Data Analytics Report for the United Kingdom Government

London School of Economics in partnership with FourthRev By Ted Malumbe 5th May 2022

Table of contents

Background:

The purpose of this report is to provide insights and recommendations to the United Kingdom Government. To improve their vaccination campaign through advertising in order to promote the COVID-19 vaccine.

The objective of the Government:

The UK Government is planning on launching a marketing campaign within its territories and wants to achieve the objectives summarized in appendix 1.

Data Analytics Process:

To be able to fulfill the objectives of the Government, I followed the Data Analytics process well explained by (Kazil & Jarmul, 2016) summarized in image 1. Although (Nelli, 2015), equally provided an adequate model as shown in appendix 2, the focus on model validation and deployment were not applicable to this scenario. Another model provided by (Rattenbury, et al., 2017) in appendix 3 was succinct, but lacked pragmatic steps of implementation unlike (Kazil & Jarmul, 2016) in image 1.

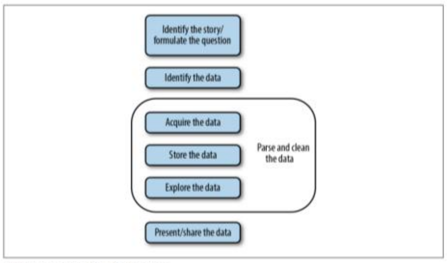
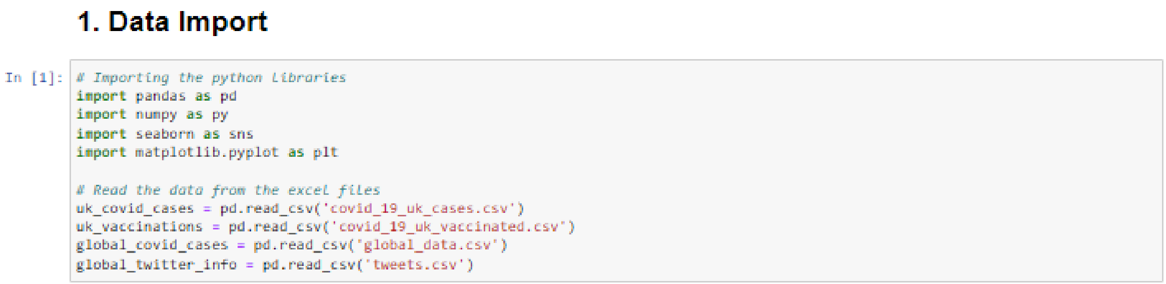


Figure 1-1. Data handling process

Image 1: Data Handling Process (Kazil & Jarmul, 2016, p. 3)

Data Collection:

To meet the ultimate objective of delivering business insights, the first step carried out was the collection of data through importing libraries and reading the right files. On this occasion, since no live data was required, the use of API`s and web scrapping was not necessary. Nevertheless, using the below function, python was able to read the data from a CSV file.



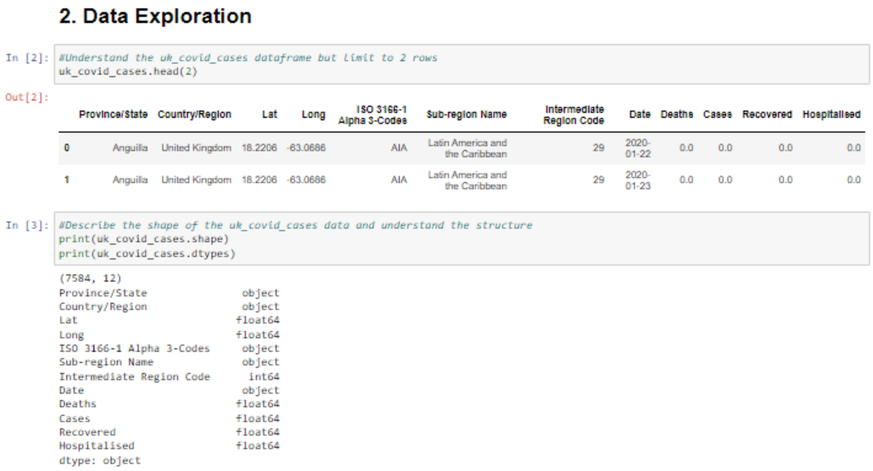
1. Data Import

```
In [1]: # Importing the python libraries
        import pandas as pd
        import numpy as py
        import seaborn as sns
        import matplotlib.pyplot as plt

        # Read the data from the excel files
        uk_covid_cases = pd.read_csv('covid_19_uk_cases.csv')
        uk_vaccinations = pd.read_csv('covid_19_uk_vaccinated.csv')
        global_covid_cases = pd.read_csv('global_data.csv')
        global_twitter_info = pd.read_csv('tweets.csv')
```

