Homework-5

**Out Date:** 10/11/2019 (Friday)

**Due Date:** 10/30/2019 (Sunday) 11:59PM

Team#: \_17\_

Team Member-1:\_\_\_\_\_ Ying-Yu Huang \_\_\_\_\_\_\_\_\_ Member’s Contribution (in %) 50%

Team Member-2:\_\_\_\_\_\_ Chao Zhang \_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) 50%

**Submission**

1. Work on the Problme-1.
2. Prepare your Python file for Problem-1 (e.g., HW5\_P1\_Team#.py).
3. Upload the files to blackboard.

**Problem Statement:** You are consulted by a health insurance company to analyze its insurance dataset. The goal is produce a set of descriptive statistics. The dataset is in the txt file format (**insurance.txt)** and is available under the homework folder.

The file includes 1,338 examples of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year. The features are:

* **age:** An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
* **sex:** The policy holder's gender, either male or female.
* **bmi:** The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9. A person with a BMI value within the range of 25 to 29.9 is considered overweight. A person with a BMI value above 30 is considered obese.
* **children:** An integer indicating the number of children/ dependents covered by the insurance plan.
* **smoker:** A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
* **region:** The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.
* **expense:** total medical expenses charged to the plan for the calendar year

Using the **numpy library** analyze the data. In particular, read the data file (numpy.loadtxt()) **[5 points]**, produce the following analysis and store the results into a text file (numpy.savetxt()) **[5 points]**:

1. Mean, standard deviation and median of age. **[5 points]**
2. Mean, standard deviation and median of BMI. **[5 points]**
3. Mean, standard deviation and median of BMI grouped by sex. **[5 points]**
4. Mean, standard deviation and median of BMI for smokers and non-smokers. **[5 points]**
5. Mean, standard deviation and median of BMI grouped by region. **[5 points]**
6. Mean, standard deviation and median of BMI of those who have more than 2 children. **[5 points]**

How do the following factors affect BMI? Justify your comments with supporting descriptive statistics (mean, standard deviation and median).

1. Smoking habit **[10 points]**
2. Region **[10 points]**
3. Children **[10 points]**

What are the primary reasons for the top 20% of the expenses? In particular, sort the data by expense, and compute the mean, and standard deviation of BMI and the mode of smoker and region. How do these values differ from the rest 80% of the population? **[10 points]**

Please make sure your code follows the Python programing style guide available here: <https://www.python.org/dev/peps/pep-0008/> **[10 points]**.

Please make sure the code is well-commented **[10 points]**