Homework #1 All the questions carries 10 points each except the last question in part B (progrm) carries 30 points

Goal: Gain familiarity with Python, Pandas, and Numpy. Use these tools to create, load, clean, and play with data.

Tags: Python, Preprocessing, descriptive statistics

Due: Thursday September 17th @ 11:59pm

How: submit your written report and code to icollege.

Prep:

- Install python, preferably version > 3
- Install Jupyter notebook
- Install pandas
- Install numpy
- Install matplotlib
- Import sklearn

Written Questions (Part A)

1. In your own words, write a paragraph or 2 explaining the value and importance of domain knowledge in data science.
2. In your own words, 2-3 paragraphs, explain why preprocessing is so important. What are some of the common reasons we need to perform preprocessing? What are some approaches for cleaning data properly. Are there any potential complications that would arise from any of these techniques?
3. Broadly speaking: what are the 2 types of analysis we are interested in doing on a particular dataset. (hint: suppose we know what we are looking for, suppose we don't) What differentiates the 2 approaches and what tools do we have at our disposal for each.
4. What are the 2 primary categories of prediction that we are usually concerned with. Give examples and describe scenarios for each category. What complications may arise in either scenario?
5. In your own words, explain the difference between linear and non-linear data. Can we perform regression on non-linear data? Why or How? Can we perform classification on all types of data? Why or why not? Are all linear models useful? Why or why not? If data is not linearly interpretable or separable, what might this tell us about the dataset or our underlying prior assumptions about the data? Do we need to rethink what we are looking for, or do we simply need to find new ways of visualizing the data. Explain your answers in detail, examples would be helpful.

Programming (Part B)

1)

Create a python file called hw1.ipynb in jupyternotebook
Load the iris dataset like so:

```
from sklearn.datasets import load_iris
```

Print the dataset
    Copy the output from your terminal and paste it into your report.
Print the shape of the dataset
    Copy and paste the output to your report
Print the feature names (columns) of the dataset
    Copy and paste to report
Print the targets of the dataset
    Copy and paste to report
Print the shape of the targets
    Copy and paste to report

2)

Import pandas into your script
Convert the dataset into a Pandas Dataframe
Assign the iris feature names to the columns in your Dataframe
Add a column called “CLASS” to your DataFrame which contains the target values from the dataset
Now print the whole dataFrame
    Copy and paste the output to your report
Select and print only the column called “petal length (cm)”, and only rows 5-10
    Copy and paste the output to your report
Select and print only the column called “sepal width (cm)” and only rows 7-19
    Copy and paste the output to your report
Using Pandas, Print the mean and standard deviation of the column called “sepal length (cm)”
    Copy and paste the output to your report
Using the describe() function from Pandas. Print the output of your Dataframe like so: df.describe().
    Copy and paste the output to your report

3) This last question carries 30 points

Download the dataset “nba.csv” from icollege.
Load the dataset with pandas.
Print the number of missing values are there in the “College” column.
    Copy and paste this to your report
Are there any missing incorrect values in the “Age” column? If so, how many?
Use df.describe on the “Age” column. Is there anything unusual?
    Describe your observations in the report.
What can you do to account for the anomalies you find?
    Explain 2 options in your report.
How would each of the options you describe affect the statistics of the “Age” column?

Implement both techniques, and use df.describe after each technique to see how the underlying statistics have changed.

Print the covariance matrix using the “Age” and “Salary” columns. Can we learn anything useful from this? (Don’t forget missing values) (and, its ok if it’s not very useful) Include this analysis in your report.