Image 2: Data Import Extract

The next step I followed was to explore the data to understand attributes such as its shape, data types and summary statistics. This step was very useful as directed by (Rattenbury, et al., 2017) to get an understanding of the "structure, accuracy, temporality and scope" of the data.



2. Data Exploration

```
In [2]: #Understand the uk_covid_cases dataframe but limit to 2 rows
        uk_covid_cases.head(2)
```

Out[2]:

	Province/State	Country/Region	Lat	Long	ISO 3166-1 Alpha 3-Codes	Sub-region Name	Intermediate Region Code	Date	Deaths	Cases	Recovered	Hospitalised
0	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-22	0.0	0.0	0.0	0.0
1	Anguilla	United Kingdom	18.2206	-63.0686	AIA	Latin America and the Caribbean	29	2020-01-23	0.0	0.0	0.0	0.0

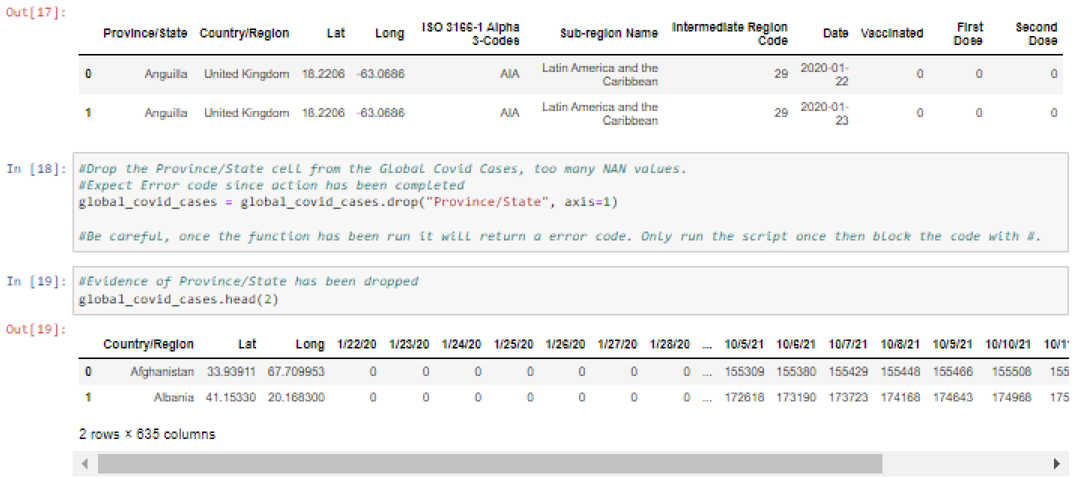
```
In [3]: #Describe the shape of the uk_covid_cases data and understand the structure
        print(uk_covid_cases.shape)
        print(uk_covid_cases.dtypes)
```

```
(7584, 12)
Province/State                object
Country/Region                object
Lat                          float64
Long                         float64
ISO 3166-1 Alpha 3-Codes      object
Sub-region Name               object
Intermediate Region Code       int64
Date                          object
Deaths                       float64
Cases                        float64
Recovered                    float64
Hospitalised                 float64
dtype: object
```

Image 3: Data Exploration Extract

Data Cleaning and Wrangling:

Once collection was completed, I went through the process of cleaning, transforming and preparing the data before I could begin my exploratory data analysis. To begin my data cleaning, I followed the guidelines in (Kazil & Jarmul, 2016) and (Embarak, 2018) in handling missing data through replacing NAN values, missing or generic data and dropping missing values. These values often distort the data analysis results and cleaning them provided the foundation of good conclusions.



```
Out[17]:
   Province/State  Country/Region     Lat     Long  ISO 3166-1 Alpha  Sub-region Name        Intermediate Region  Date     Vaccinated  First  Second
                                                            3-Codes                                        Code                       Dose    Dose
0        Anguilla  United Kingdom  18.2206  -63.0686             AIA  Latin America and the                   29  2020-01-         0      0       0
                                                                                   Caribbean                        22
1        Anguilla  United Kingdom  18.2206  -63.0686             AIA  Latin America and the                   29  2020-01-         0      0       0
                                                                                   Caribbean                        23
```

```
In [18]: #Drop the Province/State cell from the Global Covid Cases, too many NAN values.
         #Expect Error code since action has been completed
         global_covid_cases = global_covid_cases.drop("Province/State", axis=1)

         #Be careful, once the function has been run it will return a error code. Only run the script once then block the code with #.
```

```
In [19]: #Evidence of Province/State has been dropped
         global_covid_cases.head(2)
```

```
Out[19]:
   Country/Region       Lat       Long  1/22/20  1/23/20  1/24/20  1/25/20  1/26/20  1/27/20  1/28/20  ...  10/5/21  10/6/21  10/7/21  10/8/21  10/9/21  10/10/21  10/1
0     Afghanistan  33.93911  67.709953        0        0        0        0        0        0        0  ...   155309   155380   155429   155448   155466    155508   155
1         Albania  41.15330  20.168300        0        0        0        0        0        0        0  ...   172618   173190   173723   174168   174643    174968   175

2 rows × 635 columns
```

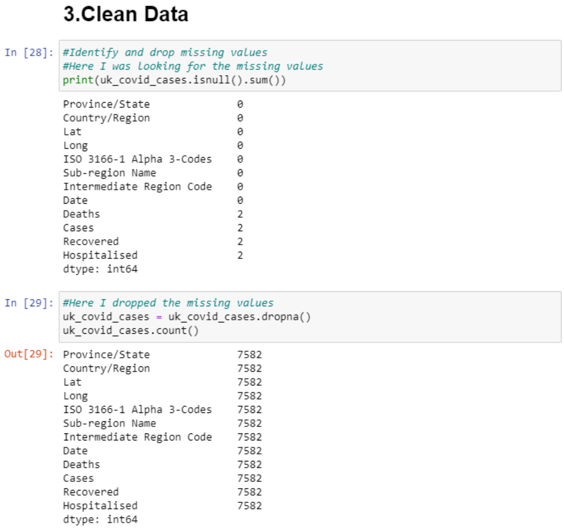
Image 4: Data Wrangling Extract

Table 1 shows an extract of the functions I used during my data cleaning and data wrangling.

Cleaning Functions	
Replace missing data	**Count missing values**
Data.fillna(x)	data.isnull().sum()
Remove duplicate rows	**Drop rows with any N/A/null data**
Data.drop_duplicate()	Data.dropna()
Wrangling Functions	
Summarise data	**Join data**
Data.describe()	Data.concat([x,y])

Table 1: Data Cleaning and Data Wrangling Functions

Take for example, I used the df.dropna() function to remove missing values, that were not significantly necessary in the COVID19 (C19) cases dataset.



3.Clean Data

```
In [28]: #Identify and drop missing values
         #Here I was looking for the missing values
         print(uk_covid_cases.isnull().sum())

         Province/State              0
         Country/Region              0
         Lat                         0
         Long                        0
         ISO 3166-1 Alpha 3-Codes    0
         Sub-region Name             0
         Intermediate Region Code    0
         Date                        0
         Deaths                      2
         Cases                       2
         Recovered                   2
         Hospitalised                2
         dtype: int64

In [29]: #Here I dropped the missing values
         uk_covid_cases = uk_covid_cases.dropna()
         uk_covid_cases.count()

Out[29]: Province/State              7582
         Country/Region              7582
         Lat                         7582
         Long                        7582
         ISO 3166-1 Alpha 3-Codes    7582
         Sub-region Name             7582
         Intermediate Region Code    7582
         Date                        7582
         Deaths                      7582
         Cases                       7582
         Recovered                   7582
         Hospitalised                7582
         dtype: int64
```

Image 5: Summary of Data Cleaning Function and Outputs

By carrying out the above data cleaning, I went on to use the data wrangling functions highlighted by (McKinney, 2017, p. 191). I transformed the data by concatenation to create a new dataset that I could use for analysis without joining the two data frames together on each occasion. In addition, by concatenating the data frame for Vaccinations and Cases, I could leverage on this single data frame to create visualizations. I also detected outliers within the data. However, unfortunately due to the size of the dataset, charts such as box plots and scatter plots provided incomprehensible outputs that would not add value.

