Homework #1

CIS 585 – Advanced Artificial Intelligence
University of Michigan - Dearborn
John P. Baugh, Ph.D.

Objectives

- To learn how to effectively use Python with data analysis and AI applications
- To practice using modules such as Pandas and Seaborn

Instructions

For this homework assignment, you will practice using Pandas and Seaborn functionality to obtain insight into a small file of sales for a given year. The file is named **sales.xlsx**. You should not change its format (e.g., such as to *csv* or some other file format).

Using markdown

You will use markdown to help organize your tasks (see the next subsection) in Jupyter Notebooks. To do this, you create a cell using the **+** symbol in the upper left (below the File, Edit, etc.) Then, you change the type from Code to Markdown using the dropdown menu. Markdown allows you to create different levels of headings and other interesting styles of text to help organize the notebook visually.

E.g.,

```
# Some Level 1 Heading
## Some Level 2 Heading
### Some Level 3 Heading
```

Note the use of the # symbol, as well as the space after the hash symbol(s), and the heading information itself.

Required tasks

You must do the following, creating a single Jupyter Notebook, giving the file the title ***name*_cis_585_hw1_w23** (see Deliverables section near the bottom of this assignment document for more details)**.** Then, you should separate them by appropriate, clear markdown in Jupyter Notebooks. I've provided bullet points that look like checkboxes to help you. The major checkbox (first level bullets) correspond to sections that you should use markdown on, preferably level 2 heading markdown (For example, ## Task 1).

- ☐ **Task 1 - Fundamentals**
  - In the first cell, import Pandas using the pd alias and Seaborn using the sns alias, and run that cell
  - In the second cell
    - create a variable to hold the URL of the sales.xlsx file
    - followed by a variable named **sales_data** that should hold the resulting data frame returned from Pandas (use functionality to load an Excel workbook into memory from file)
    - Finally, print the sales_data

- ☐ **Task 2 – Statistical data**
  - In the first cell, use the describe method to print out a table *describing* the count, mean, std deviation, min, etc.
  - In the second cell, display the result of a query to show those rows where the Revenue was greater than 200,000
  - In the third cell, display the result of a query to show only the data from Detroit
  - In the fourth cell, display the result of a query to show only the data from Detroit, Chicago, or Los Angeles that also have Revenue greater than or equal to 250,000
  - In the fifth cell, group the data by the Store, and display the sum
  - In the sixth cell, group the data by the Store, and display the median

- ☐ **Task 3 – Visualization**
  - In the first cell, use the general method for a relational plot with the following parameters:
    - The data used should be the sales data loaded by pandas previously
    - The kind of plot should be a line plot
    - The x axis should be for the Quarters
    - The y axis should be for Revenue
    - The hue grouping should be by the Store
  - In the second cell, use the general method for a relational plot with the following parameters:
    - The data used should be the sales data loaded by pandas previously
    - The kind of plot should be a scatter plot
    - The x axis should be for the Quarters
    - The y axis should be for Revenue
    - The hue grouping should be by the Store
  - In the third cell, use the general method for a categorical plot with the following parameters:
    - The data used should be the sales data loaded by pandas previously

        - You should filter the data with a query to only include the stores in Detroit, Chicago, and Tokyo
        - The kind of plot should be a horizontal bar plot
        - The x axis should be for the Revenue
        - The y axis should be for Store
        - The hue grouping should be by the Quarter
        - The aspect ratio should be 2:1 (width = 2 * height)
    - In the fourth cell, use the general method for a categorical plot with the following parameters:
        - The data used should be the sales data loaded by pandas previously
        - You should filter the data to only include data where the Revenue is greater than or equal to 250,000
        - The kind of plot should be a box and whisker plot (a.k.a., “box plot”)
        - The x axis should be for the Quarter
        - The y axis should be for Revenue
        - The aspect ratio should be 2:1 (width = 2 * height)

Deliverables

You should upload your Jupyter Notebook, named as ***name*_cis_585_hw1_w23**, where *name* is your name in the following format: *firstname_lastname*, such as *john_baugh*. (firstname = given name, lastname = family name).