Name: Diana Jaganjac

Student Number: 13170192

Date: 16/03/2020

Academic Declaration: "I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software."

Multi-Dimensional Analysis and Dimensionality Reduction

Phase 1:

1.1 Introduction

The Global Power Plant Database (World Resources Institute, 2020) is comprised of 24 dimensions, 6 of which are key to understanding the distribution of the database (Jaganjac, 2020). Of these 6 dimensions, I identified 3 as being suitable 'predictor' dimensions and 1 as a suitable 'predicted' dimension (see Figure. 1) The dimensions identified as suitable 'predictor' dimensions were longitude, latitude and capacity_mw. The dimension identified as suitable for the 'predicted' dimension was primary_fuel.

	capacity_mw	latitude	longitude	primary_fuel
0	33.00	32.3220	65.1190	Hydro
1	66.00	34.5560	69.4787	Hydro
2	100.00	34.6410	69.7170	Hydro
3	11.55	34.4847	70.3633	Hydro
4	42.00	34.5638	69.1134	Gas

Figure. 1: Table showing first five entries of three 'predictor' dimensions and one 'predicted' dimension.

1.2 Data Quality

The 'predictor' dimensions were picked as they met the relevant criteria. For instance, all three dimensions fall within the 6 key dimensions previously identified and are comprised of consistent data values with no missing data-points. As mentioned in Part I, (Jaganjac, 2020), there were a few issues regarding data quality which needed to be considered. Of these issues (see Part 1), the most prominent data quality issue was the lack of a full set of data points for the commissioning year and estimated_generation_gwh dimensions. I made the decision to remove these rows of data in order to complete the feature selection stage for PCA and dimensionality reduction. Although other methods for tidying data exist, simply deleting the incomplete rows of data was sufficient to carry out successful feature selection and to complete the analysis.

1.3 Feature Selection

Features were selected using a correlation matrix, to measure the amount of variance among dimensions (see Figure. 2) The 6 key dimensions were used for the correlation matrix, along with the 'country_long' dimension, which was also identified in Part. I (Jaganjac, 2020) as another potentially useful dimension. The correlation matrix was useful in identifying which features to use for principle-component analysis (PCA) based on increased variance.

As a result of this analysis, I identified longitude, latitude and capacity_mw as useful 'predictor' dimensions, and primary_fuel as a useful 'predicted' dimension. This allowed for the analysis to become categorical i.e. which data points fall into which primary_fuel category. In addition, I found the three 'predictor' dimensions themselves each as being suitable in predicting the 'predicted' dimension of primary_fuel (see Figure. 3). For instance, longitude and latitude correspond well as identifiers for primary_fuel across the globe. As mentioned in (BP, 2020), it is more likely that coal is used as a fuel source in China and oil in the United States. Nevertheless, it is difficult to generalise fuel source usage globally as most countries use a variety of different fuel sources, which includes cogeneration, nuclear, renewables and other fuel sources (Bloomberg, 2020).

By including capacity_mw as a third 'predictor' dimension, we can narrow down results a little further. As capacity_mw measures the capacity of electrical generation in megawatts, we can very generally differentiate world regions on which type of primary_fuel is mostly widely used when comparing capacity_mw and latitude/ longitude. For instance, Figure. 4 (bottom-left) shows that latitudes between 20° to 50°, mostly use coal and/or solar and waste as a primary fuel source with capacity megawatt generation of between approximately 2000 to 7000 megawatts. However, when comparing these findings to a world-map, we can see that 20° to 50° spans much of the northern hemisphere, meaning it is almost impossible to more specifically determine which region of the world these findings correspond to. Nevertheless, Figure.4 (top), provides a nice visual representation of primary_fuel source of longitude vs. latitude. This reproduces a world-map with points colour-coded to the primary_fuel dimension and provides a good overview of the data.

Overall, and based on this analysis, it seems as though the selected 'predictor' dimensions of longitude, latitude and capacity_mw could theoretically prove useful in generally categorising data to the 'predicted' dimension of primary_fuel source. The most important reasons for this include low correlation and high variance of the dimensions (Figure. 2) and useful visualisations of the dimensions as shown by the scatterplots (Figure. 3 and Figure. 4).

