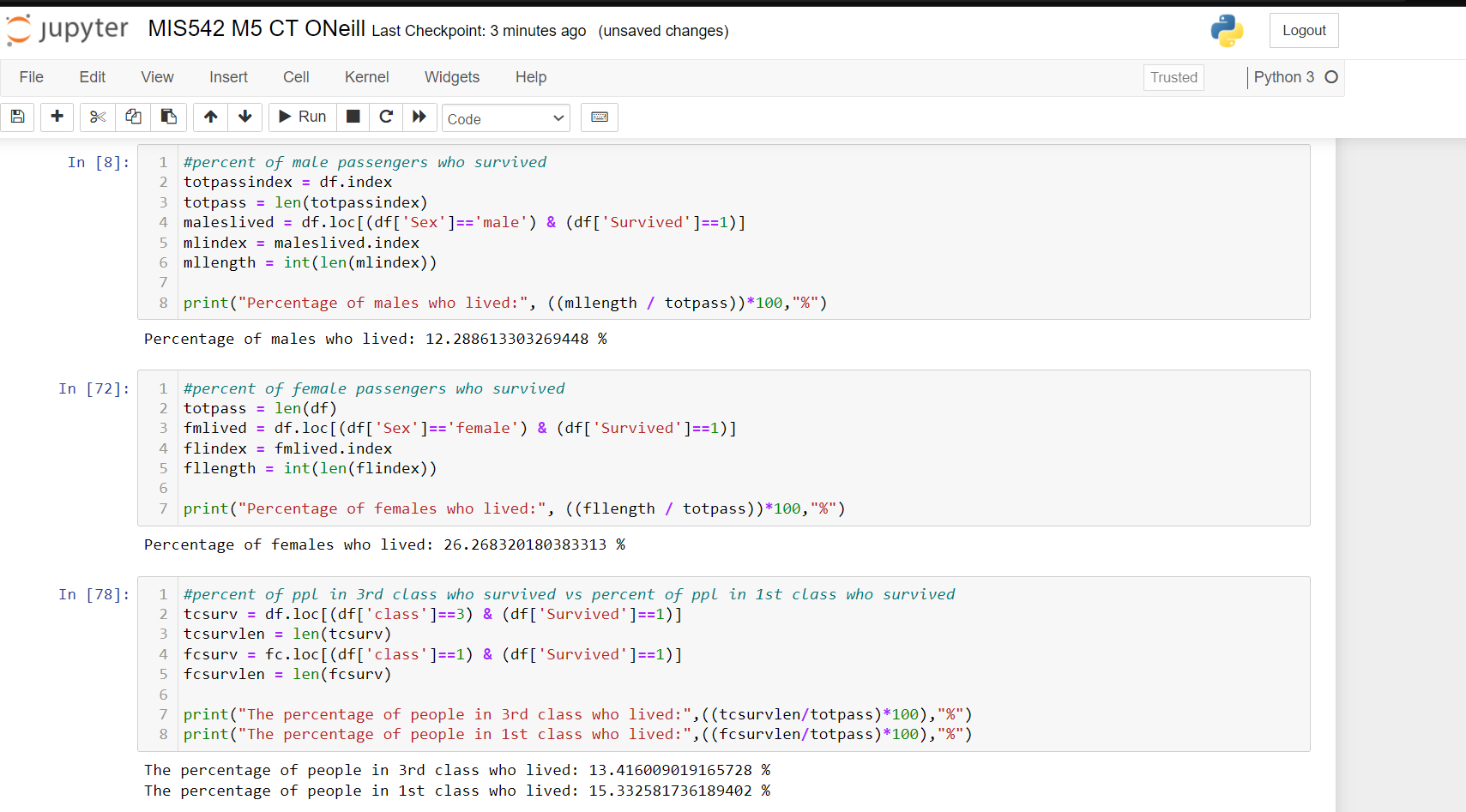
*Average price of 1st class fare; passengers over 20 with siblings; average age of those who perished*

**

This section covers three chunks of code. In the first, data is separated into first class passengers, and then the mean is calculated from that slice. The result is multiplied by 1.28 to convert the British pound into U.S. dollars. In the second, the original assignment instructions were followed, to slice out only the passengers who were over the age of 20 with more than 1 sibling/spouse listed, to qualify as sibling rather than spouse. The third code chunk separates out the non-survivors, where the Boolean status ‘survivor’ == 0. It then calculates the median value of the column ‘age’ for that slice of passengers.

