|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
|  | | **Heart Data analysis using Python for Data Science** | | |  | |
|  |  | | | | |  |
|  | | |  |  | | |
|  | | | **1911004- Nidhi Bhanushali 1911016-Kashish Harisinghani 1911053- Disha Shah** |  | | |
|  | | | 27-10-2020  —  Python for data science  — |  | | |

INTRODUCTION

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

The chosen data set helps to predict various factors which affect heart conditions over a range of age and gender.

Data Understanding

* Data set Quantities description

Initially, the dataset contains 76 features or attributes from 303 patients; however, published studies chose only 14 features that are relevant in predicting heart disease. Hence, here we will be using the dataset consisting of 303 patients with 14 features set.

1. age (Age in years)
2. sex : (1 = male, 0 = female)
3. cp (Chest Pain Type): [0: Typical Angina, 1: Atypical Angina, 2: Non-Anginal Pain, 3: Asymptomatic]
4. trestbps (Resting Blood Pressure in mm/hg )
5. chol (Serum Cholesterol in mg/dl)
6. fps (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes]
7. restecg (Resting ECG): [0: normal, 1: having ST-T wave abnormality , 2: showing probable or definite left ventricular hypertrophy]
8. thalach (maximum heart rate achieved)
9. exang (Exercise Induced Angina): [1 = yes, 0 = no]
10. oldpeak (ST depression induced by exercise relative to rest)
11. slope (the slope of the peak exercise ST segment)
12. ca [number of major vessels (0–3)]
13. thal (Thallium heart scan): [1 = normal, 2 = fixed defect, 3 = reversible defect]
14. target: [0 = disease, 1 = no disease]

