Course Project – Data Visualization

Shaun Pritchard

Rasmussen College

QMB4400

Tina Harclerode

May 10, 2021

**Course Project – Summary Document**

Looking at the data file we have 14 category features that give information on the patent, the patient's age, the type of element or disease they have uses schedule data, the neighborhood they are located, the patients age, gender, whether or not they received an SMS text, and whether or not they showed up for the appointment or not. From observing the data, I found that there are 87531 no shows in total out of 110528 features totaling 22997 no shows or 79.19% in the "No show" feature column. 677 Null values in the "No-show" feature column would have to be cleaned. Also, other dataset feature columns need to be cleaned to make sure that the analysis is correct Justified.

The data provides some useful feature columns that could determine useful correlations. understanding the correlation between the age of the patients not showing up could give insight it is whether or not it was younger patients more prone to be no-shows or older patients? The data potentially show us insights on the type of disease the patient is dealing with which might have some reasoning as to why they no show. Maybe physical limitations or dependencies? The data gives us locations that can be used to identify correlations of no-shows specific to neighborhoods.

The data shows potential correlations between those patients who have received SMS text or not. Missing SMS text could be a contributing factor two missing the appointments. The data shows the potential to determine and if more males or females are prone to being no-shows. We could also use the data to do Inferential analysis on multiple feature variables that could determine probability up scheduling, specific times of appointments that are not being met concerning the patient gender, location, and medical issue. In-turn the data can be used for forecasting or identifying specific outliers? Actions could be met to facilitate internal better be accustomed to patient care and less missed appointments. Overall, the data should be sufficient enough to tell us what determinant factors or underlying trends for the high quantity of no-shows.

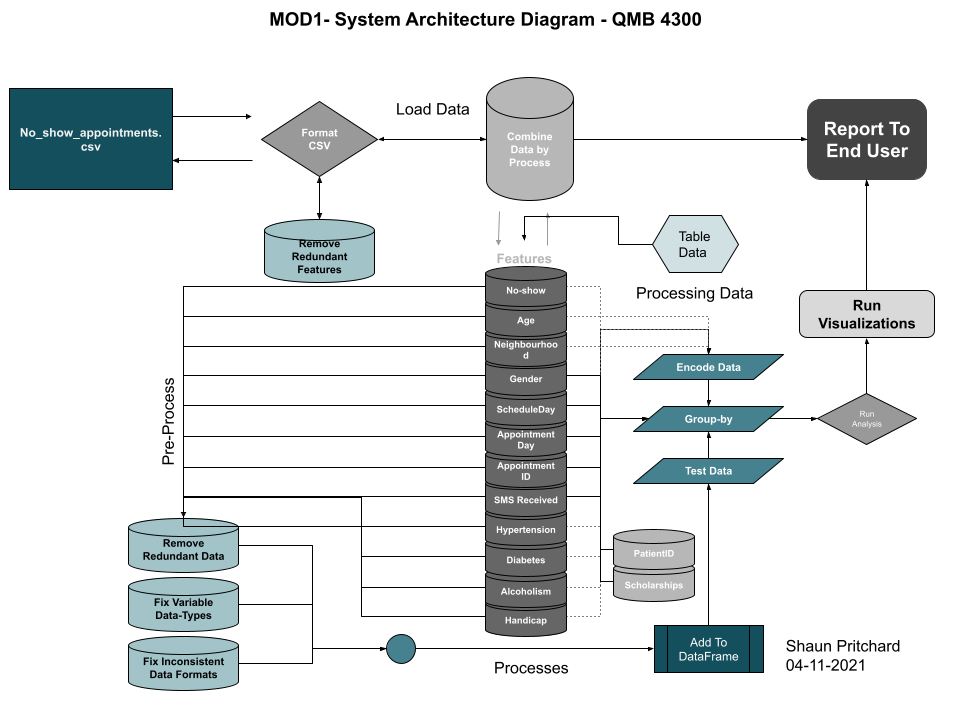
**Missing data:**

Yes, I would say there are several features is missing from the data set. For starters, there is no name associated with the patient ID. By having the patients name, we could determine if there are any patterns or Trends with specific patients who are missing appointments and being inconsistent with their patient care. Having data that describes features for the types of appointments that are being missed would be something of Great Value to this project. This would help the under the correlations between missing appointments based on the type of appointments being received by the patients. We do not have the contextual name of the appointments, but we do have identifying appointment numbers that could be used to determine which appointment types are more frequent no-shows. Also, that more detailed information surrounding the patient medical issue would provide a more feasible depiction of the issue. Also, having a phone number or contact information would be important to this dataset. It could potentially be a factor caused due to inaccurate contact information creating conflicts with scheduling or notification of prior scheduled appointments.