**Figure 4**

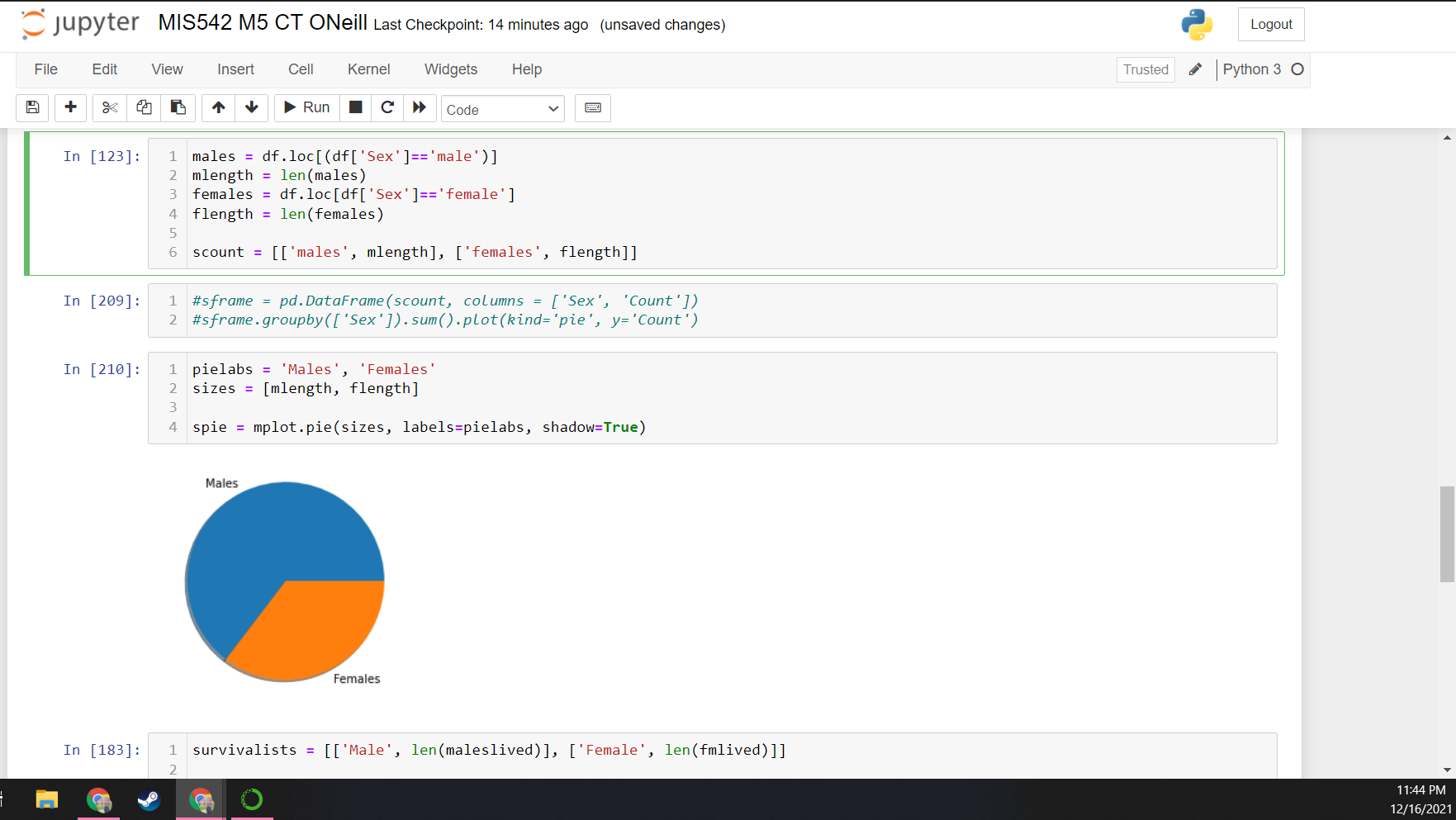
*Answering 3 additional questions: Percentage of male / female passengers who survived; percentage of 3rd vs 1st class who survived*



The three statistical questions asked regarded the percentage of surviving male passengers vs. female passengers, and then the comparing the percentage of people in 3rd vs 1st class who survived. Sadly unsurprisingly, a higher class of 1st class passengers survived. In agreement with what well-known accounts say about men putting ‘women and children first’ on the sinking Titanic, a much higher percentage of females survived than males.

