**Fuel Price Prediction**

**In New South Wales, Australia**

Contents

[**I.** **Background and Motivation** 3](#_Toc169633554)

[**II.** **Materials for Analysis** 3](#_Toc169633555)

[**1.** **Data Overview** 3](#_Toc169633556)

[**2.** **Data Quality** 4](#_Toc169633557)

[**III.** **Analytic Approach** 7](#_Toc169633558)

[**1.** **Clustering**: 7](#_Toc169633559)

[**2.** **Forecasting** 7](#_Toc169633560)

[**IV.** **Findings** 7](#_Toc169633561)

[**1.** **Clustering** 7](#_Toc169633562)

[**2.** **Forecasting** 9](#_Toc169633563)

[**V.** **Discussion** 11](#_Toc169633564)

[**VI. Recommendation** 12](#_Toc169633565)

[**VI.** **References** 13](#_Toc169633566)

[**VII.** **Appendix** 14](#_Toc169633567)

# **Background and Motivation**

Fuel price prediction is crucial for the energy sector, as it helps stakeholders, including petrol sellers, buyers, and government officials, anticipate market trends and make informed decisions. This project focuses on developing a predictive model for forecasting U91 fuel prices in New South Wales (NSW) given a dataset of prices within 2023. The model aims to forecast pump prices in a close future together with their prediction ranges at 95% level of confidence.

The project's scope involves collecting and analysing relevant datasets, such as historical U91 fuel prices, crude oil prices, exchange rates, USD/AUD rates, input cost, and demand-supply factors. By employing both unsupervised and supervised analytic techniques, the project attempts to generate accurate predictions for U91 fuel prices, taking into account both internal and external market factors.

Using ARIMA and Damped Trend Exponential Smoothing, this paper focuses on time series forecasting techniques and leverage the best model for price prediction based on evaluating a variety of performance indicators. By doing so, the project seeks to supplement the understanding of NSW’s retail petrol market and finally arrives at actionable insights for stakeholders in the energy sector.

# **Materials for Analysis**

## **Data Overview**

The data exploration for fuel price prediction, initially, integrates datasets from reputable sources such as the Australian Bureau of Statistics (ABS), Energy Information Administration (EIA), NSW Government website Fuel Check, Federal Reserve Economic Data (FRED), Reserve Bank of Australia (RBA), and the Australian Institute of Petroleum (AIP). These datasets include variables such as address, suburb, postcode, brand, Terminal Gate Prices (TGP), exchange rate, USD rate, and various regional and locational demographic features, and subsequently are merged based on NSW’s statistical areas (SA4) and dates ranging from January 1, 2023, to December 31, 2023.

To understand the factors affecting fuel prices in NSW's areas, the analysis focuses on connecting prices with locational demographic features. This involves generating new variables covering the total number of stations and examining retail price variations, including minimum, maximum, quantiles, mean, median, standard deviation, skewness, and kurtosis, for each SA4 location. The new variables, together with other locational features (which were earlier acquired from ABS website), characterise the full picture of influential factors and price dispersion in a single SA4 area, which will be used for later clustering analysis.

Details of the newly created variables can be found in Appendix.

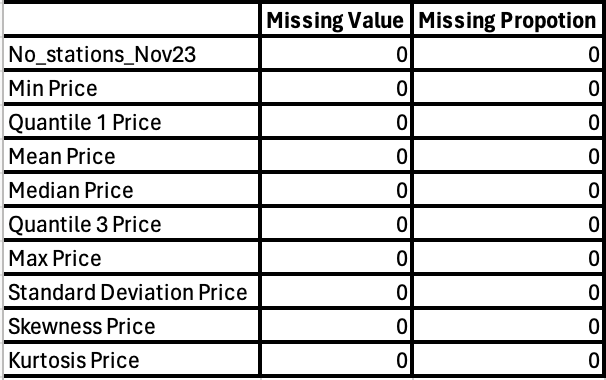
This comprehensive exploration aims to uncover insights into the factors influencing fuel prices by leveraging a wide array of socio-economic, geographical, and transportation-related variables. Through this process, patterns, correlations, and trends within the data will be identified, providing valuable inputs for the subsequent development of the fuel price prediction model.

| # | New variable | Brief description |
| --- | --- | --- |
| 1 | NUMBER\_STATIONS\_SA4 | The number of petrol stations in each SA4 location as of November 2023. Because the number of stations could be changed during the time window (2023) due to closure or new opening, the cutoff of Nov-2023 is to ensure only the recent stations are taken into calculation. |
| Features of petrol price variability: | | |
| 2 | Min | The minimum price of U91 petrol observed within each SA4 region during the specified time period. |
| 3 | Max | The maximum price of U91 petrol observed within each SA4 region during the specified time period. |
| 4 | Quantile 1 | The first quartile (25th percentile) of U91 petrol prices observed within each SA4 region during the specified time period. |
| 5 | Mean | The average price of U91 petrol observed within each SA4 region during the specified time period. |
| 6 | Median | The median price of U91 petrol observed within each SA4 region during the specified time period. |
| 7 | Quantile 3 | The third quartile (75th percentile) of U91 petrol prices observed within each SA4 region during the specified time period. |
| 8 | Standard deviation | The standard deviation of U91 petrol prices observed within each SA4 region during the specified time period. |
| 9 | Skewness | The skewness of the distribution of U91 petrol prices observed within each SA4 region during the specified time period. |
| 10 | Kurtosis | The kurtosis of the distribution of U91 petrol prices observed within each SA4 region during the specified time period. |

