Module 3 Assignment - Problem Statements (80 points)

Note that the answers to each of these questions should be the direct result of running appropriate Python code in Jupyter notebook and not involve any manual processing of dataset files. Answers without either the code or code output will not receive any grade.

Make sure you have all packages installed and imported as below:

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

1. The dataset “weights.tsv” contains weight values (in pounds) for 20 people: 164, 158, 172, 153, 144, 156, 189, 163, 134, 159, 143, 176, 177, 162, 141, 151, 182, 185, 171, 152.

    1.1. Import the data from the file weights.tsv into a Pandas Series object in Python. **(4 points)**

    1.2. Create a new series object with weights converted to kilograms from pounds (1 pound = 0.453592 kilograms). Round the results to two decimal places. **(4 points)**

    1.3. Find the mean, median, and standard deviation of both series objects using Pandas functions. **(4 points)**

    1.4. Plot a histogram of weight (in kilograms) using matplotlib library with 10 bins. **(4 points)**

2. For this problem statement, you are given a dataset named “boston.csv”. This dataset contains information collected by the US Census Service concerning housing in the areas of Boston, Mass. The data was originally published by Harrison, D., & Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management,5,81–102.

    Here is the description of variables/columns in this dataset:

Column name	**Description**
CRIM	per capita crime rate by town

ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
NDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
LSTAT	% lower status of the population
MEDV	median value of owner-occupied homes in $1000's

2.1. Import the dataset "boston.csv" into a Pandas dataframe and obtain the number of rows and columns for the dataframe. **(3 points)**

2.2. What is the owner-occupied home value (MEDV) for the lowest nitric oxide concentration (NOX) from the dataframe? **(3 points)**

2.3. Create a boxplot of per capita crime rate (CRIM) using Matplotlib. Obtain the interquartile range for crime rate (CRIM) using Pandas functions. **(4 points)**

2.4. Subset all columns of the dataframe for rows with outliers of crime rate into a new dataframe. Compare the averages of AGE between the two dataframes with respect to crime rate, what do you interpret? (Hint: Outliers exist 1.5 times of interquartile range above third quartile and below first quartile) **(4 points)**

2.5. Create scatterplot between distances to employment centers (DIS) and nitric oxide levels (NOX). Obtain correlation index between the two columns and interpret their relation. **(4 points)**

2.6. Similarly, create a scatterplot between highway accessibility index (DIS) and property tax rate (TAX). Obtain correlation index, compare it to the scatter-plot

and interpret the relation between DIS and TAX. Take appropriate action on the data based on your observation. **(6 points)**

3. We will be using the “tips” dataset from seaborn package for this problem statement. This dataset contains information about restaurant bills and tips made by people classified by their gender along with few other attributes which are self explanatory. You can import this dataset into a pandas dataframe as follows:

```
tips_df = sns.load_dataset('tips')
tips_df.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

3.1. Calculate percentage of tip amounts for bill totals, rounded to two decimal places and create a new column “tip_percent” in the same dataframe. **(3 points)**

3.2. For what days in the week do we have the data, and which day on average has the highest bill? (Hint: lookup for “groupby” in pandas documentation) **(3 points)**

3.3. Are there more dinners or lunches? Create a dataframe with this data. Are there more smokers during lunches or dinners? Create another dataframe with this data. Join the two dataframes by time of day and calculate the percent of smokers at lunch and dinner. Compare the results. **(6 points)**

3.4. Using the boxplot function from seaborn package, create plots on “tip” column for Male and Female from “sex” column. Compare the boxplots and provide your interpretation on outliers between males and females. **(4 points)**

3.5. Create the same boxplots as above for “tip_percent” and “sex”, for tip percent below 70. Now compare the boxplots between male and female, which boxplot has more outliers and which one is more symmetric? **(4 points)**

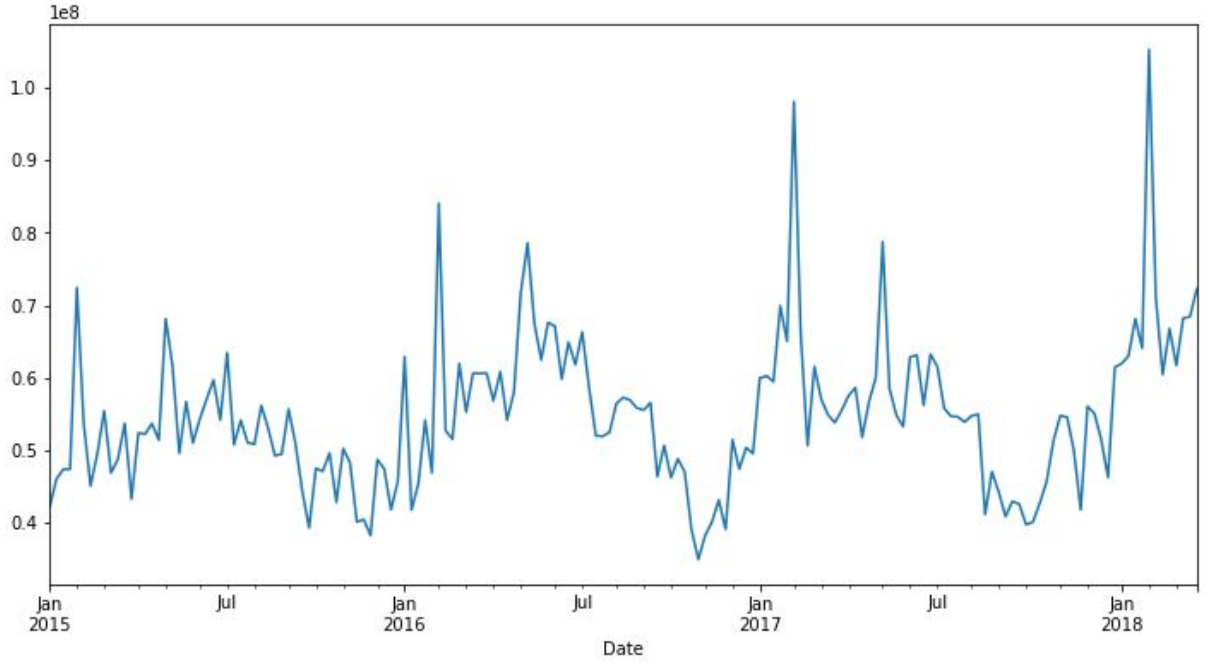
4. For this last problem statement, you will work on the “avocado.csv” dataset which contains information related to avocado sales across multiple regions/cities over the years 2015 to 2018 organised by date. The data contains 10 columns which are self explanatory.

    4.1. Import the dataset file into a Pandas dataframe and identify the count of missing values per column. Handle missing values based on column type and explain your reasons behind selecting appropriate techniques. **(8 points)**

    4.2. Convert the fields Type, Year and Region to categorical data type and subset the dataframe to exclude region “TotalUS” and sort the dataframe by date in ascending order. Is the average price of an avocado higher in 2017 compared to 2016? **(4 points)**

    4.3. Sum up the total volume of avocado sales by region and create a horizontal bar plot using Matplotlib. Which state from the region has the highest sales of avocados by volume? Subset the data for that state, create a histogram of average price and interpret it. Obtain the correlation index between average price and total volume for that state, what do you find? **(6 points)**

    4.4. Provide you observations of the following timeline plot of avocado sales by volume. Which month consistently has the highest volume of sales every year? In general, what could be some possible reasons driving this surge in sales? **(2 points)**


1e8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Jan
2015
Jul
Jan
2016
Jul
Jan
2017
Jul
Jan
2018
Date