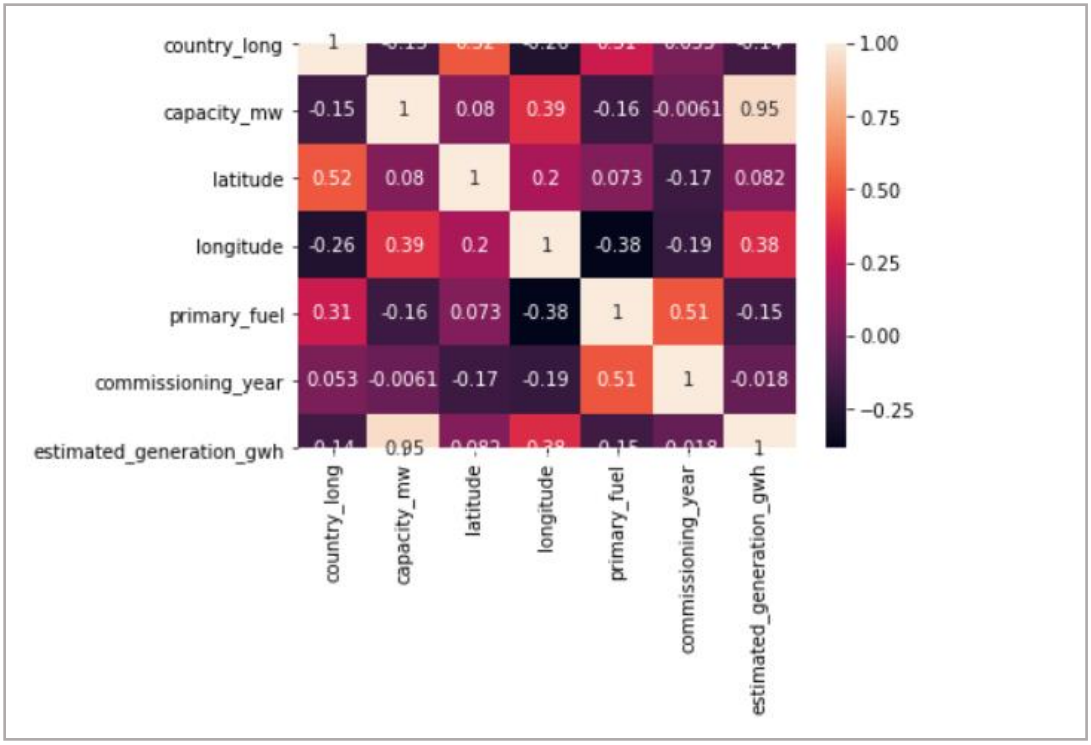

Figure. 2: Correlation Matrix for 7 dimensions of the GPPD.

