**FORECASTING THE PRICE OF LETTUCE IN THE UK**

**INTRODUCTION**

Lettuce (Lactuca sativa) is one of the world's most economically important leafy vegetables. It is one of the most popular salad crops and is well-known for its delicate, crispy texture and slightly bitter taste, as well as its fresh condition of milky juice. It is among the most widely grown salad vegetable crops because of its high demand. Lettuce is rich in vitamins as well as minerals such as calcium and iron. It is commonly served alone or with dressing as a salad with tomatoes, carrots, cucumbers, or other salad vegetables.

Lettuce has been cultivated for over two thousand years, and many different varieties have evolved. According to Marks (Marks, 2021), four major types of lettuce are grown and consumed in the United Kingdom (UK), with numerous varieties within each type. These four types of lettuce are Romaine, Butterhead, Looseleaf, and Iceberg lettuce.

This project aims to forecast the price of iceberg lettuce in the UK using data collected by the Office of National Statistics (ONS). Price forecasting is the prediction of a commodity or product price based on various factors such as its characteristics, demand, seasonal trends, the prices of other commodities, offers from various suppliers, and so on.

Price forecasting could be a feature of consumer-facing apps that are used to increase customer loyalty and engagement. At the same time, other businesses may benefit from price forecasting information. Entrepreneurs may need to determine the best time to buy a commodity to adjust the prices of products or services that require a commodity (lettuce) or to assess the investment appeal of fixed assets (Petersson, 2019).

**BACKGROUND OF THE PROJECT**

For decades, the cost of living in the UK has been rising. In April 2022, inflation hit its highest level on record, with the ONS estimating that it is currently higher than it has been since roughly 1982, hurting the affordability of goods and services. Increases in the cost of consumer goods, which have been caused by high consumer demand and supply chain bottlenecks, have contributed to inflation. Over the years, several events have happened that have directly affected the economy of the UK. Among these are the economic recession of 2008, the Brexit referendum of 2016, and the COVID-19 pandemic of 2019.

The ONS, the UK's main statistics body, has gathered data for what it calls its Consumer Price Index, which essentially averages out what people (rather than businesses) pay for everyday items and converts it into a single figure that shows how prices have increased or decreased (Leslie, 2017). In this project, we intend to use this public data from the ONS to forecast or predict the price of lettuce in the future.

This project can be expressed as a regression problem. Regression analysis is a statistical technique used to estimate the relationship between a dependent/target variable (in this case, the price of lettuce) and one or more independent (interdependent) variables, also known as predictors, that influence the target variable.

**AIMS AND OBJECTIVES**

The aims of this project are:

1. Forecasting Lettuce prices in the UK to help retailers set their prices by creating a time series model.

**METHODOLOGY**

For this project, the team employed the typical data science lifecycle, which is made up of 6 uniquely important steps. These are:

1. Gathering data
2. Exploratory Data Analysis (EDA)
3. Data Wrangling and Cleaning
4. Feature Engineering
5. Modeling
6. Model Evaluation and Interpretation

The following sections describe each of these steps in detail to reveal our thought process throughout this entire project.

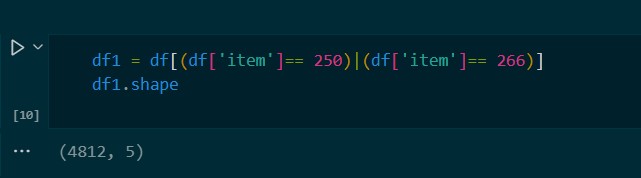
**Gathering Data:**

Getting datasets to work with on this project proved to be a difficult task. We eventually came across ONS, which stands for Office for National Statistics, and is the executive office of the UK statistics authority. The dataset contains information about price indices, percentage changes, and weights for the different measures of consumer price inflation. The data has been recorded and is being updated monthly since February 1988.

The dataset had five columns and over 1.7 million rows. The columns were date, shop code, region, price, and item. The date column contained details of the date each data was recorded, the shop code column contained codes that were used to represent different stores across the UK, the region column referred to different areas in the UK, e.g. London and Northern Ireland, the price column contained details of the retail price of each item, and the item column contained codes used to represent the different retail items captured in the dataset.

Lettuce - which we’re interested in, was represented in the dataset as item 250 and item 266. We extracted the information relating to just lettuce by selecting only those rows with item number 250 or item number 266 and saved it to a new dataframe df1. This dataframe had 4812 observations.

The image below shows the sample code we used to select the data relating to lettuce.

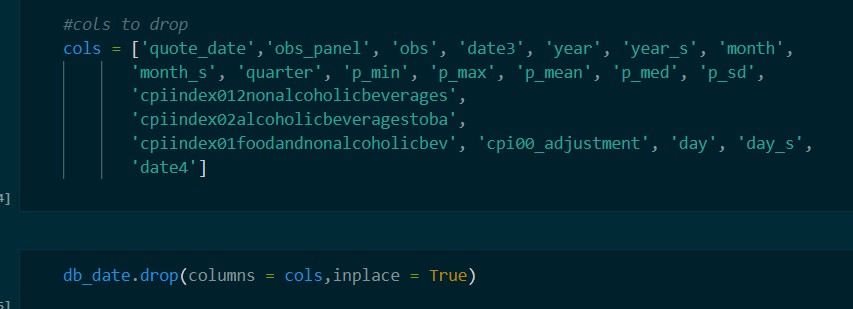


**Exploratory Data Analysis (EDA):**

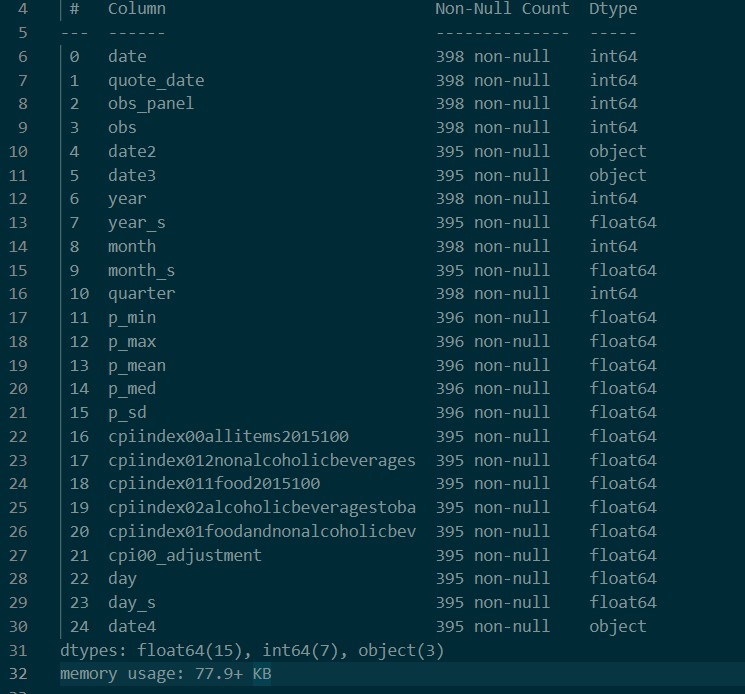
Exploratory Data Analysis refers to exploring the data that we have gathered to find patterns and obtain insights that will inform our next steps in building the solution to forecast the prices of lettuce.