## **Data Quality**

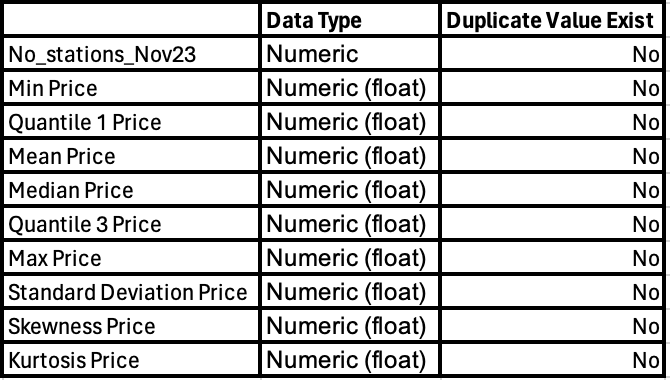
Below is New Variable Level Analysis:

* **Completeness**: Checked for missing values in the dataset and calculated the proportion of missing data for each variable.



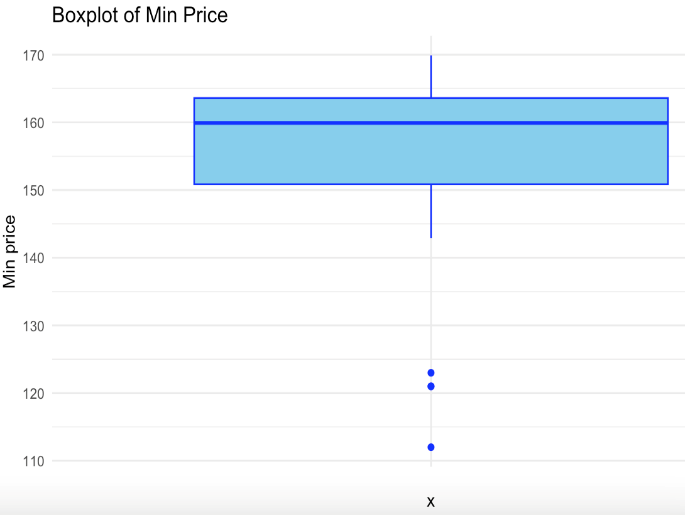
(Table 1. Missing Values)

* **Consistency**: Look for inconsistencies in data formats or units across variables, will check unique data types of each variable and check for duplicate records within the dataset. The dataset was gathered from various reputable sources such as Australian Bureau of Statistics (ABS), Energy Information Administration (EIA), NSW Government websites, FuelCheck, Federal Reserve Economic Data (FRED), Reserve Bank of Australia (RBA), and the Australian Institute of Petroleum (AIP) who have unique data types and no duplicate value exist.

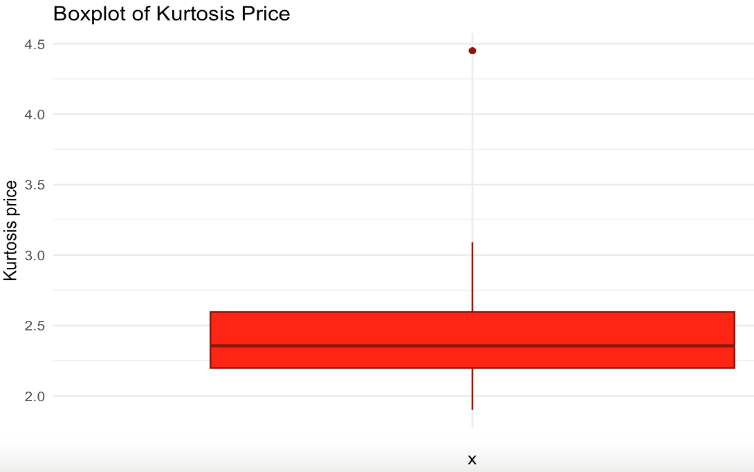


(Table 2. Unique data type and check duplicate value)

* **Accuracy**: For accuracy, check outliers in quantitative variables which are No\_stations\_Nov23, and patrol price variability features. The presence of outliers in min, max and kurtosis price (Figure 1 & 2), which represents sudden and significant fluctuations in the market, can be attributed to various factors: Events such as wars, political instability, or trade disputes can disrupt global supply chains and lead to sudden changes in fuel prices. Changes in supply and demand for oil and petroleum products, influenced by factors like economic growth, seasonal demand variations, or changes in consumer behaviour, can lead to price fluctuations. Changes in government policies, regulations, or taxes related to the energy sector can impact fuel prices by affecting production costs or market competitiveness.

(Figure 1. Boxplot for Min Price & Max Price)



(Figure 2. Boxplot for Kurtosis Price)

* **Timeliness:** The data within 2023 was used to make predictions. For clustering task, the combined data from 2020 to 2022 was leveraged to group regions based on their population density, demand, and average values. This helps to understand how these factors affect different groups of regions.
* **Validity:**  From above all points a trustworthy dataset for fuel prediction models can be obtained.