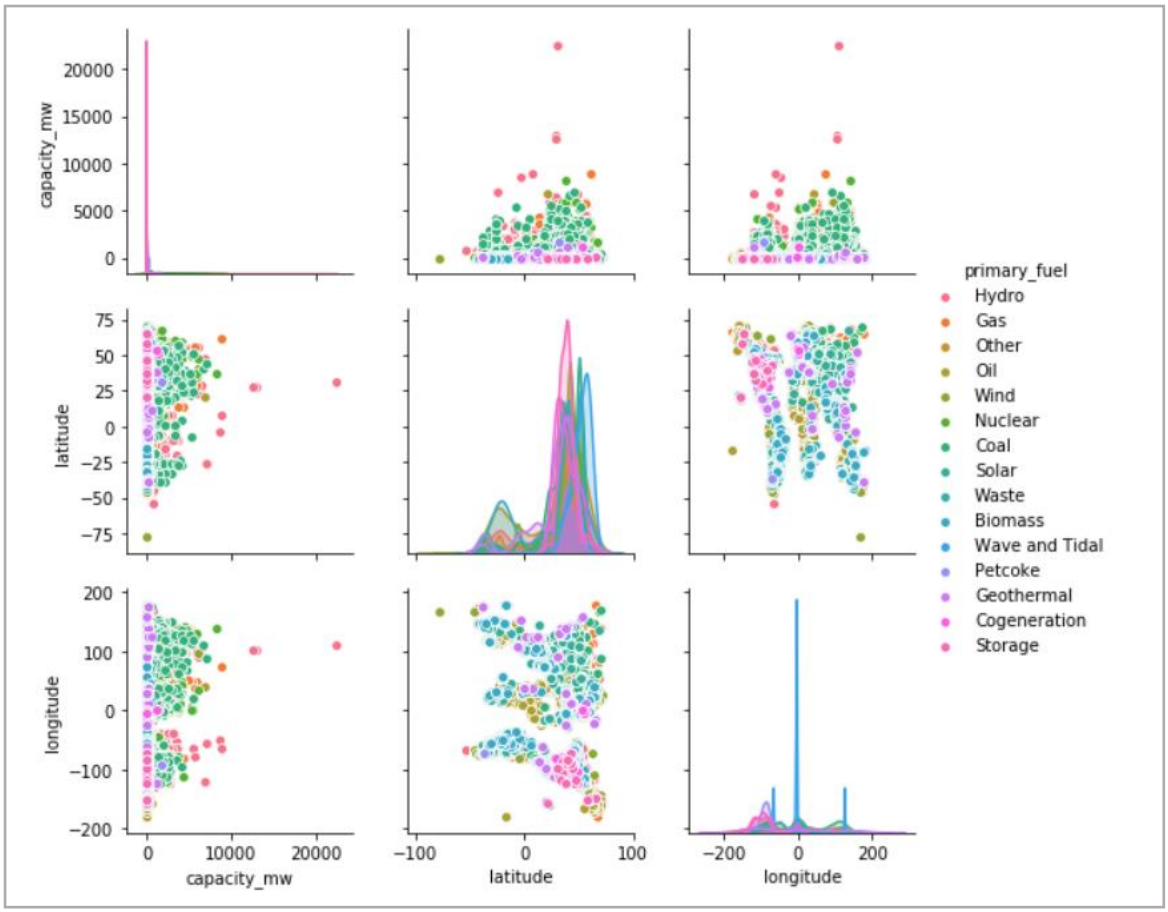

Figure. 3: Scatter Matrix for Longitude, Latitude and Capacity_mw dimensions. Colour-coded by Primary_fuel.

