CSC 4780/6780
Fall 2022
Homework 1

Priyanka Mahendra Ghare
pghare1@student.gsu.edu

August 28, 2022

This homework is due at 11:59 pm on Sunday, Aug 28. It must be uploaded to iCollege by then. No credit will be given for late submissions. A solution will be released by noon on Monday, Aug 29.

It is a very bad idea to put this off until the last minute.

This homework is worth 10 points.

1 Purpose

In every data science job, you get two things on the first day: an email address and a laptop. Your first task is to get the laptop ready for work.

The second thing you do is get a table of data that you are expected to understand. Here we will be using Pandas and SQLite to explore some data.

2 Study

Read pages 71 - 133 of *Practical Data Science with Python*.

Optional: We will be using pandas all semester. You should get comfortable with it. Here is a good video tutorial: `https://youtu.be/PcvsOaixUh8`

3 Install python3

If you have not already, install python on your computer. You will be using this all semester. Python version 3.10 has been released, but anything after 3.7 is fine. You can check your version on the command line like this:

```
python3 --version
```

4 Install some python tools and libraries

```
pip3 install pandas
pip3 install numpy
pip3 install scikit-learn
pip3 install matplotlib
pip3 install black
```

(If you prefer to use conda, that is OK with me.)

5 JupyterLab

Jupyter notebooks allow an author to mix text (in Markdown) and code (in Python) in a single document that you can work with in a browser. I will use them from time to time to send you annotated examples.

Install JupyterLab on your machine.

```
pip3 install jupyterlab
```

Your homework directory includes `main.ipnb`. In that directory, start jupyter-lab:

```
jupyter-lab
```

In the pane on the your browser, you should see `main.ipnb`. Study and run the code in each cell.

Replace the question marks in this sentence: The mean waist size is 1.21 meters. (2 points)

6 Apply for GitHub Copilot

As a student, you get free access to Copilot (which is used from an extension in VS Code). You should apply: Copilot will make your life easier, but it is also an astonishing example of what machine learning can do.

`https://copilot.github.com`

7 Write some python using pandas

There is a file called `make_report.py`. When you fill in the missing lines, you will be able to run it like this:

```
python3 make_report.py employees.csv
```

Then, it will print a report like this:

```
*** Basics ***
Rows: 10,000
Columns: 6

*** Columns ***
employee_id: int64
        Range: 1712 - 9998838
gender: object
        Missing in 82 rows (0.8%)
                4917: m
                4907: f
                36: F
                23: M
                19: male
                16: female
height: float64
        Range: 1.34 - 2.07
        Mean: 1.71
        Standard deviation: 0.11
        Median: 1.71
waist: float64
        Range: 0.47 - 2.18
        Mean: 1.21
        Standard deviation: 0.23
        Median: 1.19
salary: float64
        Missing in 70 rows (0.7%)
```

```
        Range: 297.0 - 140902.0
        Mean: 63033.98
        Standard deviation: 20093.83
        Median: 63078.50
dob: object
        Range: 1945-01-01 - 1984-12-21
death: object
        Range: 1960-03-20 - 2022-06-12
```

DO NOT LOOP THROUGH ALL 10,000 ROWS. Let pandas do that for you.

Do this work by yourself. Stackoverflow is OK. A hint from another student is OK. Looking at another student's code is *not* OK.

Include the completed `make_report.py` in the zip file. Also copy your `series_report` function here: (4 points)

```
def series_report(
        series, is_ordinal=False, is_continuous=False, is_categorical=False
):
print(f"{series.name}: {series.dtype}")
if is_ordinal and is_continuous and not is_categorical:
        missingKey = series.isnull().sum()
        if missingKey > 0:
                print(f"\tMissing in {missingKey} rows ({missingKey / 100}%)")
        print(f"\tRange: {series.min()} - {series.max()}")
        print(f"\tMean: {series.mean():.2f}")
        if is_ordinal and is_continuous:
                print(f"\tStandard Deviation: {series.std():.2f}")
        print(f"\tMedian: {series.median():.2f}")
if is_categorical:
        missingKey = series.isnull().sum()
        print(f"\tMissing in {missingKey} rows ({missingKey/100}%)")
        print(f"\t{series.value_counts().to_string()}\t\t")
if is_ordinal and series.dtype == object:
        print(f"\tRange:{series.min()}-{series.max()}")
elif is_ordinal and not is_continuous and not is_categorical :
        print(f"\tRange:{series.min()}-{series.max()}")
```

In my solution, this function is 18 lines long.

8 Write some SQL

There is an `employees.db` file containing similar data in sqlite3 format.

In one SQL query, get the mean height of all employees who have a salary greater than $35,000.

Write your SQL query and the result here: (2 points)

```
sqlite >select avg(height) from Employee where salary >= 35000;
Result - 1.70632927449103
```

9 Install LaTeX and build a PDF

There are a lot of ways to install a TeX/LaTeX processing system. I use TeX Live (`https://www.tug.org/texlive/`).

After installing it, you will be able to render this document into PDF like this:

```
pdflatex report.tex
```

Open `report.pdf` to make sure it looks good. (Did you put your name and email in the `author` section?)

Include that pdf in the zip file you turn in. (2 points)

10 Tidy up

Before you zip up this directory and submit it, clean things up for the graders:

- Rename the folder. First name "Derek"? Last name "Zoolander"? The folder should be `HW01_Zoolander_Derek`.
- Reformat your code with black: `black make_report.py`.
- Delete intermediate files from pdflatex: `report.aux`, `report.log`, `report.synctex.gz`.

When you zip this directory, it should be called `HW01_Zoolander_Derek.zip`

Our amazing TAs have to check homeworks from a lot of students, so this sort of tidiness is very important for *every assignment*. If your code doesn't immediately run as-is, you will get points off. If we can't find your name on the folder or the PDF, you will get points off.

The most common problem is that in your code you have the path to your data file is something like `"C://home/zoolander/gsu/hw1/employees.csv"`. Leave the data file in the directory and use a relative path like `"employees.csv"`.

11 Looking ahead

Want to look ahead a little? We will be doing data visualization with matplotlib next. Here is a good video tutorial: `https://youtu.be/UO98lJQ3QGI`