To begin, we loaded the dataset “5 pct\_sample” - which is a 5 percent sample of the original data - into a pandas dataframe to enable us to utilize all the amazing features of the python library: pandas to explore, wrangle, and clean our data. Initially, our dataset contained information on the date, shop code, region, price, item, and CPI for various items. The date, item, and region columns were displayed as codes on the dataset, and we needed to find an explanation for these codes. After examining the information contained in the dataset, we realized there were a few columns we didn’t need and others that we needed but didn’t have. Therefore, we dropped the unnecessary columns from the dataset and looked for another dataset that contained the new information that we needed.

The image below shows the unnecessary columns we dropped from the dataset.



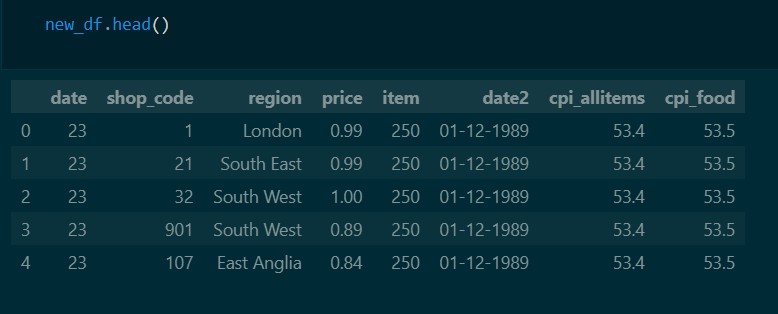
We got a dataset called “db\_date” that holds relevant data about the date. This dataset had 25 columns, as shown in the image below.



We cleaned up this dataset by dropping irrelevant columns and converting the date column from an int 64 to a DateTime object.

After cleaning, we merged the datasets “df1” and “db\_date” into a new dataframe called “new\_df”. This new dataframe had 7 columns: date, shop\_code, region, price, item, cpi\_allitems, and cpi\_food. The dataframe also had 4812 observations on lettuce. From the date column, we created additional features we needed to plot visualizations to enable us to gain significant insights into the data.

Screenshot of the new\_df dataset



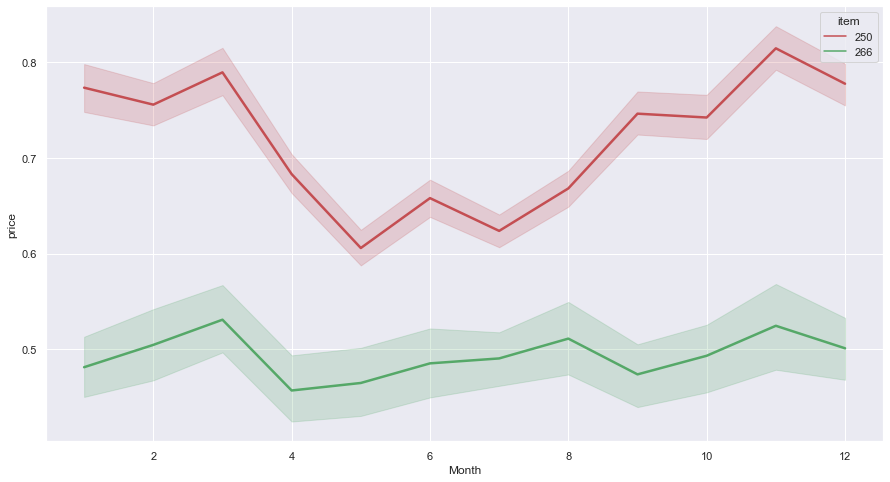
**Column Description**

From the figure above, we can see that the dataset has 8 significant columns.

1. The date column contains codes that represent the actual date the data was recorded.
2. The shop code column contains codes that represent various stores in different regions in the UK.
3. Region refers to different areas in the UK. This dataset captures 12 regions - The 9 regions of England plus Wales, Northern Ireland, and Scotland. The regions are:
   1. North East has major cities like Newcastle, Sunderland and Durham.
   2. Northwest has major cities like Manchester, Liverpool, and Blackpool.
   3. Yorkshire has major cities like Sheffield, Leeds, and York.
   4. East Midlands has major cities like Leicester and Derby.
   5. West Midlands has major cities like Birmingham and Wolverhampton.
   6. South East has major cities like Brighton and Portsmouth.
   7. South West has major cities like Bristol and Bath.
   8. East of England (East Anglia) has major cities like Cambridge and Norwich.
   9. London is England’s capital city and is home to over 8 million people.
   10. Wales
   11. Scotland
   12. Northern Ireland
4. The price column refers to the retail price of lettuce recorded on that particular day.
5. The item column holds codes that are used to represent the different retail items. 250 and 266 represent lettuce.
6. The “Date2” column holds the actual date in the correct format.
7. The “cpi\_allitems” column refers to the price index of various retail items.
8. The “cpi\_food” column refers to the price index of food items.

**Insights from the dataset**

One of the main advantages of lettuce is that it is available throughout the year, but one of the questions we asked was how seasonal changes affected the prices of lettuce in the UK. The plot below shows the distribution of lettuce through the different seasons of the year.