Although there are many other functions in data wrangling such as renaming columns, changing the index or adding and removing columns. These were not necessary since the data was already structured in a clean and orderly format. In addition, to avoid over cleaning and wrangling the data, I selected the above functions because of their applicability to the case study.

Further data wrangling will be needed from twitter information data to extract value from the dataset and follow the sentiments of users as the vaccination continues. Overall, although well structured, the LONG and LAT features did not provide much value to be able to draw map charts.

Data Analysis:

Data visualisation is a key tool in the data analytics process that helps elicit key patterns and trend previously unseen in data in its tabular form. The process that underpins data visualisation is called Exploratory Data Analysis that is the process of finding patterns and anomalies in the dataset. Different visualisation techniques exist and are summarized in appendix 4, however more importantly is the use of the right visualisation technique for the intended purpose summarized in appendix 5.

To fulfill the business objective I referred to the guide by (Mukhiya & Ahmed, 2020, p. 67) to select the correct chart for analysis. Appendix 1: Government objectives and chart selection, shows how each business question had a corresponding chart supported by appendix 5.

Data Visualizations and Presentation:

The below summary will focus on two of the objectives aimed by the UK Government. Others can be seen on the Jupyter notebook added to this submission.

Objective 1: Identify the total vaccinations for a region:

The visualisation chosen was the Distribution Chart: Bar Chart. It provided insights that the region Latin America and the Caribbean had the highest amount of vaccination between the regions. Using bar charts quickly answers the objective to understand total vaccinations per region. However, using the bar chart can only provide a general overview since drilling down creates too many regions. The trends show Latin America and the Caribbean have a better vaccination rate than Northern America, guiding the Government to focus more on the Northern America market to target their marketing campaign.

```
Out[30]: Text(0.5, 1.0, 'Total Vaccination Per Region')
```

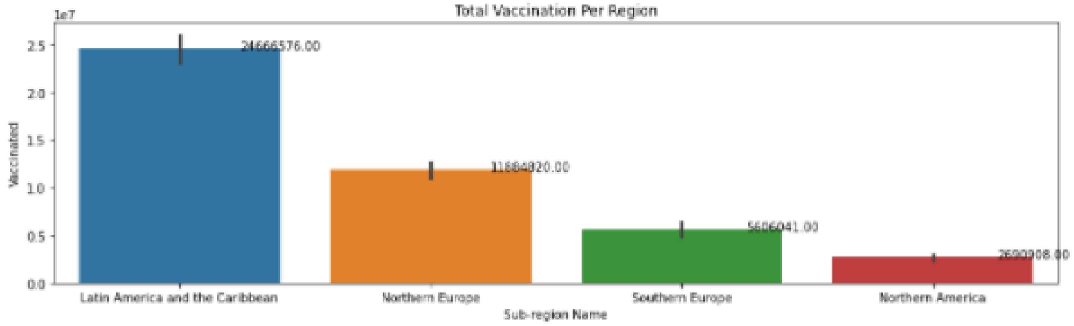
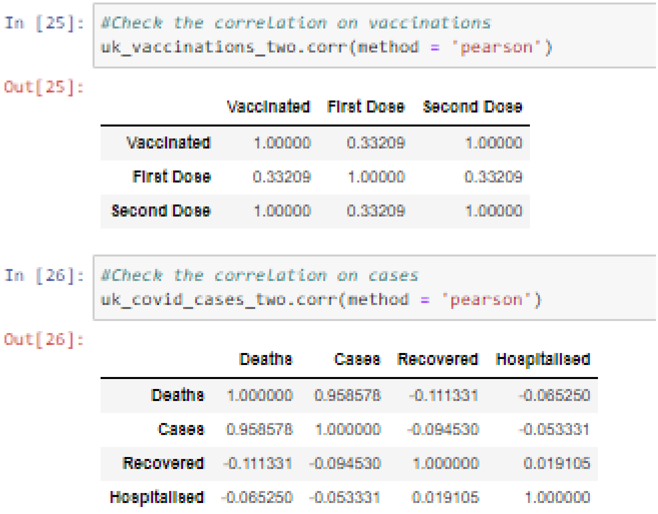


Image 6: Bar Chart of Total Vaccination per Region

Objective 2: Where should they target their first marketing campaign?:

To determine which area should be the target in the first campaign was a challenge in deciding which factor between high deaths or high number of cases is more significant. Using correlation, I saw that death and cases had a very high correlation as shown in Table 7.



```
In [25]: #Check the correlation on vaccinations
         uk_vaccinations_two.corr(method = 'pearson')
```

Out[25]:

	Vaccinated	First Dose	Second Dose
Vaccinated	1.00000	0.33209	1.00000
First Dose	0.33209	1.00000	0.33209
Second Dose	1.00000	0.33209	1.00000

```
In [26]: #Check the correlation on cases
         uk_covid_cases_two.corr(method = 'pearson')
```

Out[26]:

	Deaths	Cases	Recovered	Hospitalised
Deaths	1.000000	0.958578	-0.111331	-0.065250
Cases	0.958578	1.000000	-0.094530	-0.053331
Recovered	-0.111331	-0.094530	1.000000	0.019105
Hospitalised	-0.065250	-0.053331	0.019105	1.000000

Image 7: Data Analysis: Correlation carried out in python

To confirm this, I used a multiple regression model to understand which variable had the greatest influence on the data. As such without the independent variable of cases the r2 score dropped to 0.0097433185772281, however with it the r2 score was 0.926722837088056. I therefore concluded that Northern Europe with the highest cases should be the area they should target their first marketing campaign because they will directly lead to more deaths.

```
Text(0.5, 1.0, 'Total Average Cases Per Region')
```

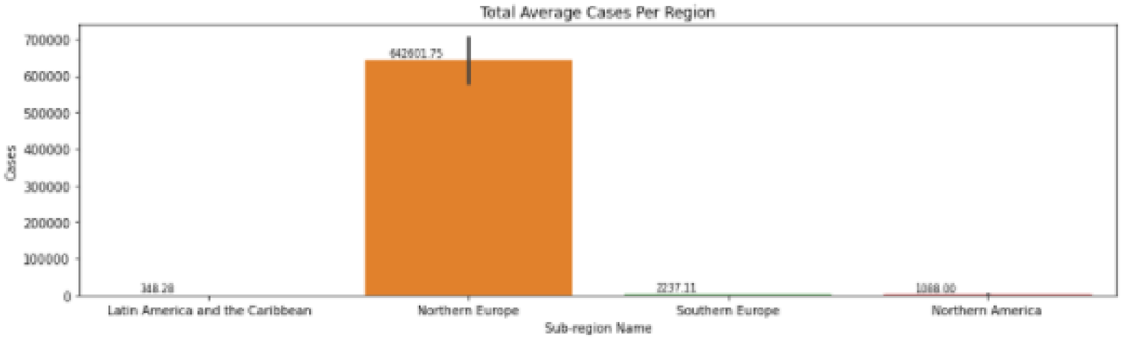


Image 8: Highest Amount of C19 Deaths

Recommendation and Conclusion:

I would recommend that the Government focuses on Northern America that has the lowest amount of vaccination between each region. Whilst Latin America can be used as a benchmark amongst all the other regions.

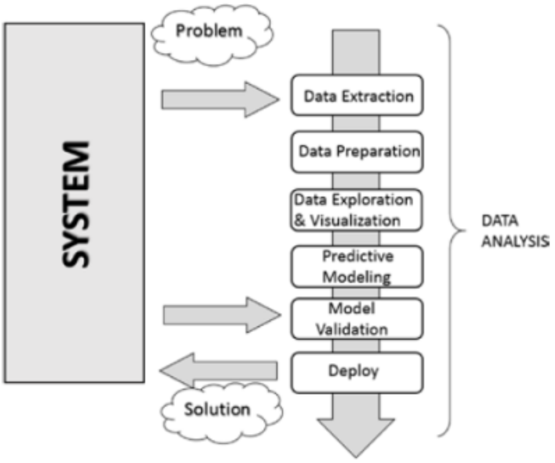
Using linear regression, I was able to confirm the significant positive correlation between death and cases and as such the Government should focus on areas with high cases because these have the highest detrimental effect on the population. This would suggest Northern Europe as the best country to pursue the first marketing campaign.

However, Northern Europe also shows the region with the highest recoveries amongst the rest. Deaths have been increasing over time but the peak in hospitalization has reduced across all regions. As the data suggest the pandemic is on the decline therefore the Government should focus on the prevention rather than cure.