**Justification for additional data:**

By having the patients name data, we could do a detailed descriptive analysis to find out which patients are missing more than others. This would be able to pinpoint specific patients who might be having issues with the times being scheduled. This could lead to the provider implementing solutions to schedule appointments contingent on the specifications of the patient. Having more detail on the actual patient medical issue could be a determining factor in scheduling specific patients with specific issues at the wrong time. Inaccurate contact information can lead to patients missing appointments. Therefor making sure that the contact information is readily available and correct to show correlations between patients who might need to be updated. More detailed location information could help to see if there are any correlations in patients who m9is from specific areas.

**System architecture diagram:**

****

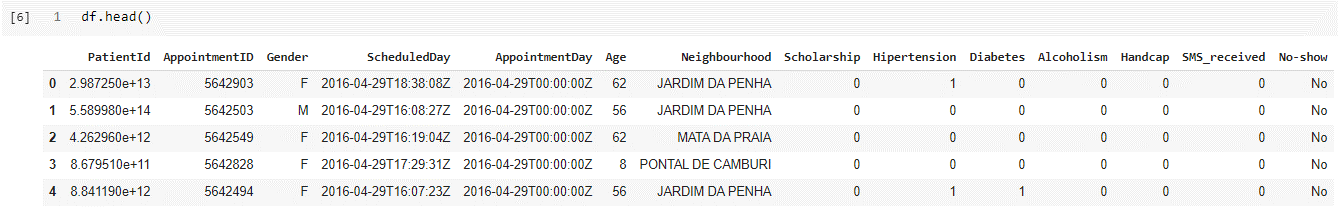
**Course Project – Numpy and Pandas**

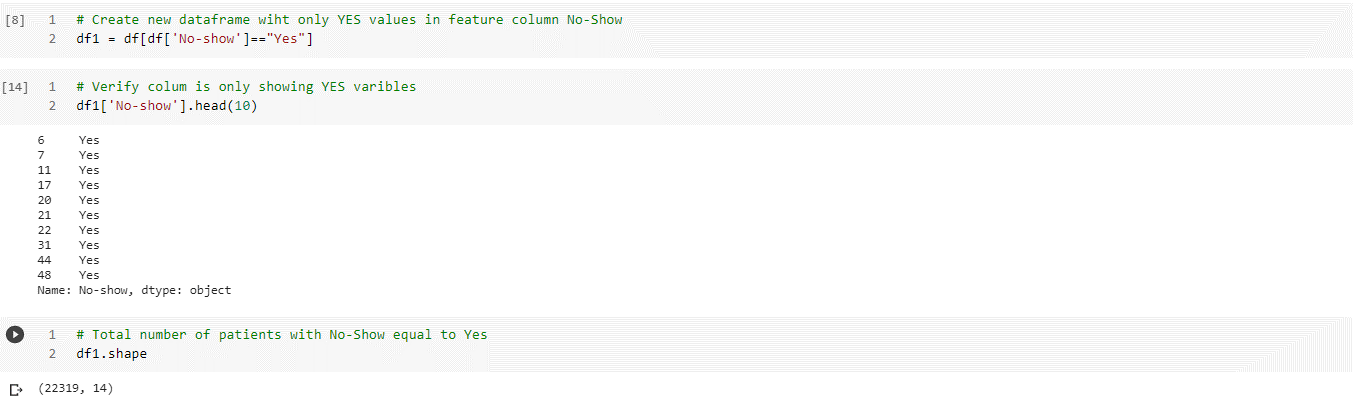
For this course project assignment, I started by importing all my library such as pandas and numpy I then created a dataframe label *df* and imported the ***no-showappointments.csv*** file. After this I verified my data using the head function. I then created a new dataframe labeled *df1* to filter out the “**No-show”** column values based on the values of yes. Then I verified the column using the head function from the new dataframe. Then I used the *shape* function to get the total number of patients with no shows equal to yes. I then used to ***Value\_counts()*** function to drop any null values and to get an accurate count of the “**No-show”** values based on no and output the dataframe to a new dataset file called ***no\_show\_dataset.csv***.

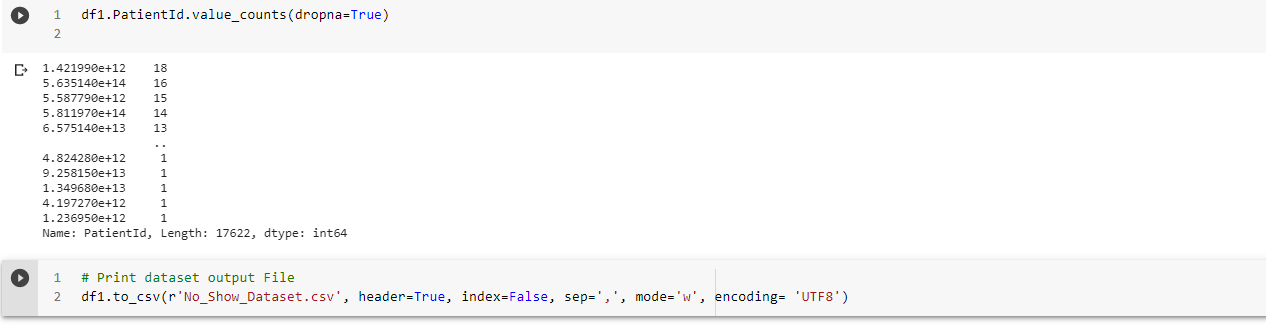
All files for the project are included with my work including a link to my Google Collab working script.

