1/31/2020

Suseel K, Rohit

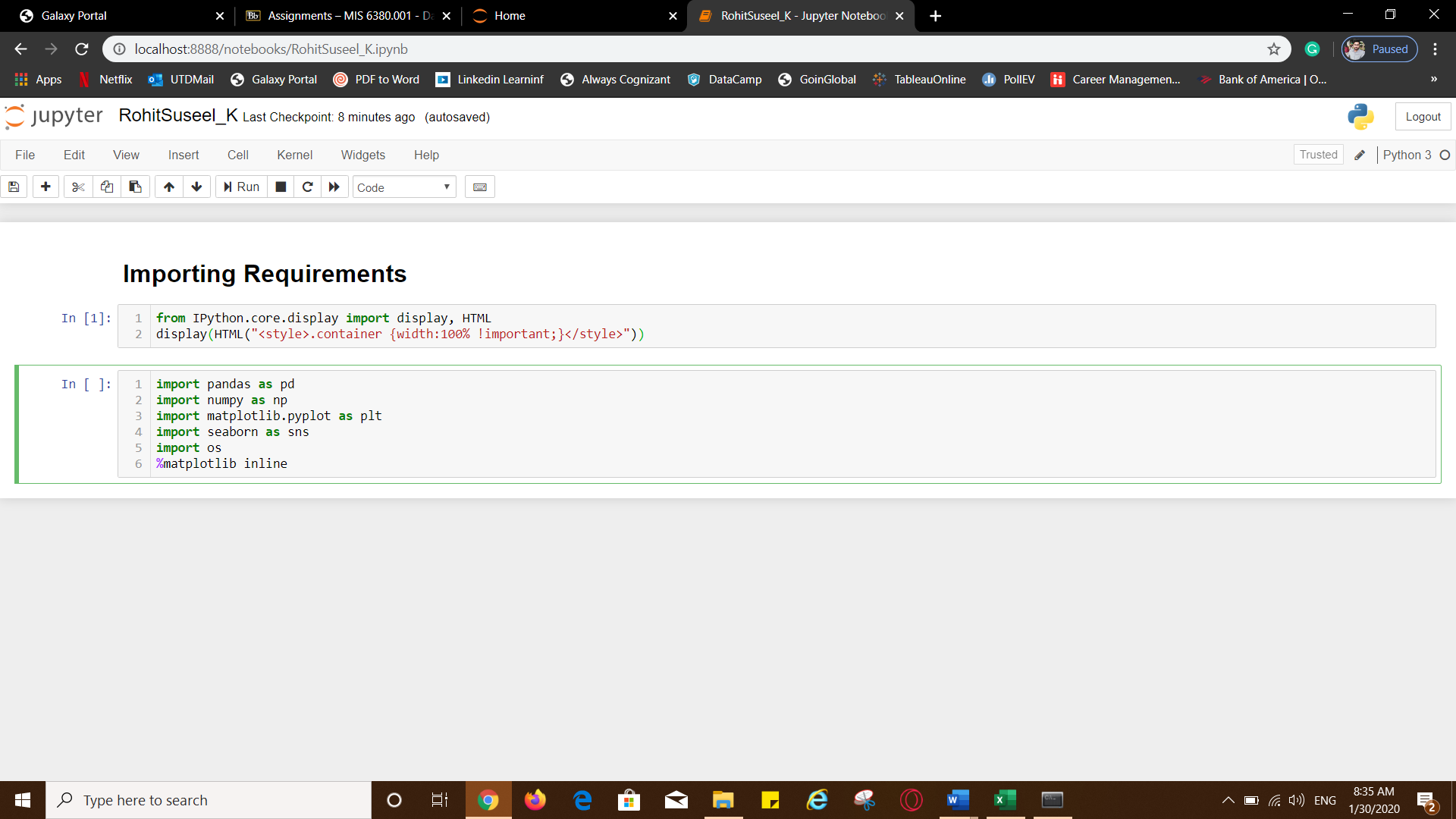
RXS190016

Data Cleaning

Assignment 3

**Question:**

1. **Take a screen shot of your page, showing page Title (your name) and page background.**



**b. How many Sheets are present, also name them?**

There are two sheets. They are Observations and Products.