# **Analytic Approach**

1. **Clustering**:

* Data preparation for clustering involves selecting relevant features, including locational demographic features (e.g., population density, business density, station density, public transport usage, petrol demand, commute distances) and petrol price variability features (e.g., min, max, quantiles, mean, median, standard deviation, skewness, kurtosis).
* The selected features are normalized or standardized to ensure similar scales and avoid biasing the clustering algorithm. The optimal number of clusters (k) for the k-means algorithm is determined using techniques like the elbow method.
* The k-means clustering algorithm is then applied by initializing k centroids randomly or using a specific initialization method (e.g., k-means++). Each SA4 region is assigned to the nearest centroid based on the Euclidean distance between the region's feature values and the centroid's coordinates. The centroids are updated by calculating the mean feature values of the regions assigned to each cluster.
* This process is repeated until the centroids no longer change significantly or a maximum number of iterations is reached. The elbow method suggests that dividing the 28 locational areas of NSW into 2 or 3 clusters is optimal based on the selected features. However, for this analysis, 2 clusters will be used. [Refer to Appendix]

1. **Forecasting**

* Based on the results of clustering, average prices for each cluster are calculated and are used as input for subsequent predictive models. In each cluster, price patterns (trend, cycle, and seasonality) are later explored, aiming to identify relevant forecasting techniques.
* Focusing on methods that can deal with time series variables, the following 2 techniques have arrived, including ARIMA (autoregression integrated moving average) and Damped Trend Exponential Smoothing (ES). While ARIMA offers various applications to time series prediction thanks to its built-in trend parameters, Damped Trend ES has claimed as “the benchmarking forecasting model for all others to beat” (Fildes et al., 2008). With each technique, a model will be built per single cluster, incorporating the parameters established on time components of prices.
* To ensure that constructed models can work effectively on unseen data, the dataset is partitioned into a training set (data from 01-Jan to 24-Dec-2023) and a test set (series from 25-Dec to 31-Dec-2023, i.e., 7 days = 1 week). The fitted models will be used to forecast prices in the next 7 days so that the obtained prediction can be contrasted against the actual prices in the test set.
* An array of error metrics (e.g., MAE, RMSE, MAPE), acquired from testing different models, will be used to evaluate their predictive performances and conclude at the best method for each cluster.

In overall, by exploring and assessing the strength of forecasting methods, the modeller will be able to estimate price’s future trajectory, then evaluate the feasibility of this project given the data collected.

# **Findings**

1. **Clustering**

The Euclidean distance between each SA4 region and its assigned centroid provides insights into how well the region fits into its cluster. Regions with smaller distances are more similar to their cluster's centroid, while regions with larger distances might be outliers or have unique characteristics. Analysing petrol prices aggregated by day and cluster can help identify distinct pricing patterns or behaviours across different clusters. The clustering analysis reveals two clusters, with prominent differences illustrated in the table below.

A green and white sheet with numbers and text

Description automatically generated

* **Cluster 1: Predominantly Regional Areas**

Characteristics: Lower population and business density, lower public transport usage, and moderate to high petrol price variability.

Examples: Capital Region, Central Coast, and Central West.

Insights: These areas have less population and business activity, leading to higher dependence on private vehicles and potentially longer commuting distances, resulting in higher petrol price variability. On average, Cluster 1 (Predominantly Regional**)** has lower petrol prices (i.e., min price) during a downtrend with smaller variability (SD), meaning the price disparity among regions in this Cluster is less severe than in the other Cluster.

* **Cluster 2: Urban and Suburban Areas**

Characteristics: Higher population and business density, higher public transport usage, and moderate to low petrol price variability.

Examples: Sydney - Blacktown, Sydney - City and Inner South, and Sydney - Eastern Suburbs.

Insights: These areas are more urbanized with a higher reliance on public transportation. Better infrastructure and competition somewhat stabilize petrol price variability. Generally, compared to Cluster 1, petrol prices in Cluster 2 (Urban and Suburban**)** are higher when prices go into a downtrend, the price dispersion over the time is wider. In addition, there seems to be fierce competition during a downtrend with price markdown happening in many locations.

The discrimination of 28 statistical areas (SA4) into clusters is detailed in the following tables:

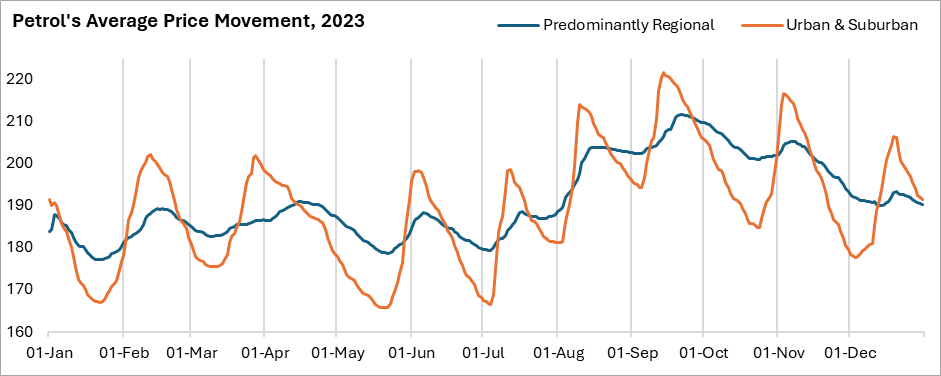
| **#** | **SA4\_Name** | **Cluster** |  | **#** | **SA4\_Name** | **Cluster** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Capital Region | 1 |  | 19 | Sydney - Blacktown | 2 |
| 2 | Central Coast | 1 |  | 20 | Sydney - City and Inner South | 2 |
| 3 | Central West | 1 |  | 21 | Sydney - Eastern Suburbs | 2 |
| 4 | Coffs Harbour - Grafton | 1 |  | 22 | Sydney - Inner South West | 2 |
| 5 | Far West and Orana | 1 |  | 23 | Sydney - Inner West | 2 |
| 6 | Hunter Valley exc Newcastle | 1 |  | 24 | Sydney - North Sydney and Hornsby | 2 |
| 7 | Illawarra | 1 |  | 25 | Sydney - Northern Beaches | 2 |
| 8 | Mid North Coast | 1 |  | 26 | Sydney - Parramatta | 2 |
| 9 | Murray | 1 |  | 27 | Sydney - Ryde | 2 |
| 10 | New England and North West | 1 |  | 28 | Sydney - South West | 2 |
| 11 | Newcastle and Lake Macquarie | 1 |  |  |  |  |
| 12 | Richmond - Tweed | 1 |  |  |  |  |
| 13 | Riverina | 1 |  |  |  |  |
| 14 | Southern Highlands and Shoalhaven | 1 |  |  |  |  |
| 15 | Sydney - Baulkham Hills and Hawkesbury | 1 |  |  |  |  |
| 16 | Sydney - Outer South West | 1 |  |  |  |  |
| 17 | Sydney - Outer West and Blue Mountains | 1 |  |  |  |  |
| 18 | Sydney - Sutherland | 1 |  |  |  |  |

1. **Forecasting**

**Model construction**

Adopting the results of clustering, the modeller calculated the average petrol prices of the 2 clusters (**Predominantly Regional** and **Urban & Suburban**), and plotted the aggregated prices into one graph as in below graph.

Prices in both the regions showcase multiple cycles of peaks and troughs throughout 2023. In other words, they share the same upward or downward patterns. However, price ranges between peaks and troughs are much larger for Urban & Suburban (orange line), making the trend of prices more obvious in this region.



From the above graph, the period of each cycle is different from one to another, hereby, seasonal component (or the frequency of cycle) is not detected. Recalling 2 methods used in this section (ARIMA and Damped Trend ES), they are able to handle data following a trend without seasonality, which makes these applied methods well-suited to predict prices in this scenario.

The variables for modelling are “**Date**” and “**Average price**”, which are required by the design of time series forecasting methods. Technically, it is assumed that price itself is autocorrelated, meaning the price of the past is the predictor of it in future. The use of macroeconomic (crude oil price) and industrial factor (terminal gate price) as covariance for time series forecast can complicate the analysis, therefore, is not adopted.

Using R and SAS Enterprise Miner, the training set, containing average petrol prices from 1-Jan to 24-Dec-2023, are fitted into 2 models for each of the 2 regions. The optimal parameters of **4 models** are shown in the following table. Notably, within the scope of this analysis, technical details of the models’ construction will not be presented.

A screenshot of a computer

Description automatically generated

* For ARIMA: (p, d, q) represent 3 components of the model, which are autoregression (p), integration (d), and moving average (q).
* For Damped Trend ES: (a, b, f) speak for 3 model parameters, including level smoothing (a), trend smoothing (b), and autoregressive-damping (f).

**Model performance**

After the models’ training is completed, the fitted models are employed to forecast average petrol prices between 25-Dec and 31-Dec-2023 (i.e., 7 days = 1 week). Then, the obtained forecasts are compared against the actual prices (i.e., test data) to evaluate the performances of those models. The discrepancy between forecasts and real-world values is used to calculate a variety of error rates, as summarised in the following table. In this regard, model with smaller error rates is considered to produce more accurate results.

A green and white text

Description automatically generated

*(\*)* ***ME:*** *Mean Error;* ***RMSE:*** *Root Mean Square Error;* ***MAE:*** *Mean Absolute Error;* ***MAPE:*** *Mean Absolute Percentage Error;* ***MASE:*** *Mean Absolute Scaled Error.*

Based on the above table, in both clusters (regions), ARIMA method outperforms Damped Trend ES in most of the provided error metrics (except for ME). Because ME is less commonly used for model comparison than the other ones, in this analysis ARIMA stand as a better technique to predict petrol prices in NSW. Besides, between the 2 ARIMA models, the error rates (except for ME) by the model in Urban & Suburban cluster are smaller than that of the other ARIMA, indicating that the prediction produced in Urban & Suburban areas are more accurate.

The forecast values, including Mean, Lower bound and Upper bound at confidence interval 95%, are exhibited along with the actual prices in the below table. It is noted that prices for prediction represent average petrol prices in each cluster. Thus, forecast’s mean value is interpreted as the expected future value of average petrol price, while lower and upper bounds form prediction range of that expected future value.