**Figure 5**

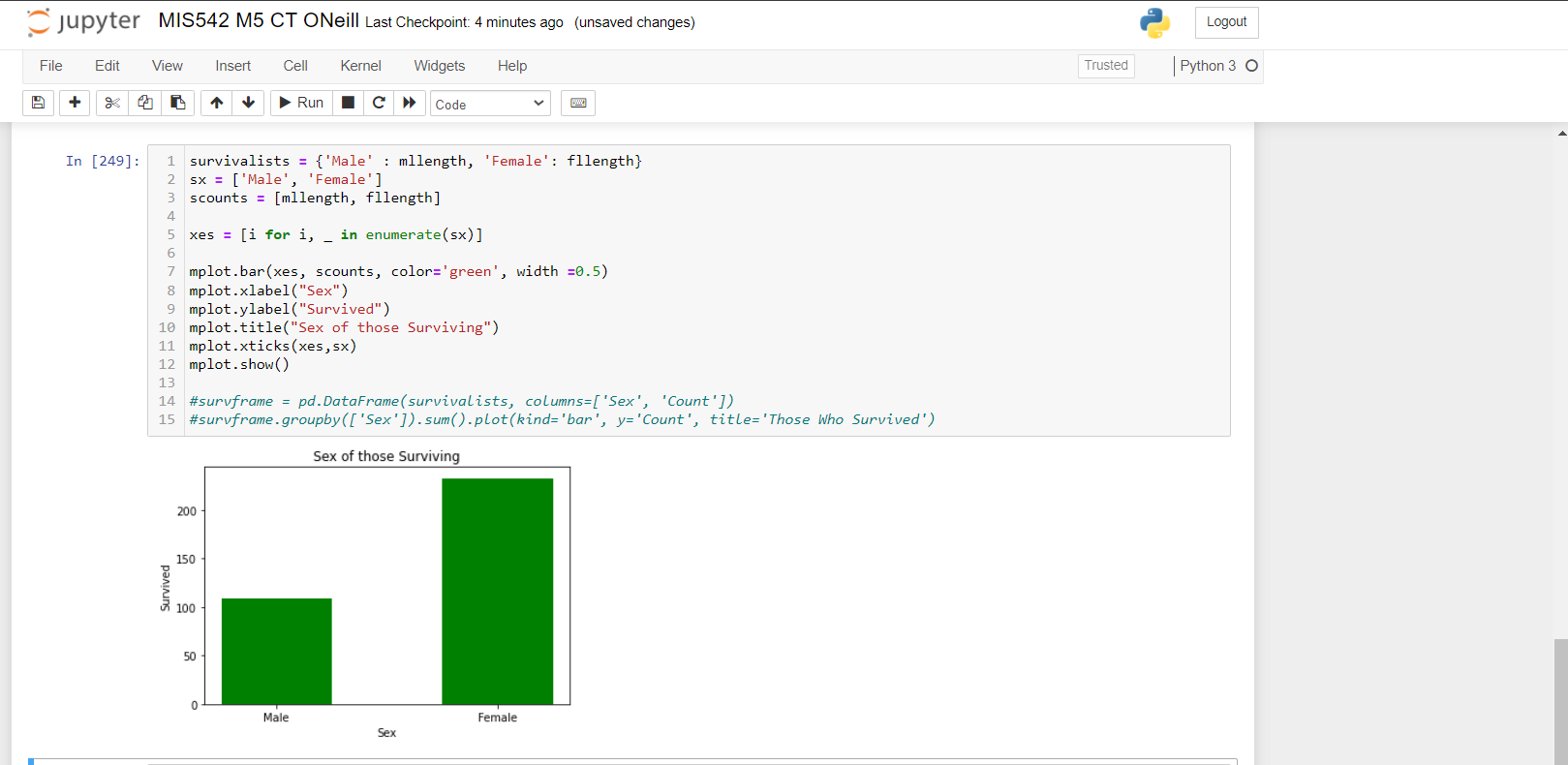
*Pie chart showing males vs females*



The Python package Matplotlib allows for the creation of charts which can help visualize data. In order to create an effective pie chart, it was required to find the count of columns which matched a specific criterion; in this case, “male” vs. “female”. A drop shadow was added to the chart for extra flavor.