**c. How many non-null rows are present in the data set?**

Observations sheet : 931766 non-null rows are present in the dataset. Since Price column has the greatest number of null values, taking non null values in Price column will give the total non-null rows present in the data set.

Products sheet : 30 non-null rows

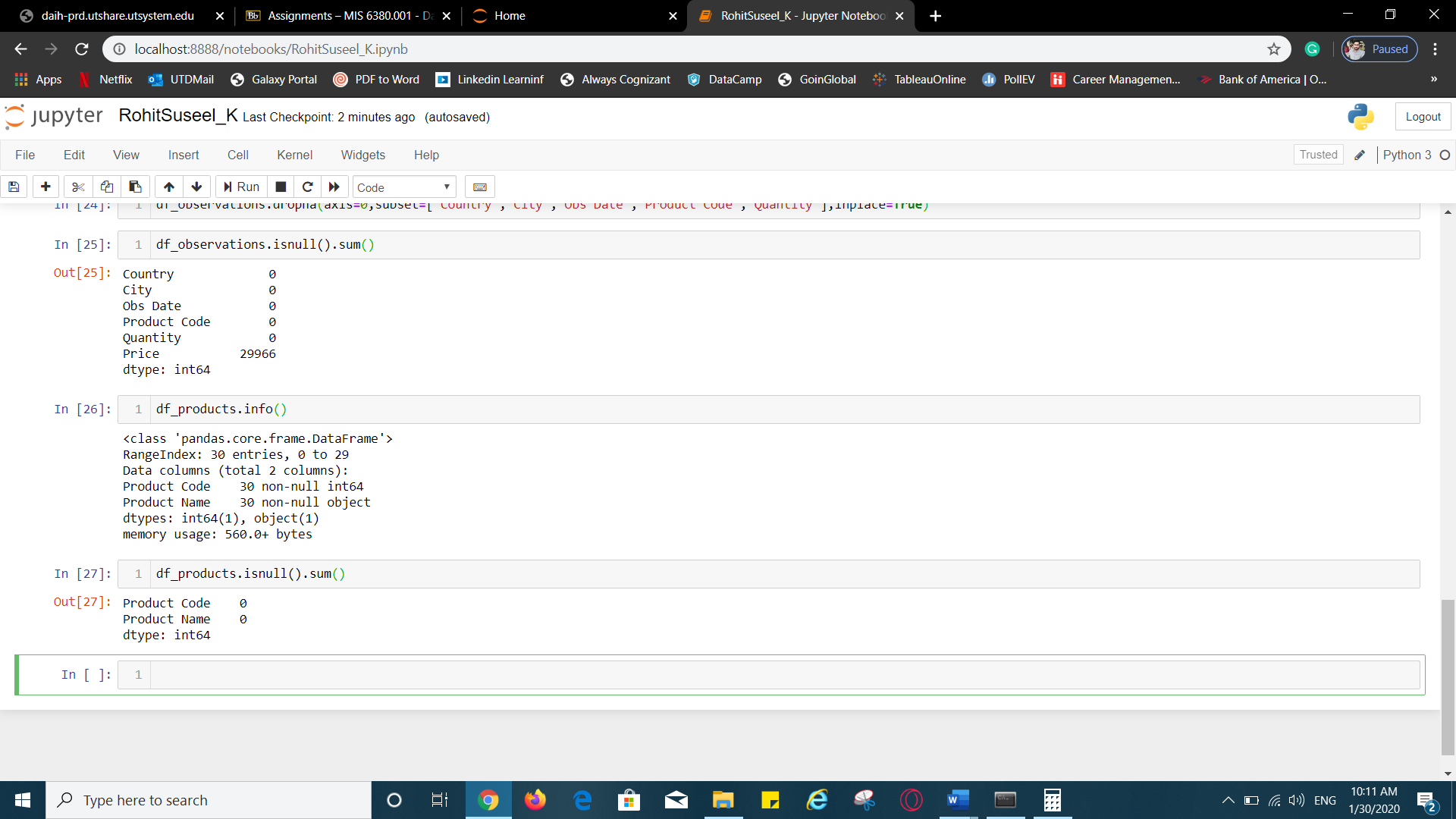
**d. How many empty rows were present in the data set before we removed them?**

