# **Technical task course work**

Using datasets from the supplied reference sites, select import and manipulate more than one connected data set to support or demonstrate an interesting trend or facet of the data; build a presentation from this data, showing the business opportunity and/or challenge and how the data fails or supports this proposition.

# **1. Installation Guides**

Before starting, install the tools that will be used throughout the coursework. (please use your DTSQUARED email to apply or register for tool use where applicable)

## **Python**

A recommended python IDE can be the following.

PyCharm IDE: <https://www.jetbrains.com/help/pycharm/quick-start-guide.html>

Python: <https://www.python.org>

Google collab [Welcome to Colaboratory - Colaboratory (google.com)](https://colab.research.google.com/)

## **Microsoft SQL**

Use the following example to download Microsoft SQL server database;

<https://www.youtube.com/watch?v=QsXWszvjMBM>

<https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15>

Internal Snowflake logins

## **Power BI**

Download Microsoft Power BI from the following link: <https://powerbi.microsoft.com/en-us/downloads/>

Tableau Dashboards

# **2. Data resources**

You can use your own source to find data sets or use some of the following open source data sets:

Financial data sets: <https://www.quandl.com>

Data sets: <https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/>

Government data: <https://data.gov.uk/search?filters%5Bformat%5D=HTML&q=free+traffic>

# **3. Overall Approach**

(steps with extra kudos are shown in red)

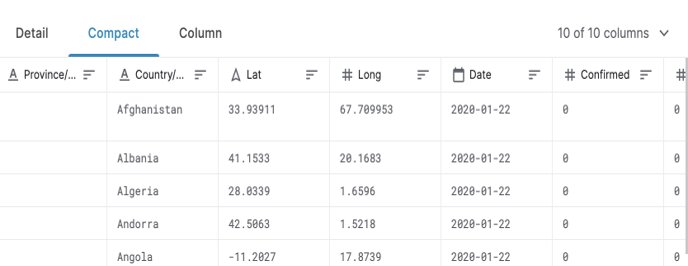
1. Review available data sets
2. Decide on course work subject
3. Profile data sets and record findings (Statistics on the data)
4. Clean the data sets you have picked. This can be done using Pandas and Numpy for example (Give reasoning for actions taken to change any data points)
5. Ingest data sets and automate process for visualizations (database ie SQL server)
6. Analyse & query the data to draw the business/social reasoning for your results
7. Create a presentation to explain your findings
8. Deliver the presentation with a drill down into live data that adds to the presentation

**4. Skeleton Approach**

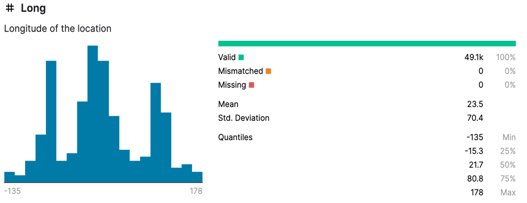
The following is a **basic** skeleton approach to solving the problem. This is to give an example of how this problem can be approached, you do not have to follow this guideline.

1. Pick a data set as a pair, keeping in mind the business problem or the trends you want to look at. Don’t just look at it from a 1-dimensional view but combine your business and technical knowledge to plan.

We picked a Covid-19 data set, we wanted to look at the relationship between the UK and Germany Covid-19 cases and compare the recovery rates. The following is a sample screenshot of the dataset we had for all countries

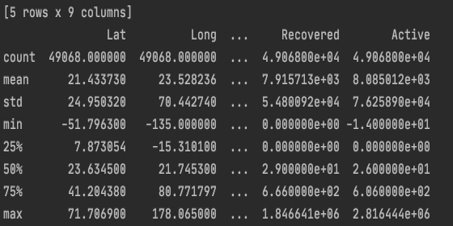


1. Once we have decided the data set, we then stored this data set in Microsoft SQL.
2. The next step will be to clean the data set. This was done using Python pandas and by pulling the data from SQL. The tool highlighted some errors in the data set, these were fixed. The diagnostics tool was run to show the result off the cleaning process, the following shows the result of post cleaning for one column.



The following is a good introductory resource with regards to cleaning data: <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

1. The next step was to profile the data. This was to give insight and give more understanding of the dataset. The following diagram shows the results of our structural profiling of the dataset.



Also, any changes to the data set were made at this stage. We removed certain columns and rows and updated the database accordingly. This is a good opportunity for you as a pair to work together to do a deeper dive into the data set, and look at the data from both a business and technical prospective. The following link is a good introductory resource: <https://panoply.io/analytics-stack-guide/data-profiling-best-practices/>

1. Finally, we used Power BI by connecting it to our Microsoft SQL server with our data set and we were able to draw insights and illustrate our hypothesis from the data set.

# **5. Useful links**

* <https://www.geeksforgeeks.org/python-pandas-series/?ref=lbpu> – Complete data analysis
* <https://panoply.io/analytics-stack-guide/data-profiling-best-practices/>  - To understand the different types of data profiling that can be completed.