Appendix 1: Government objectives and chart selection

Question	Government Objective	Chart Selection
1	Identify the total vaccinations for a region	Distribution
2	Recommend where the Government should begin their first campaign	Distribution
3	Illustrate the areas with the largest number of people that have received one vaccination does and not a second	Distribution
4	Illustrate which areas have the greatest number of recoveries to priorities the marketing campaigns	Distribution
5	Illustrate whether deaths have been increasing across all the regions over time or if the peak has reached	Distribution
6	Identify trends and patterns that can inform the marketing approach to increase the number of vaccinated people	Correlation
7	Identify what other twitter data points and tweets have both #coronavirus and #vaccinated hashtags	Composition
8	Illustrate which regions have experienced a peak in hospitalizations number and if other regions have not reach it either	Change
9	Suggest potential future outcomes based on the trends in the hospitalization rates of different regions	Change

Appendix 2: Data Analytics Process



Appendix 3: Data Analytics Process

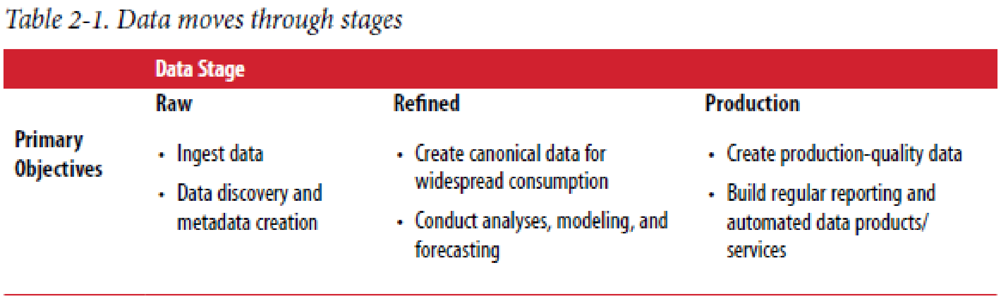


Table 2-1. Data moves through stages

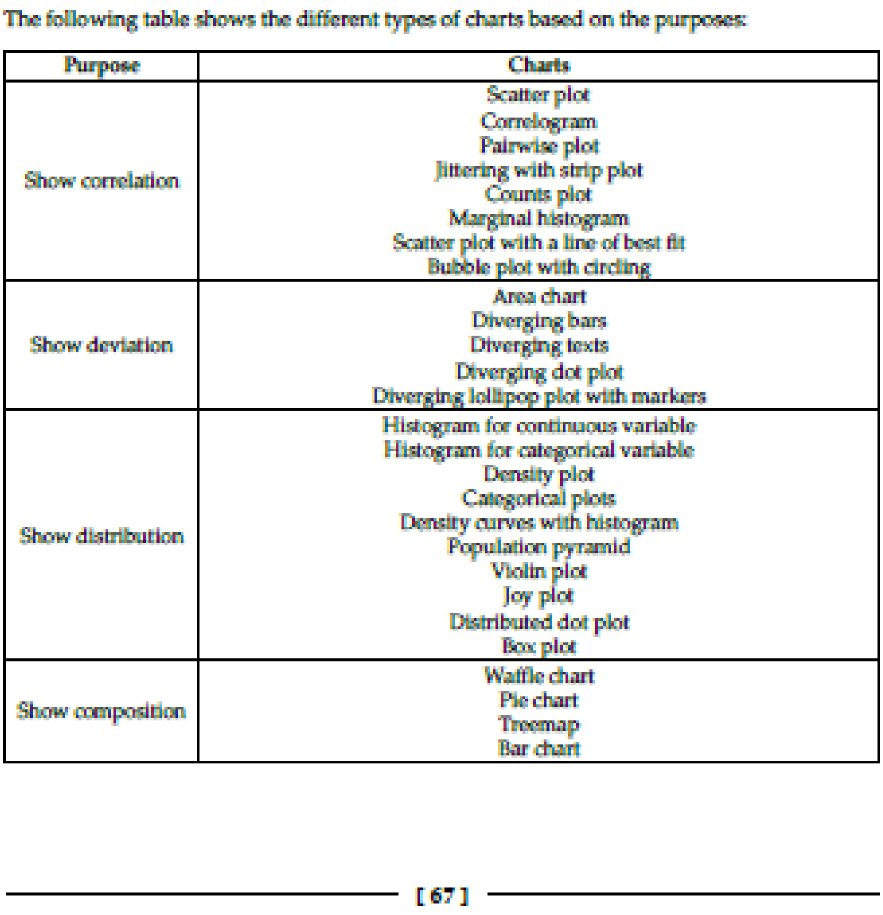
	Data Stage		
	Raw	Refined	Production
Primary Objectives	• Ingest data • Data discovery and metadata creation	• Create canonical data for widespread consumption • Conduct analyses, modeling, and forecasting	• Create production-quality data • Build regular reporting and automated data products/ services