Observations sheet : 65

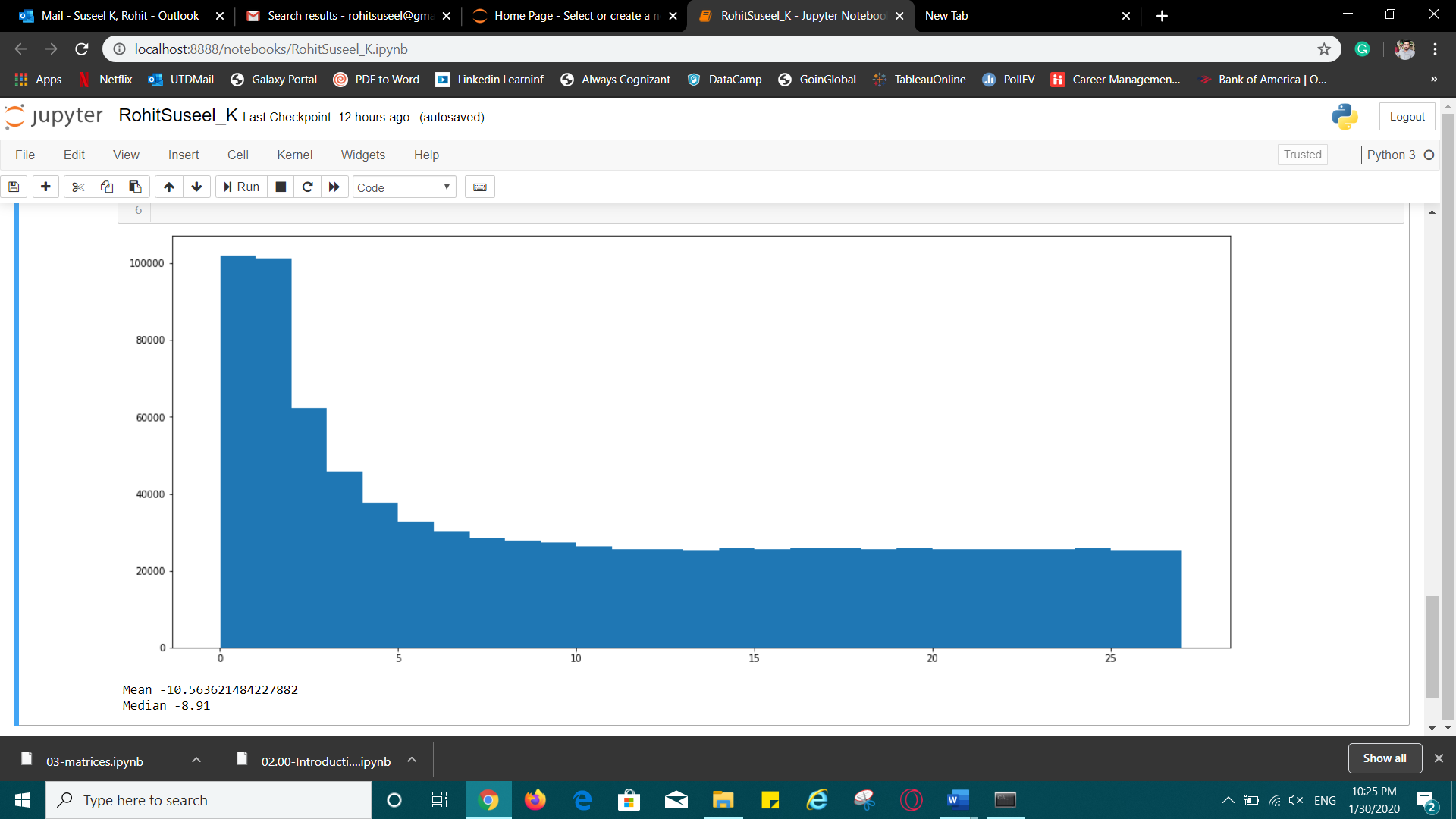
Products sheet : 0

**e. How many empty rows are present in the Products sheet? Paste the Screenshot of the output showing page title (your name).**