A table with numbers and a few lines

Description automatically generated with medium confidence

To better observe the forecast prices, I plot them along with the real-world values, which are shown in the below graphs. As can be seen, the orange dots are forecast’s mean value, while grey area represents its 95% confident range. Between the 2 plots, the ARIMA model for Urban & Suburban cluster appears more accurate since its forecasts lie very close to the actual prices. In Predominantly Regional cluster, although the model correctly captures the downtrend of petrol prices, its yielded prediction is conspicuously larger than the real values.

A graph showing the growth of the weather

Description automatically generatedrr

A graph of a graph

Description automatically generated with medium confidence

# **Discussion**

The project focuses on the method of Clustering and Predictive Modelling to forecast the prices of petrol in numerous zones of New South Wales (NSW) in Australia. The main purpose is to understand and capture the price movement for a given calendar year 2023, which will provide knowledge to the stakeholders and help in better decision making.

1. **Clustering**

The dataset used for this project consists of daily petrol prices, and the region’s attributes as input variables. Thanks to the clustering analysis, 28 statistical areas of NSW are systematically grouped into classes of two using k-means clustering. This kind of approach helps identify regions having similar socio-economic and geographic profiles, which can then be used to understand how these influence petrol price variations.

**Cluster 1: Predominantly Regional** consists of 18 regions characterised by lower population and business densities. Because of larger area, public transport system is not widespread, causing residents to travel in longer journeys. As a result, the choice of petrol stations is limited, leading to less fluctuations of prices over the time.

**Cluster 2: Urban and Suburban** includes 10 regions with higher population density, more business activities and larger number of petrol stations. The areas scatter around the capital city of NSW; therefore, public transportation is more popular, and people commute in shorter trips in general. The petrol prices in these locations have wider variability, suggesting there be numerous highs and lows together with more competition in retail prices.

When petrol price data are split into 2 subsets in accordance with the 2 clusters, price movement in each cluster demonstrates a distinct pattern during the time window. The number of highs and lows is similar between 2 clusters, suggesting that when external factors cause petrol prices to move, the direction of that movement (up or down) is simultaneously observed in all the locations.

However, for stations in Predominantly Regionalcluster, price trajectories are flatter, i.e., the differences between peak and bottom prices are relatively narrow. This is derived from the fact that there are fewer stations (less competition), scarce population, limited business activities and public transportation. The drivers for significant price changing, therefore, are not strong according to the results of this analysis.

Meanwhile, stations in Urban and Suburbancluster encounter more intense factors which impact their retail prices. This includes competition from many surrounding stations and easier accessibility to public transport. Generally, they pursue a pricing strategy that involves extremely low prices for increasing market share, together with critically high prices for maximising profit.

1. **Forecasting**

The two forecasting techniques used in this analysis, including ARIMA and Damped Trend Exponential Smoothing, perform quite well on the data assigned to them. The error rates obtained from both methods are relatively close to one another. With better outcomes in all clusters (measured on RMSE, MAE, MAPE, and MASE), the ARIMA method proves to be more effectively capture price patterns and accurately predict future values.

However, the ARIMA method is not equally good at forecasting future petrol prices. The larger error rates from modelling the Predominantly Regionalcluster indicate less accurate prediction compared with the model of Urban and Suburbancluster. Looking at the graphical presentation of prices, it is obvious that the cyclical patterns in Urban and Suburban locations have relatively similar shapes, with prices moving down slowly but increasing in quicker pace. The ranges between peaks and bottoms are quite consistent among cycles. Meanwhile, in Predominantly Regional areas, the speed of price up and down, the length of each downturn and upturn, and the gap between highest and lowest are different in each price cycle. The time components of prices in those remote locations, therefore, are more difficult for the time series model to capture and project forecasts. As a result, predictions for prices in Predominantly Regional areas are less precise.

Both models were able to depict a general decline in the prices of petrol towards the end of December 2023 across both clusters with the detailed modelling showing a 95% window for each day's estimated price. The analytical approach used in this project creates a robust platform for monitoring and forecasting petrol price dynamics in NSW. The clustering and forecasting models showed not only regional disparities but also have insights for decision-making and operative responses that could be actionable by multiple parties in their operations and policies to petrol prices under market changes. To further improve the accuracy of forecasts, it is important to continue enhancing the data inputs and modelling techniques in the future.

# **VI. Recommendation**

1. This analysis emphasises on forecasting average petrol prices in 2 groups of geographical areas. The obtained forecasts are indicative of expected values of retail prices in all the areas within a certain cluster. This approach, therefore, disregards price patterns of a particular brand, suburb, or service station. Further experiment with relevant data granularity can be employed to acquire broader knowledge of price variations for the area of interest.
2. In this analysis, the 2 groups of NSW’s regions are agreed, based on k-means clustering method and use this result for subsequent predictive models. Another experiment can start with different clustering techniques, such as hierarchical, density-based, and grid-based methods, to explore their results and impacts on petrol price’s final forecasts.
3. The applied forecasting methods encompass “Date” and “Price” as input variables under the assumption that prices are autocorrelated and historical prices are able to predict the future. This implementation puts aside other factors which are believed to influence retail prices (such as wholesale prices). Further research can embrace ensemble techniques which treat concerned factors as covariance, to ultimately improve modelling performance.
4. To fit ARIMA (as well as other time series methods), model parameters and coefficients are estimated upon the time components of historical observations. For better and more regular use, predictive models should be frequently re-trained when there are more data observed and recorded. By doing so, predictive models can better capture new patterns arising from recent data and as a result, produce more accurate and reliable forecasts of petrol prices.

