# Exploratory Data Analysis

## Introduction

To find a dataset for this task I read though Jupiter’s UK Growth Fund Annual Report and Accounts, and found a company to research, cleanse their data and provide valuable insights. I decided to take the first investment in the portfolio statement, Anglo American.

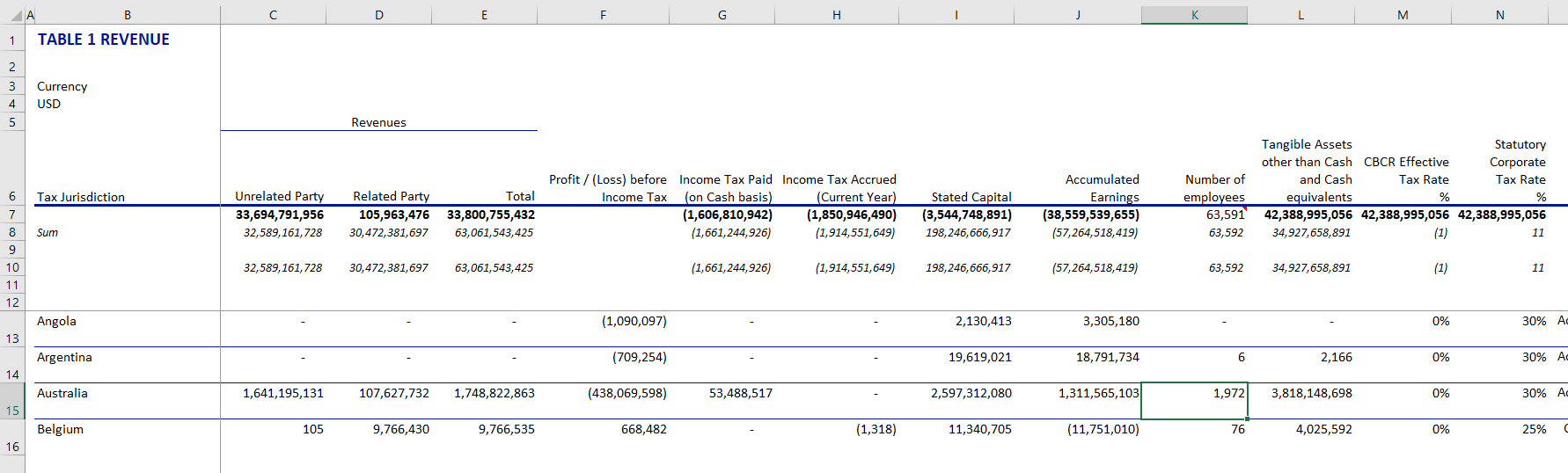
I found their country-by-country reporting data for 2021, 2020 and 2019 in .xlsx and xlsm format. Upon opening the 2021 file there was a lot of noise in the spreadsheet with some fields merged. Although we could make the necessary changes in excel, I ingested the raw file and cleansed the data in python. Doing so means we can repeat the process more easily for future reports and maintain an audit trail of changes if required.

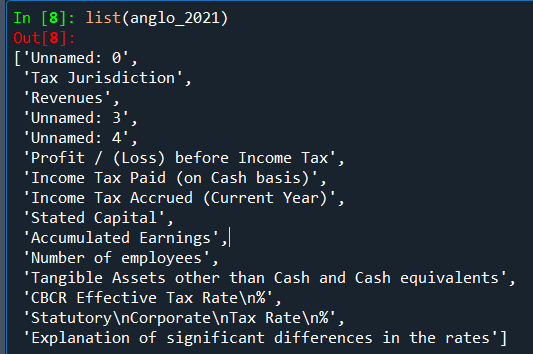
## Content

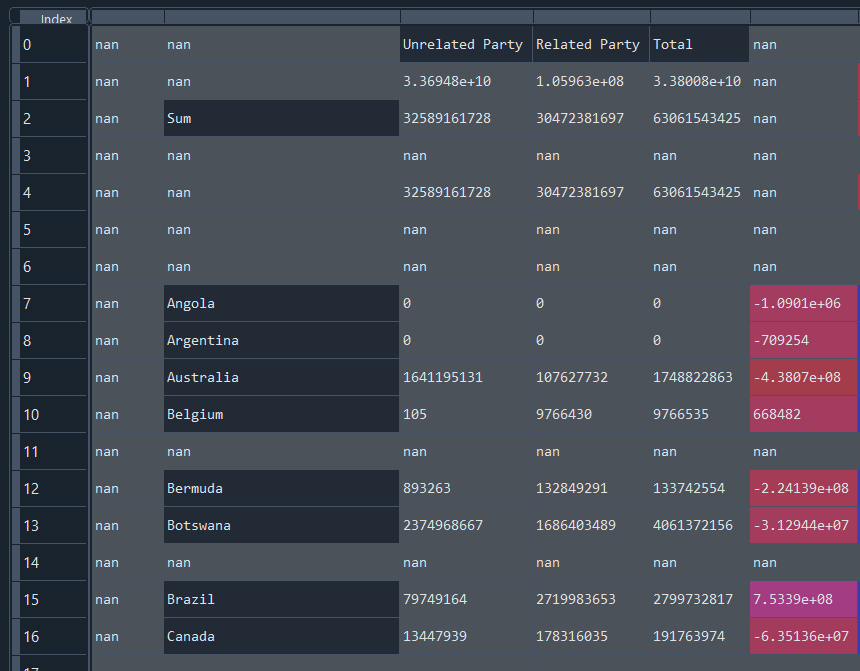
**Python script**: 20210122\_LF\_Task1\_Exploratory\_Data\_Analysis.py  
**Input files**: anglo-american-country-by-country-reporting-data-2020.xlsx, anglo-american-country-by-country-reporting-data-2019.xlsm, anglo-american-country-by-country-reporting-data-2018-01.xlsm   
**Output file**: 20220123\_anglo\_american\_analysis.xlsx

## Process

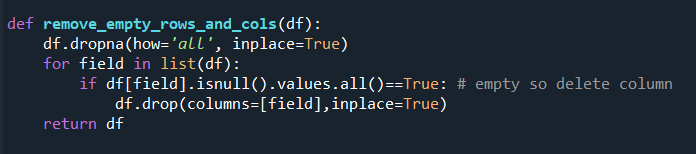
1. The headers first appear in row 5 so I ingested this sheet as a pandas df from this row.



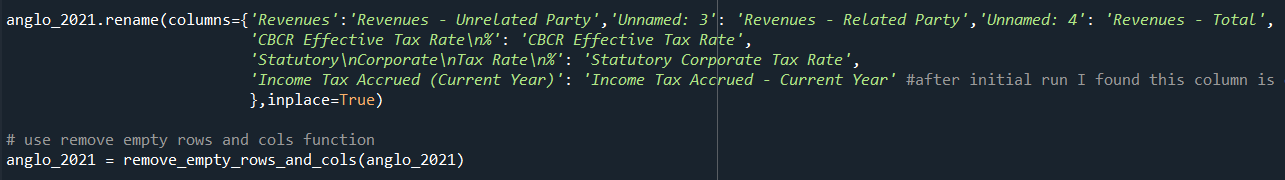
Columns 1, 4 and 5 are not coming through due to merged/hidden cells. Also, there are line breaks in some columns.

Opening the dataframe in the variable explorer we can see that the first row includes field names that make up the revenue. Also, the first column might be NULL. 

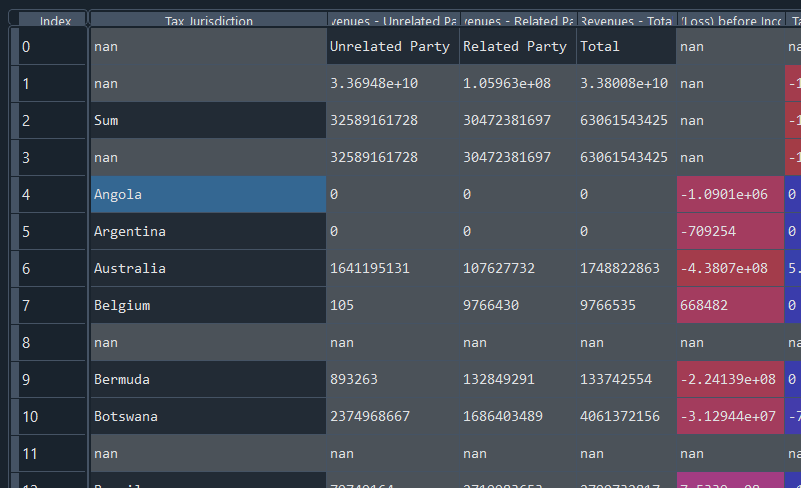
I first wrote a function to remove NULL columns and rows:



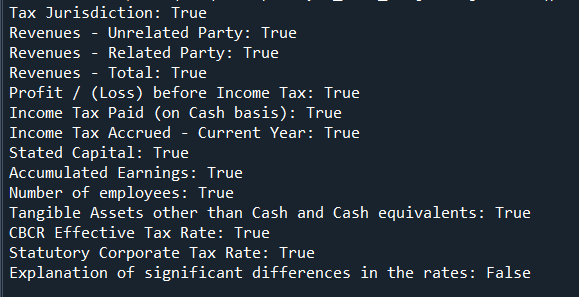
and renamed the fields:



Looking through the Tax Jurisdiction column we can see that data starts on the 5th row, the rows above are the summations. I saved these rows as a separate df in case we need them later then removed them from the existing one. Also, there appears to be more null rows (at index = 8 and 11) so we will investigate these.

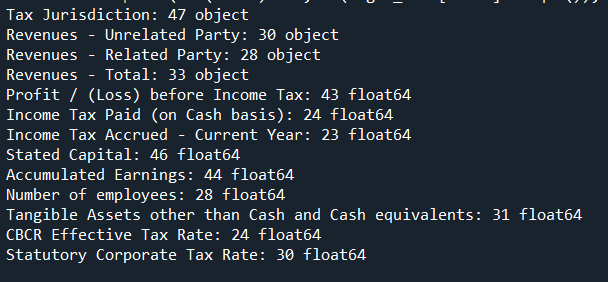


I filtered the dataframe for Tax Jurisdiction is NULL and then checked which columns had all NULL values for this subset. Only the last column has values populated.



This column contains explanatory values relating to the country above it. We will remove this column for now but save a copy of it as a dataframe along with the country code for the Key.

I then removed NULL rows again from the original df and checked whether the Tax Jurisdiction contained any NULL values which it did not. Finally, I checked the datatypes for each column and the number of unique values to spot any obscurities.



Nothing unusual here so we can move on to 2020.

Both the 2020 dataset and the 2019 dataset were cleansed using a similar logic as above. All three datasets are now in the same format, so I concatenated them together. If I was planning on ingesting this data into a database table I would export the data frame as a CSV file but in this instance I decided to export the data as an excel spreadsheet. I wrote some functions to format the data as an excel file and exported it.