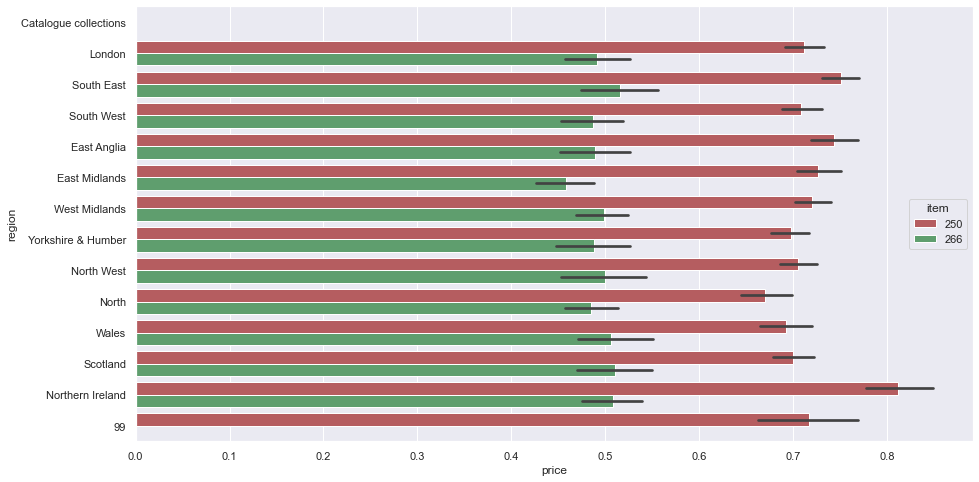


From the above chart, we can see that there is an upward spike and a very deep low in spring (March and May), while there is also a significant increase in autumn (November). From this information, it is not quite clear yet what the effect of seasons will have on the prices of iceberg lettuce in the UK.

We pressed further to investigate this dataset and tried to see the distribution of prices over the years.



This chart shows the distribution of prices within the range of our dataset. Prices fluctuate over the years, and particular years were of interest to us as these might offer us an explanation for the rise and fall of lettuce prices.

Then we proceeded to look at how the prices of iceberg lettuce are distributed across various regions.

This chart shows that the prices were almost evenly distributed across the 12 regions, with the highest recorded price being in Northern Ireland.

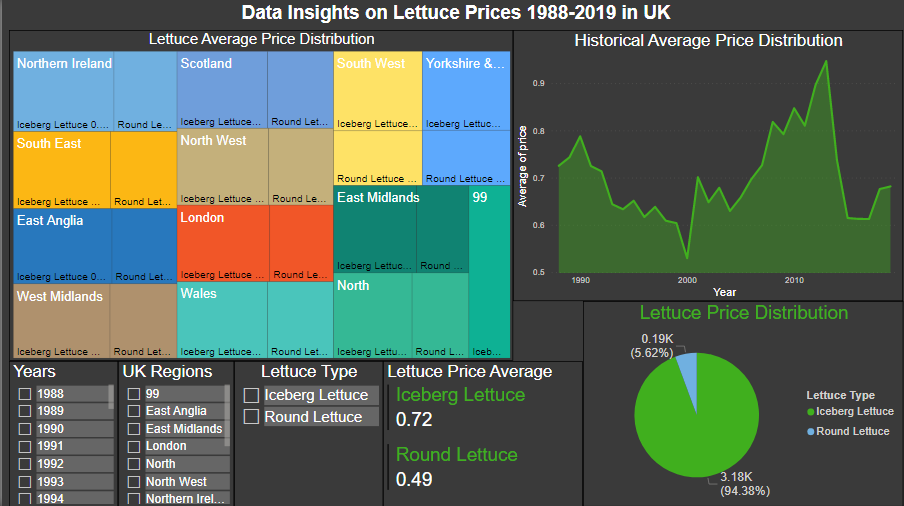
**Special Years of Interest**

Continuing our investigations, we tried to explore the datasets to see whether major world events affected the prices of iceberg lettuce in the UK. The events we investigated were the global economic recession in 2008, the Brexit referendum in 2016, and the COVID-19 pandemic in 2019.

**DASHBOARD**

This dashboard was created using Microsoft Power BI - A powerful business intelligence tool used in creating amazing interactive dashboards, find trends in data and display intuitive visualizations. After cleaning our data in Python, we imported the “new\_df” dataset into Power BI to create our visuals.

The image below shows a screenshot of the interactive dashboard we created to further explore the data and display useful insights and visualizations.



The *treemap* shows the average price by region and item, which enables clients to compare average prices in different regions for both Iceberg and Round lettuce. The round lettuce is the second type of lettuce recorded in our dataset, and it was saved as item 266.

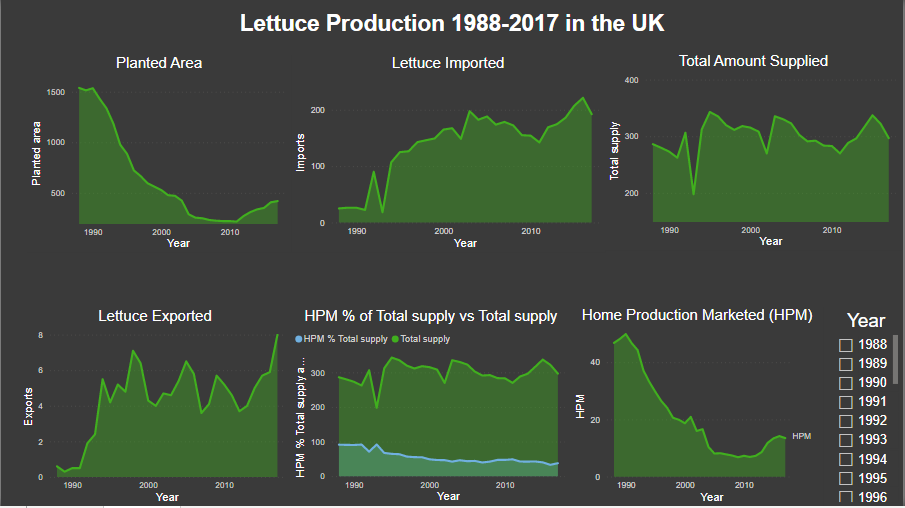
The clustered column chart displays the average price by item. This helps in assessing the item which is more expensive than the other based on the prices. The higher the average price, the more expensive the lettuce is, implying that iceberg lettuce is the most expensive in the UK.

The area chart shows the average price over the years from 1988 up to 2019. The prices have been fluctuating, with the year 2000 recording the lowest price. Also, item 266 (round) lettuce was captured between 2002 and 2012.

There are two measures of the average prices for item 250 (iceberg) with an average price of 0.79 and item 266 (round) lettuce with an average price of 0.69.

The region, year, and item slicers will aid in filtering the information the client wants.

The image below shows the production of lettuce since 1988.

