**Assignment 4**

**Problem 1. Please perform the following steps on Assignment4\_data.csv:**

1. Explore the datasets.
2. Find and handle missing values are in the data. (It is your choice how you handle the missing data.)
3. Explore the variable column and Convert the "variable” column to dummy variables and join the dummies to the data.
4. Convert the "one” column into 3 bins.

**Problem 2**

**Use the following speech by the Rev. Dr. Martin Luther King, Jr:**

s = “I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation. Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity. But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languishing in the corners of American society and finds himself an exile in his own land. So we have come here today to dramatize a shameful condition."

1. Find out how many unique words in s.
2. Which word appears the most?
3. How many words start with ‘t’.

**Problem 3**

1. Load and read Registration.csv and Course\_info.xlsx using Pandas.

2. Explore and clean Registration data.

3. Explore and clean Course\_info data.

4. Which course has the highest registration?

5. Inner join two datasets.

Create a data frame with student name as the index, course numbers as columns, and if the student registered a course as values(0, 1).

Below is a part of the final data frame:

|  |  |  |
| --- | --- | --- |
|  | ARTS400 | ARTS401 |
| ABella Abzug | 0 | 0 |
| Al Gore | 0 | 0 |
| Al Hirt | 0 | 0 |
| Al Roker | 1 | 0 |

**Problem 4**

**Consider the in class activity problem 8 and do the following:**

1) Set the class labels ‘benign’ as 0 and ‘malignant’ as 1 in the dataset using dictionary

2) create another variable ‘cell\_type\_label’ which has two categorical values ‘normal’ and ‘abnormal’ depending on the following conditions:

‘Normal’: if sell\_size\_uniformitty > 5 and cell\_shape\_uniformity > 6

Else : ‘abnormal’

Find the subset of the that has ‘abnormal’ cell\_type\_label