DATA CLEANING

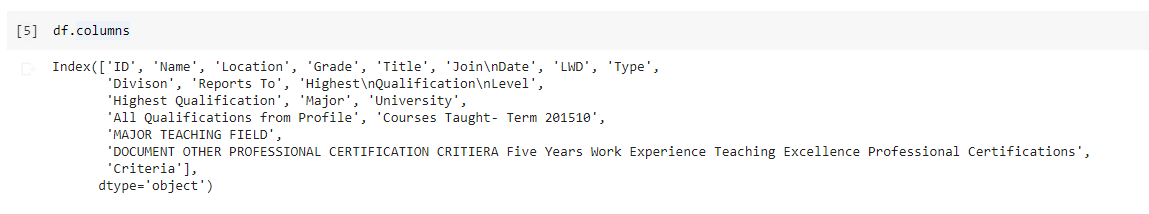
As a data scientist we are going to encounter with poor quality/incomplete datasets. To generate high-quality analysis we need to ensure that we are cleaning data in order to accurately represent the dataset. Pandas offer a diverse range of built-in functions that can be used to clean and manipulate datasets prior to analysis. It can allow you to drop incomplete rows and columns, fill missing values and improve the readability of the dataset through category renaming.

Once you get the data, we need to do the below step before we proceed to training:

**STEP 1: Print columns**

Printing the columns give an insight to the data. We can rename the column appropriately, like we can shorten the name of the column if the column name is very big. For example we can change the column name “DOCUMENT OTHER PROFESSIONAL CERTIFICATION CRITIERA Five Years Work Experience Teaching Excellence Professional Certifications” to “CertificationInLastFiveYear”.