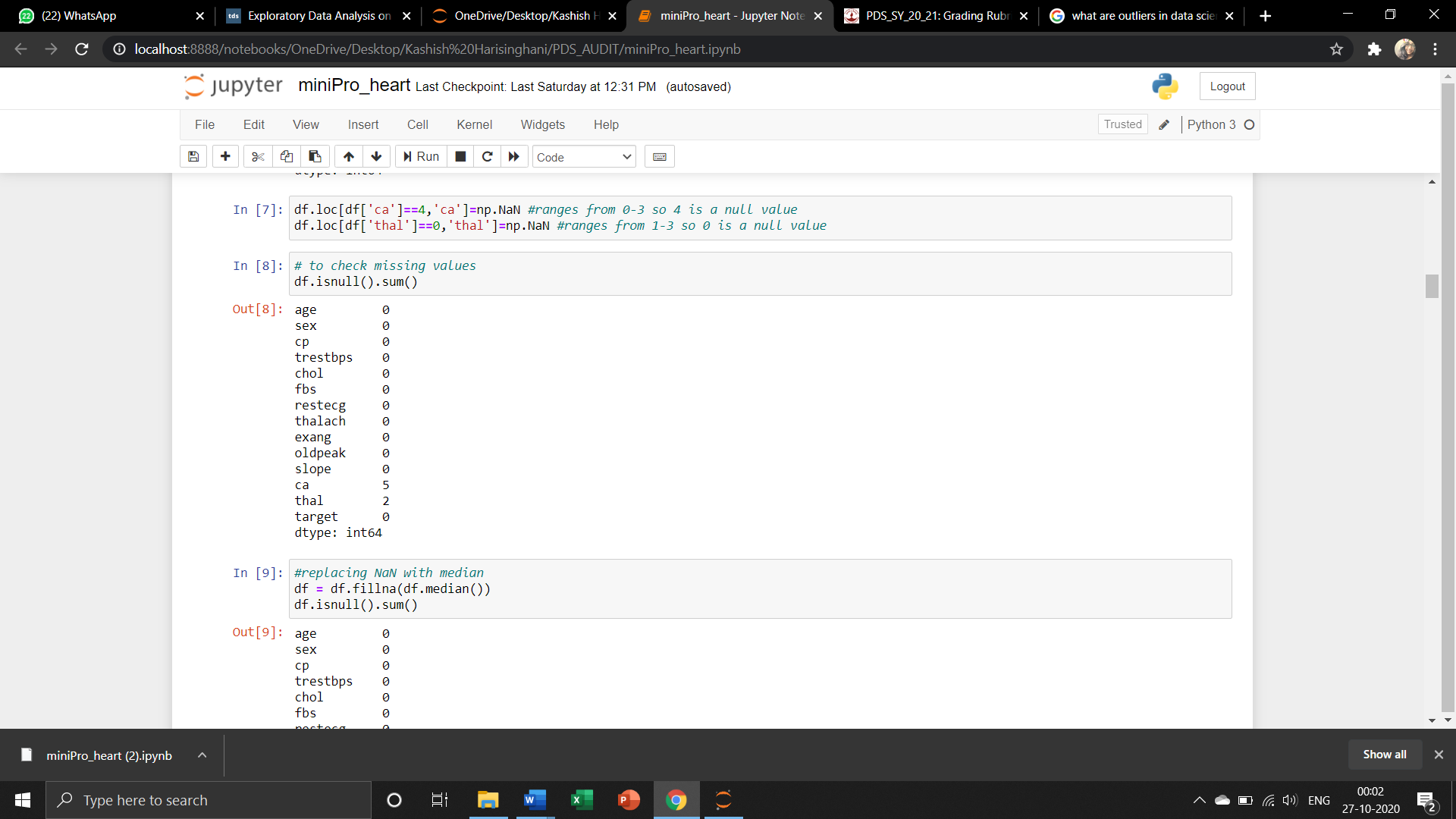
* Dataset Cleaning

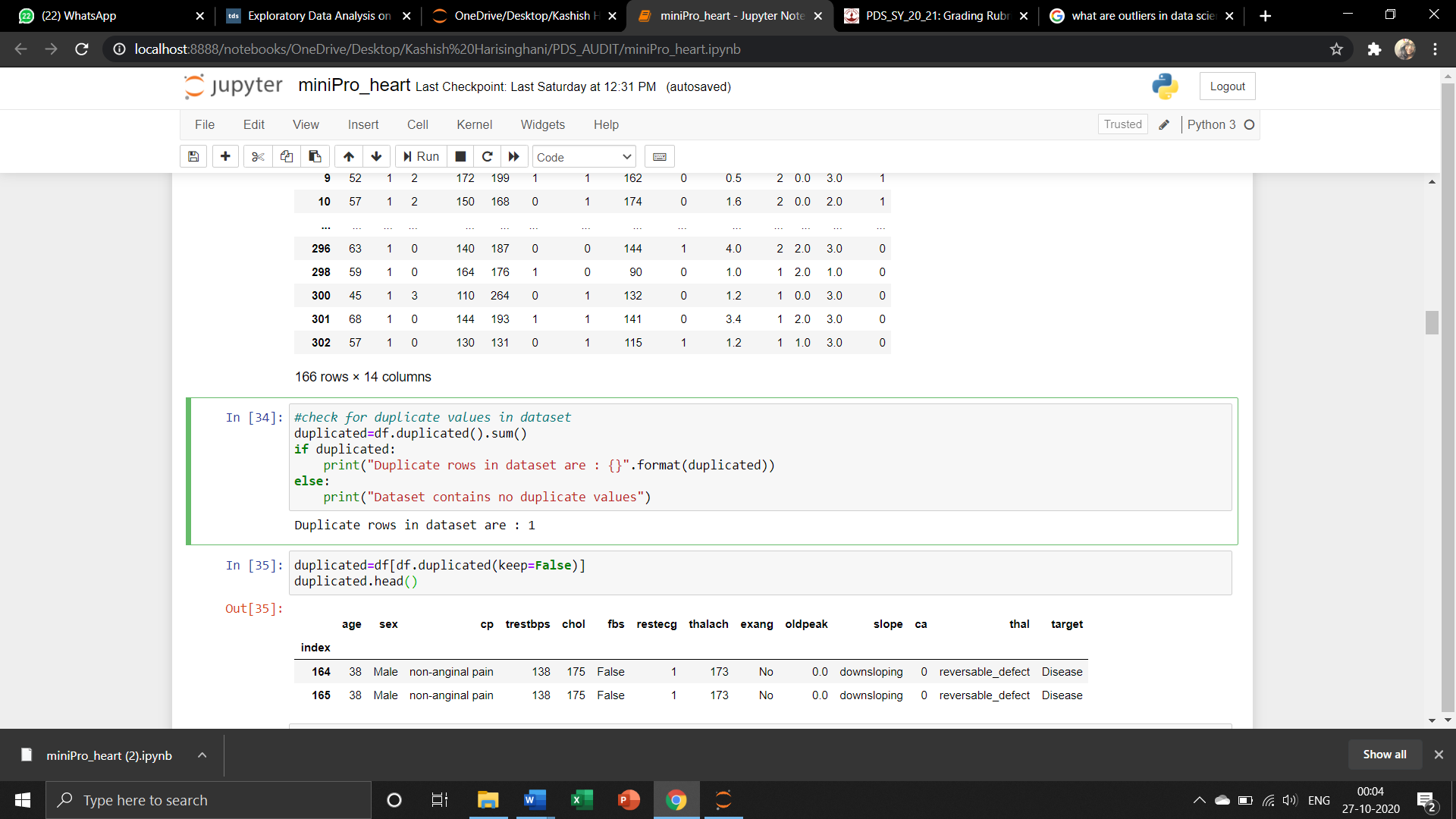
The variables types are

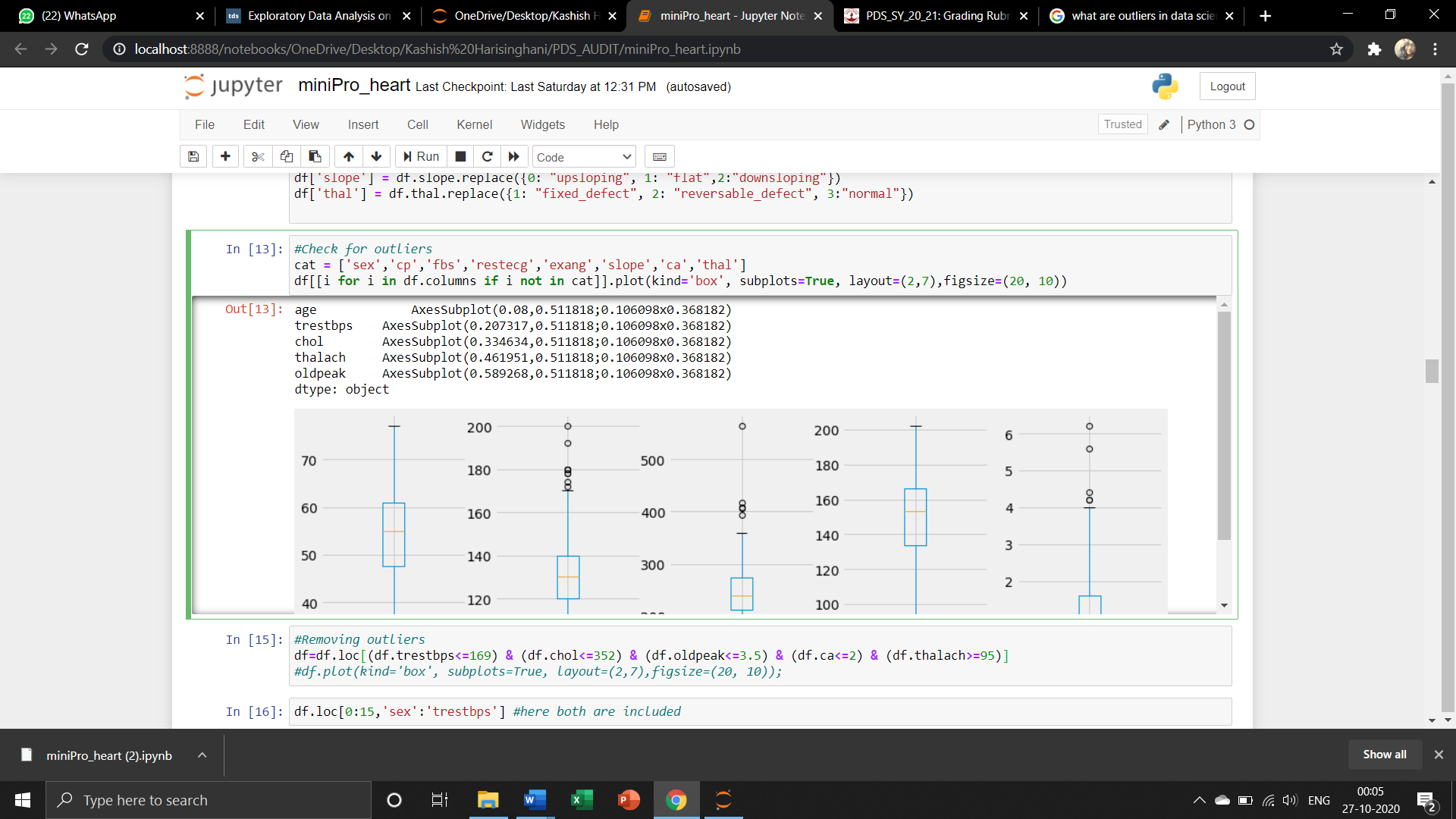
* Binary: sex, fbs, exang, target
* Categorical: cp, restecg, slope, ca, thal
* Continuous: age, trestbps, chol, thalac, oldpeak

We check if the data character mistakes.

1. feature **‘ca’** ranges from 0–3, however,  df.nunique() listed 0–4. So lets find the ‘4’ and change them to NaN.
2. Feature**‘thal’** ranges from 1–3, however,  df.nunique() listed 0–3. There are two values of ‘0’. So lets change them to NaN.

We also check for missing values in all columns to replace them with NaN and find none. Next check for duplicate values is performed and rectified. We also check for outliers in our dataset, which are values that differ drastically from the rest of the data set. Such values need to be removed to visualise the given data correctly.



* Dataset Visualisation
  1. **target variable distribution: After observing the graph we see that there are more diseases patients than healthy ones.**
  2. **Age variable distribution: After observing the graph we see that m**ost of the patients are in the age between 50s to 60s.
  3. **Sex variable distribution: After observing the graph we see that m**ost patients are males.

The distribution graphs give us a brief idea about categories of patients that are suffering from heart diseases in general. This information can be very useful for prediction models.

To get a clearer understanding of the data through its visualisation we have also created some graphs using **seaborn** like a countplot to get the disease classes according to Sex, Exercise Induced Angina and fbs as shown below:

