

**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

Persons in cultural employment per million inhabitants: Number of people who have a cultural or non- cultural occupation in cultural ventures or who have a cultural occupation in a non-cultural industry communicated as million occupants matured 15 years and over.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

We imported pandas, numpy, scipy, seaborn, missngno, matplotlib.pyplot and plotnine

To assist with our data analysis.

* We use count() to determine the number of missing points in all the columns.
* We used sum() to add all the missing values so we can find the percentage of data meaning throughout the data frame.
* From missingno we used matrix to plot (visualize) our missing data
* We used drop() to delete columns that had a lot of missing data namely:

1. Flag codes with 33814 out of 36799
2. Flags with 33814 out of 36799

* We used countplot() and plot from matplotlib.pyplot to plot out graph
* We used geom\_point(), ggplot() to visualize our data

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

We found that the missing data takes at least 21% of the dataset which is pretty much a lot. The missing values come from the following columns:

* Flag codes with 33814
* Flags with 33814
* Value with 2961,

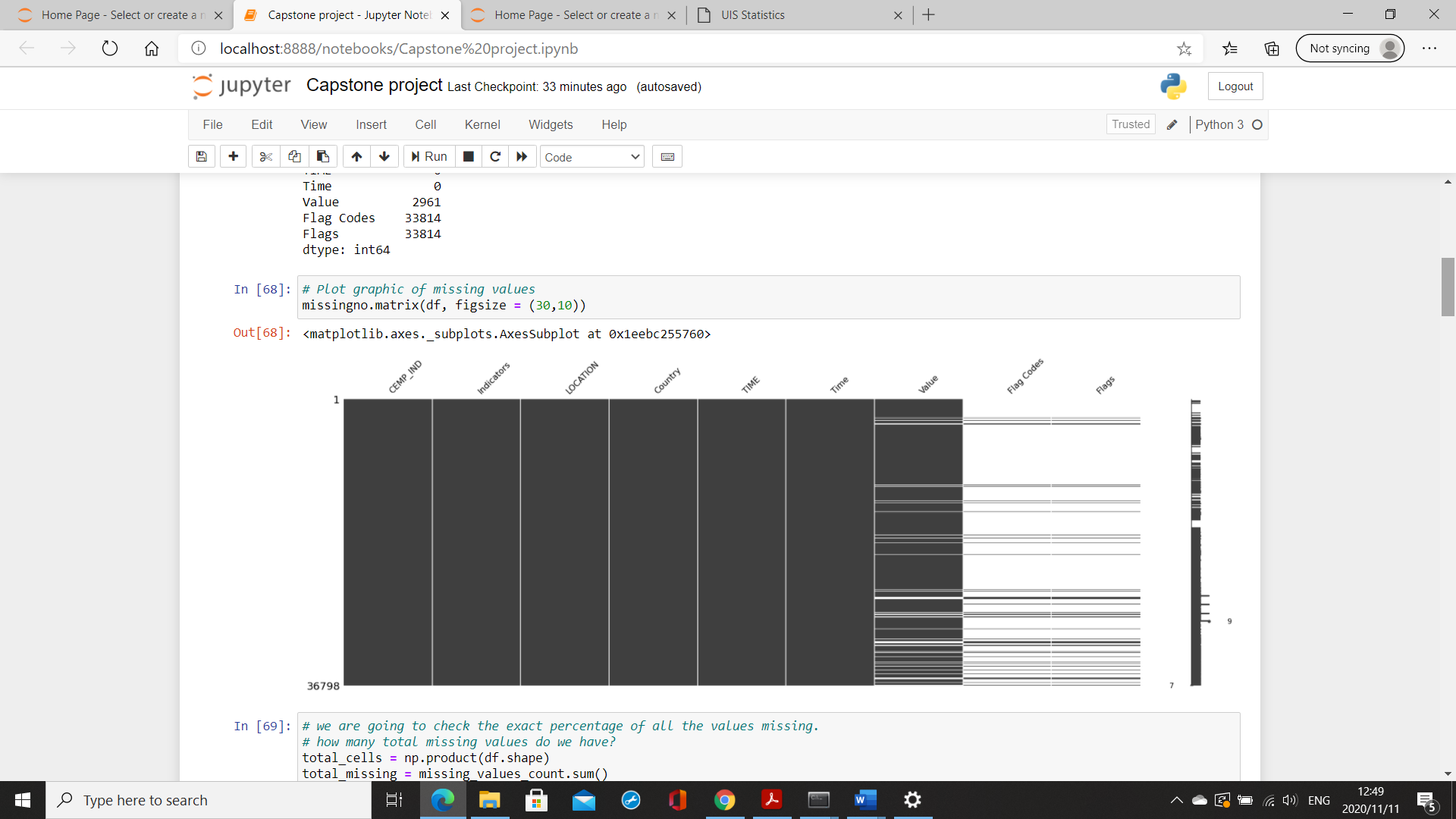
these missing values of the mentioned columns take up majority of each column except the Value and thus deciding to drop those columns because they don’t add value.

* All these missing values in the dataset where Identified as NaN.

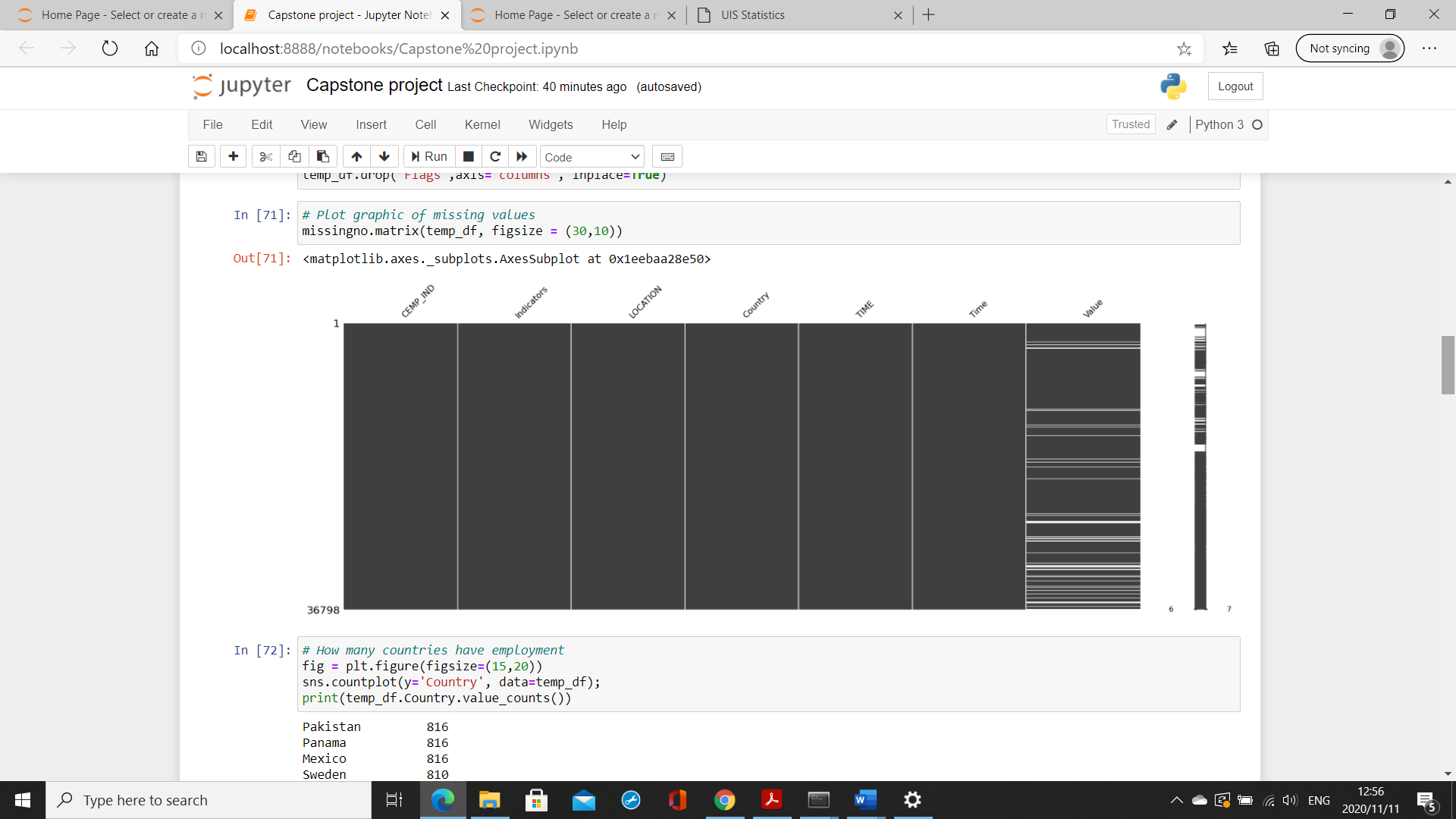
DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

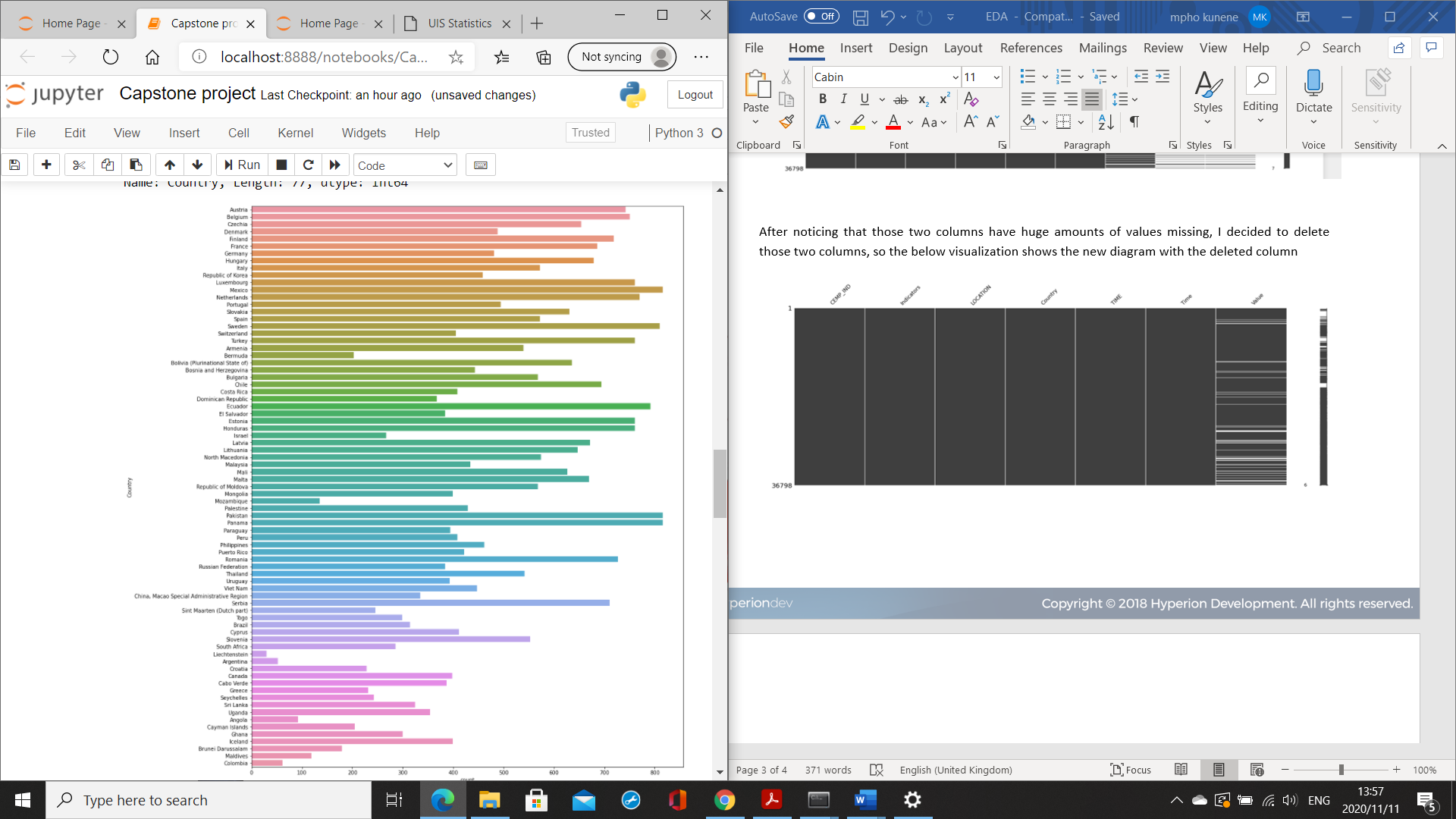
Judging from the below visualization the diagram shows that there is a lot of data missing In Flag Codes and Flags



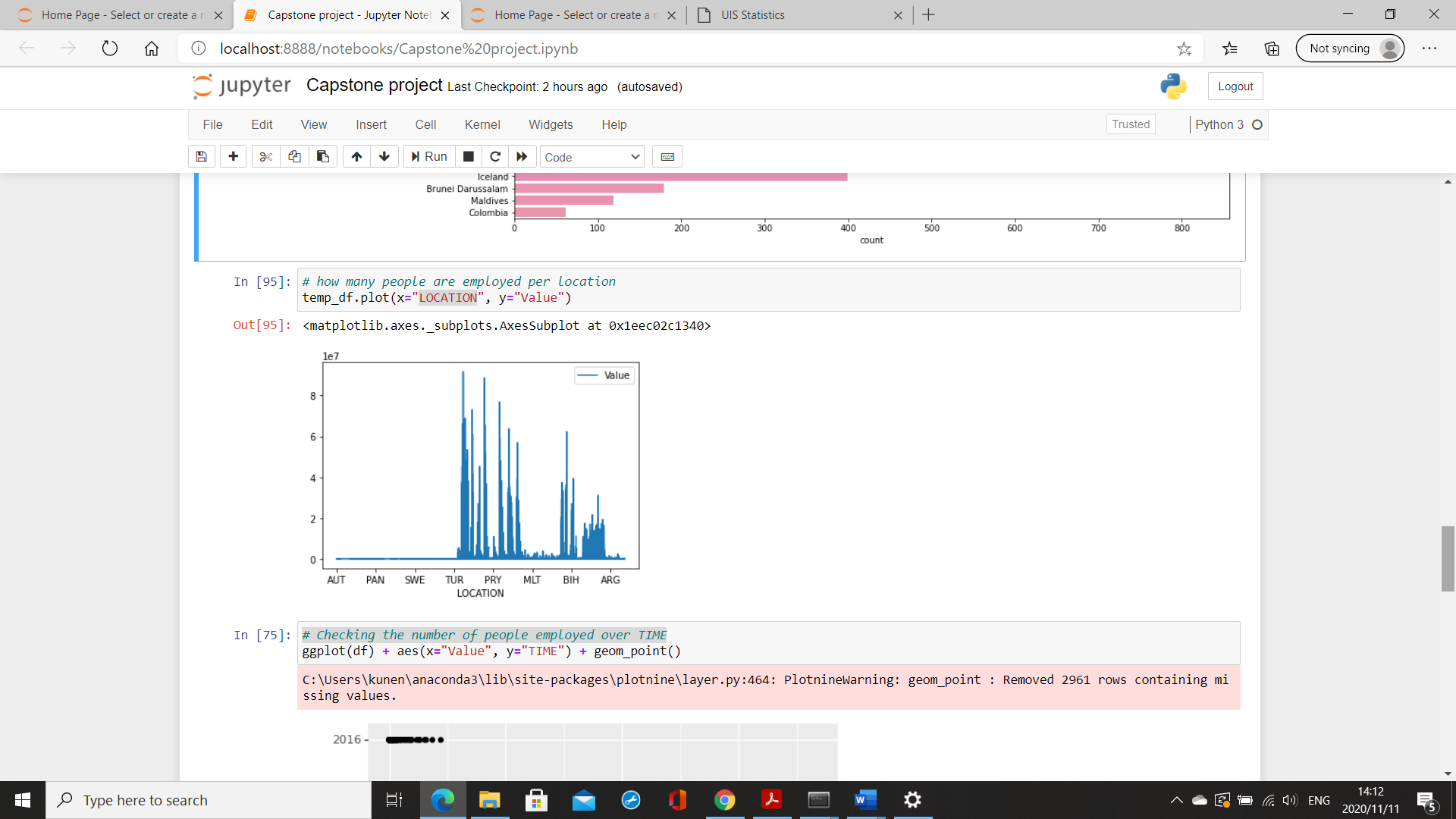
After noticing that those two columns have huge amounts of values missing, I decided to delete those two columns, so the below visualization shows the new diagram with the deleted column



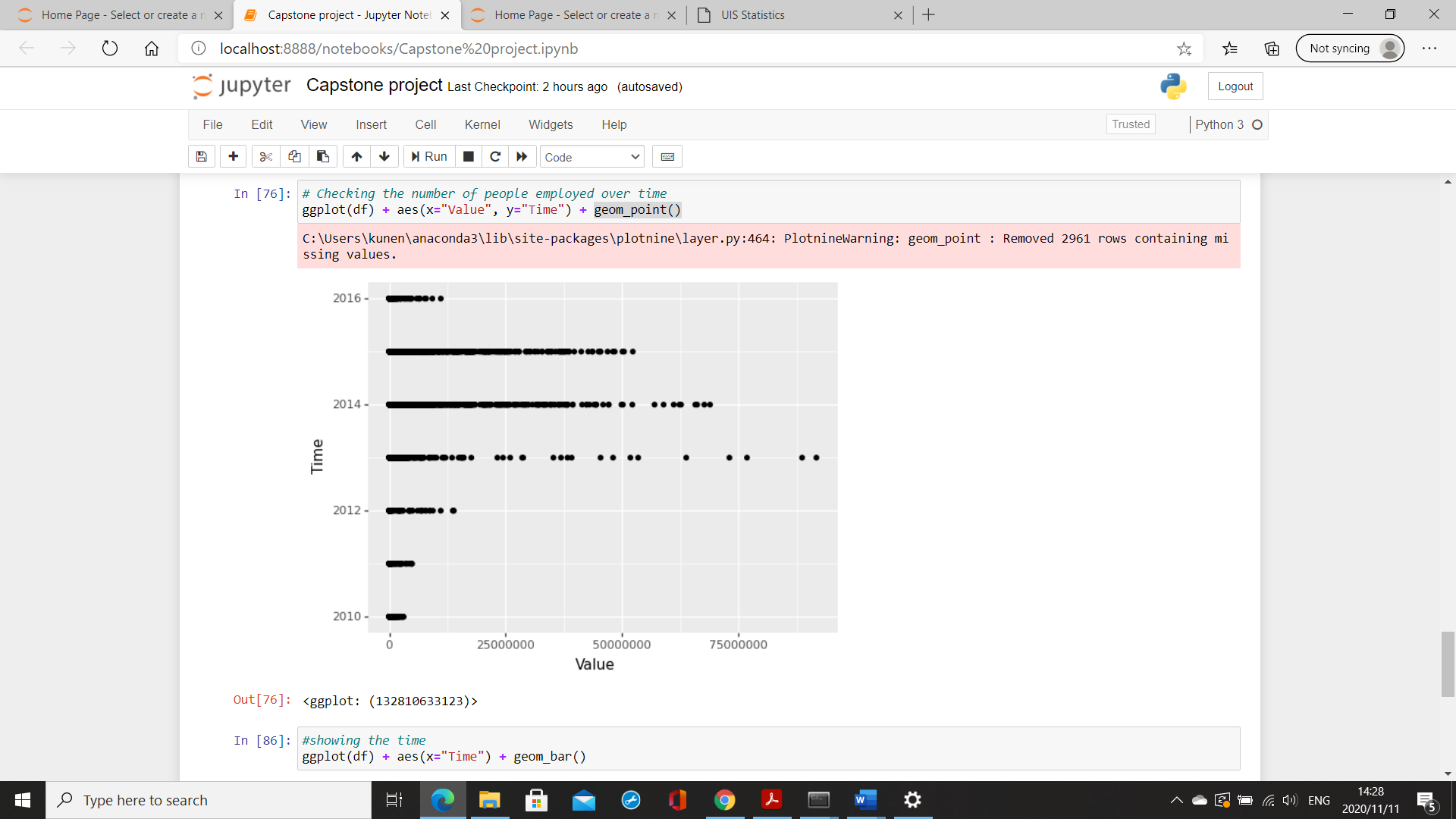
The below diagram shows employment rate in these countries, we see that Pakistan, Panama and Mexico have the highest rate of employment by 816 while the lowest is Liechtenstein by 30.



We decided to plot a diagram in which the employment is based on the location and the value to see how many people from that location have been employed and as we can see that from the first three locations namely AUT, PAN, SWE their employment rate is at 0 after that it increases rapidly on TUR and slowly drops even on PRY but it picks up again and goes sky high but not passing the TUR employment rate. Following the diagram we that MLT and ARG stay under 2 meaning they also have the lowest employment rate.