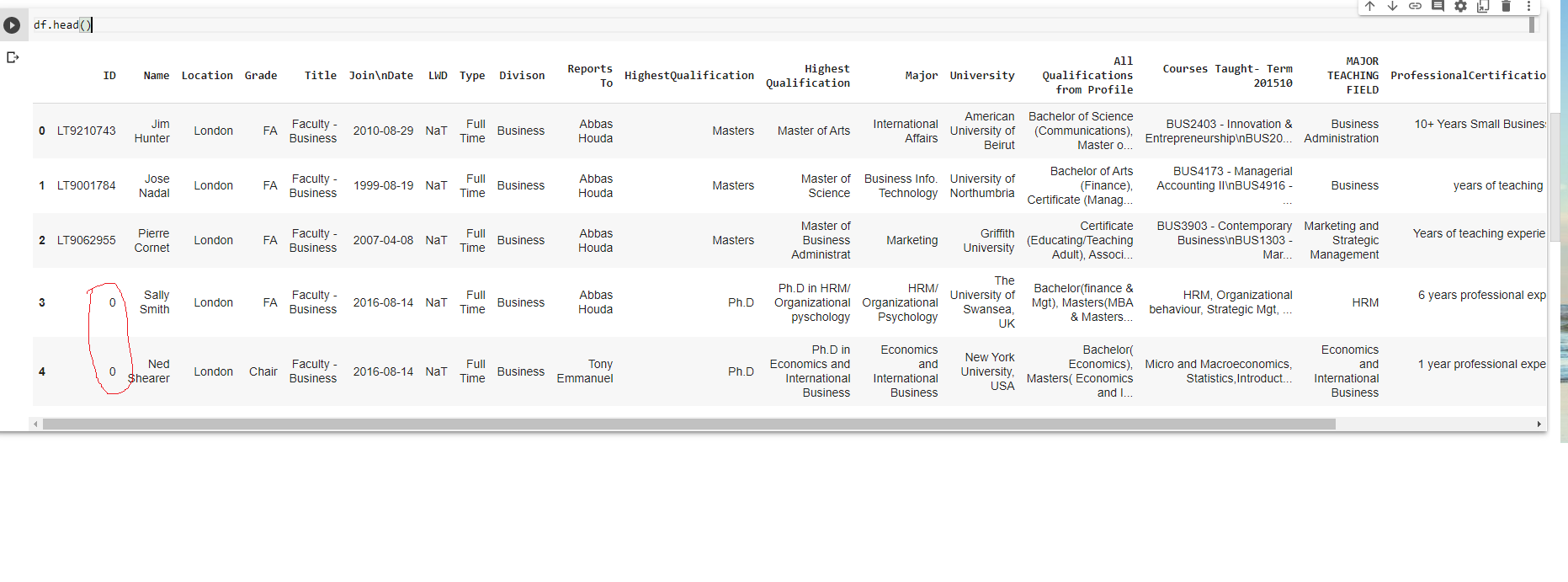
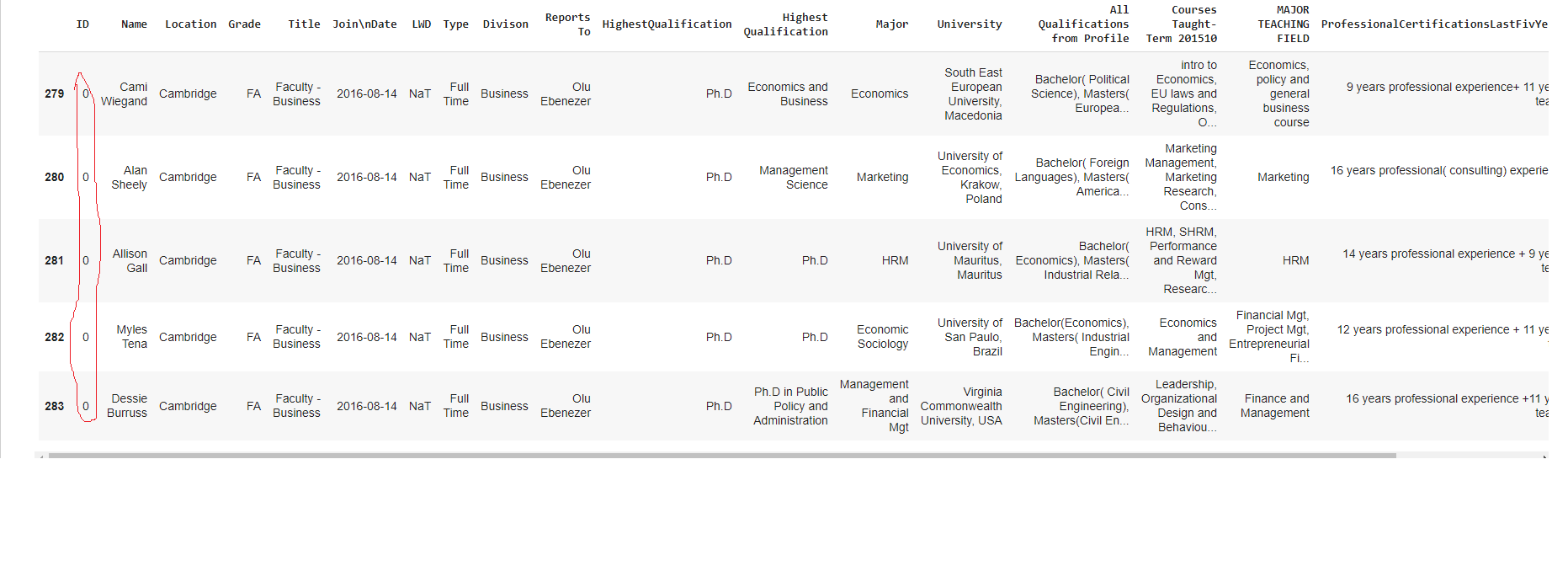
Also we can remove whitespaces from column name. Like Highest\nQualification\nLevel has 2 ‘\n’, which we can remove.



**STEP 2: Head and Tail of dataset**

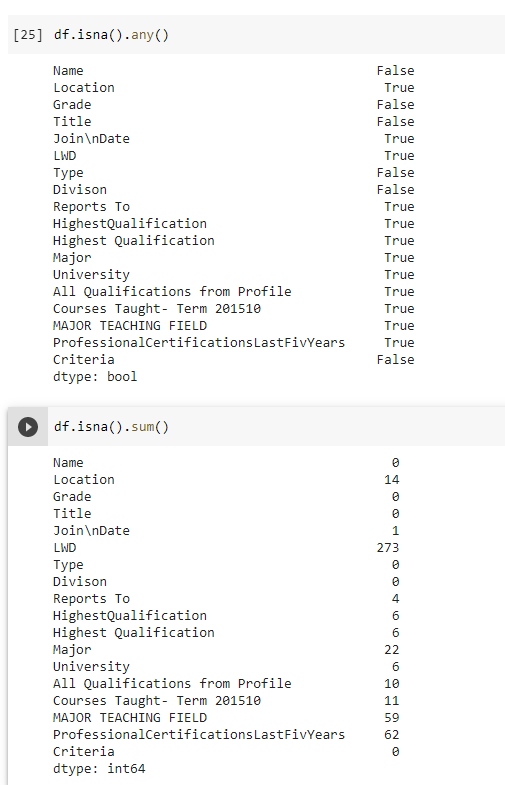
Printing head and tail of data set help you understand the data, which column to retain, which column need a transformation, which to remove etc.