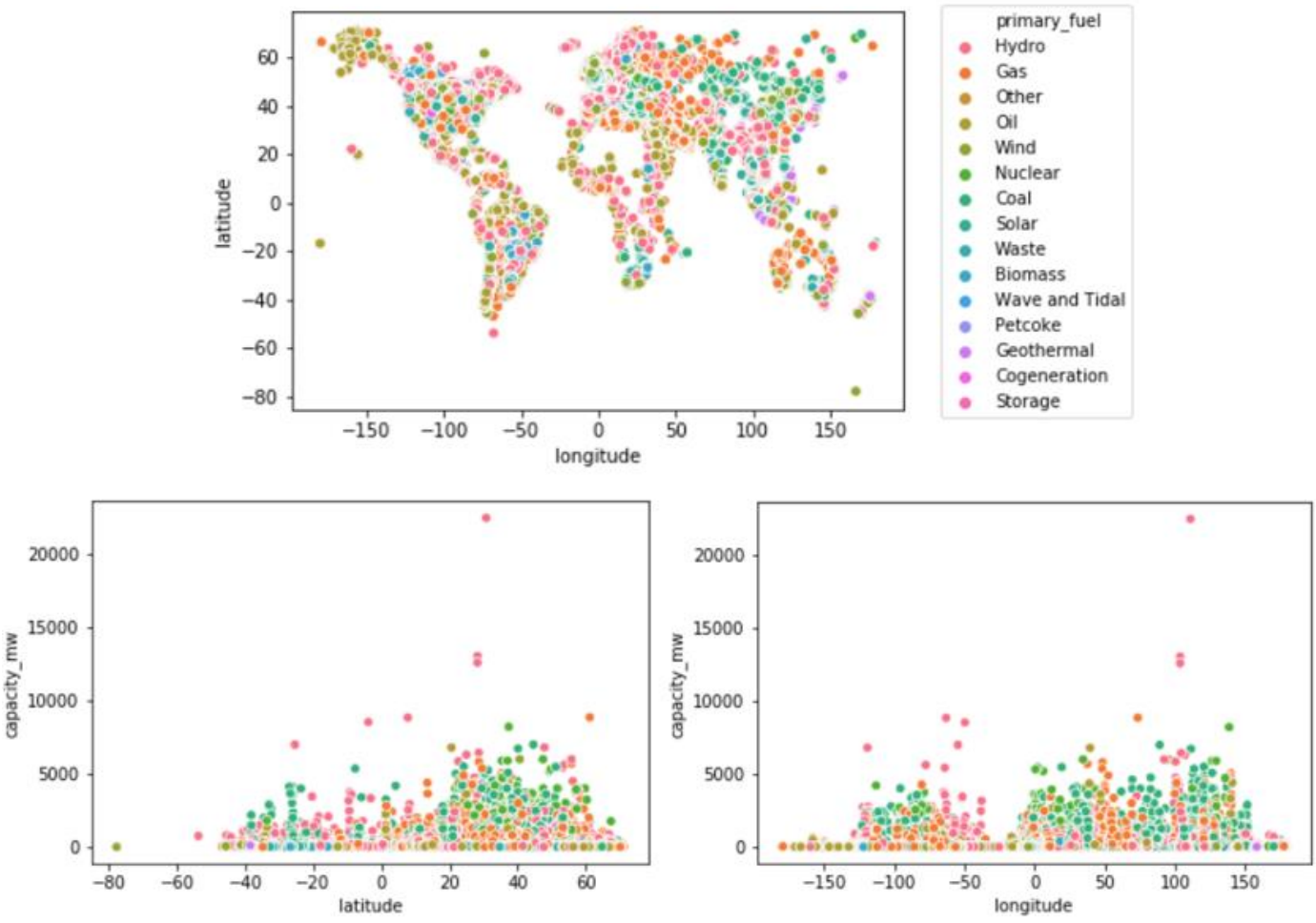

Figure. 4: Scatter-graphs of longitude vs. latitude (top), latitude vs. capacity_mw (bottom-left) and longitude vs. capacity_mw (bottom-right).

Phase 2:

2.1 Principle Component Analysis using Scikit-Learn

Principle component analysis (PCA) is a technique which reduces dimensionality in highly dimensional data whilst retaining important patterns which are key to understanding the distribution of the data. PCA is used to find the most important principle components, with the intention of providing an insightful summary of the data.

In my technical implementation of PCA, I closely followed the method used by Gaël Varoquaux which is available on the Scikit-Learn webpage (Varoquaux, 2020). Scikit-Learn is a machine learning library available in Python, and contains functions to perform many statistical techniques, including PCA.

Whilst implementing PCA, I realised that my 'classes' for my 'predicted' dimension primary_fuel were too high in number and need to be reduced. For example, I was performing PCA to reduce dimensionality from 3 to 2 dimensions, and therefore the 'classes' for my 'predicted' dimension also needed to be equal to 2. I therefore decided to add a new dimension to my dataset which would separate the 'primary_fuel' data in to renewable (class label = 1), or non-renewable (class label = 0) fuel sources. The renewable fuel sources included biomass, geothermal, hydro, solar, storage, wave and tidal and wind fuel sources, whilst non-renewable fuel sources included coal, cogeneration, gas, nuclear, oil, other, petcoke and waste. I then used this new dimension, named "R_NR", to perform PCA and dimensionality reduction on the data. Section 2.2 further illustrates my use of Scikit-Learn for PCA (see Figure. 5).