Products sheet : 0 empty rows

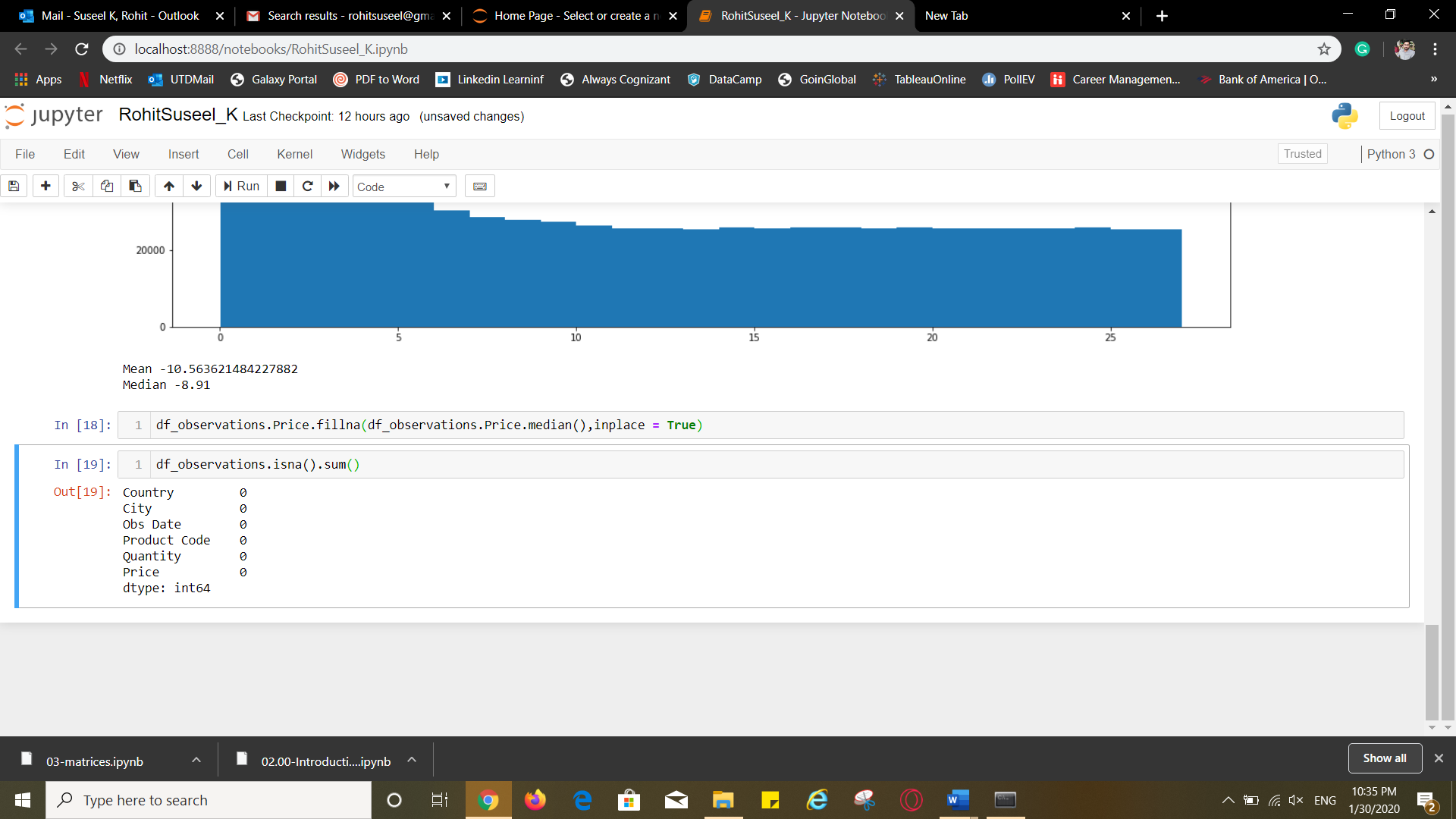


**f. Paste a screenshot of the histogram and explain the methodology chosen for imputation.**



The histogram appears to be skewed and hence we should adopt Median as the method for imputation. Median here is 8.91.

**g. Paste a screenshot of the screen ( display name ) with the imputed method code showing the result of the ‘ .isna().sum ’ function. How many null values are there?**



There are no null values now in Observations sheet.