**Figure 6**

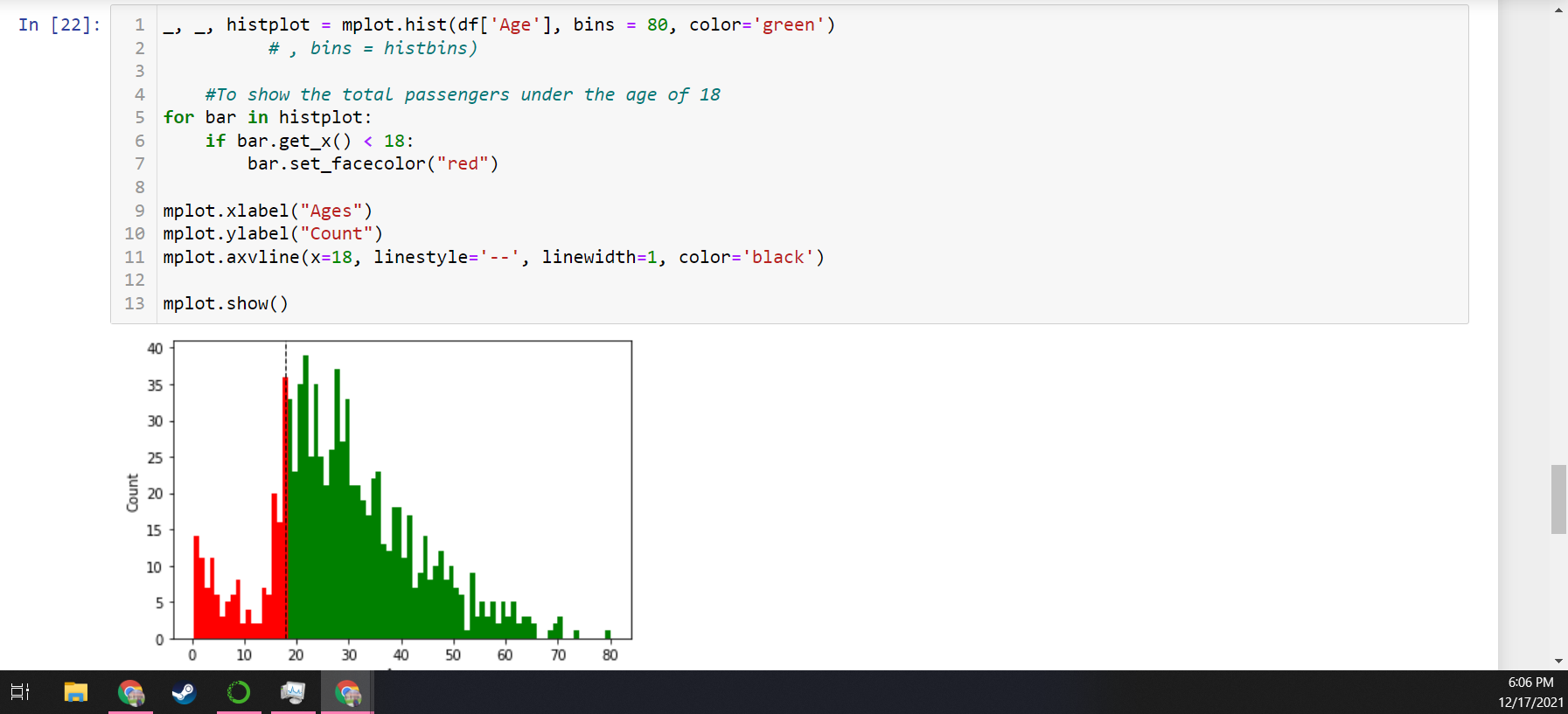
*Showing by sex those who survived using MatPlotLib*



This visualization helps those who may not easily visualize percentages to see the vast disparity between surviving men and surviving women. In order to create this bar chart, two separate arrays were created in order to create the x and y axes.