2.2 Code Segments

```
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from matplotlib.cm import gist_rainbow
from sklearn.preprocessing import StandardScaler

#I select the features
features = ['capacity_mw', 'latitude', 'longitude']
x = df.loc[:, features].values

#I scale features to standardise results
x = StandardScaler().fit_transform(x)

#I plot the 2D axes
fig = plt.figure(1, figsize=(8, 6))

# I set the 'R_NR' dimension to the colours variable
colours = df["R_NR"]

# I perform PCA
X_reduced = PCA(n_components=2).fit_transform(x)

#I produce a 2-D graph of the PCA with all relevant labels
graph = plt.scatter(X_reduced[:, 0], X_reduced[:, 1], \
                    c = colours, cmap = plt.cm.summer, edgecolor='k', s=40)

plt.title("First two PCA directions", fontsize = 15)
plt.xlabel('First Principle Component', fontsize = 10)
plt.ylabel('Second Principle Component', fontsize = 10)

cbar = plt.colorbar(graph, ticks=[0, 1])
cbar.ax.set_yticklabels(["Non-Renewable", "Renewable"])

plt.show()
```

Figure. 5: Code segment illustrating how Principal Component Analysis was carried out in Jupyter Notebook (Jaganjac, 2020).

Phase 3:

3.1 Results of "R_NR" Scatterplots

Following the addition of the new dimension "R_NR" I completed some further analysis to see how this dimension would classify the three 'predictor' dimensions 'capacity_mw', 'longitude' and 'latitude'. Figure. 6 illustrates the interaction of these three dimensions with the new dimension "R_NR" when plotted against each other.

Figure. 6 illustrates that the data is a lot easier to interpret, compared to Figures 3 and 4. By aggregating the 15 previous classes to now only 2 classes, the data is visually more meaningful. Furthermore, by aggregating the data to a useful partition, i.e. renewable and non-renewable fuel sources, we can learn even more than when the data was in its original format. For example, Figure. 6 (top), nicely presents the distribution of the data globally, and allows for new interpretations of the data. Nevertheless, some elements do retain the same interpretability as in Figure. 4. For example, Figure. 6 (bottom-left) identifies a similar region as in Figure. 4 as belonging to non-renewable fuel sources, and spans from approximately 20° to 60° latitude. Despite appearing useful on this scatter-graph, when comparing this finding to 20° to 60° latitude on a world-map we of course find that corresponds to much of the northern hemisphere and so it is difficult to pin-point this finding to exact countries/ regions.

Nevertheless, by aggregating the 15 classes of the 'primary_fuel' dimension to just 2 classes in the "R_NR" dimension, the data is easier to interpret, and allows for PCA dimensionality reduction to be performed. Results for the Principal Component Analysis and dimensionality reduction are further explained in section 3.2 and Figures. 7 and 8.