**h. List the country with duplicate entry/ spell check error.**

['Kenya',

'Pakistan',

'India',

'Bangladesh',

'Indonesia',

'Nigeria',

'Philippines',

'Brazil',

'Brazel',

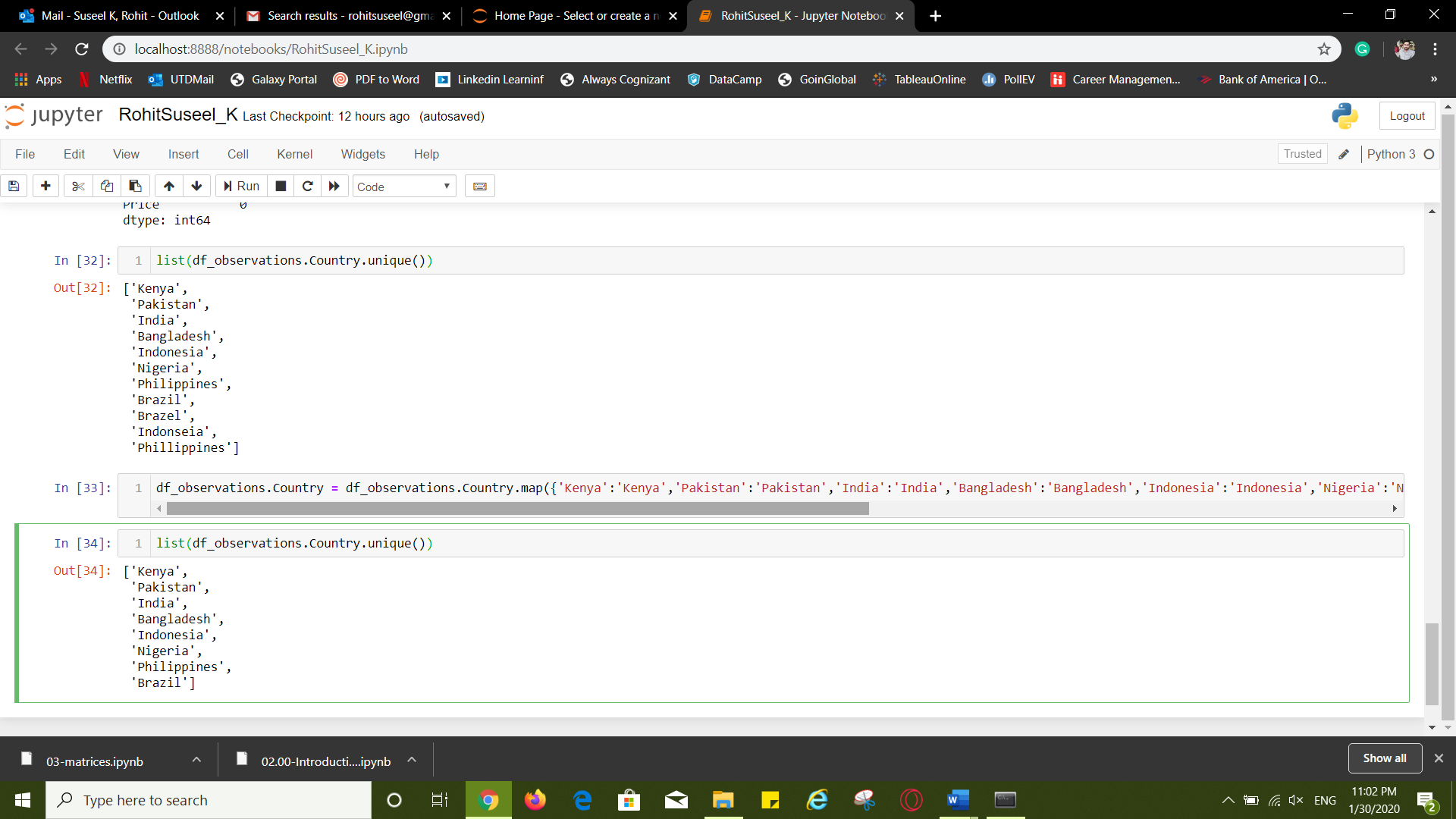
'Indonseia',

'Phillippines']

Brazel,Indonseia and Phillippines are three countries with spell check error and

these are duplicates of Brazil, Indonesia and Philippines.