For the given data set we can see ID is mostly coming zero. So we can ignore this column.

****

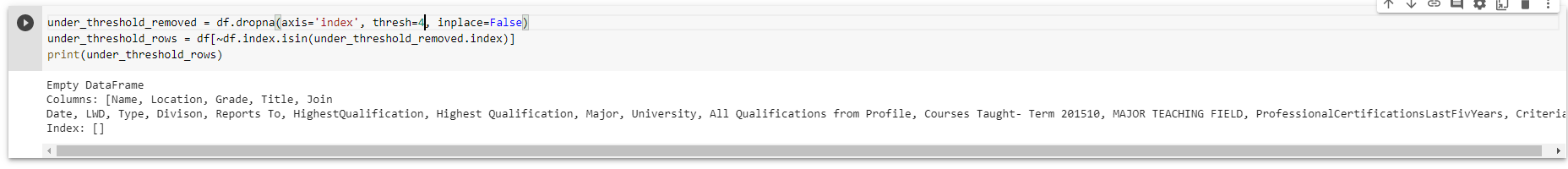
**STEP 3: Identifying whether columns has null values and getting the count of null values**

df.isna().any() will return Boolean saying whether a data frame contain null values. And df.isna().sum()will give the sum of nulls.



**STEP 4: Dropping rows if most of the columns are null**

We can check see if there are any rows that do not contain enough usable elements. If we want to remove rows that have less than four elements by using .dropna(thresh=4) as below:



Here no such rows are there.

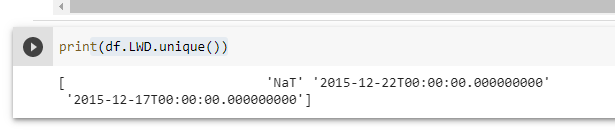
## STEP 5: Filling Missing Values

## In certain circumstances, we may want to retain rows that contain missing values and instead give them a default value when missing

## For example in the above data set LWD has 273 null values.

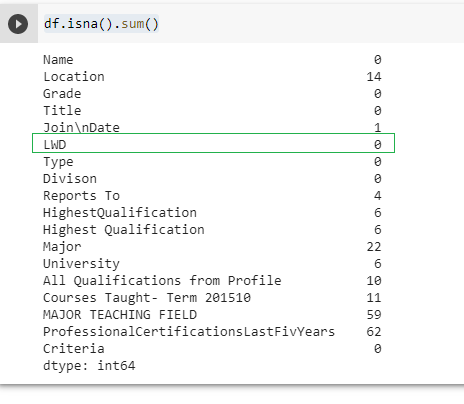
****

And the distinct value of that columns are:



Here we can fill NaT for missing values using .fillna(value=None, method=None, axis=None, inplace=False, limit=None, downcast=None)

After performing df.fillna(value={'LWD': 'NaT'}, inplace=True)if we perform df.isna().sum(), there will zero null values for the column LWD.



## STEP 6: Improving readability

## The stakeholders receiving the data may not understand the meaning of certain data. To improve the readability of the analysis we can rename the column value to something more meaningful.

## For example here Grade has 2 values. ‘FA’ and ‘Chair’.

## 

## Here FA is Faculty and ‘Chair’ is Chairman. We can replace this as below.

## 

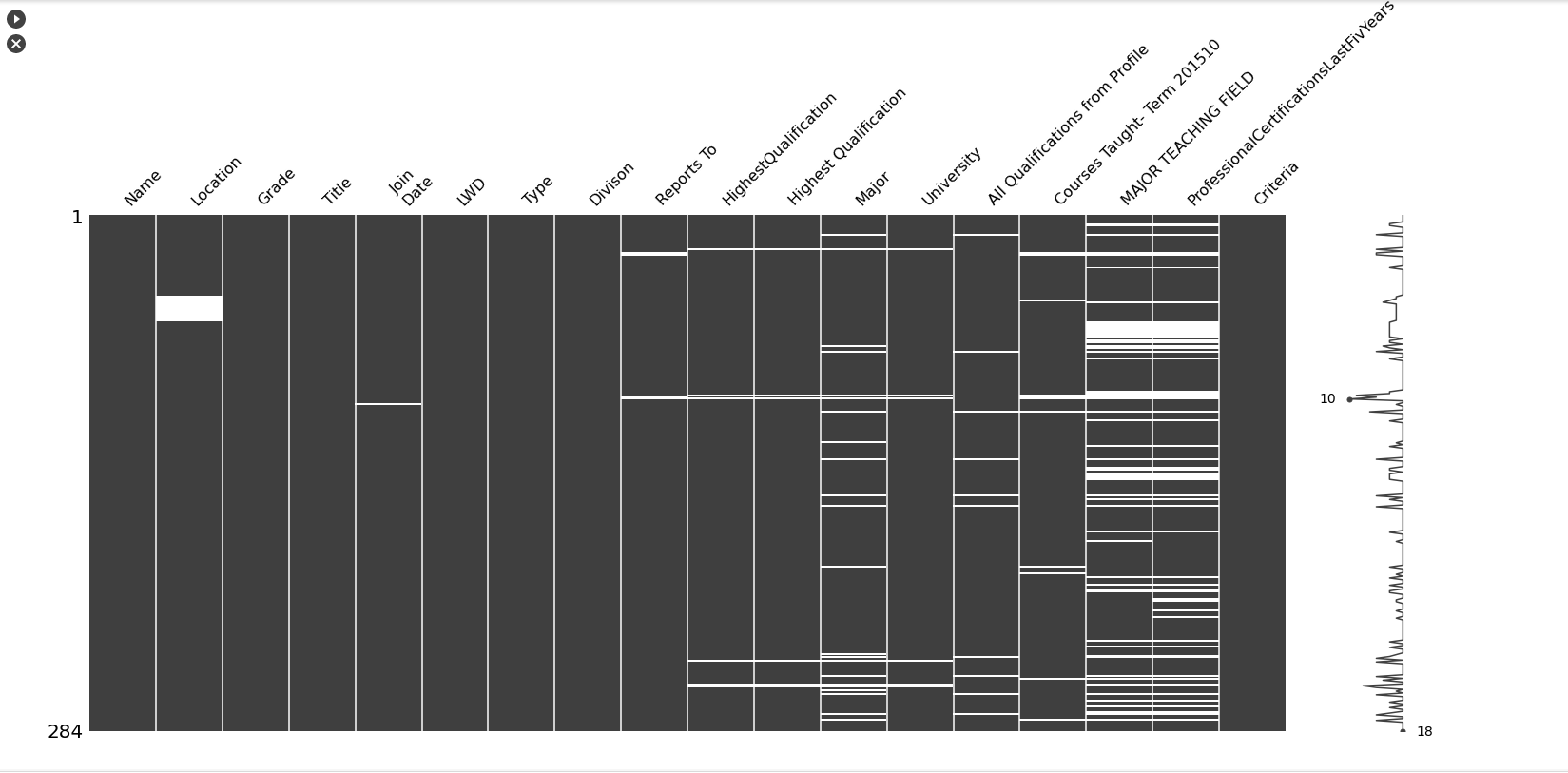
## STEP 7: Data Visualization

## Missingno is library which will provide the missing values in the dataset by informative visualizations.

import missingno as msno

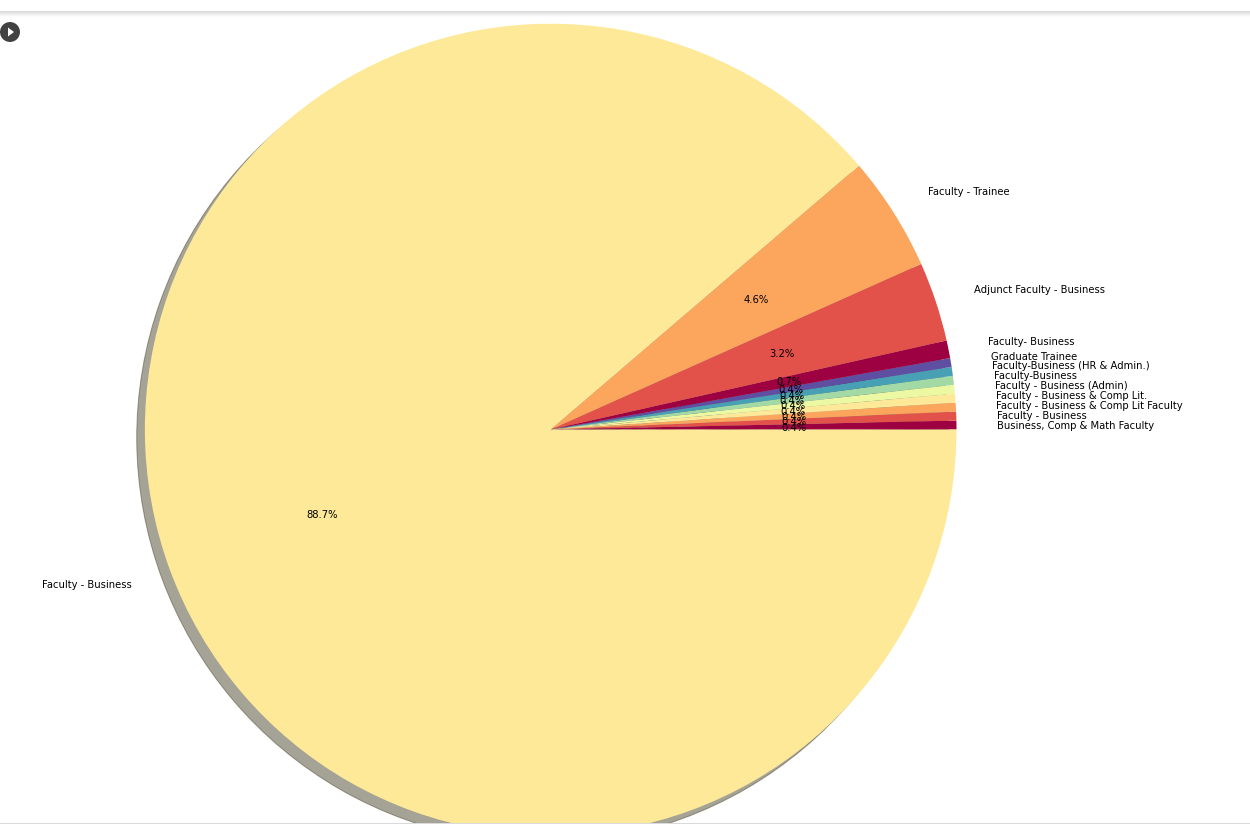
msno.matrix(df)

The above command gives us visualization of missing values.



We can draw bar chart (msno.bar(df)) and heat map (msno.heatmap(df)) of missing values as well.

Also we can use Pie chart to understand which how many faculties each department has. From the image it is clear that 88.7 faculties are for Business and 4.6 for Trainee.

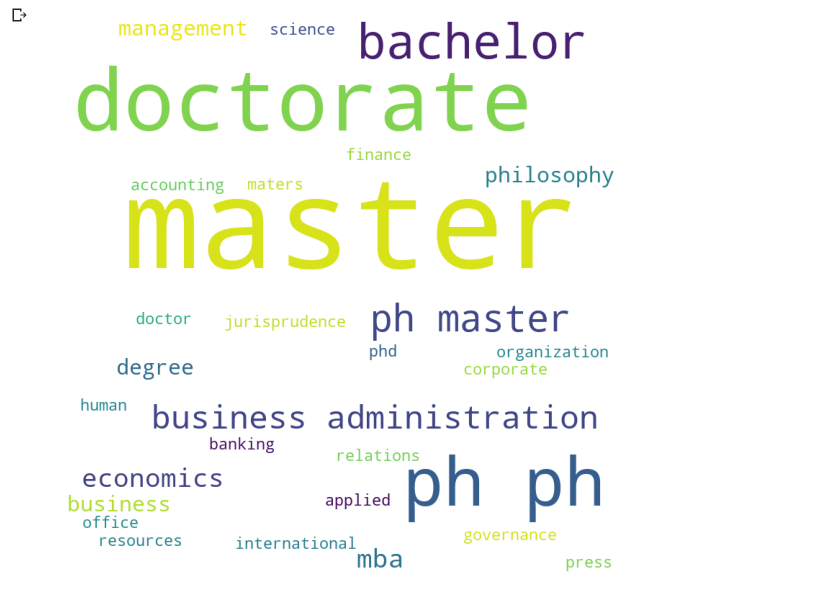


Apart from that we can visualize data, especially integer data using Bar Chart, Line Chart, scatter chart etc. Also we can use Charity API to visualize data.

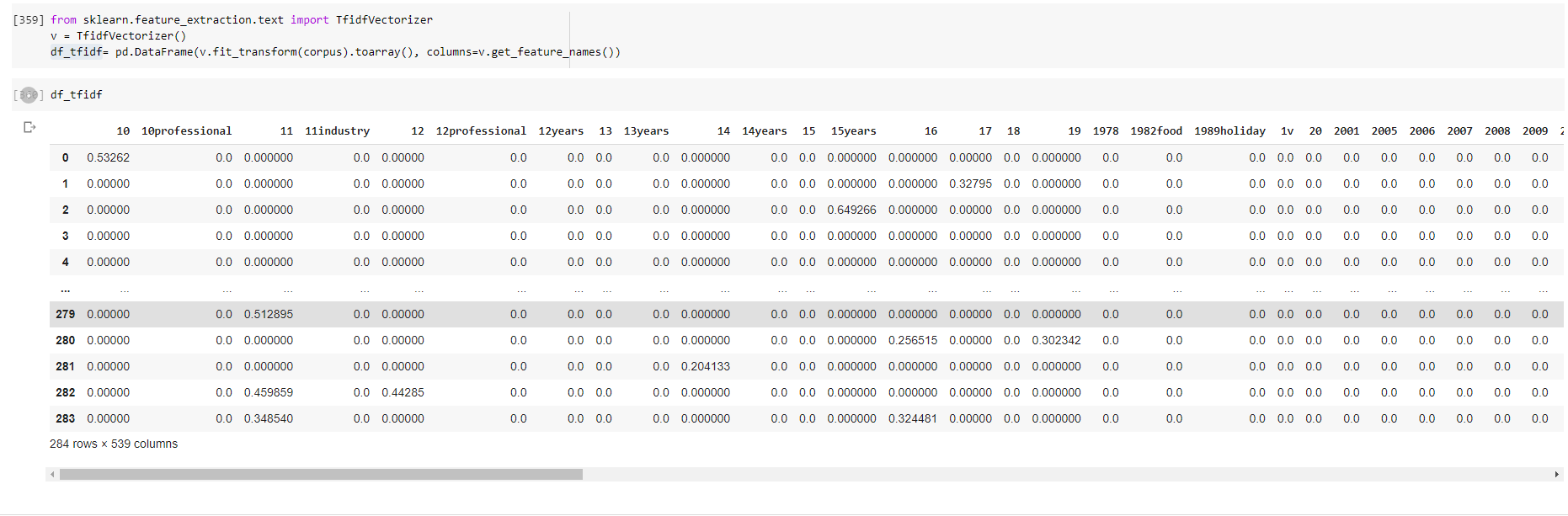
## STEP 8: Analyzing text data

## There are a lot of libraries for analyzing text data.

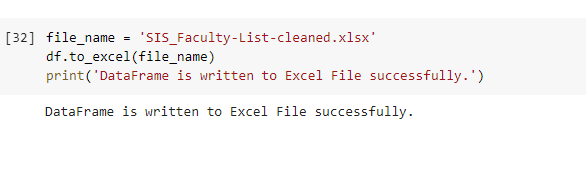
We can use wordcloud library to display important keywords of any data frame. For example word cloud for Highest Qualification level looks like below. We can use NLTK stopwords library to stop displaying English stop words and we can extend the same as well.



Also we can vectorize the text data into numerical data as most of the algorithm works for numerical data. There are a lot of vectorization methods today. TF-IDF (Term Frequency – Inverse Document Frequency) and Word2Vec are 2 of the popular methods we use today.



**Exporting cleaned Data Frame to xlsx**



**Summary**

In data science identifying and cleaning data is very important. For that we need to print and visualize the data. There are plenty of libraries present in python to clean the data and here we did the some of them with few lines of code and generated cleaned xlsx ☺