GSCM 410/510 – Summer 2020
Homework Week 1

Directions

It is OK to work together on the Homework. Of course, make sure that everyone working together on an assignment is learning the material. Each person still turns in their own homework even if exactly the same as that of other group members. Learning this material well enough to do the homework problems is how you succeed in this course.

Short-Answer Problems

These concepts can appear on the short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. *Discuss with an example: The core of machine learning is pattern recognition.*

   Life and the world are not random collections of atoms. There are patterns everywhere, which form, for example, the basis of science, and also the basis of human decision making. Many of these patterns are not easily detected by humans. The usefulness of machine learning is to detect these patterns and generalize them to future events, the basis of forecasting.

2. *Why is Python a good language to use for machine learning?*

   Python offers a framework for machine learning, so if you run one machine learning analysis with one algorithm, it is easy to re-run the same code with another algorithm just by changing the name of the algorithm.

3. *What are some advantages and disadvantages between running Python with the Anaconda distribution on your own computer versus Google Colab? Include an assessment of costs for both alternatives.*

   Running Python on your computer provides the best security as no one else has access to your files. Plus, no Internet connection is needed to do the work. A negative is that for larger data files a more powerful computer is needed, so more expense. For Colab, access is free with a good amount of available computer power, though ultimately some limits for free access. Moreover, the environment is already set up, ready to go as soon as logged in. Two major disadvantages of Colab is that Google has access to your files, and an Internet connection is needed.

4. *Briefly explain the concept and purpose of a Jupyter notebook.*

   The Jupyter notebook allows the user to interactively program and run analyses with full documentation.

5. *What is a package manager? Why do we use a package manager for our Python work, and which package manager do we use?*

   Base Python by itself makes an excellent programming language. But one does not want to program everything, but use other, developed software, such as for machine learning. This additional software is organized by packages of related functions. A package manager is used to download, install, and update these different packages without much work on the part of the user.

6. *What are the two types of cells in a Jupyter notebook? What is the purpose of each?*

   A code cell is for entering and running Python code. A markdown cell is for documentation.

7. *Explain the concept of a current working directory.*

   The Jupyter notebook needs a starting point for referencing files to read and write. That reference point is the current working directory. All file references are relative to this directory.

8. *What is the relationship of Python and Pandas?*

   Pandas is a package of functions that add pre-built data analysis capabilities to Python. The primary Pandas data structure is the data frame.

9. *What is a data table and how are the data values organized?*

   A data table is a rectangular table of the data values subject to analysis. The first row contains the variable names, each other row contains the data for a single unit of analysis, such as a person or company. Each column contains the data values for a single variable.

10. *What is a csv file? What are its properties, its primary advantage and its primary disadvantage* (compared to an equivalent Excel file)?

    A csv file is a comma separated values file, pure text, so readable by virtually every application that can read text. Its primary advantage is near-universal readability. Its primary disadvantage is unlike a worksheet, its columns are not aligned when viewing the file as text, remedied by opening the file in a worksheet app such as Excel.

11. *What is the distinction between categorical and continuous variables? Provide an example variable of* each along with some sample values.

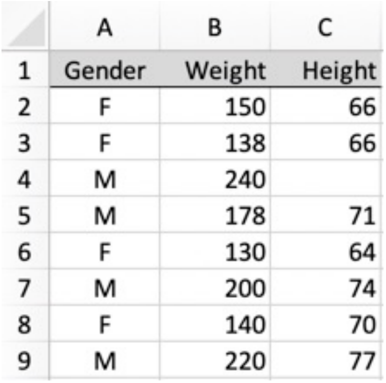
    The values of categorical variables are non-numeric categories, even if with integer values, and there are relatively few unique values. Continuous variables are always numeric and have many possible values.

Worked Problems

1. Create an Excel data table

Consider the data in Figure 1, randomly selected from a data file of the body measurements of thousands of motorcyclists.

Manually enter each data value from Figure 1 into a worksheet (such as Excel).



	A	B	C
1	Gender	Weight	Height
2	F	150	66
3	F	138	66
4	M	240	
5	M	178	71
6	F	130	64
7	M	200	74
8	F	140	70
9	M	220	77

Figure 1: Gender, Weight and Height of eight motorcyclists.

a. List each of the variable names in Figure 1 and classify each as continuous or categorical.

Gender – Categorical

Weight – Continuous

Height – Continuous

b. Every data table, whether in Python or Excel, has a name. Excel stores a data table in a worksheet, and every worksheet has a name. What is the name of that worksheet (located on a tab toward the bottom left corner) on which you entered your data? Compare the name of the data table to the names of the variables defined in that worksheet. (This is the same distinction in any analysis system, such Excel or Python.)

The default name of the first worksheet is `Sheet1`. □

Read these data directly from the Excel file you created on your computer system into a Python data table. Here the data file is in a folder called `data` that is right next to the notebook.

```
d = pd.read_excel('data/employee.xlsx')
```

c. Verify that the data values for the variables in the data table are stored within the analysis system in the intended format, that is, character string or numeric.

```
d.dtypes
```

Looking at the output, the intended format matches the way in which Python read the data.

d. Display the data from within Python.

Just enter the name of the data table.

```
d
```

e. From the previous answers, compare the data stored in Excel and then compare to the representation of the data stored in Python. What does `NaN` refer to in the Python data table?

The data are the same. Though Panda data frames use `NaN` as the missing data code for *Not a Number*. Missing data in an Excel file is indicated by an empty cell.

2. Large data set

Data: http://web.pdx.edu/~gerbing/data/Siegel/Donations.csv

Codebook: http://web.pdx.edu/~gerbing/data/Siegel/DonationsDefs.pdf

Prelude:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

a. Read the data into Python.

```
d = pd.read_excel('http://web.pdx.edu/~gerbing/data/Siegel/Donations.csv')
```

b. How many rows of data? Columns of data?

```
d.shape
```

c. List the first ten rows of data and the variable names.

```
d.head(10)
```

d. Use Python to identify numeric and character variables.

```
d.dtypes
```

e. All character variables are necessarily categorical variables. Are any numerical variables categorical?

Variable `CatalogShopper` is coded as 0 and 1 to indicate if the respondent shops by catalog.