# **References**

Australian Competition and Consumer Commission. (n.d.). What affects fuel prices. <https://www.accc.gov.au/consumers/petrol-and-fuel/what-affects-fuel-prices>

Australian Institute of Petroleum, Commissioned Report March 2020 <https://www.aip.com.au/>

Australian Bureau of Statistics

<https://www.abs.gov.au/statistics/industry/tourism-and-transport/sales-new-motor-vehicles/latest-release>

BYRNE, D. P., NAH, J. S. & XUE, P. 2018. Australia Has the World's Best Petrol Price Data: FuelWatch and FuelCheck. *Australian Economic Review,* 51**,** 564-577.

Fildes, R., Nikolopous, K., Crone, S., & Syntetos, A. (2008). Forecasting and operational research: A review. Journal of the Operational Research Society, 59, 1–23.

Ghazanfari, A., 2022. What Drives Petrol Price Dispersion across Australian Cities? *Energies*,*15*(16), p.6025

GHAZANFARI, A. 2020. *Analysis of the retail petrol price patterns in Western Australia.* RMIT University.

Isakower, S & Wang, Z 2014, ‘A comparison of regular price cycles in gasoline and liquefied petroleum gas’, *Energy Economics*, vol. 45, pp. 445–454.

Mitchell, JD, Lian Ong, L & Izan, HY 2000, ‘Idiosyncrasies in Australian petrol price behaviour: evidence of seasonalities’, *Energy Policy*, vol. 28, no. 4, pp. 243–258.

OCZKOWSKI, E., WONG, A. & SHARMA, K. 2018. The impact of major fuel retailers on regional New South Wales petrol prices. *Economic Analysis and Policy,* 57**,** 44-59.

Shafiee, S., & Topal, E. (2010). "A long-term view of worldwide fossil fuel prices." Applied Energy, 87(3), 988-1000. ISSN 0306-2619. [Online] Available at: <https://doi.org/10.1016/j.apenergy.2009.09.012>.

VALADKHANI, A., ANWAR, S., GHAZANFARI, A. & NGUYEN, J. 2021. Are petrol retailers less responsive to changes in wholesale or crude oil prices when they face lower competition? The case of Greater Sydney. *Energy Policy,* 153**,** 112278.

Valadkhani, A, Ghazanfari, A, Nguyen, J & Moradi-Motlagh, A 2021, ‘The asymmetric effects of COVID19 on wholesale fuel prices in Australia’, *Economic Analysis and Policy*, vol. 71, pp. 255–266.

Valadkhani, A., 2013. Seasonal patterns in daily prices of unleaded petrol across Australia. *Energy policy*, *56*, pp.720-731.

# **Appendix**

1. **Input Variable for Clustering Analysis**

Below are the dataset incorporating demographic features of 28 statistical areas (SA4) and their price variability.