**Figure 7**

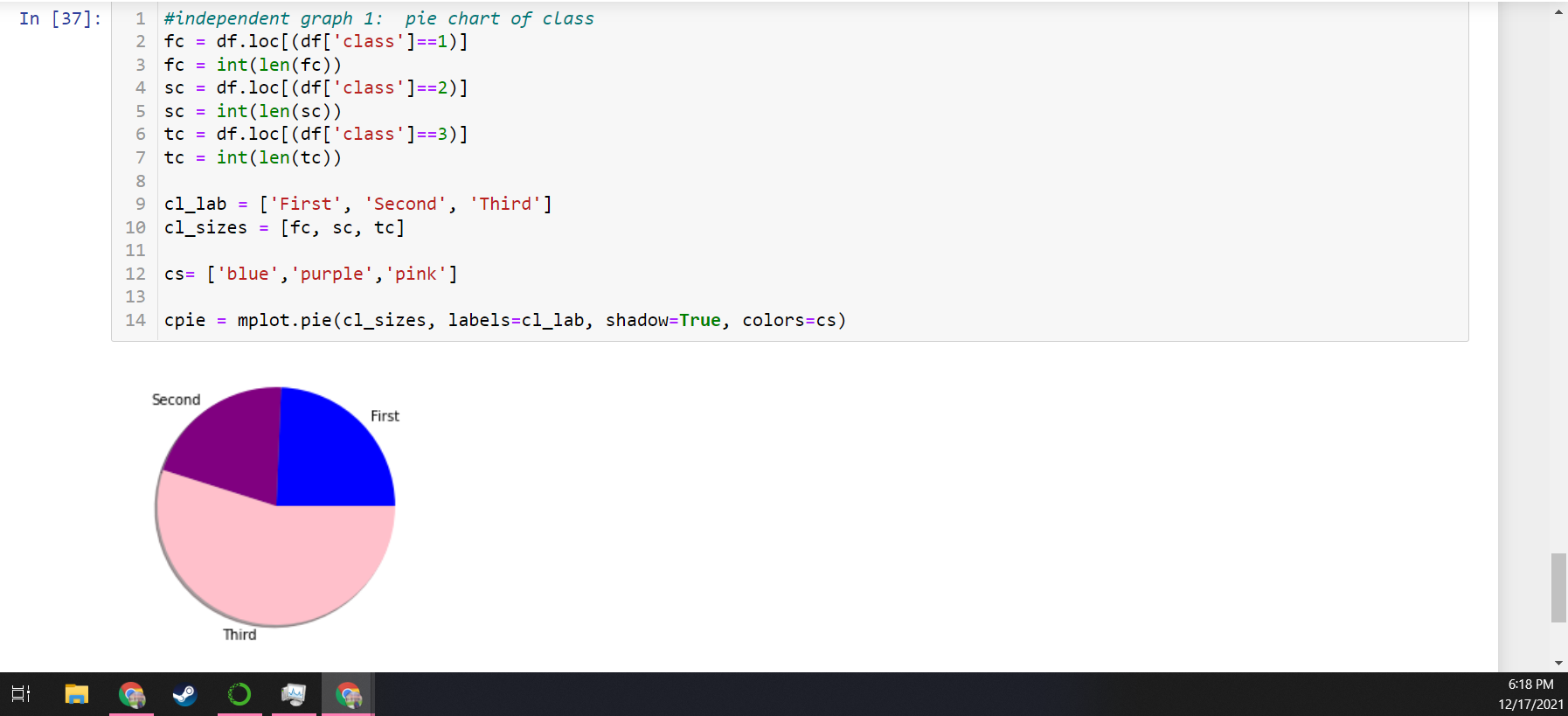
*Showing a histogram of ages of people on the Titanic*



This histogram displays the ages of the Titanic’s passengers; the peak count of the passengers appear to be between the ages of 20 and around 35. Anything displayed in red to the left side of the dotted line was coded to represent the passengers under the age of 18. There appared to be a very small amount of people between the ages of 60 and 80.

**Figure 8**

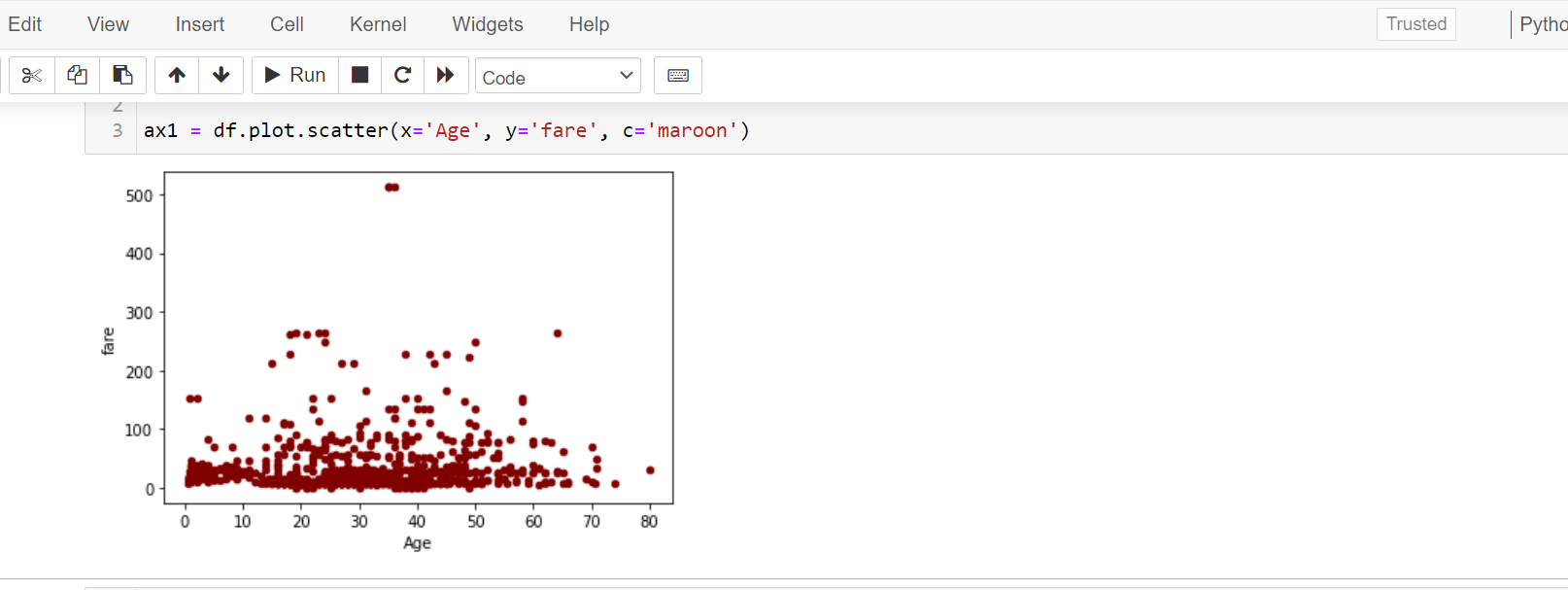
*Independent graph 1: a pie chart showing the classes of the passengers*.



Unsurprisingly, this pie chart shows a vast majority of third-class passengers. It was surprising to see that there were nearly an equal amount of first and second class passengers.

