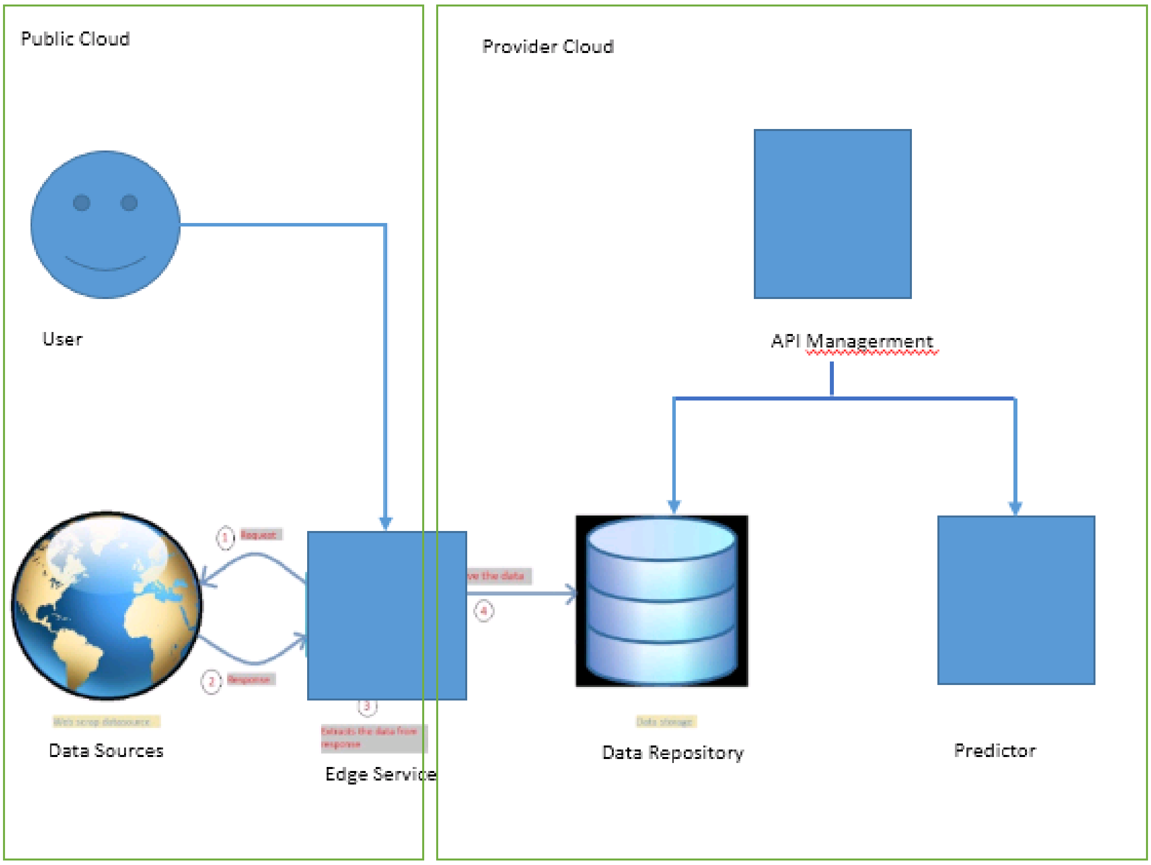
Architectural Decisions Document for Oil Target Price Prediction

1 Architectural Components Overview



1.1 Data Source

Scrapping of oil price via https://fred.stlouisfed.org/series/DCOILBRENTEU

1.2 Data Repository

1.2.1 Technology Choice

Data is stored in the cloud since it is a small dataset.

