**Exercise 3.2: Preparing for Exploratory Data Analysis using Python**

**By: Saima Rahmanzai**

**DSC530**

Complete the following exercises:

* Page 11: 1-1 (In the repository you downloaded, you should find a file named chap01ex.ipynb, which is an IPython notebook. You can launch IPython notebook from the command line…)

Answer: I used Gitbash and ran the following commands to move the GitHub location from the book to my local Github repository on my computer:

saima@DESKTOP-BQ0ENVG MINGW64 ~/class\_assignments/dsc530

$ git clone https://github.com/AllenDowney/ThinkStats2

Cloning into 'ThinkStats2'...

remote: Enumerating objects: 2826, done.

remote: Counting objects: 100% (5/5), done.

remote: Compressing objects: 100% (5/5), done.

remote: Total 2826 (delta 0), reused 4 (delta 0), pack-reused 2821

Receiving objects: 100% (2826/2826), 172.42 MiB | 718.00 KiB/s, done.

Resolving deltas: 100% (1609/1609), done.

Updating files: 100% (485/485), done.

saima@DESKTOP-BQ0ENVG MINGW64 ~/class\_assignments/dsc530

$ ls -al

total 8

drwxr-xr-x 1 saima 197609 0 Nov 29 21:57 ./

drwxr-xr-x 1 saima 197609 0 Nov 29 21:41 ../

drwxr-xr-x 1 saima 197609 0 Nov 29 22:02 ThinkStats2/

Next, I created a new conda environment for this class in the Anaconda Prompt such as:

Conda create –name srahmanzaiDSC530

I then loaded all libraries necessary including pandas, scipy, Jupyter, and others using the command prompt. Jupyter Notebook kicked off the server where I opened the folder where the chap01ex.ipynb was downloaded from github and completed all the assignments. I saved both the chap01ex.ipynb and the pdf version and submitted via Blackboard. Files are named as:

* Chap01ex\_Rahmanzai.ipynb (the ipynb file with completed exercises)
* Chap01ex\_ipynb\_output\_Rahmanzai.pdf (the output of the ipynb file for chapter 1

I also submitted it via my Github repository. The link is: [GitHub - srahmanzai/DSC530: srahmanzai\_EDA](https://github.com/srahmanzai/DSC530)

* Page 11: 1-2 (Create a file named chap01ex.py and write code that reads the respondent file, 2002FemResp.dat.gz. You might want to start with a copy of nsfg.py and modify it…)

Answer: I was able to open the nsfg.py and created a chap02ex.py and copied the code. I ran it and it ran with no errors. All tests passed. I downloaded the pdf version and saved the .py version which is submitted via Blackboard. Files are named as:

* chap01ex.py\_output\_Rahmanzai.pdf (output of the .py file when ran)
* chap01ex\_Rahmanzai.py (the actual .py file required to complete exercise 1-2)

I also submitted it via my Github repository. The link is: [GitHub - srahmanzai/DSC530: srahmanzai\_EDA](https://github.com/srahmanzai/DSC530)

* Page 25: 2-1 (Based on the results in this chapter, suppose you were asked to summarize what you learned about whether first babies arrive late…)

Answer: The difference between means of Firsts and Others of Pregnancy length was 0.078 and the Cohen's difference was not very significant (0.0288). Due to the low Cohen’s difference and especially when the firsts numbers are also smaller than the others group that could also effect the results, the evidence is not conclusive that the first babies arrive late. See the details of commands and exercises completed in files below:

* chap02ex\_ipynb\_output\_Rahmanzai.pdf
* chap02ex\_Rahmanzai.ipynb
* chap02ex\_Rahmanzai.py

See the files uploaded to Github at: [GitHub - srahmanzai/DSC530: srahmanzai\_EDA](https://github.com/srahmanzai/DSC530)

* Page 25: 2-4 (Using the variable totalwgt\_lb, investigate whether first babies are lighter or heavier than others…)
  + You can follow along with the solution file for this exercise, or you can use a different package to do this work like NumPy – if you want to use a different package, make sure you include histograms as part of your analysis.

Answer: The difference between means of Firsts and Others of Pregnancy length was -0.1247 which employs that on an aggregate, the mean of totalwgt\_lbs is smaller than that of others but only slightly. However, similar to the pregnancy length’s between the two groups (firsts and others), the Cohen's difference was not very significant (0.0554). Due to the low Cohen’s difference and especially when the firsts numbers are also smaller than the others group that could also effect the results, the evidence is not conclusive that the first babies are lighter or heavier than others. See the exercise performed in files below which are uploaded on both Blackboard and Github.

* chap02ex\_ipynb\_output\_Rahmanzai.pdf
* chap02ex\_Rahmanzai.ipynb
* chap02ex\_Rahmanzai.py

See the files uploaded to Github at: [GitHub - srahmanzai/DSC530: srahmanzai\_EDA](https://github.com/srahmanzai/DSC530)

I also created a visual representation using the hist in pandas (see histogram below). The frequencies don’t appear to be very significantly apart. See details of commands in the chap02ex\_ipynb\_output\_Rahmanzai.pdf document.

Chart, histogram

Description automatically generated