<https://colab.research.google.com/drive/1P_ecnbxSOWuFYE-SrOIAfI0GzDo7eGYc?usp=sharing>







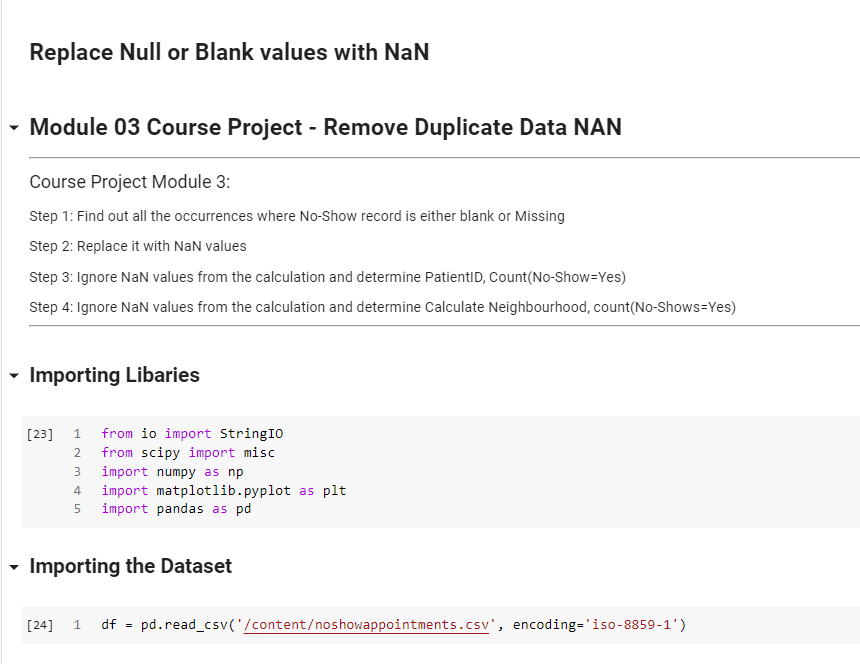


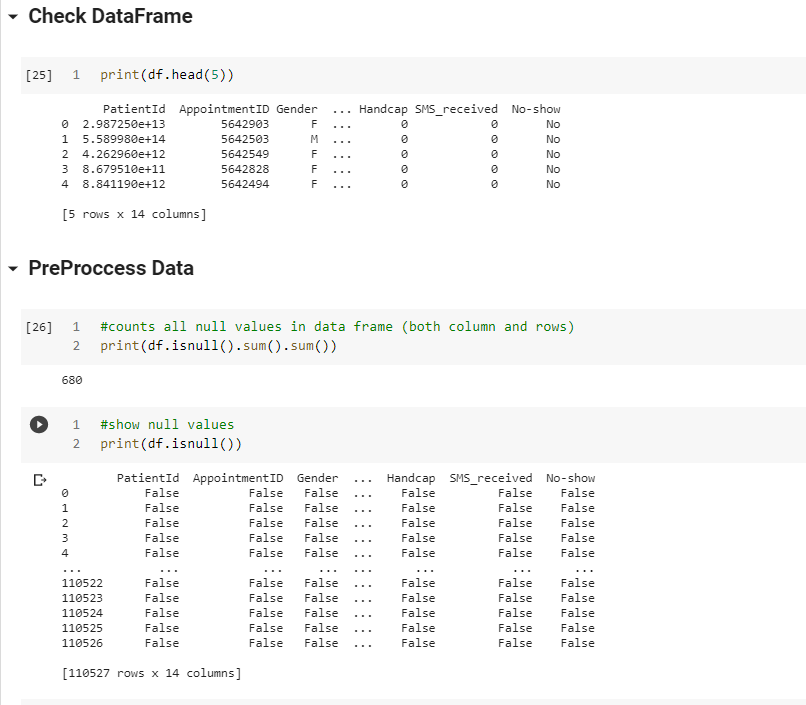
**Course Project – Values**

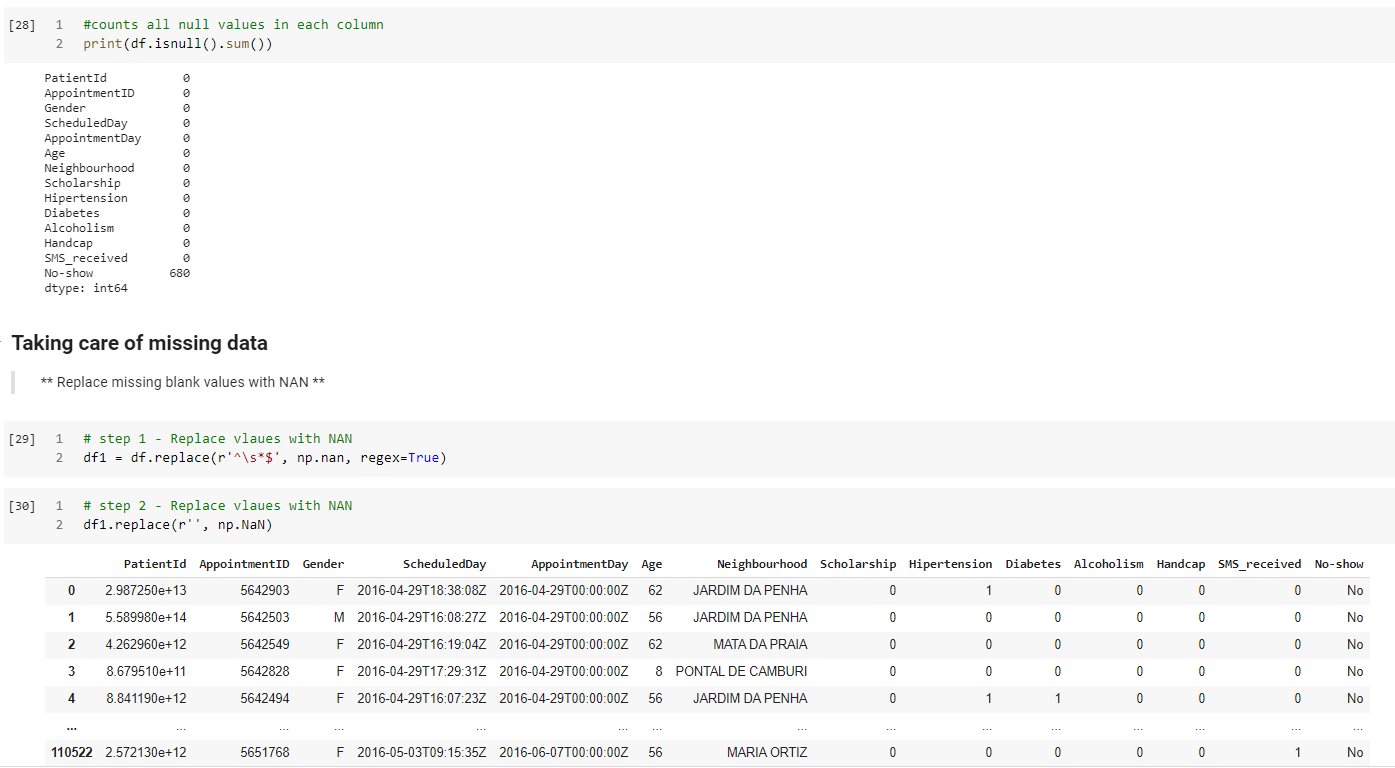
For this course project I first important my python libraries and then imported my data set ***noshowappointments.csv***. After importing the data, I pre-process the data by checking the data frame to verify that they that integrity was good. Then I processed the Null data variables by checking the null values determining which columns had no values and listing those no values in the dataset with descriptive statistics. Then I targeted the *Null* values and use several methods to replace the Null values with NaN. I then proceeded to verify the data replacements took place and then created a new dataframe ***df1*** and ***df2*** targeting the *no-show* column setting all of its variables removing any Null values. I've then verified the output of that data and proceeded to Ignore NaN values from the calculation and determine PatientID, Count(No-Show=Yes) and Ignore NaN values from the calculation and determine Calculate Neighbourhood, count(No-Shows=Yes)..

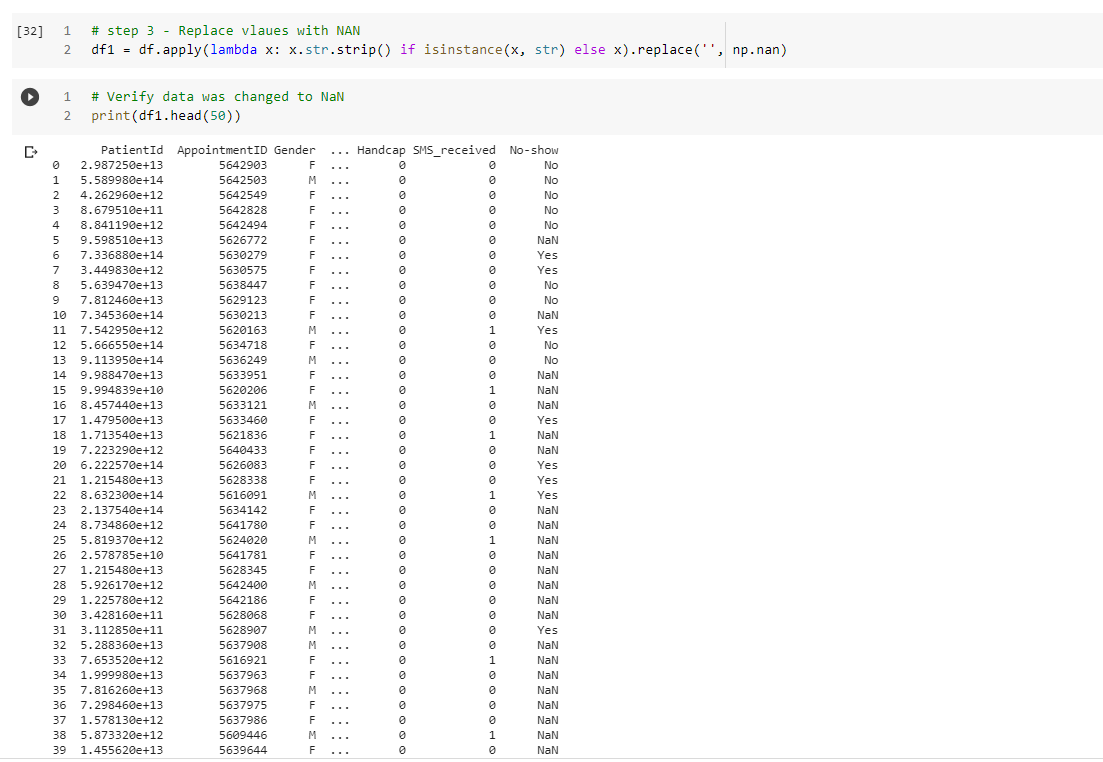
I have included a full jupyter notebook file of my python script and output as well as provided a link below to the Google colab file with my work as follows.