| Cluster | Location (SA4\_Name) |  | Locational demographic features | | | | | | | | No\_  stations  \_Nov23+ | Petrol price variability features | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Pop\_density  \_2022 | Biz\_density  \_2022 | Station\_  density | Public\_  trans | Petrol\_  demand | Commute\_  dist\_avg\_1 | Commute\_  dist\_med\_1 | Commute\_  dist\_avg\_2 | Commute\_  dist\_med\_2 | Min | Qu. 1 | Mean | Median | Qu. 3 | Max | SD | Skewness | Kurtosis |
| 1 | Capital Region | 4.60 | 0.45 | 0.00 | 0.00 | 0.57 | 20.80 | 10.60 | 17.20 | 5.80 | 95 | 123.00 | 184.90 | 192.85 | 191.90 | 199.90 | 259.70 | 10.48 | 0.06 | 4.45 |
| 1 | Central Coast | 207.70 | 15.36 | 0.05 | 0.01 | 0.58 | 28.30 | 14.40 | 15.40 | 10.40 | 78 | 160.90 | 182.90 | 194.78 | 193.90 | 205.90 | 229.90 | 15.04 | 0.32 | 2.28 |
| 1 | Central West | 3.00 | 0.31 | 0.00 | 0.00 | 0.54 | 17.40 | 4.70 | 17.30 | 4.70 | 95 | 163.50 | 181.90 | 191.25 | 190.90 | 199.90 | 227.90 | 11.85 | 0.17 | 2.35 |
| 1 | Coffs Harbour - Grafton | 11.20 | 0.91 | 0.01 | 0.00 | 0.55 | 14.30 | 6.80 | 13.80 | 6.80 | 70 | 167.70 | 182.90 | 191.22 | 189.90 | 199.90 | 232.50 | 11.07 | 0.52 | 3.09 |
| 1 | Far West and Orana | 0.30 | 0.04 | 0.00 | 0.00 | 0.48 | 16.30 | 3.80 | 16.60 | 3.80 | 72 | 163.90 | 183.90 | 193.56 | 190.90 | 203.90 | 256.90 | 12.04 | 0.40 | 2.61 |
| 1 | Hunter Valley exc Newcastle | 13.90 | 0.95 | 0.00 | 0.00 | 0.55 | 24.50 | 16.20 | 23.50 | 14.50 | 102 | 142.90 | 177.90 | 188.67 | 186.90 | 198.90 | 229.90 | 14.55 | 0.46 | 2.70 |
| 1 | Illawarra | 206.40 | 13.88 | 0.04 | 0.01 | 0.57 | 22.10 | 11.10 | 13.60 | 8.50 | 68 | 159.90 | 183.90 | 196.15 | 195.90 | 208.90 | 229.90 | 16.09 | 0.16 | 2.19 |
| 1 | Mid North Coast | 12.30 | 0.96 | 0.00 | 0.00 | 0.58 | 15.40 | 6.60 | 14.50 | 6.40 | 88 | 162.90 | 179.10 | 187.97 | 187.90 | 195.90 | 228.90 | 10.80 | 0.41 | 2.78 |
| 1 | Murray | 1.30 | 0.13 | 0.00 | 0.00 | 0.47 | 15.30 | 5.80 | 15.70 | 6.20 | 68 | 169.90 | 183.90 | 192.08 | 189.90 | 199.90 | 223.00 | 11.14 | 0.24 | 2.21 |
| 1 | New England and North West | 1.90 | 0.21 | 0.00 | 0.00 | 0.49 | 15.80 | 4.70 | 16.10 | 4.70 | 109 | 165.90 | 181.50 | 190.10 | 188.00 | 199.00 | 229.70 | 10.89 | 0.38 | 2.27 |
| 1 | Newcastle and Lake Macquarie | 455.20 | 33.75 | 0.11 | 0.01 | 0.59 | 16.90 | 9.10 | 15.50 | 9.40 | 99 | 121.00 | 177.50 | 189.89 | 186.80 | 201.90 | 229.90 | 16.05 | 0.51 | 2.38 |
| 1 | Richmond - Tweed | 25.20 | 2.57 | 0.01 | 0.00 | 0.55 | 17.30 | 9.30 | 15.60 | 8.50 | 86 | 169.90 | 187.90 | 197.55 | 196.90 | 205.90 | 229.90 | 12.13 | 0.29 | 2.59 |
| 1 | Riverina | 2.90 | 0.30 | 0.00 | 0.00 | 0.51 | 13.40 | 5.20 | 13.40 | 5.20 | 96 | 158.90 | 179.90 | 188.45 | 185.90 | 197.50 | 225.00 | 10.93 | 0.37 | 2.38 |
| 1 | Southern Highlands and Shoalhaven | 24.20 | 2.15 | 0.01 | 0.00 | 0.61 | 23.60 | 8.70 | 18.10 | 8.20 | 60 | 164.50 | 179.90 | 188.14 | 186.90 | 196.70 | 221.90 | 10.61 | 0.46 | 2.48 |
| 1 | Sydney - Baulkham Hills & Hawkesbury | 83.50 | 10.06 | 0.01 | 0.01 | 0.56 | 18.20 | 15.20 | 14.60 | 9.40 | 26 | 112.00 | 178.90 | 192.74 | 191.50 | 205.90 | 229.90 | 16.98 | 0.16 | 2.44 |
| 2 | Sydney - Blacktown | 1,723.30 | 128.58 | 0.25 | 0.03 | 0.49 | 18.50 | 15.50 | 16.50 | 11.80 | 60 | 159.90 | 177.90 | 191.67 | 189.90 | 204.90 | 231.90 | 17.19 | 0.21 | 2.17 |
| 2 | Sydney - City and Inner South | 5,145.10 | 1,368.27 | 0.44 | 0.05 | 0.46 | 7.10 | 4.40 | 18.00 | 12.20 | 29 | 156.80 | 175.90 | 189.25 | 186.90 | 200.70 | 229.90 | 16.33 | 0.44 | 2.36 |
| 2 | Sydney - Eastern Suburbs | 4,565.20 | 629.28 | 0.45 | 0.03 | 0.46 | 8.20 | 6.40 | 11.70 | 5.60 | 26 | 164.50 | 185.90 | 199.73 | 201.90 | 211.90 | 239.90 | 16.83 | -0.24 | 2.06 |
| 2 | Sydney - Inner South West | 3,706.50 | 404.88 | 0.67 | 0.03 | 0.50 | 12.70 | 11.80 | 12.60 | 8.20 | 110 | 143.50 | 172.90 | 184.94 | 181.90 | 194.90 | 231.90 | 15.79 | 0.68 | 2.80 |
| 2 | Sydney - Inner West | 4,741.40 | 579.90 | 0.56 | 0.05 | 0.48 | 10.20 | 8.90 | 14.10 | 9.50 | 36 | 154.80 | 174.60 | 188.74 | 185.90 | 201.90 | 229.90 | 17.76 | 0.45 | 2.20 |
| 2 | Sydney - North Sydney & Hornsby | 1,545.90 | 218.07 | 0.20 | 0.03 | 0.52 | 11.30 | 8.50 | 16.80 | 11.80 | 54 | 160.60 | 178.90 | 195.93 | 197.90 | 210.90 | 230.90 | 18.10 | -0.03 | 1.90 |
| 2 | Sydney - Northern Beaches | 1,035.80 | 130.46 | 0.12 | 0.02 | 0.56 | 12.40 | 10.60 | 11.40 | 5.70 | 30 | 154.90 | 179.90 | 194.91 | 195.90 | 209.90 | 229.90 | 17.82 | -0.01 | 1.99 |
| 1 | Sydney - Outer South West | 238.10 | 17.59 | 0.04 | 0.02 | 0.55 | 25.60 | 21.90 | 18.00 | 11.10 | 57 | 121.00 | 176.90 | 191.16 | 189.90 | 201.90 | 229.90 | 17.24 | 0.20 | 2.63 |
| 1 | Sydney - Outer West and Blue Mountains | 84.70 | 6.35 | 0.02 | 0.01 | 0.57 | 22.90 | 17.90 | 16.00 | 10.20 | 76 | 157.90 | 180.90 | 194.60 | 194.50 | 207.90 | 229.90 | 17.13 | 0.07 | 2.13 |
| 2 | Sydney - Parramatta | 3,062.30 | 360.79 | 0.53 | 0.03 | 0.50 | 13.80 | 11.80 | 17.00 | 13.50 | 87 | 148.90 | 173.50 | 187.61 | 183.90 | 199.90 | 229.90 | 17.65 | 0.59 | 2.40 |
| 2 | Sydney - Ryde | 2,955.80 | 311.46 | 0.32 | 0.03 | 0.51 | 12.10 | 11.00 | 16.60 | 12.60 | 22 | 159.90 | 173.90 | 188.80 | 187.90 | 199.90 | 229.90 | 17.05 | 0.42 | 2.29 |
| 2 | Sydney - South West | 899.30 | 75.35 | 0.15 | 0.01 | 0.51 | 17.50 | 14.60 | 16.60 | 11.30 | 81 | 151.50 | 175.80 | 189.28 | 185.90 | 201.90 | 231.90 | 17.75 | 0.44 | 2.21 |
| 1 | Sydney - Sutherland | 781.80 | 76.78 | 0.08 | 0.02 | 0.57 | 17.20 | 15.80 | 12.80 | 6.90 | 25 | 162.90 | 180.90 | 196.18 | 196.90 | 209.90 | 229.90 | 17.32 | 0.01 | 2.00 |