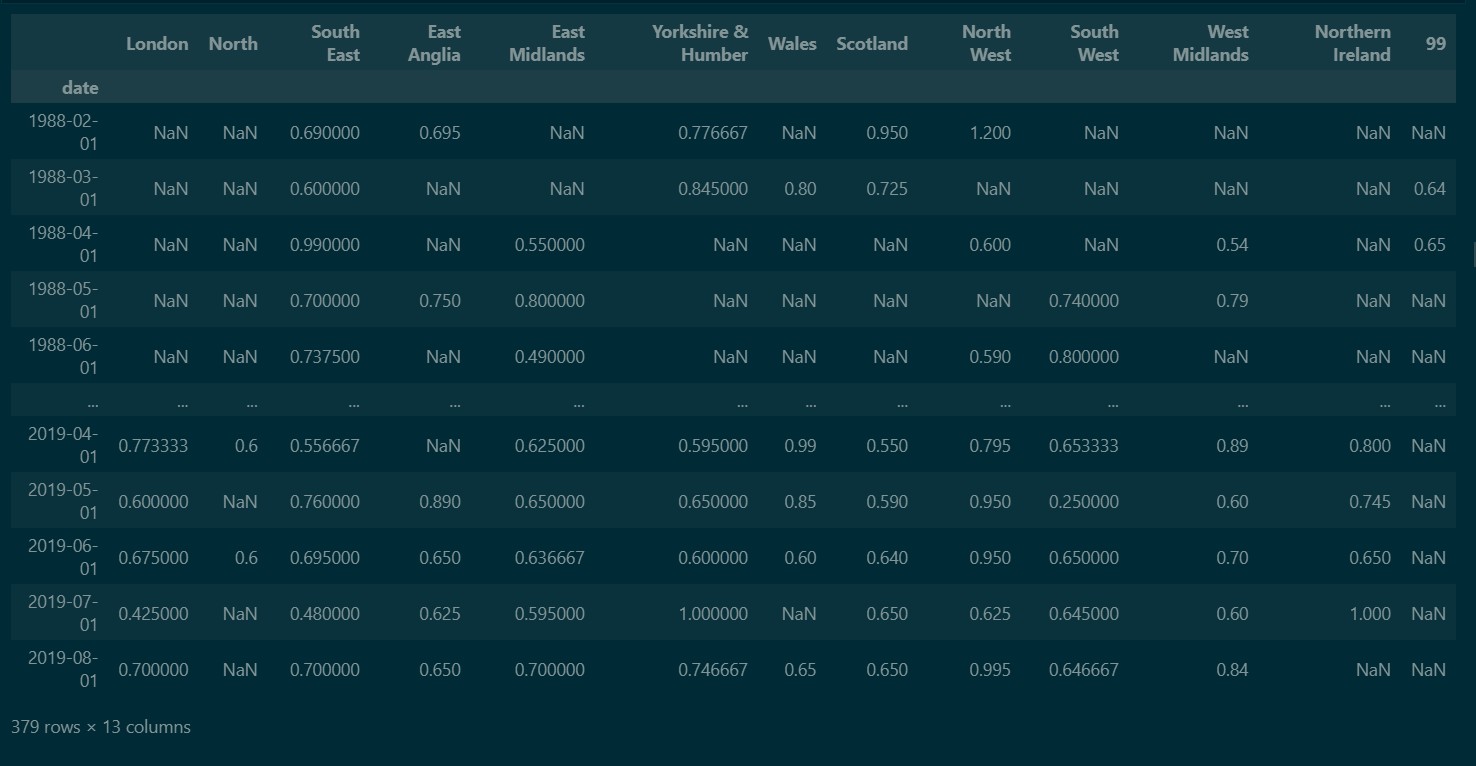


**MODELLING**

After exploring and cleaning our data, we proceeded to create models to enable us to forecast the prices of lettuce. We decided to build a time series model to implement lettuce price prediction using our clean data set. Time series forecasting is the process of making scientific predictions based on time-stamped historical data. It entails creating models through historical analysis, using them to draw conclusions and inform future strategic decision-making.

Before building our model though, there were some extra features that we needed to add to our dataset. Firstly, we created a new dataset called “df\_ts” that had the date as the index and the regions as columns, while the rows had the prices.

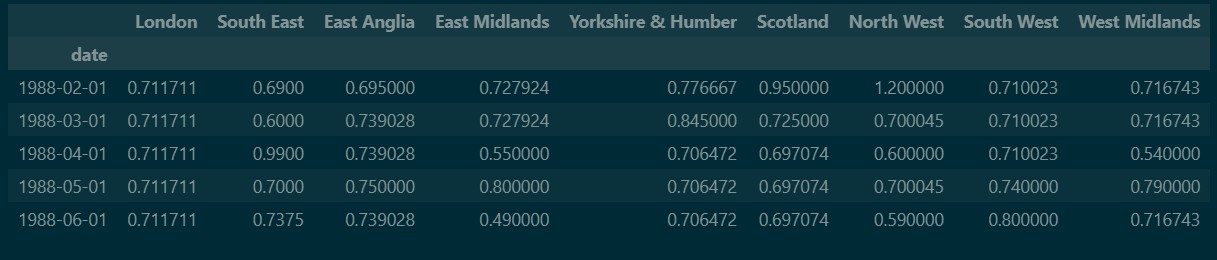
The image below shows the screenshot of the initial “df\_ts” dataset



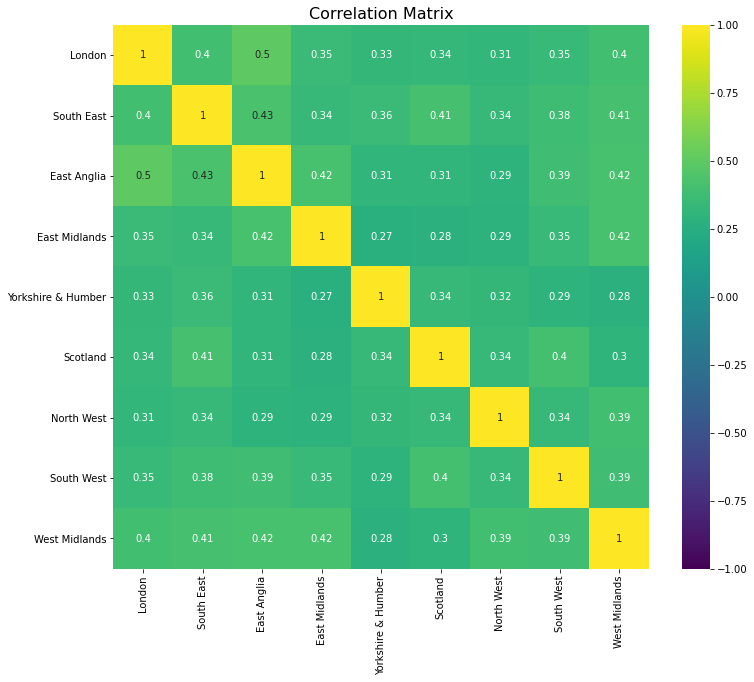
From the above, we observed that many of the regions recorded a number of null values (these are represented as NaN), and we needed to fix that or else our model would not perform optimally. We proceeded to drop the columns with too many null values. There were 4 of these - North, Wales, Northern Ireland and "99".