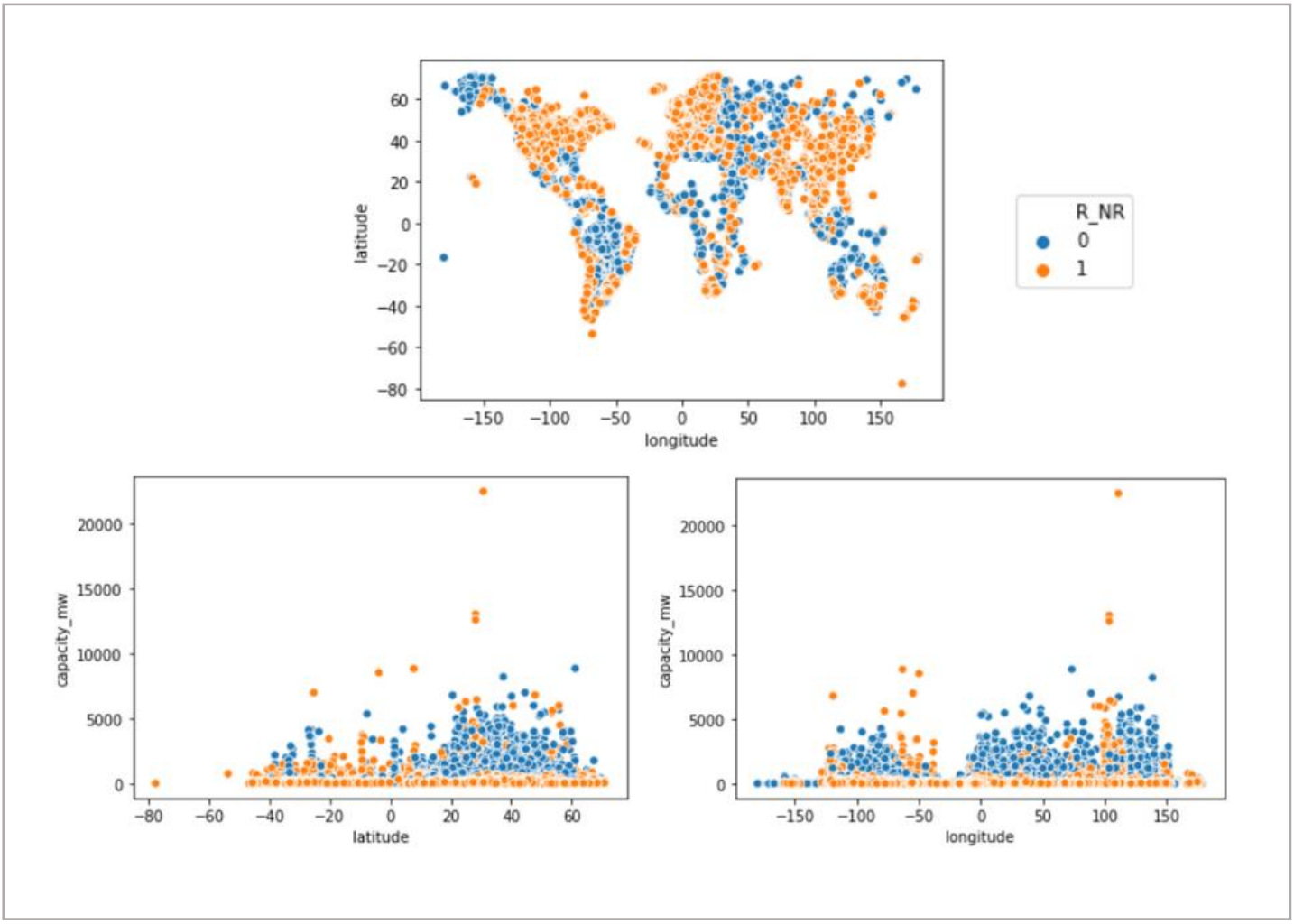

Figure. 6: Scatter-graphs of longitude vs. latitude (top), latitude vs. capacity_mw (bottom-left) and longitude vs. capacity_mw (bottom-right).

3.2 Results of Principal Component Analysis & Dimensionality Reduction

Figure. 7 illustrates very generally that non-renewable energy sources tend more towards the higher-end of both the first and second eigenvector, and that renewable energy sources tend more towards the lower-end of both eigenvectors. The two classes overlap, and it is difficult to separate them, even after input data-values have been standardised. Overall, this makes it more difficult to interpret the data visually.

Figure. 8 aids our understanding of this, as it explains the level of variance of each principle component. For example, Figure. 8 illustrates that PCA feature 1 has an explained variance of 1.2 and PCA feature 2 an explained variance of 0.95. As these features have explained variance, which is quite close in range, this means that the eigenvectors may not separate the data as well. If the gap between explained variance were higher, the eigenvectors may have separated the data slightly more.

Figure. 8 also aids in our understanding of the intrinsic dimensions of the data. The intrinsic dimension refers to the number of PCA features needed to represent the intrinsic patterns in the dataset. As I am only using 3 features, and the explained variance of these features is within a similar range, the data tends to overlap and there is no clear-cut distinction between intrinsic and non-intrinsic variables which is the ideal. To improve this, it would have been useful to use data which had more variance between PCA features and clear-cut intrinsic variables. For this particular dataset, it may also have been useful to explore more of the features, and to complete PCA dimensionality reduction with more than 3 features.