(Rattenbury, et al., 2017)

Appendix 4 : Visualisation techniques (Mukhiya & Ahmed, 2020)

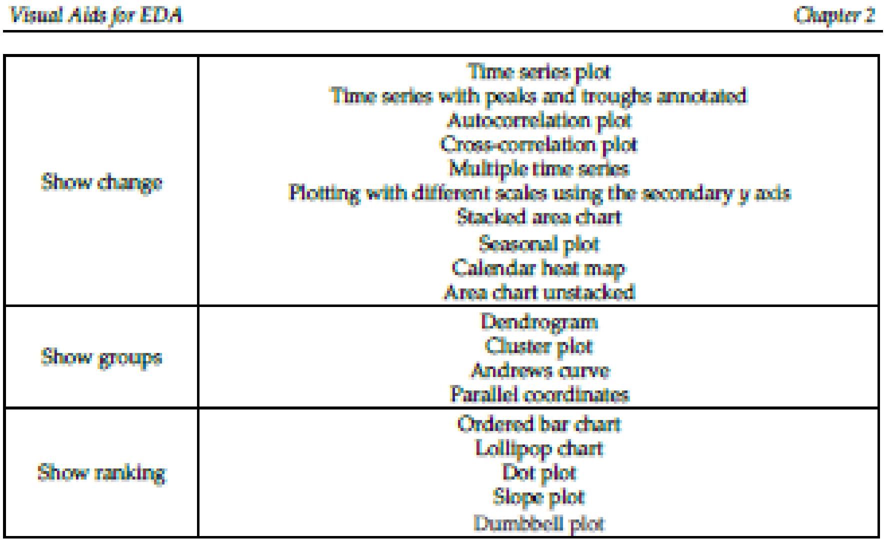
- Line chart
- Bar chart
- Scatter plot
- Area plot and stacked plot
- Pie chart
- Table chart
- Polar chart
- Histogram
- Lollipop chart

Appendix 5: Effective visualisation



The following table shows the different types of charts based on the purposes:

Purpose	Charts
Show correlation	Scatter plot Correlogram Pairwise plot Jittering with strip plot Counts plot Marginal histogram Scatter plot with a line of best fit Bubble plot with circling
Show deviation	Area chart Diverging bars Diverging texts Diverging dot plot Diverging lollipop plot with markers
Show distribution	Histogram for continuous variable Histogram for categorical variable Density plot Categorical plots Density curves with histogram Population pyramid Violin plot Joy plot Distributed dot plot Box plot
Show composition	Waffle chart Pie chart Treemap Bar chart



Show change	Time series plot Time series with peaks and troughs annotated Autocorrelation plot Cross-correlation plot Multiple time series Plotting with different scales using the secondary *y* axis Stacked area chart Seasonal plot Calendar heat map Area chart unstacked
Show groups	Dendrogram Cluster plot Andrews curve Parallel coordinates
Show ranking	Ordered bar chart Lollipop chart Dot plot Slope plot Dumbbell plot

(Mukhiya & Ahmed, 2020, p. 67)

Bibliography

Embarak, O., 2018. *Data Analysis and Visualization using Python.* 1 éd. New York: Apress.

Kazil, J. & Jarmul, K., 2016. *Data Wrangling with Python. Tips and tool to make your life easier..* 1st éd. New York: O'Reilly Media.

McKinney, W., 2017. *Pythin for Data Analysis. Data Wrangling with Pandas, Numpy and Ipython.* 2 éd. Boston: O`Reilly.

Mukhiya, S. K. & Ahmed, U., 2020. *Hands-on Exploratory Data Analysis with Python.* 1 éd. Birmingham: Packt.

Nelli, F., 2015. *Python Data Analytics: Data Analysis and Science using pandas, matplotlib and the Python Programming Language.* New York: Apress.

Rattenbury, . T. et al., 2017. *Principles of Data Wrangling: Practical Techniques for Data Preparation.* 1 éd. O'Reilly Media: New York.