Then, we handled the rest of the dataframe by replacing the missing (null) values with the mean (average) of the observations in the column. We also used the “bfill” method to ensure that the frequency of the data is monthly - this was done to ensure that no months were skipped while filling in the data.

The image below shows the final “df\_ts” dataframe with complete values.



In order to prevent overfitting of our model to the data, we needed to check the correlation between the columns. The image below shows the correlation heatmap between the columns.



To understand correlation, it’s important to talk about the coefficients 1, 0, and -1. When the correlation coefficient between two columns is 1, it means there is a positive correlation between them and as one increases, the other increases. A correlation coefficient of -1 refers to a negative correlation where one decreases as the other increases, while a coefficient of 0 means that there is no correlation between the columns. From the correlation matrix above, we can see that the columns are too correlated, therefore we can proceed with building our model.

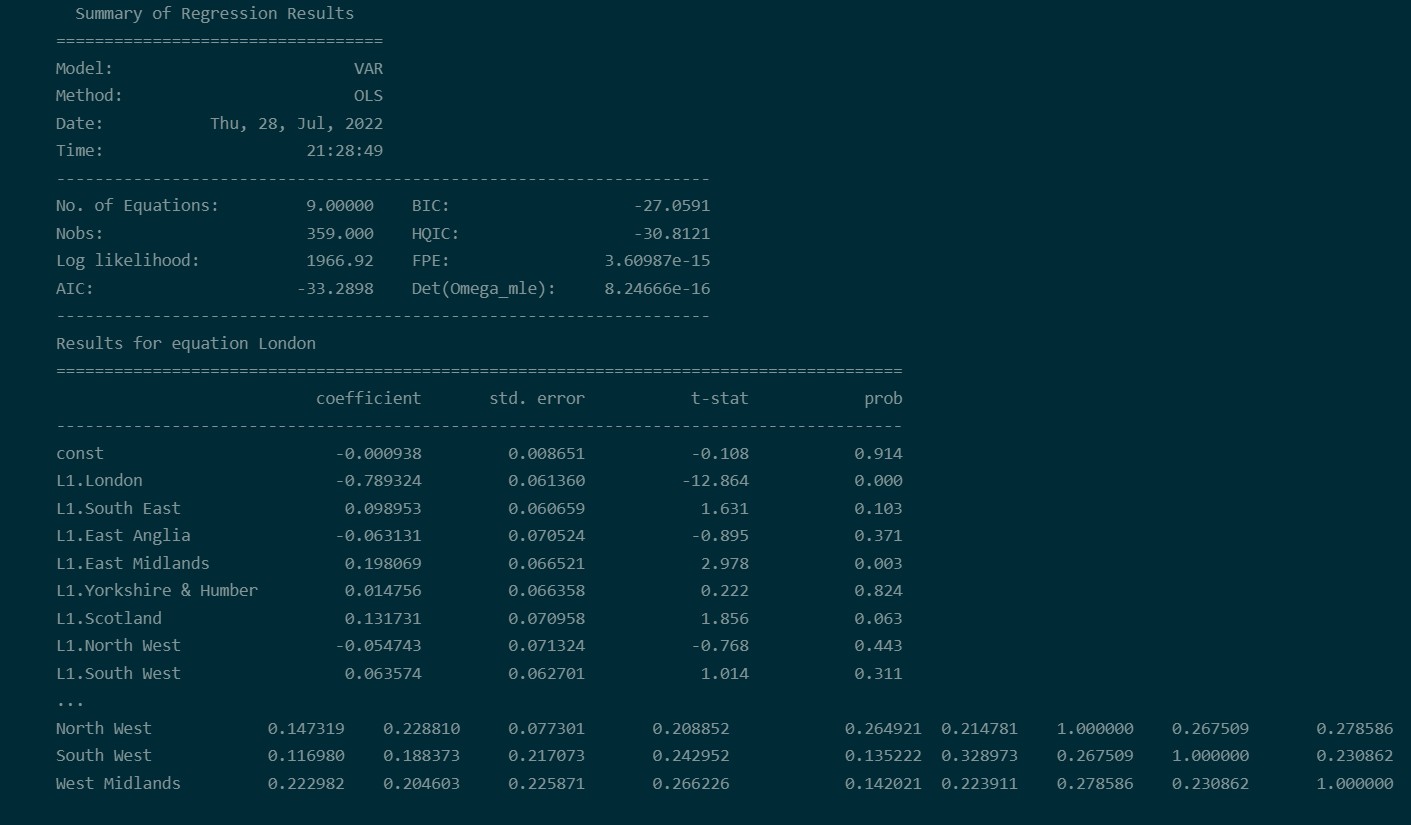
For this project, since it is a time series project, we needed a time series model capable of taking our data and delivering excellent results. We explored a couple of options, including XGBoost and ARIMA, which are single-variable autoregressivemodels and are very good models on their own, but didn’t quite fit our data. We therefore settled for the Vector Autoregressive (VAR) model. VAR models generalize the single-variable autoregressive model by allowing for multivariate time series.

The image below shows a function to check the stationarity of our data.



When building a time series model, it is imperative to check for stationarity by running a stationarity test. A stationary time series is one whose characteristics like mean and variance do not change over time. Since the VAR model requires the time series you want to forecast to be stationary, it is customary to check all the time series in the system for stationarity. If a series is found to be non-stationary, you can make it stationary by differencing the series once and repeating the test until it becomes stationary. Here we used the Augmented Dickey - Fuller Test (ADF test) to confirm the time series is stationary.