Overall, it seemed as though the chosen ‘predictor’ dimensions would perform a good analysis as there was low correlation among the ‘predictor’ dimensions (see. Figure 2). Despite this, it seems as though the level of variance, when performing PCA and dimensionality reduction is not so high, and therefore data overlaps and does not form distinct and easily interpretable patterns. What’s more, a reduction from 3 to 2 dimensions seems to not be sufficient to encompass a clear-cut distinction between intrinsic and non-intrinsic dimensions (see Figure. 8) and use of more PCA features may have been beneficial. In general, my selection of dimensions and PCA perform a good dimensionality reduction, however it could be improved by exploring the rest of the dataset.

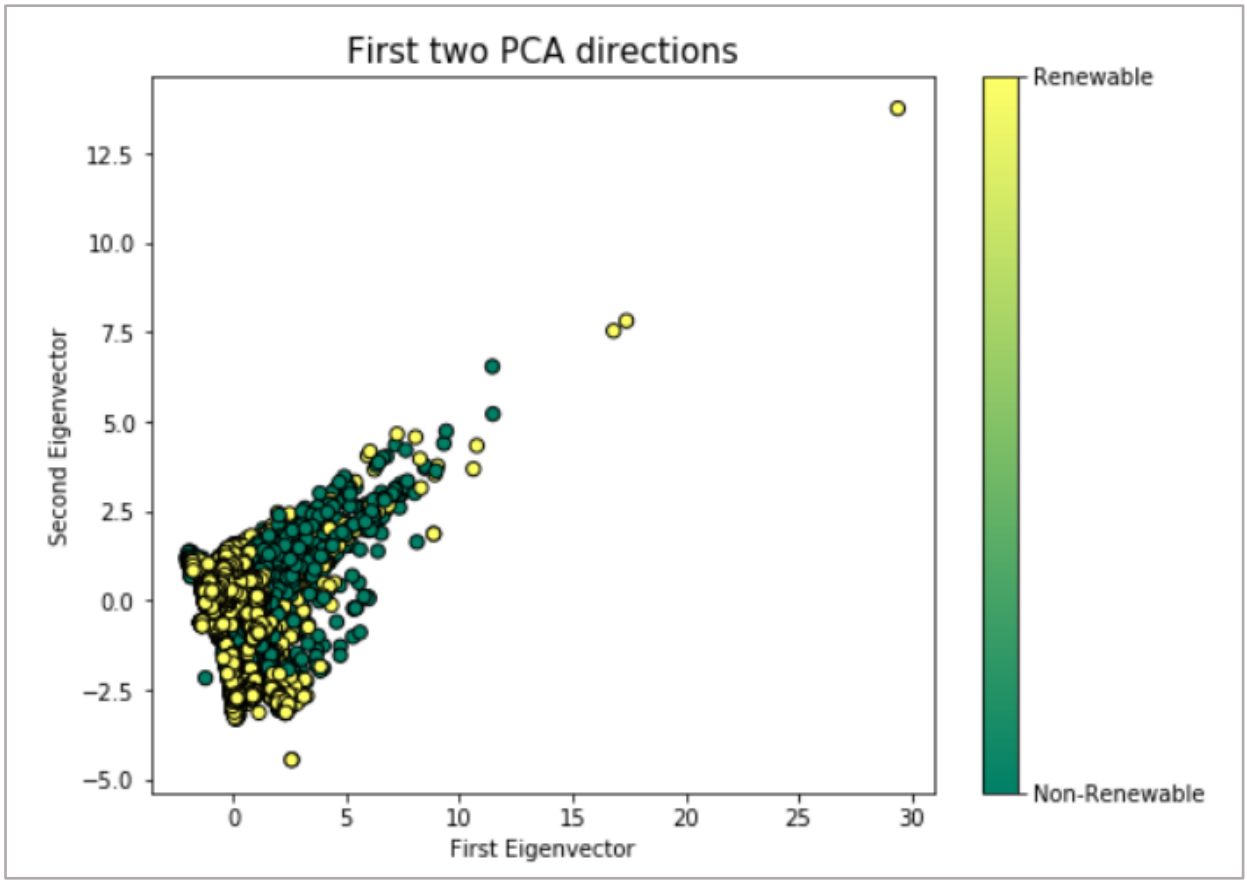

Figure. 7: Scatterplot of First Eigenvector and Second Eigenvector after PCA dimensionality reduction has been performed.

