**ASSIGNMENT 3**

Topic: Data Cleaning and Preparation & Data Wrangling

Due: 1 week from announcement.

**TASK 1:**

1. Read the “*Diabetes Dataset.csv*” and save the data into a DataFrame.
2. What is the shape of the DataFrame?
3. Get the number of missing values for each column.
4. Fill in all the missing values in column *No\_Times\_Pregnant* with value 1.
5. Drop all rows which contains less than 4 observation values.
6. Fill in all the missing values in below columns with the given values or methods:

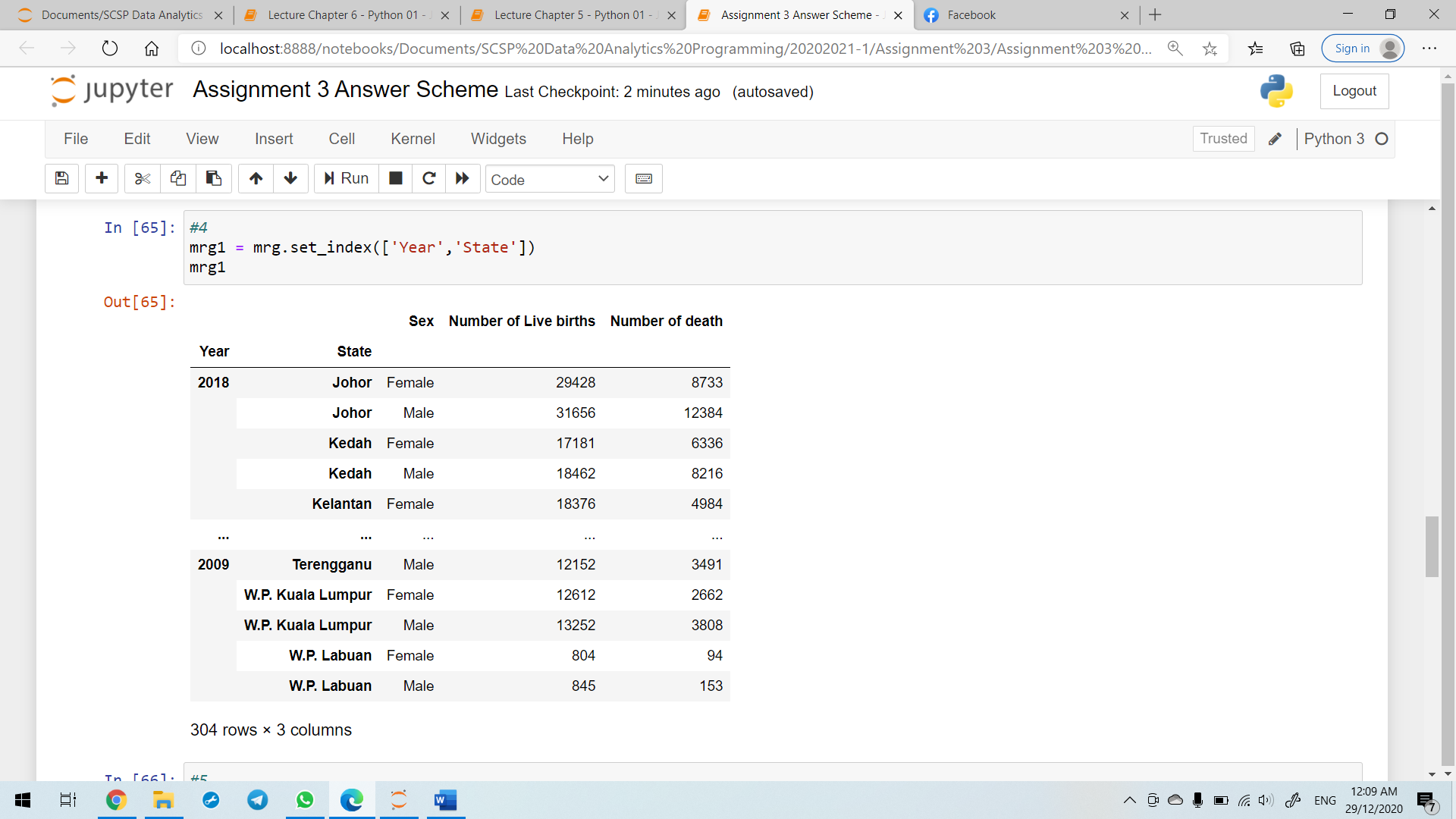
|  |  |
| --- | --- |
| **Column Name** | **Value/Method** |
| *Plasma\_Glucose* | Mean |
| *Diastolic* | Forward Fill |
| *Triceps* | Backward Fill |
| *Insulin* | Mean |
| *BMI* | Median |

1. Check back the number of missing values for each column. It should now be 0 for all the columns.
2. Bin the age into below categories. How many patients with diabetes for each category?