**i.Again, run the unique value command and paste the Screenshot of the whole screen. Explain what happened.**



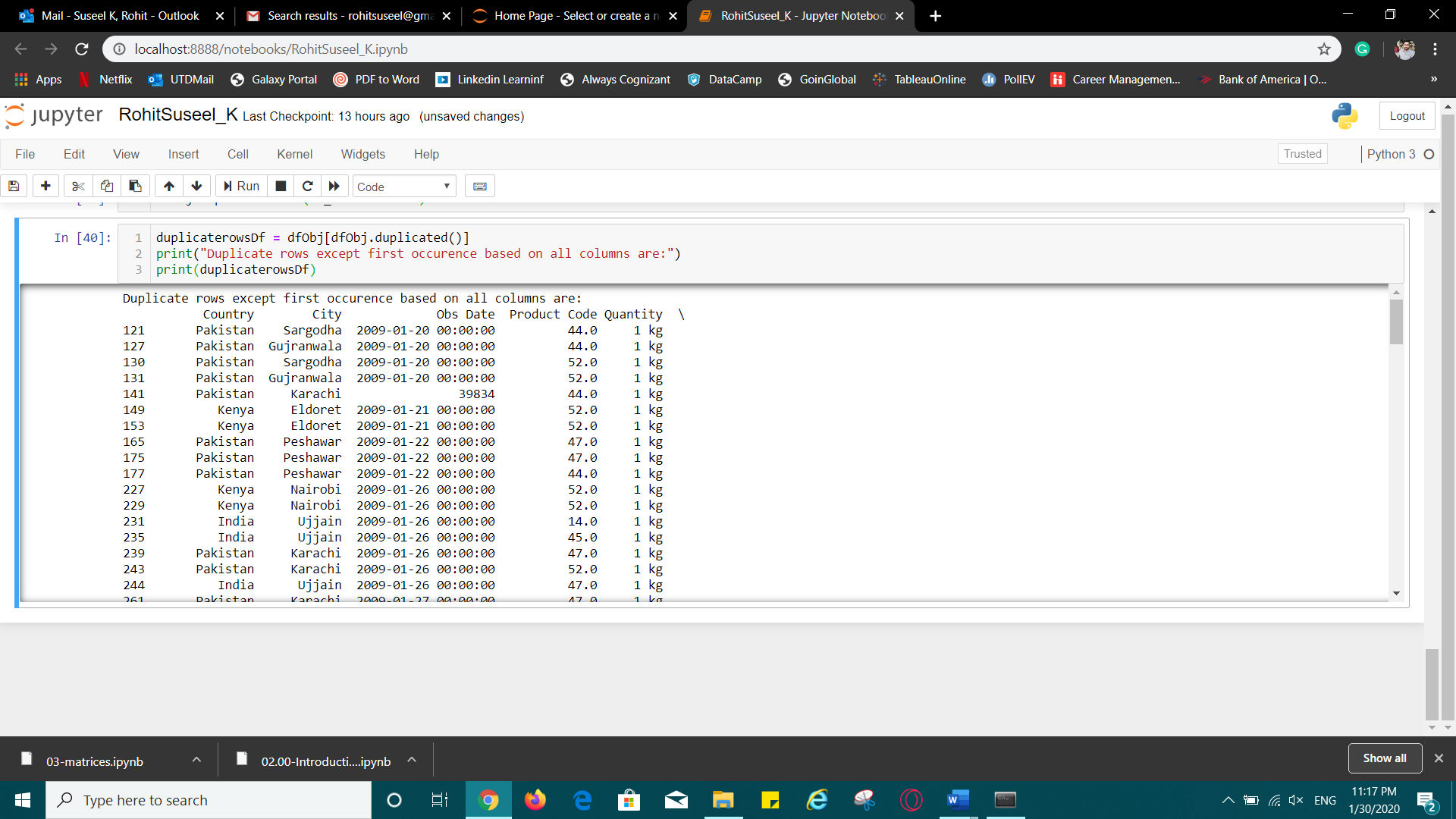
The map function mapped the values in the Country column to the values provided in the code. Essentially, the misspelt countries are removed when they were mapped with the correctly spelt country names.