After performing differencing and ensuring that all variables passed the stationarity test, we proceeded to fit our data to the VAR model. We split our data into 2 parts, one part for training our model and the other part to test how our model will perform on new or unseen data.

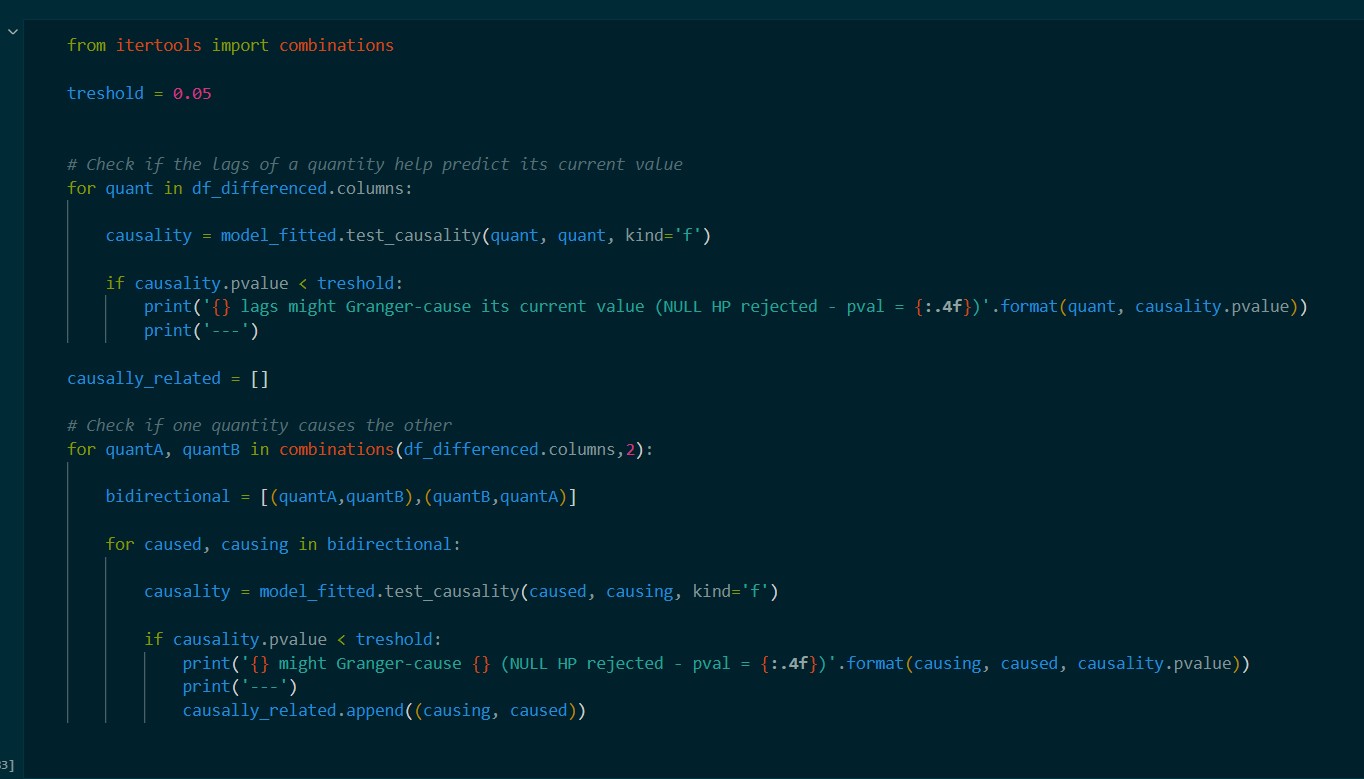
The image below shows the summary after fitting the model to our data.

Next, we proceeded to check whether there was any autocorrelation in the residuals. The presence of autocorrelation indicates there might be a pattern that our model does not capture. In this case, there is no autocorrelation in different variables, so all residuals in our variables are random processes. Serial correlation of residuals is used to check for any leftover pattern in the residuals (errors).

If there is any correlation left in the residuals, then there is some pattern in the time series that is still left to be explained by the model. In that case, the typical course of action is to either increase the order of the model, or induce more predictors into the system, or look for a different algorithm to model the time series.

Then we run the Granger causality test to check if the previous values of a region can be used to predict the future values of that region or if the values of a region can be used to forecast for another region.

The image below shows the function to run the Granger causality test.



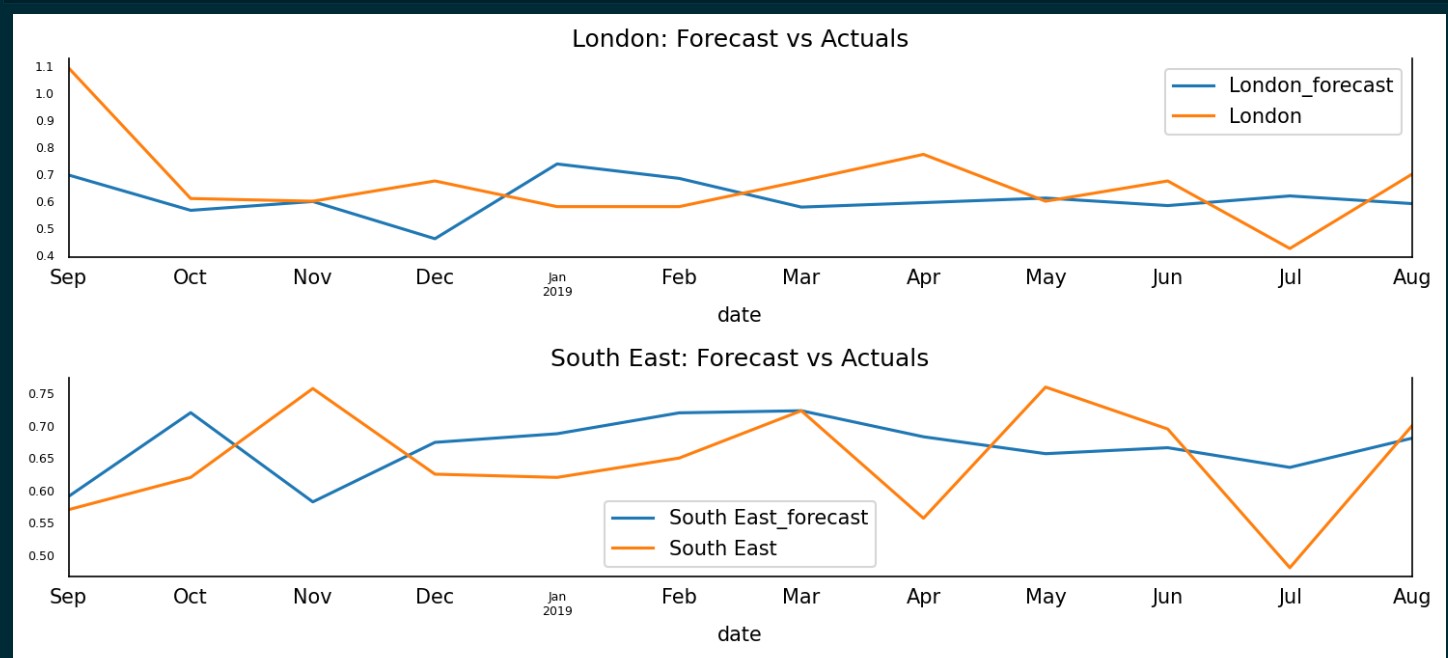
In order to forecast, the VAR model expects to see the lag order number of observations from the past data. This is because the terms in the VAR model are essentially the lags of the various time series in the dataset, so we need to provide as many of the previous values as indicated by the lag order used by the model.