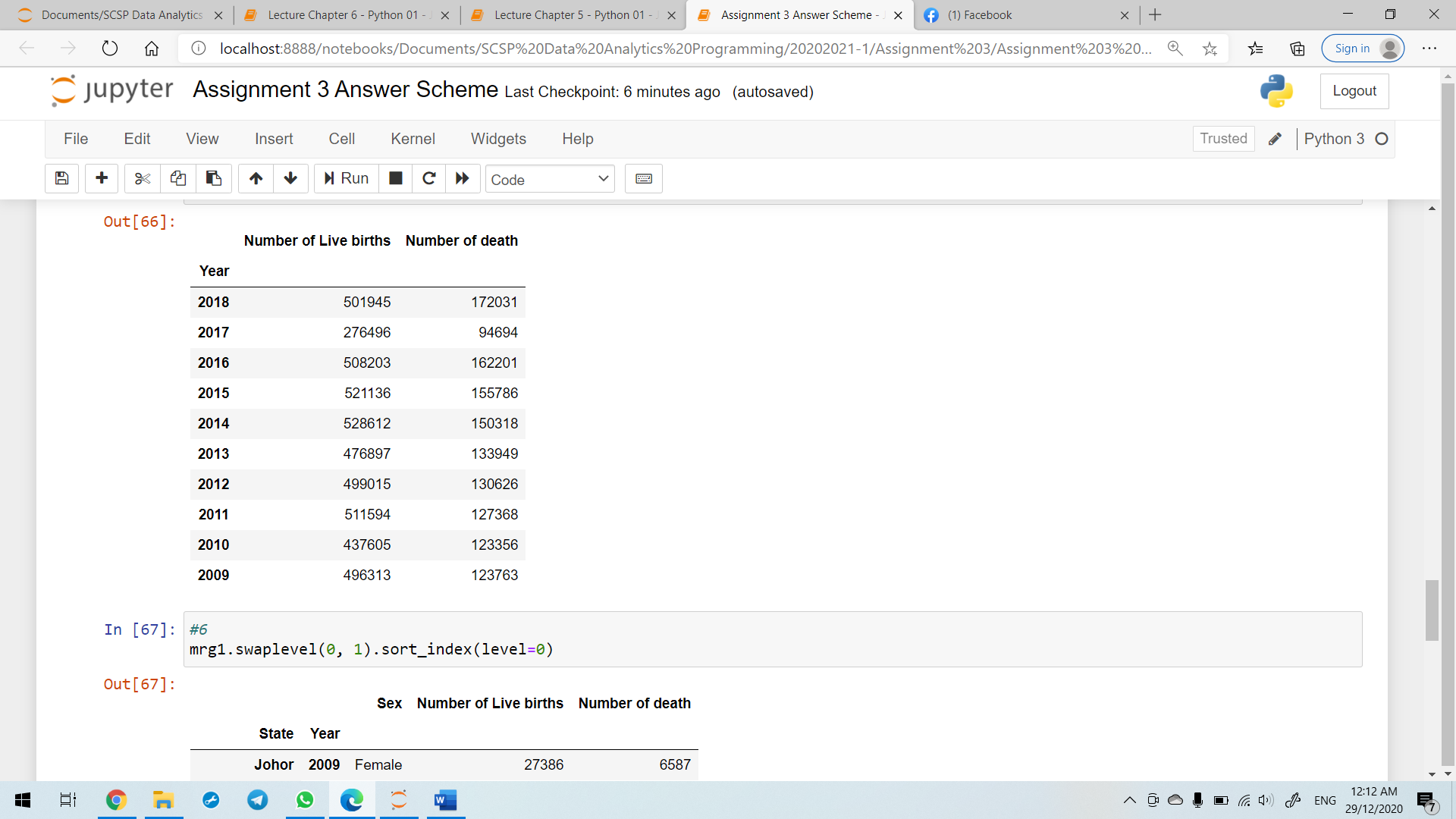
|  |
| --- |
| **Age Category** |
| 20 ≤ Age < 30 |
| 30 ≤ Age < 40 |
| 40 ≤ Age < 50 |
| 50 ≤ Age < 60 |
| 60 ≤ Age < 70 |
| 70 ≤ Age < 80 |
| 80 ≤ Age < 90 |

