In recent years, Spain has experienced an ever growing demand for electricity ,forcing the government to seek alternative options to increase production and sustain this demand.

Spain's population has grown by 8% over the last decade and industries have increased by 15% over the same period according to data from the Spain’s Bureau of Statistics. This is already putting a lot of pressure on the existing fossil fuel energy that is limited and also not sustainable.

We were approached by the government of Spain to research on the trends and patterns of the country's renewable and fossil fuel energy generation with the view of providing recommendations to prevailing scarcity challenge that exists in Spain.

This presentation covers data analysis,data engineering,model building and model performance

Under EDA we looked at a number of things.

First : We looked at the dataset, which contained 48 features on weather attributes as can be seen on this slide [with information wind, rain, humidity, cloud, pressure, snow and temperature]. The data was collected from 5 cities in Spain for the period January 2015 to December 2017

**Second:** We analysed the data shape and types i.e. float or object, numerical or categorical as this determined the visualizations to be used in our presentation

We also calculated the data statistics like mean, median and looked at feature distributions to check for data symmetry

**Feature Distributions**

* Histogram showing 3 hourly shortfall
* Barplot showing the interaction between the Wind degree & 3 hourly shortfall
* Based on the correlation heatmap above, we can see the highest correlation would be with the pressure attributes of other cities with the highest correlation being with Valencia\_pressure.
* Histograms examining features

based on the datasets provided

**Third :** We checked for missing values and duplicate columns to ensure that we get accurate insights in data engineering.

Our dataset contained 23% null values under valencia pressure column and 1 duplicate column

**Last:**We checked for weather features correlation, both + and - as it determines how 2 weather attributes change together.We were looking for strong +ve or -ve correlation highlighted in orange .

As previously seen, **Valencia\_pressure** contained 2068 null values. In order to carry out the replacement, we need to determine the columns that show the highest correlation with Valencia\_pressure.

We have used Valnecia\_pressure to fill in the missing values.

1. Random forest -It builds decision trees on different samples and takes their majority vote for classification and average in case of regressionDecision Tree
2. Decision tree A decision tree is a supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes
3. **Linear Regression**  Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
4. We will evaluate our models using a metric for regression machine learning: the **mean squared error (MSE). but** MSE is highly biased for higher values. **RMSE is better in terms of reflecting performance when dealing with large error values**. As per the Model performance Test above , The Random forests model is the best performing model with the lowest RMSE score, as it is great with high dimensional data and provides the highest accuracy. The random forest technique can also handle big data with numerous variables.
5. Outcomes
6. To analyse Spain’s electricity shortfall due to reduced supply by renewable sources by analysing weather attributes.
7. Utilize data analysis & machine learning principles to predict future shortfalls
8. Developed a predictive model that can be used to determine shortfalls.
9. Questions