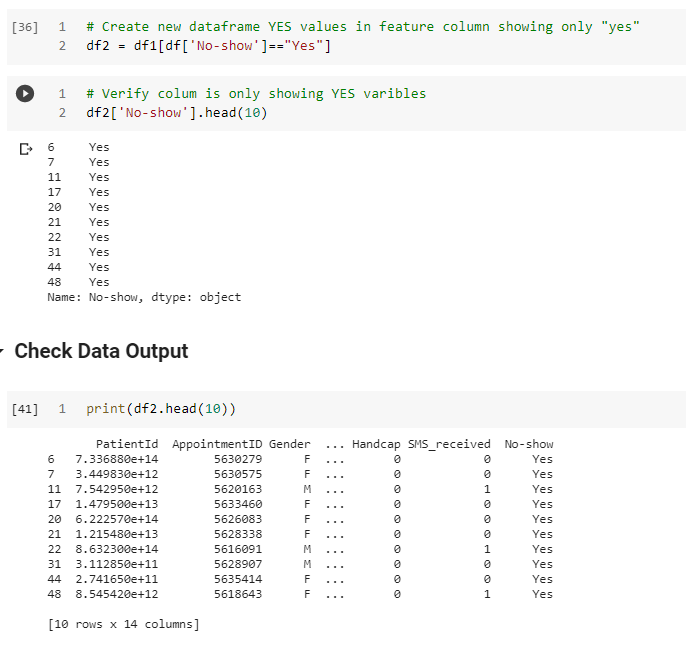
<https://colab.research.google.com/drive/1GkqjOVahCfsn69bBD8w9yVlK9FlE_ohQ?usp=sharing>