**Figure 9**

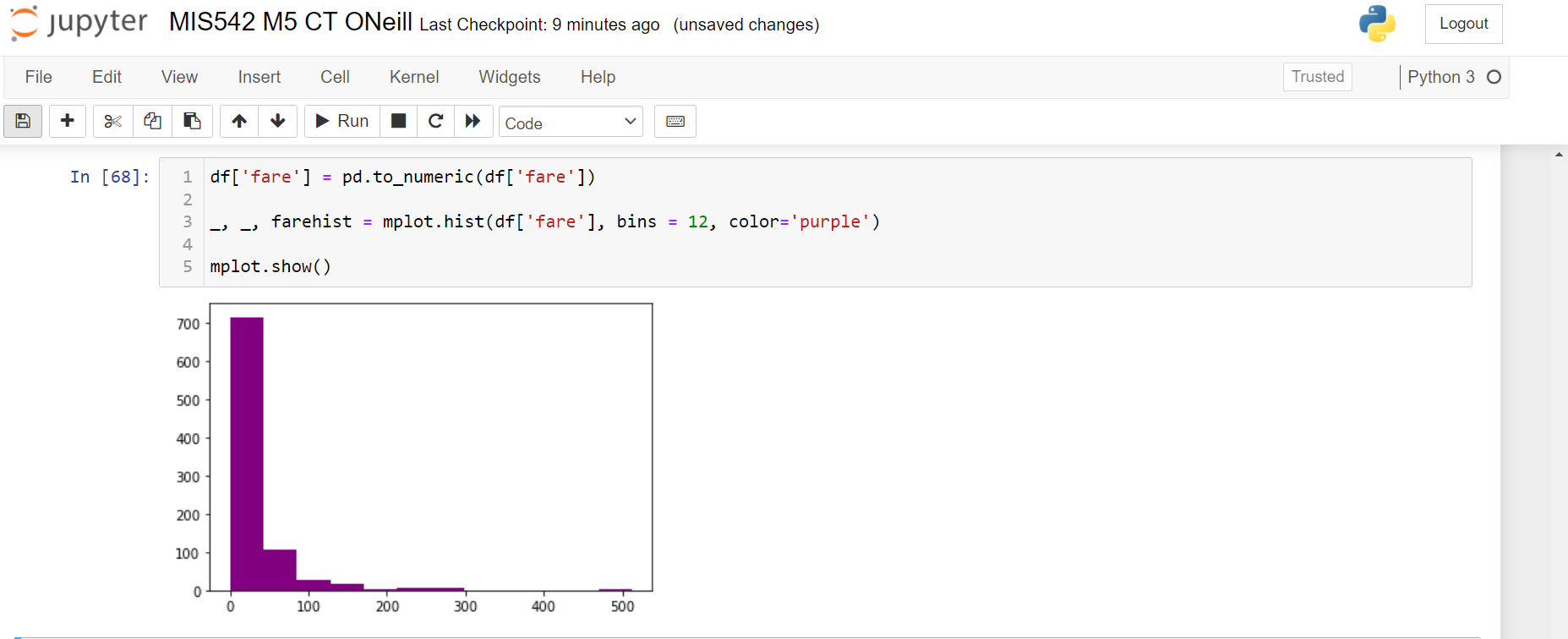
*Independent graph 2: a scatter plot showing the fare of the passengers according to age.*



This scatter plot shows the fare of passengers; after its creation, it seems evident that a histogram would probably have been more effective at displaying the various fare prices. It was surprising that some of the youngest passengers paid significantly higher fares; maybe some of the first-class passengers had children.

**Figure 10**

*Independent graph 3: A histogram of the fares.*



A plain histogram of the passenger fares shows heavy right skew, where a fare of $500 is definitely an outlier and might significantly affect any analysis done on the dataset.

**Learned / Conclusion**

Overall, I found that I should spend more time learning about the accepted arguments for matplotlib, as it seems like a very powerful tool that can be customized with legends and subplots. Data visualization is important for being able to understand patterns within data and exploring its trends, structure, and features (Unwin, 2020), but also to communicate cross-departmentally and engage upper management with data culture. Storytelling in data is essential to being able to effect change (Midway, 2020), deploy models, and make progress in an organization’s business goals.

As I was going through the various exercises, it occurred to me that next time, it might have been more effective to do all my data slicing at the beginning of the code, rather than regularly throughout the program. I also found it important to display each new dataframe as I created it, in order to ensure it was really the data I was looking for. However, while running through each exercise, while the percentages did not surprise me, it made me feel successful in uncovering insights about the dataset that I hypothesized, but did not know for sure. Data exploration is crucial to forming basic assumptions and providing the analyst with avenues of exploration and insight (Liu et al., 2020) - being able to identify outliers, trends, patterns, and potential correlation is very helpful in saving time when beginning actual hypothesis testing or building predictive models.

**References**

Liu, M., Wang, Z., Gong, Z., Yoon, J., & Wang, X. (2020, Aug 31). *2 - Data exploration*. MLCMU. <https://blog.ml.cmu.edu/2020/08/31/2-data-exploration/>

Midway, S. (2020, Dec 11). *Principles of effective data visualization*. National Center for Biotechnology Information. <https://pubmed.ncbi.nlm.nih.gov/33336199/>

Unwin, A. (2020, Jan 31). *Why is data visualization important? What is important in data visualization?* HDSR. <https://hdsr.mitpress.mit.edu/pub/zok97i7p/release/3>