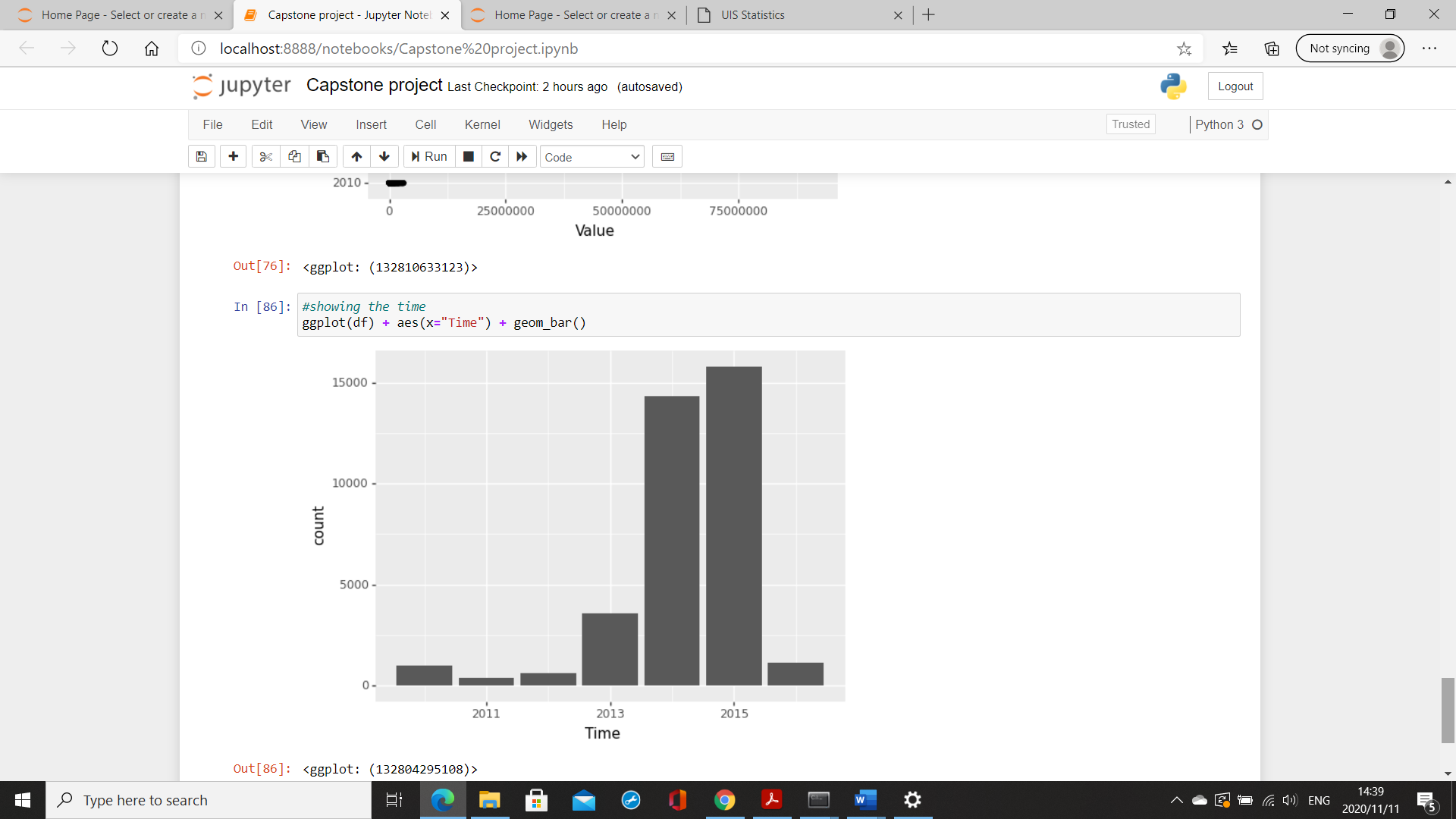


In the below diagram we are checking the rate of employment over time since there is two columns with time we decided to compare those two to see if there is any difference and we have found that there is no difference, its safe to assume that both diagrams are the same meaning “Time” and “TIME” have the same data

 Graphical user interface, text, application, email

Description automatically generated

Here we are trying to see in which year was the employment rate the highest and we can see that 2015 had the highest number of people employed compared to others but towards the end of the year the employment rate went rapidly down following 2013 which started off a bit slow but quickly raised to the occasion but 2011 was just the worst because the employment rate year was just down and it didn’t pick up at all till mid 2013



# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Mpho Kunene**