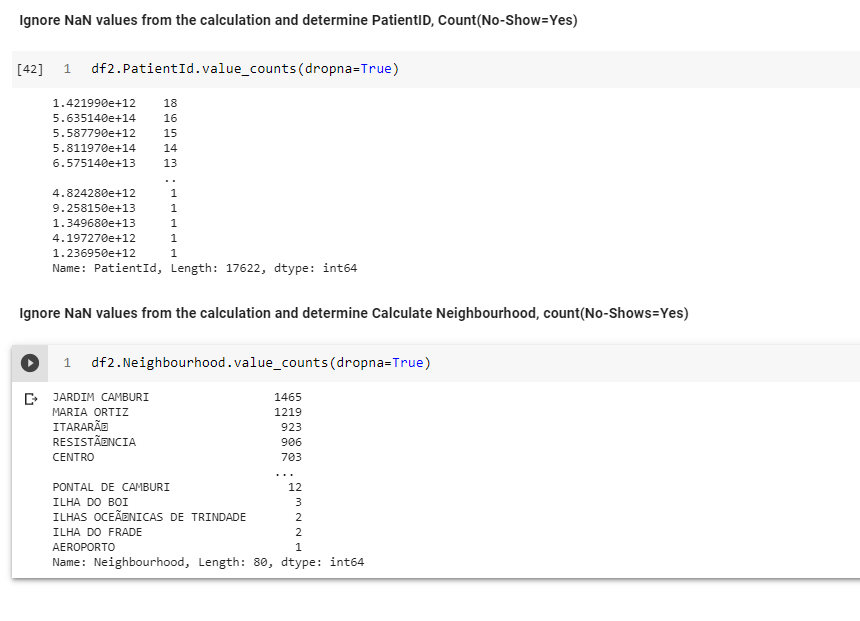










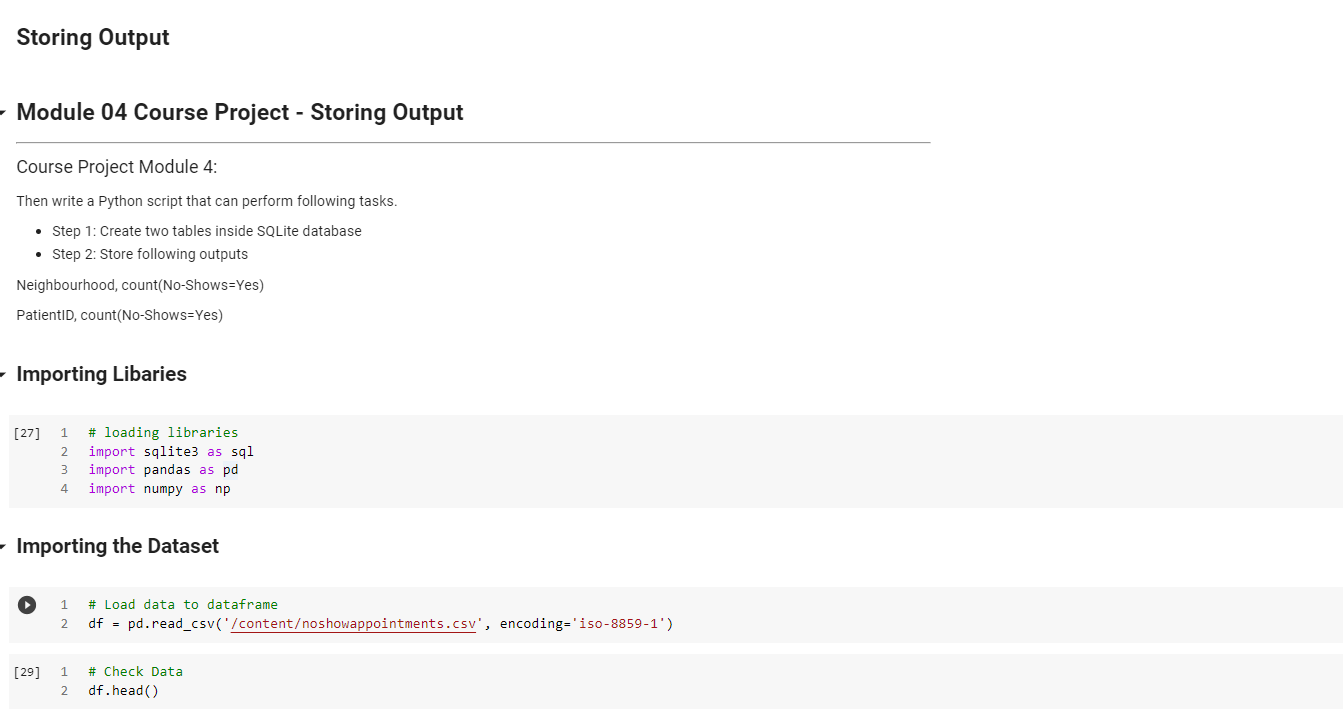


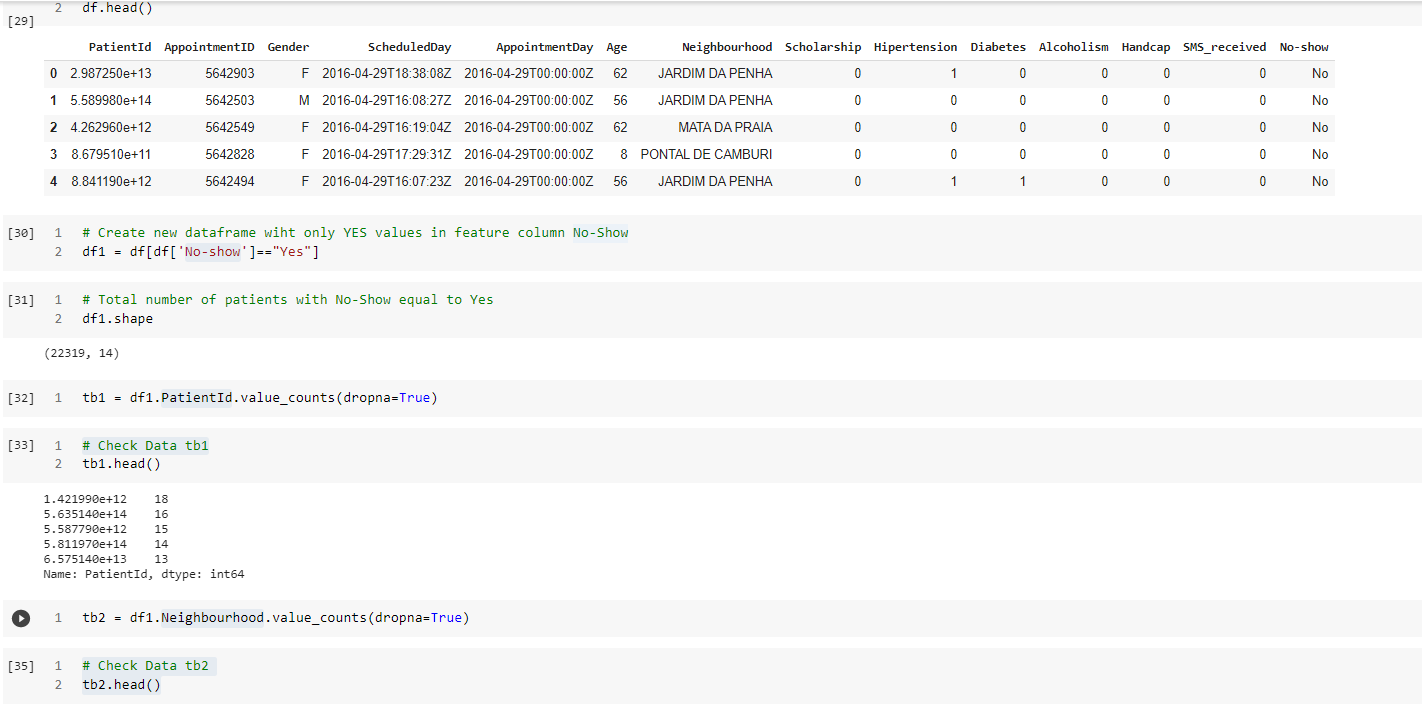
**Course Project – Storing Output**

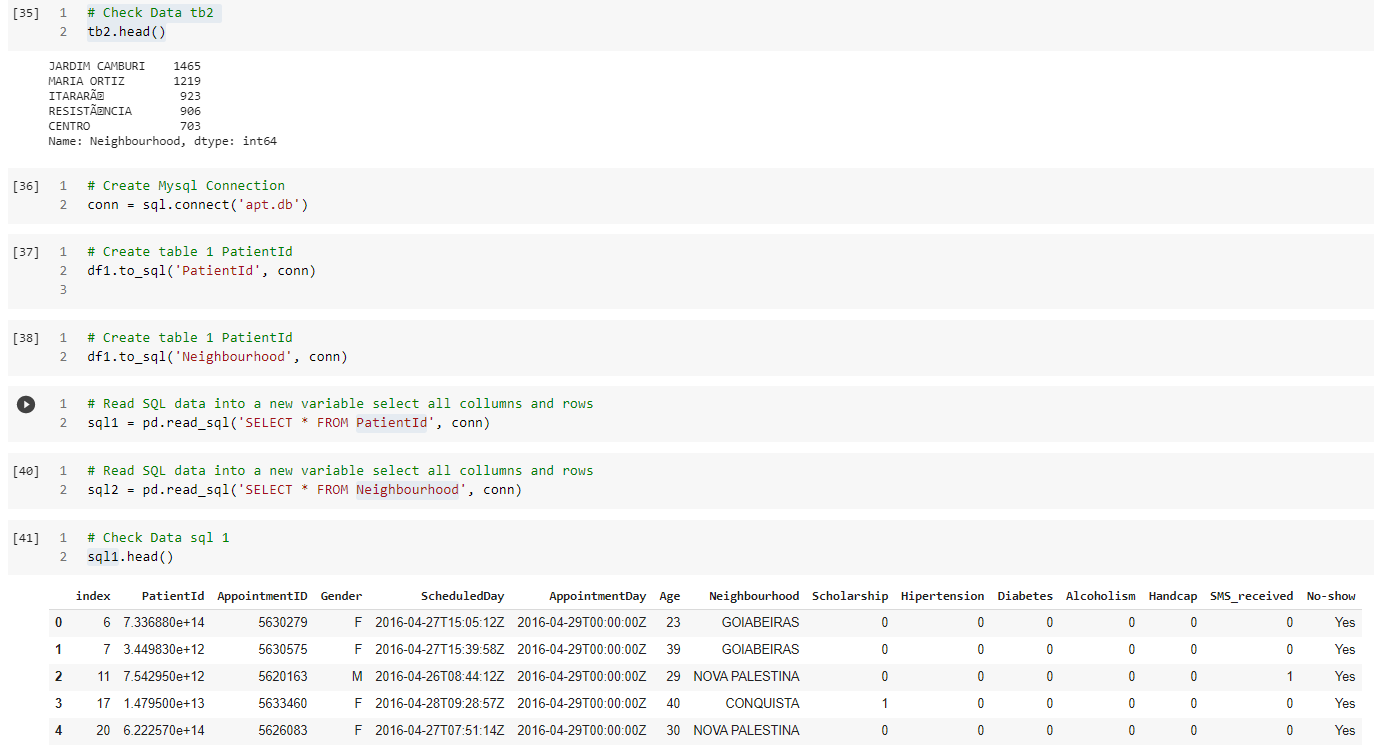
This project I loaded my python libraries, bloated the no-show CSV file into a pandas dataframe, and then check the validity of the data. After this, I created a new data frame *df1* to organize the no-show data. I then use the shape method to get the total number of patients with no shows equal to yes. I then created another variable call *tb1* for the ***PatientId*** and *tb2* for ***Neighbourhood*** value counts. I then used the head method to validate the data for both tables. I then proceeded to create a new connection database called ***apt.db***. I then created 2-new tables in sqlite database for ***PatientId*** and ***Neighbourhood*** value counts of *yes*. I then validated the data in each of these new SQL tables. I then use the ***concat*** method to join the tables for easy visualization followed by testing the output with the shape and head method followed by outputting the Joint tables to a new CSV file called ***NoShowOutput\_mod4.csv***.