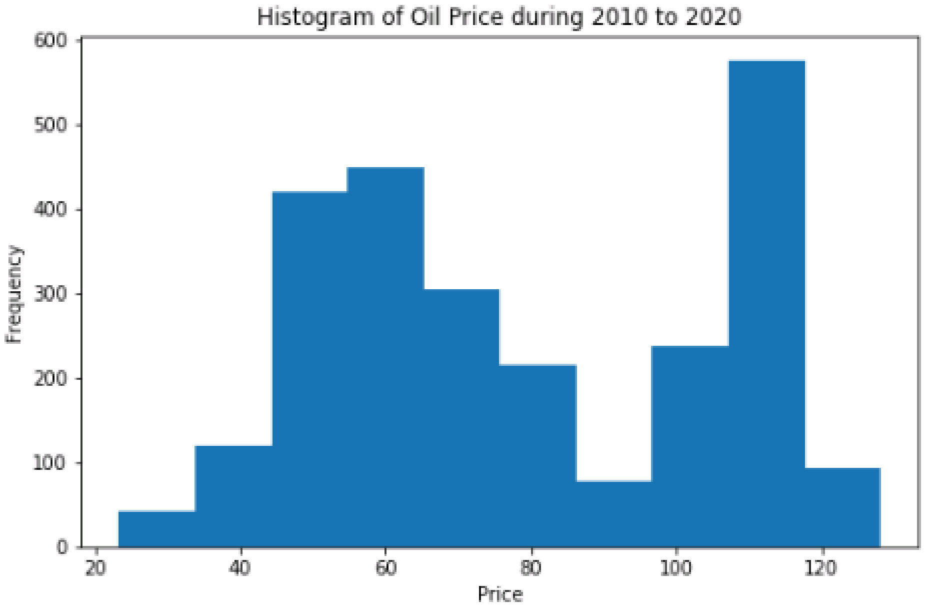
1.3 Data Quality Assessment

1.3.1 Technology Choice

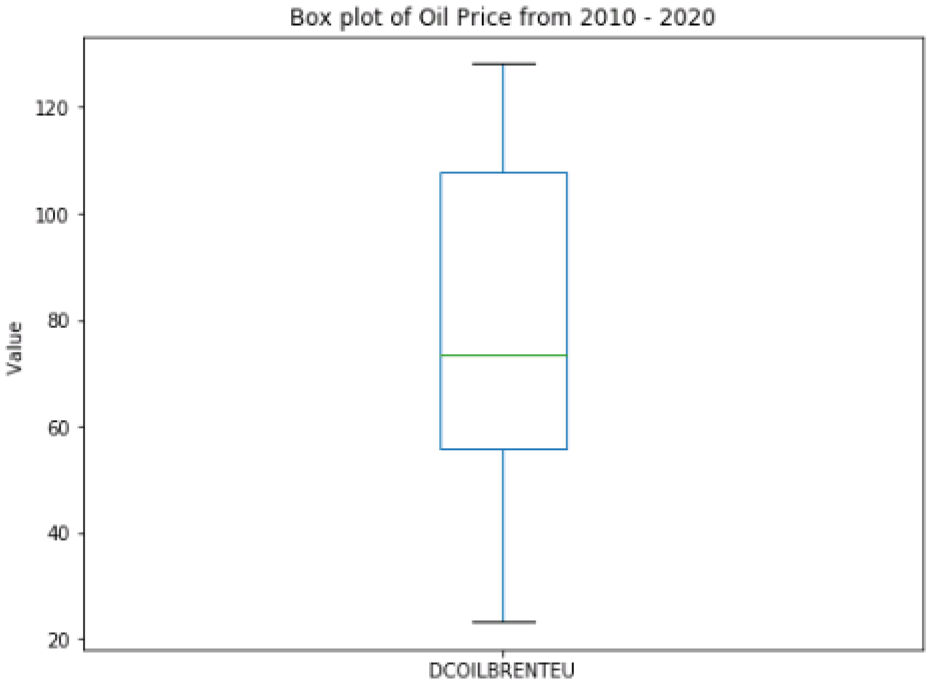
A histogram is a way of representing the *frequency* distribution of numeric dataset.
A box plot is a way of statistically representing the distribution of the data.
Time series plot.

1.3.2 Justification

A histogram is a way of representing the *frequency* distribution of numeric dataset. The way it works is it partitions the x-axis into *bins*, assigns each data point in our dataset to a bin, and then counts the number of data points that have been assigned to each bin. So the y-axis is the frequency or the number of data points in each bin. Note that we can change the bin size and usually one needs to tweak it so that the distribution is displayed nicely.