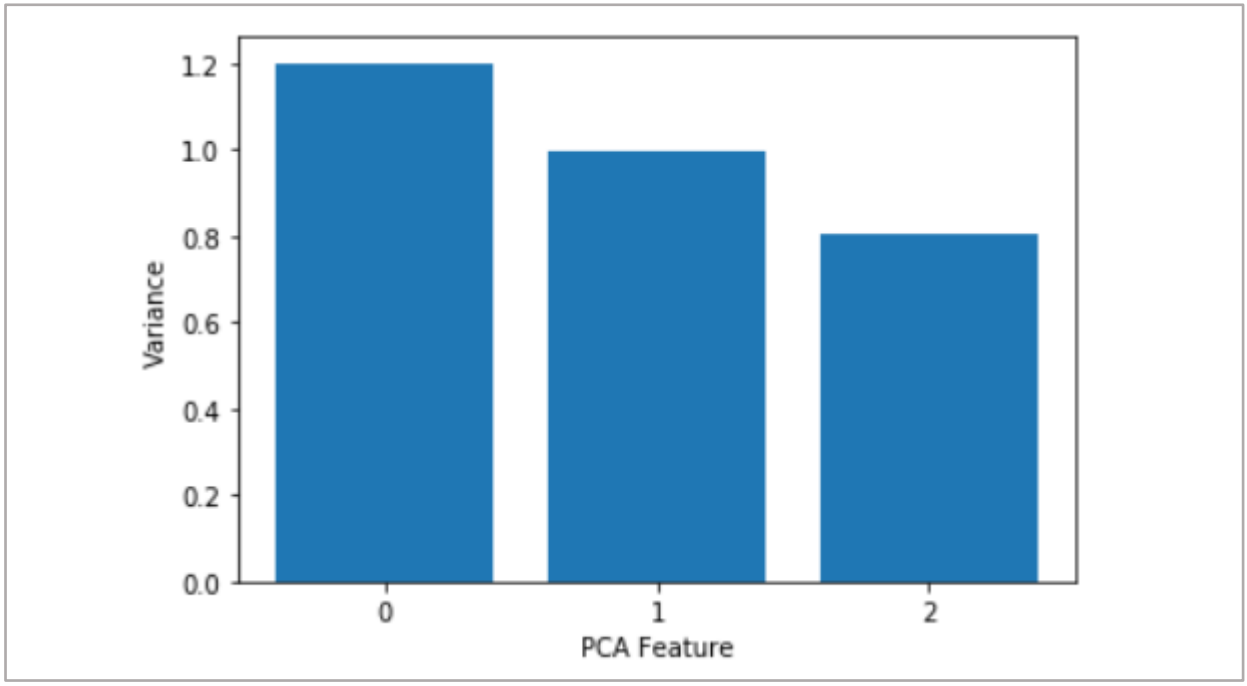

Figure. 8: Bar-chart showing the explained variance of PCA features.

Bibliography:

Bloomberg, How Each Country Contributed to the Explosion in Energy Consumption, available online at https://www.bloomberg.com/graphics/2019-international-energy-use-renewables-coal-oil/. Last accessed 16/03/2020.

BP, Energy demand by region, available online at, https://www.bp.com/en/global/corporate/energy-economics/energy-outlook/demand-by-region.html. Last accessed 16/03/2020.

Jaganjac, D., 2020. DSTA_CW1_13170192_D_Jaganjac.

Varoquaux, The Iris dataset, available online at, https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html#sphx-glr-auto-examples-datasets-plot-iris-dataset-py. Last accessed 16/03/2020.

World Resources Institute, Global Power Plant Database, available online at http://datasets.wri.org/dataset/globalpowerplantdatabase. Last accessed 25/02/2020. Last accessed 16/03/2020.