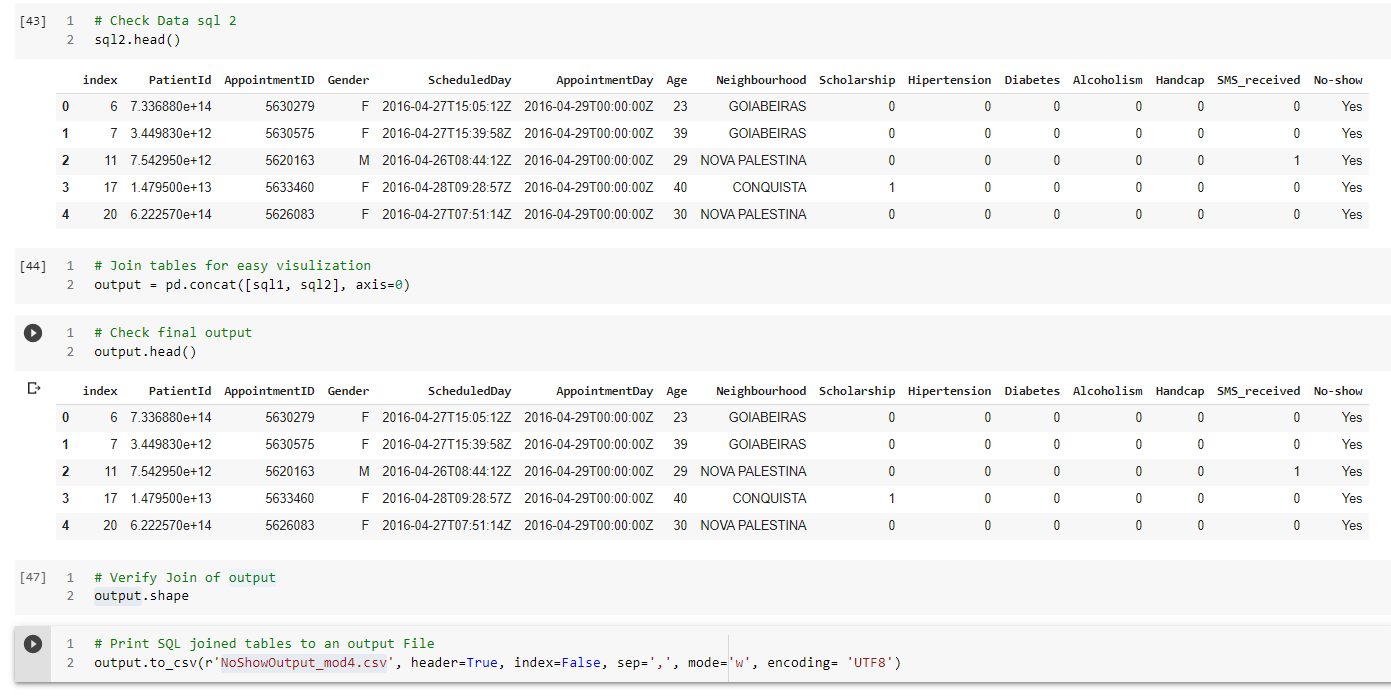
I have included all project files as required for this course project as well as a working URL to the Google code cloud script.

<https://colab.research.google.com/drive/1zq_7SEbJHDa8sp8odvecFzDDC6ktuVnw?usp=sharing>









**Course Project – Data Visualization**

For this course project I first started out by loading my libraries I loaded some extra libraries print PDFs and to customize the RC parameters the font text to fix any errors from the font symbols missing in the dataset that I was experiencing. Then I imported my data into a dataframe ***df*** then checked the validity. Next, I started to pre-process the data to a new dataframe ***df1*** to remove all the NaN values from the *no-show* column vector. After this I checked the data, continued to clean the data, and drop NaN values. I knew I wanted trendline in my plot so I encoded the ***neighborhood*** column vector and a sub variable with the neighborhood value count in two separate variables for my X & Y features. From this point I continue to process the visual data, set my trend lines, grid, x-y-labels, and my visual data point including a supporting PDF output of the chart.

I have included all the necessary requirements and files for this project also I have provided a link to the cloud Google colab working script file.

<https://colab.research.google.com/drive/1yIz-qs9D-IiyhtOuZeXFIWnvWaWG9Cdu?usp=sharing>