**TASK 2:**

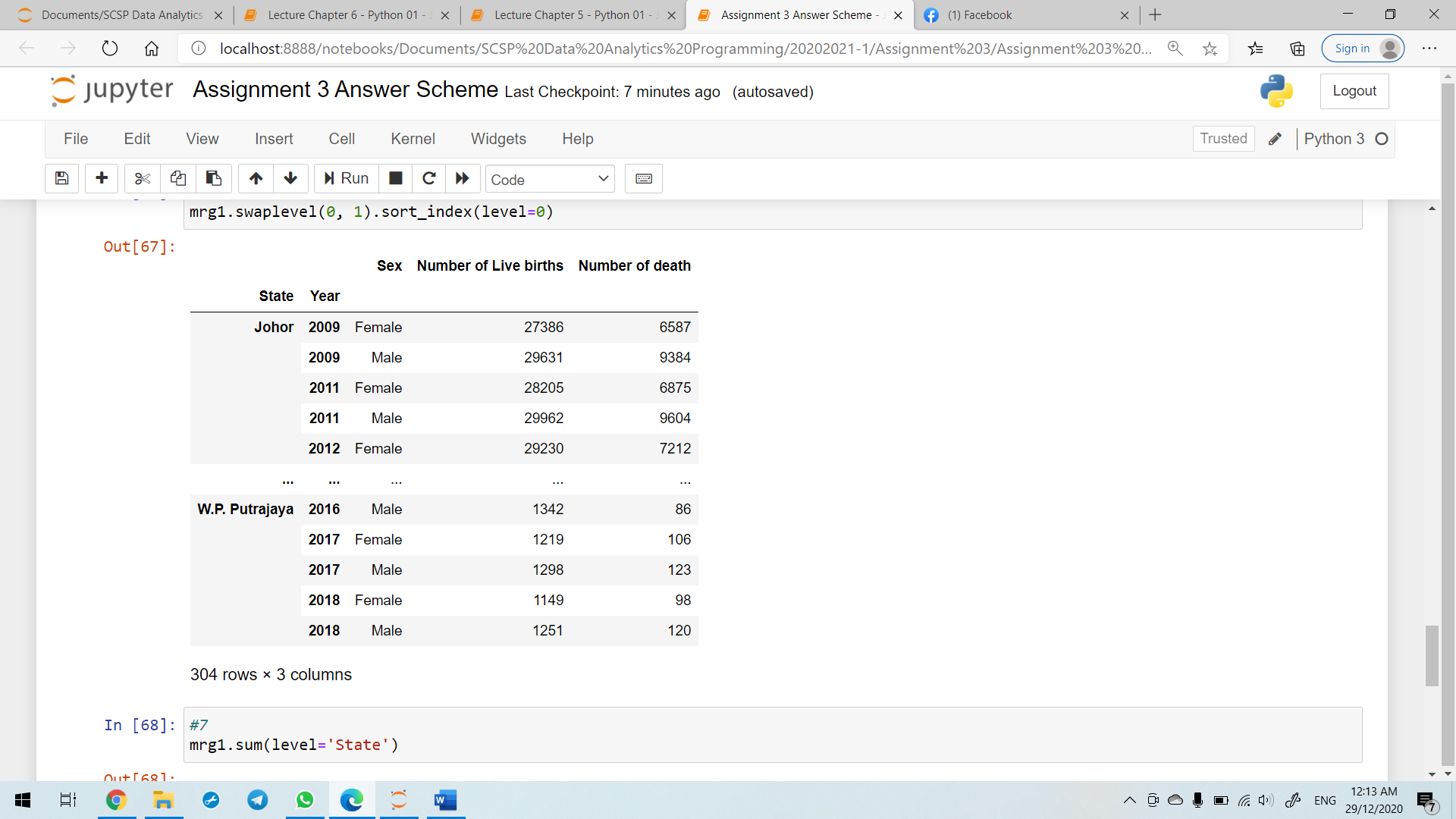
1. Read the “*Live births by state and sex.xlsx*” and save the data into a DataFrame.
2. Read another file named “*Death by state and sex.xlsx*” and save it into another DataFrame.
3. Merge both files.
4. Set the index as figure below.



1. Display the number of live births and number of deaths for all the years (2009-2018) as below.



1. Swap the level as figure below.



1. Display the number of live births and number of deaths for all states as below.

