Daniel Clark

CTEC 298-101 Spring 2021 1st 8w March 4, 2021

1. Define the terms below using the definitions found in the tutorials that were provided to the class. Each answer will consist of the definition and the tutorial where it was found.

Jupyter Notebook – a tool for interactively developing and presenting data science projects

Cell – a container for text to be displayed in a notebook or code to be executed by the notebooks kernel

Kernel – a computation engine that executes code contained in a notebook docomuent

Ipynb file – a document used to store Jupyter notebook information

JSON – a format for data in a notebook; represents objects as name/value pairs

Metadata – data that provides information about other data

Code cell – contains code to be executed in the kernel

Markdown cell – contains text formatted using Markdown

Matplotlib Plots – data visualizations created using the matplotlib library in python

1. Describe each of the 6 plots that we have used in Matplotlib
   1. The six plots we have used are stack graphs, bar graphs, histograms, scatter plots, pie charts, and simple plots.
   2. Simple plots chart given data based on an X and Y axis value. These plots are mainly used to show basic correlations and visualizations of data.
   3. Stack graphs show values of data on top of each other or next to each other in the same column. These graphs are used to convey the importance of individual pieces and the whole value.
   4. Bar graphs show the values of data as bars of different heights. These graphs are used to show which category or piece of data has more or less value.
   5. Histograms show a representation of grouped data based on certain ranges displayed as bars. These graphs have similar purpose to bar graphs, where they show which category has more or less value. But, these graphs show the value of ranges of data rather than individual data.
   6. Scatter plots are graphs that place points of data based on two important factors. These graphs are used to show potential correlations between the two selected factors. The closeness and pattern of scattered points on the graph determine the type of correlation.
   7. Pie charts are graphs that show the relative sizes of pieces in comparison to the whole. These graphs are used to convey the value of one piece over another.
2. What is a panda in this class?
   1. Panda is a library in python that provides high performance, easy use data structure and analysis.
3. How are Pandas related to Excel?
   1. Excel is a common application that is utilized for data storage and analysis. Pandas can be used to analyze or manipulate data files, including those in common Excel formats. Pandas allows for the importing or exporting of Excel format files, as well as data analysis, within python.
4. Pandas are used for what is the major data science concept.? What is a dataframe?
   1. Pandas is used for data analysis and data munging, or data wrangling. A dataframe is the main object in pandas, and represents data with rows and columns, similar to an excel spreadsheet.
5. Compare Matplotlib and Tableau for creating plots.
   1. Creating plots using matplotlib requires coding with python. Creating plots with Tableau is done through a user interface. Both are able to create similar plots with high customization and accuracy.
6. What types of input files can be used for each?
   1. Tableau accepts Excel, Text, JSON, Access, PDF, Spatial, and Statistical files as well as server compatibility. Matplotlib accepts all files that python can accept. Excel, CSV, ZIP, Text, JSON, XML, HTML, and other file types are all accepted by python.
7. What is a dataframe? Give an example.
   1. A dataframe is the the main object in pandas and represents data with rows and columns. Examples of a dataframe include an excel spreadsheet, or a table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name** | **University** | **ID** |
| **1** | Howard Boyd | Cornell University | 123456 |
| **2** | Darnell Collins | Tuskegee University | 654321 |

1. Name 3 statistics concepts that can be used with dataframes?
   1. 3 statistic concepts that can be used with a dataframe are find the maximum value of a column, find the average of data within a column, and finding the standard deviation of data within a column.
2. Describe 3 techniques for creating a dataframe using Pandas?
   1. 3 techniques that can be used to create a dataframe in pandas is importing from a file, creating it from a list of values in python, and creating it from a python dictionary.