The forecasts are generated but they are on the scale of the training data used by the model. So, to bring it back up to its original scale, we need to de-difference it as many times we had differenced the original input data. In this case, it is just once. Then we inverted the transformation to get the real forecast. We saved the forecasted prices to a new dataframe called “df\_results”.

The image below shows a screenshot of the “df\_results” dataframe with the forecasted prices for the regions.



We plotted some graphs to compare how our forecasted prices fared against the actual prices. The image below shows a screenshot of the plots for London and South East regions.



**MODEL EVALUATION**

Model evaluation is the process of analyzing a machine learning model's performance, as well as its strengths and weaknesses, using various evaluation metrics. Model evaluation is necessary to assess the efficacy of a model during the early stages of research, and it also plays a role in model monitoring. For this project, we used 5 different metrics to gauge the performance of our model. These are:

1. MAPE - The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics.
2. MAE - Mean Absolute Error calculates the average difference between the calculated values and actual values
3. MPE - The Mean Percentage Error (MPE) expresses forecasting errors as ratios, and they are, therefore, dimensionless and easy to interpret.
4. RMSE - Using RMSE, we can easily plot a difference between the estimated and actual values of a parameter of the model.
5. R-squared - is the proportion of the variance in the response variable that can be explained by the predictor variables in a linear regression model.

**PRICE FORECAST APPLICATION**

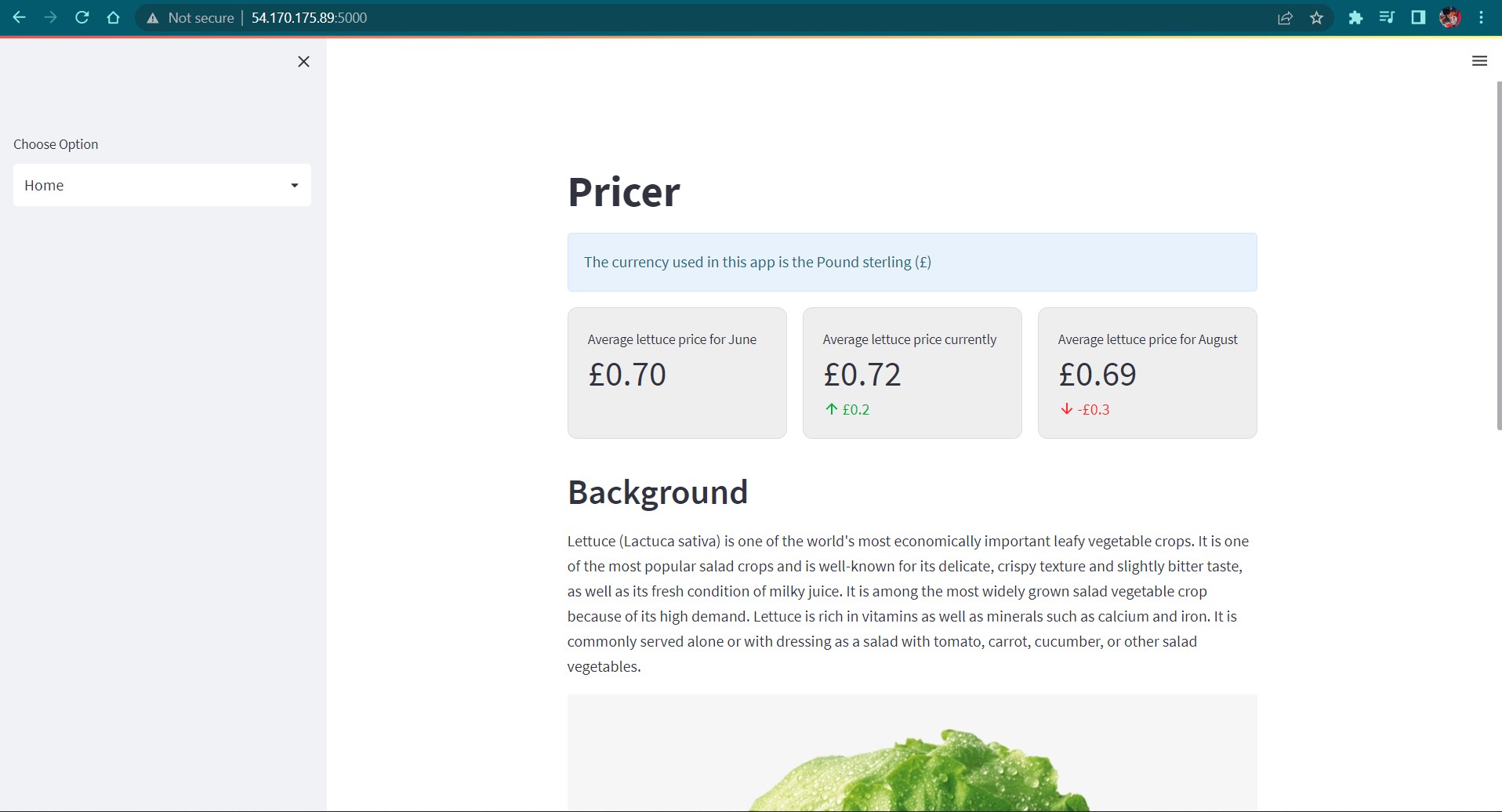
Building and training a model is a fun activity but we needed a way for other people to benefit from this amazing model we have created, a way for us to interact with the model and actually make price forecasts. This led us to building an app that uses our refined model to predict the price of lettuce.

