DAV 5400 Project 3 (Module 11) (100 Points)

Data Aggregation, Grouping, Reshaping & Analysis

***** You may work in small groups of no more than three (3) people for this Project *****

As data analytics practitioners, we are often required to work with data sets that are provided to us in formats that are not readily "usable" for computational / analytics work. When faced with such a task, we may be required to apply data reshaping, aggregation, and/or grouping methods to restructure the data into a "usable" format prior to attempting any actual computations or analysis.

A primary task for this Project is to separate a data set into two distinct "populations" of people, analyze each of those populations independently of each other, then compare and contrast the characteristics of each population for purposes of facilitating decision making within an organization that is interested in improving their engagement with both populations. However, the data set is provided in a format that is definitely not suitable for computational or analytical work. As such, prior to attempting the required analysis we must first decide upon how best to reorganize the data, and, using our Python skills, implement our recommended manipulations of the data set so that we can subsequently perform the required analysis.

The data set you are to utilize is provided by 538.com:

- https://github.com/fivethirtyeight/data/blob/master/flying-etiquette-survey/flying-etiquette.csv

The data set is comprised of the results of a survey that addressed the air travel habits and preferences of the general public, with each column within the data set containing responses to a specific survey question. The column headings contain the actual survey questions. Some of the questions are directly related to flying habits/behavior while others are clearly meant to collect demographic information on each survey participant (e.g., "How tall are you?", Gender, Age, Household Income, Education, Location).

Get started with the Project as follows:

1. Using the provided Github link, load the data set into a Pandas dataframe within your Jupyter Notebook.
2. Create a new dataframe comprised of only the demographic data for each individual that also includes the unique RespondentID. Make sure that each column heading in this new dataframe conforms to "best practice" methods for naming Pandas columns.
3. For the data values that were provided in response to the survey question "How tall are you?", convert the provided feet/inches values to centimeters.
4. Create a new dataframe comprised of only the non-demographic survey question responses along with the corresponding unique RespondentID. Make sure that each column heading in this new dataframe conforms to "best practice" methods for naming Pandas columns.
5. Convert the content of the dataframe resulting from Step 4 (above) to a "tidy" long format. (*NOTE: "tidy" data concepts are covered in Module 10*). How you go about achieving a "tidy" version of the data is up to you as an analytics practitioner to decide.

6. Add a new column to the demographic dataframe resulting from Step 2 (above) that indicates whether an individual either ***NEVER*** flys or at least sometimes flys. How you go about achieving this is up to you as an analytics practitioner to decide.
7. Using your Python skills, perform exploratory data analysis (EDA) on the dataframe resulting from Step 6 (above). This EDA work should provide a reader of your work with a thorough understanding of the demographic composition of the entire population of survey respondents. Your EDA writeup should include any insights you are able to derive from your statistical analysis of the attributes and the accompanying exploratory graphics you create (e.g., bar plots, box plots, histograms, line plots, etc.), all of which should be of publication-quality. It is up to you as the data science practitioner to decide how you go about your EDA, including selecting appropriate statistical metrics to be calculated + which types of exploratory graphics to make use of. Your goal should be to provide an EDA that is thorough and succinct without it being so detailed that a reader will lose interest in it.
8. Next, analyze and summarize the demographic characteristics of individuals who **NEVER** fly vs. those of individuals who at least sometimes fly. Be sure to discuss similarities and disparities in these two distinct sub-populations of survey respondents(i.e., flyers vs. non-flyers); for example, are there any particular characteristics of the population that appear to be the most indicative of whether an individual is likely to have flown recently? Be sure to include publication-quality graphics in support of your response.

Now, use the results of Steps 5 and 6 (above) to help you answer the following questions regarding individuals who ***DO*** fly:

1. **(3 Points)** Which age grouping is most likely to responded "**About half the time**" to the question "**Do you ever recline your seat when you fly?**" Be sure to include publication-quality graphic(s) in support of your response.

2. **(3 Points)** Are male or female flyers **most likely** to have responded "**Yes**" to the question "**Do you have any children under 18?**" Be sure to include publication-quality graphic(s) in support of your response.

3. **(3 Points)** Which income group is **least likely** to have responded "**The person in the window seat should have exclusive contro**l" to the question "**Who should have control over the window shade?**". Be sure to include publication-quality graphic(s) in support of your response.

4. **(5 Points)** How have the various 'Location" groupings responded to the question "**Under normal circumstances, does a person who reclines their seat during a flight have any obligation to the person sitting behind them?**" Be sure to include publication-quality graphic(s) in support of your response.

5. **(3 Points)** Which gender is most likely to have responded "**No, not at all rude**" to the question "**Is it rude to wake a passenger up if you are trying to go to the bathroom?**" Be sure to include publication-quality graphic(s) in support of your response.

6. **(3 points)** What is the average height (in centimeters) of the male survey respondents? What is the average height (in centimeters) of the female survey respondents? Be sure to include publication-quality graphic(s) in support of your response.

7. **(5 points)** Provide a general demographic profile of individuals who have responded "**No, not at all rude**" to the question "**In general, is it rude to knowingly bring unruly children on a plane?**" Be sure to include publication-quality graphic(s) in support of your response.

Your deliverable for this Project is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: In formatted Markdown, provide a succinct introduction / problem statement describing the work you intend to perform and why you believe it will prove to be of value to a stakeholder who works for a major airline,
2) **Data Loading + Manipulation (20 Points)**: This section should appropriate explanatory narratives (using Markdown) and include the Python code used to load the data set + complete Steps 1 through 6.
3) **Exploratory Data Analysis (25 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis. This section should include appropriate explanatory narratives (using Markdown) as well as any Python code and graphics used for the EDA.
4) **Demographic Analysis: Flyers vs**. **Non-Flyers (20 Points)**: Explain and present the analysis you performed for Step 8 (above). This section should include This section should include appropriate explanatory narratives (using Markdown) as well as any Python code and graphics used for that analysis.
5) **Flyers: Demographic "Drill-down" (25 Points total)**: This section should contain all Python code and explanatory commentary (using Markdown) related to your analysis and answers for Questions 1 through 7 regarding survey respondents who ***DO*** fly (from above)**.**
6) **Findings / Summary (5 points)**: Provide a clear summary of your findings and/or recommendations for a stakeholder who works for a major airline.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload / submit your Jupyter Notebook within the provided Project 3 Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_Project3_assn**" (e.g., J_Smith_Project3_assn_). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***