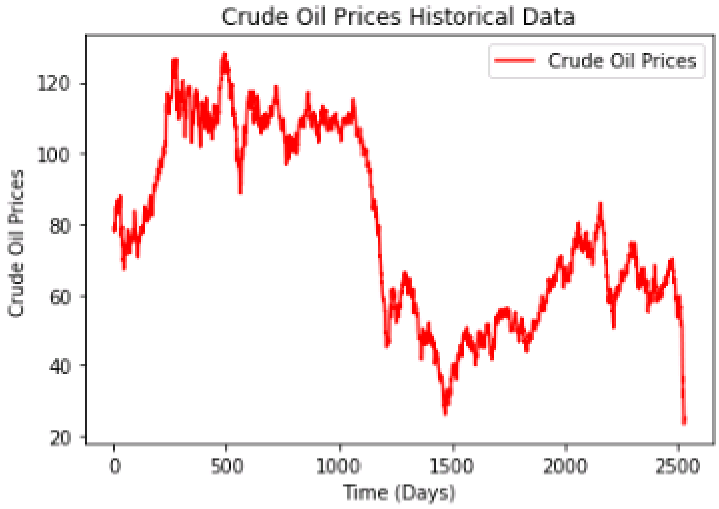


A box plot is a way of statistically representing the distribution of the data through five main dimensions:
Minimun: Smallest number in the dataset.
First quartile: Middle number between the minimum and the median.
Second quartile (Median): Middle number of the (sorted) dataset.
Third quartile: Middle number between median and maximum.
Maximum: Highest number in the dataset.



We can immediately make a few key observations from the plot above:
1. The minimum number of immigrants is around 23 (min), maximum number is around 130 (max), and median number of immigrants is around 73 (median).
2. 25% of the price of ~56 or fewer (First quartile).
3. 75% of the price of ~107 or fewer (Third quartile).
4. There is no outlier since our data source is reliable

Time series plot as follows.



1.4 Feature Engineering

We choose scaling or normilization. Because the data for the sequence prediction problem probably needs to be scaled when training a neural network, such as a Long Short-Term Memory recurrent neural network. We will scale between 0 and 1.

1.5 Actionable Insights (Select Algorithm)

1.5.1 Technology Choice

There are several Deep learning model for time series prediction such as Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs), Long-Short Term Memory (LSTM). LSTM is pretty good at extracting patterns in input feature space, where the input data spans over long sequences. Given the gated architecture of LSTM's that has this ability to manipulate its memory state, they are ideal for such problems. Thus, LSTM is according to research one of the most suitable algorithms for time series predictions compared to CNN (Convolutional Neural Networks), we hold the same view based on our results and the logic behind the usage of both algorithms, where CNN is favored for other usages.

1.5.2 Justification

Stateful vs. Stateless LSTM

1. **Stateless**: LSTM updates parameters on batch 1 and then initiates cell states (meaning - memory, usually with zeros) for batch 2
2. **Stateful**: it uses batch 1 last output cell sates as initial states for batch 2.

When to use which?

- When sequences in batches are related to each other (e.g. prices of one commodity), we should better use *stateful* mode
- Else, when one sequence represents a complete sentence, we should go with *stateless* mode

1.6 Select Framework

Because Keras easy to learn and easy to use, our data is small and simple, thus we choose Kera to implement LSTM.

1.7 Performance Indicator to Model Evaluation and Comparison

Commonly used metrics to evaluate forecast accuracy are the coefficient of variation (CV RMSE), the root mean squared error (RMSE) and the Mean Absolute Error (MAE). CV (RMSE) is the RMSE normalized by the mean of the measured values and quantifies typical size of the error relative to the mean of the observations. A high CV score indicates that a model has a high error range. Root Mean Square Error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. MAE, a commonly used metric, is the mean value of the sum of absolute differences between actual and forecasted. RMSE is another commonly used metric. It penalizes the larger error terms and tends to become increasingly larger than MAE for outliers.

In order to compare between MSE and MAE model, we normalize RSME and MEA by the mean of the measured values and quantifies typical size of the error relative to the mean of the observations (e.g, mean, median, Thirst quartile - First quartile). A high CV score indicates that a model has a high error range.

1.8 Applications / Data Products

1.8.1 Technology Choice

Cloud service to Users. Subscribed users will request the target price of next day/month via email or app or web service.