This app was built using Streamlit and was hosted on AWS. The features of the app are:

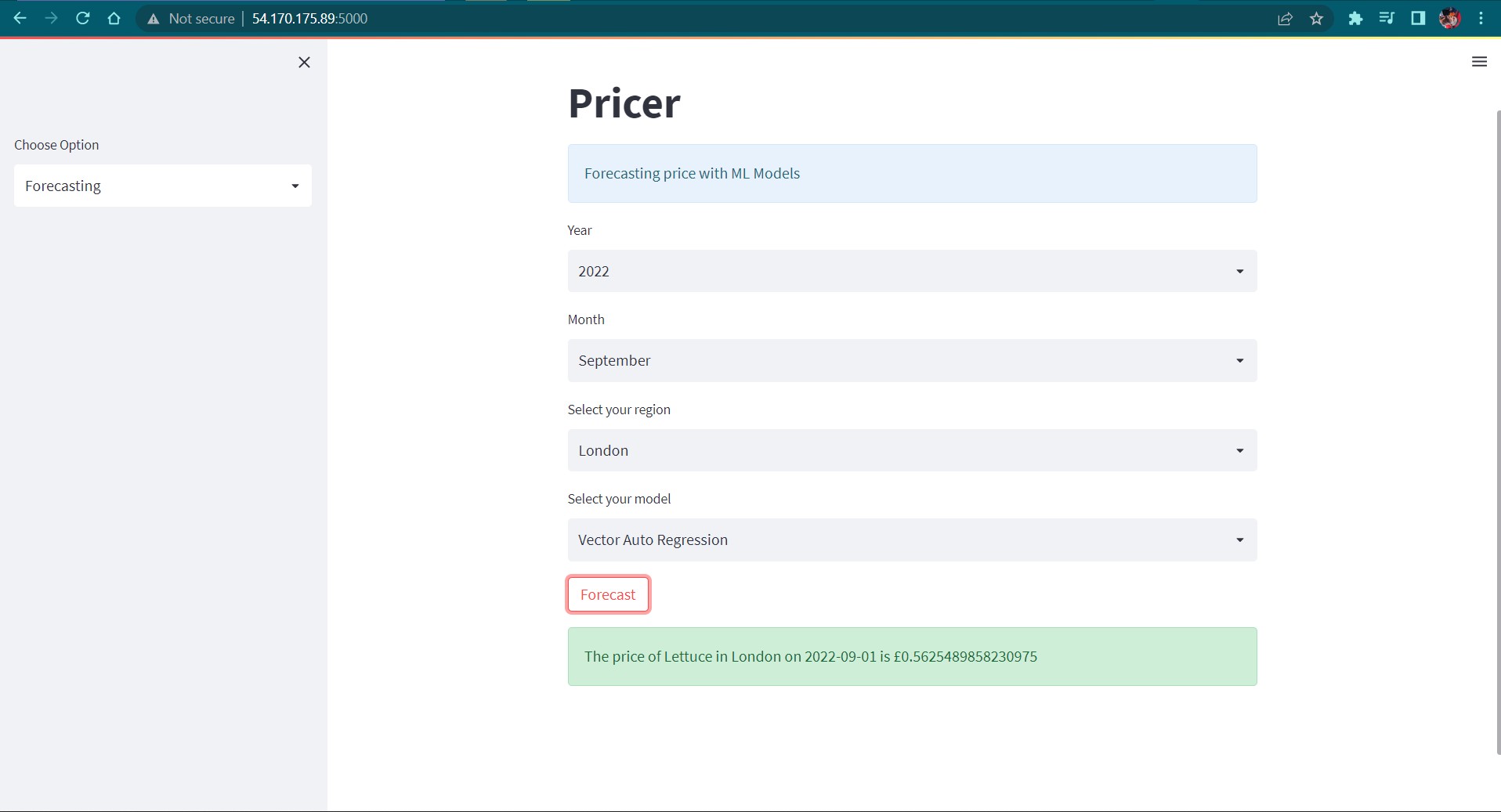
1. Eanbles users see the price of lettuce in a previous month, the present month and the next month.
2. Predict the prices of lettuce up to 4 months in the future.
3. Allows users to see useful trends and statistics related to lettuce.

The following images show sections of the app and how it is used to make predictions.

Homepage



Price forecast page



This page displays the price forecast page. To make a forecast, users will have to select the year and month of interest, select a region and a model, then click the forecast button. The Pricer app will display the result of the forecast almost immediately.

**LIMITATIONS OF THE DATA**

1. We couldn’t get data up to the present date. The dataset we used was only recorded until 2019.
2. The dataset didn’t contain information as to whether the lettuce were organic or inorganic.
3. The dataset didn’t contain any information on the volume of lettuce sold.
4. The dataset only contained prices recorded monthly so we couldn’t get predictions for particular days.
5. The dataset didn’t contain the lettuce national consumption based on official production data and population statistics.

**CHALLENGES ENCOUNTERED**

1. Finding a worthy project topic proved to be a tasking sojourn.
2. Finding datasets to work with was a really big challenge because mosts datasets we came across were not free and we were required to pay a certain amount to access them.
3. Non commitance of some team members to accomplish the project.

**REFERENCES**

Leslie, B. (2017, March 21). *The price of an iceberg lettuce in the UK increased by 67%*. Ecnmy.org. Retrieved July 3, 2022, from https://www.ecnmy.org/engage/the-price-of-an-iceberg-lettuce-in-the-uk-increased-by-67/

Marks, D. (2021, January 10). *Different types of lettuce in the UK - GardenFocused.co.uk*. Garden Focused. Retrieved July 3, 2022, from https://www.gardenfocused.co.uk/vegetable/lettuce/different-types.php

Petersson, J. (2019, February 26). *Price Forecasting: Applying Machine Learning Approaches to Electricity, Flights, Hotels, Real Estate, and Stock Pricing*. AltexSoft. Retrieved July 3, 2022, from https://www.altexsoft.com/blog/business/price-forecasting-machine-learning-based-approaches-applied-to-electricity-flights-hotels-real-estate-and-stock-pricing/