**j.How many duplicate entries are present?**

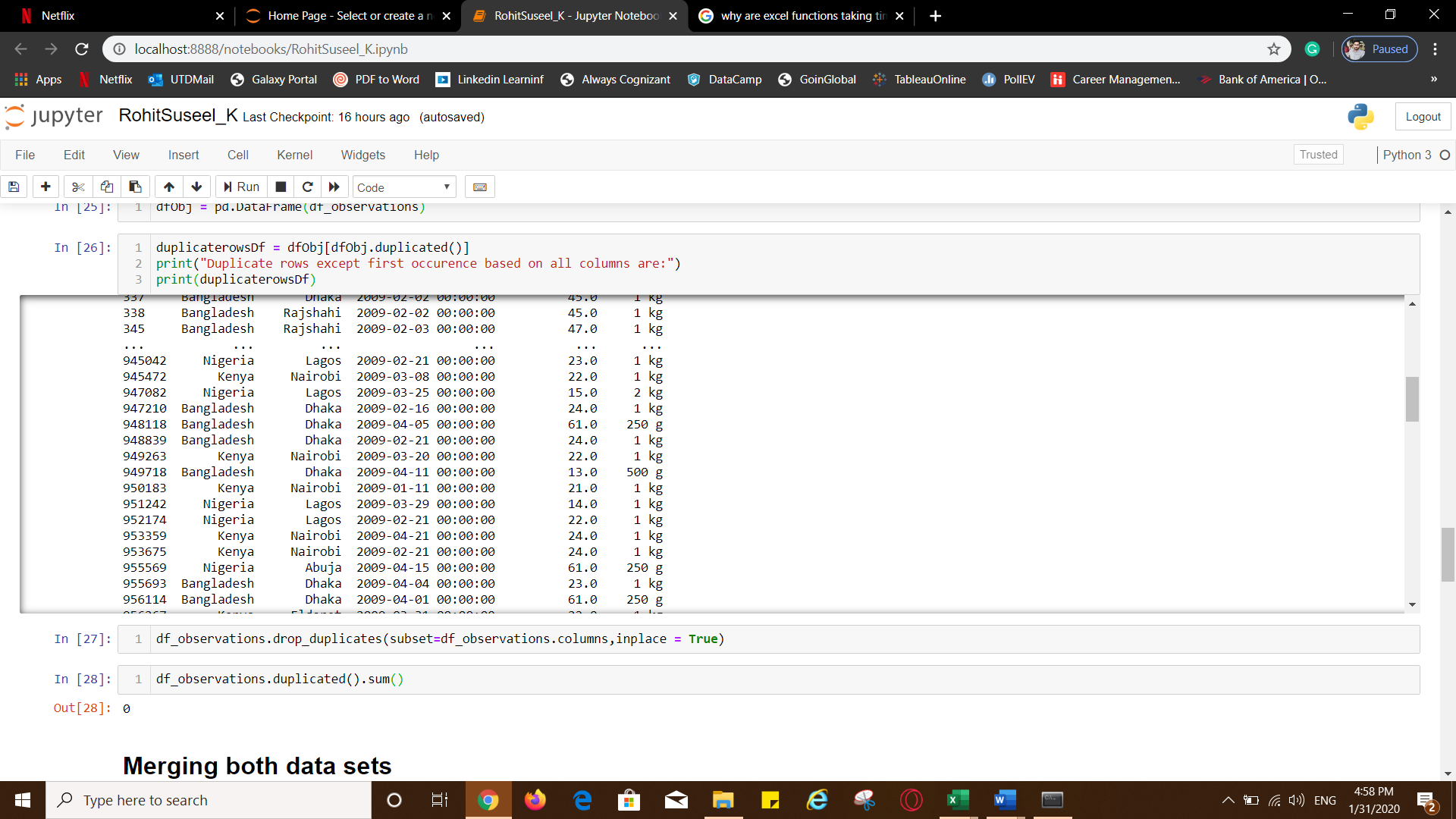
14397

**k. Explain the outcome of the executed query, also paste the Screenshot of the code and the output using whole screen.**

The query displays the duplicate rows except the first occurrence of each.



1. **Use the duplicated(). Sum() function, to check the duplicate entries present. Paste the Screenshot of the code and the output using whole screen.**



1. **Paste the screenshot of the head of df\_full .**