1. **Data Dictionary of Newly Created Variables**

Below are the engineered variables which are created from petrol prices corresponding to each SA4 location.

|  |  |
| --- | --- |
| Field Name | No\_stations\_Nov23 |
| Unique ID | NUMBER\_STATIONS\_SA4 |
| Definition | The number of petrol stations in each SA4 region as of November 2023. |
| Source | Derived from the demographic field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (Integer) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the count of petrol stations within each SA4 region, providing insight into the availability and distribution of petrol stations in different areas. This information can be used to analyze the relationship between the number of stations and petrol price variability within each region. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Min |
| Unique ID | MIN\_PRICE\_SA4 |
| Definition | The minimum price of U91 petrol observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the lowest price point within each SA4 region, providing insight into the minimum price consumers could expect to pay for U91 petrol in that area. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Max |
| Unique ID | MAX\_PRICE\_SA4 |
| Definition | The maximum price of U91 petrol observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the highest price point within each SA4 region, providing insight into the maximum price consumers could expect to pay for U91 petrol in that area. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Quantile 1 |
| Unique ID | Q1\_PRICE\_SA4 |
| Definition | The first quartile (25th percentile) of U91 petrol prices observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the price point below which 25% of the prices fall within each SA4 region, providing insight into the lower end of the price distribution. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Mean |
| Unique ID | MEAN\_PRICE\_SA4 |
| Definition | The average price of U91 petrol observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the central tendency of prices within each SA4 region, providing a baseline for comparison and analysis. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Median |
| Unique ID | MEDIAN\_PRICE\_SA4 |
| Definition | The median price of U91 petrol observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the middle value of the price distribution within each SA4 region, providing a robust measure of central tendency that is less sensitive to outliers compared to the mean. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Quantile 3 |
| Unique ID | Q3\_PRICE\_SA4 |
| Definition | The third quartile (75th percentile) of U91 petrol prices observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Represents the price point above which 25% of the prices fall within each SA4 region, providing insight into the higher end of the price distribution. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Standard Deviation |
| Unique ID | SD\_PRICE\_SA4 |
| Definition | The standard deviation of U91 petrol prices observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Measures the dispersion of prices around the mean within each SA4 region, providing insight into the variability and spread of prices in that area. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Skewness |
| Unique ID | SKEW\_PRICE\_SA4 |
| Definition | The skewness of the distribution of U91 petrol prices observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Measures the asymmetry of the price distribution within each SA4 region, indicating whether prices are skewed towards higher or lower values compared to the mean. |

|  |  |
| --- | --- |
| Field Name | Petrol price variability features - Kurtosis |
| Unique ID | KURT\_PRICE\_SA4 |
| Definition | The kurtosis of the distribution of U91 petrol prices observed within each SA4 region during the specified time period. |
| Source | Price from Australian government website. Derived from the Price field in the source database, grouped by SA4\_NAME\_2021. |
| Format | Numeric (float) |
| Scope | Applies to all SA4 regions in the dataset for the year 2023. |
| Relationships and Contexts | Measures the tailedness and peakedness of the price distribution within each SA4 region, providing insight into the presence of outliers and the concentration of prices around the central tendency. |