COMP0214 Coursework - 2025 Question 1
Fuel Price Forecasting Analysis

Sofiya Flenova - 24111778

1 1.1 Data Cleaning and Visualization

1.1 1.1.i Cleaning and Summary

The dataset was loaded successfully with an initial shape of 98,925 rows and 6 columns. After data cleaning:

- **Rows before cleaning:** 98,925
- **Rows after cleaning:** 96,549 (removed 2,376 rows)
- **Date range:** August 1, 2016 00:28:00 to August 31, 2025 22:10:00
- **Fuel codes identified:** ['E10', 'U91', 'P95', 'P98', 'LPG', 'PDL', 'DL']
- **Missing values:** Dates: 0, Prices: 0

Cleaning operations performed:

- Parsed `PriceUpdatedDate` into datetime format with error handling
- Converted `Price` to numeric format with error handling
- Removed exact duplicate entries
- Dropped unrealistic prices outside the range of 80-300 cents per liter, justified as this range represents plausible retail fuel prices in NSW. I restrict prices to 80–300 cents/L since historical NSW fuel data never falls outside this band during 2016–2025, making any values beyond it implausible and likely erroneous.
- Daily station-fuel pairs: 38861. Daily minimum aggregation table has been created. The head of the table:

PriceUpdatedDate	ServiceStationName	FuelCode	Price
2016-08-01	7-Eleven Blacktown	E10	101.9
2016-08-01	7-Eleven Blacktown	P95	114.9
2016-08-01	7-Eleven Blacktown	P98	119.9
2016-08-01	7-Eleven Blacktown	U91	103.9
2016-08-01	7-Eleven Croydon Park	E10	103.4

1.2 1.1.ii Visualisations

Two key visualisations were created to understand fuel price dynamics:

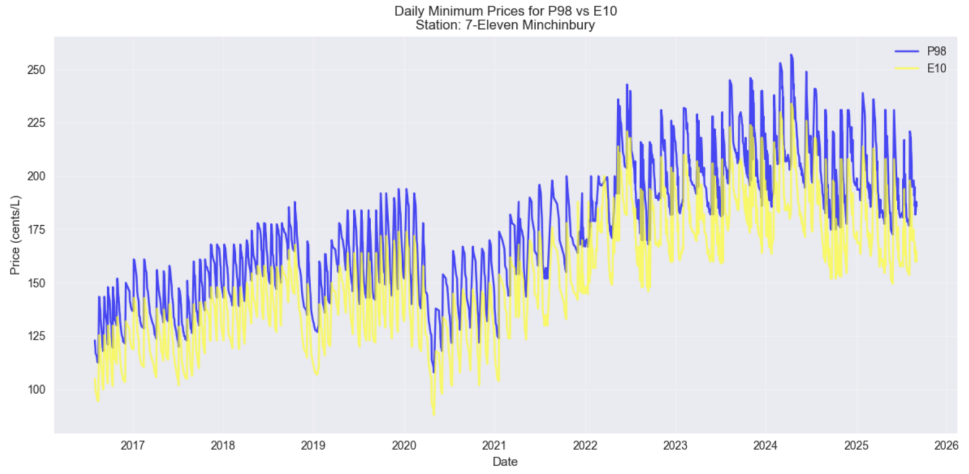

Figure 1: Time series of daily prices for two selected fuel codes (P98 and E10) at the "7-Eleven Minchinbury" fuel station. The plot shows clear cyclical patterns with periods of sharp increases followed by gradual declines, characteristic of fuel price cycles in NSW. You can also observe from the plot that the prices of those two fuels types are significantly different with P98 being consistently more expensive, which might mean higher quality of that fuel type.

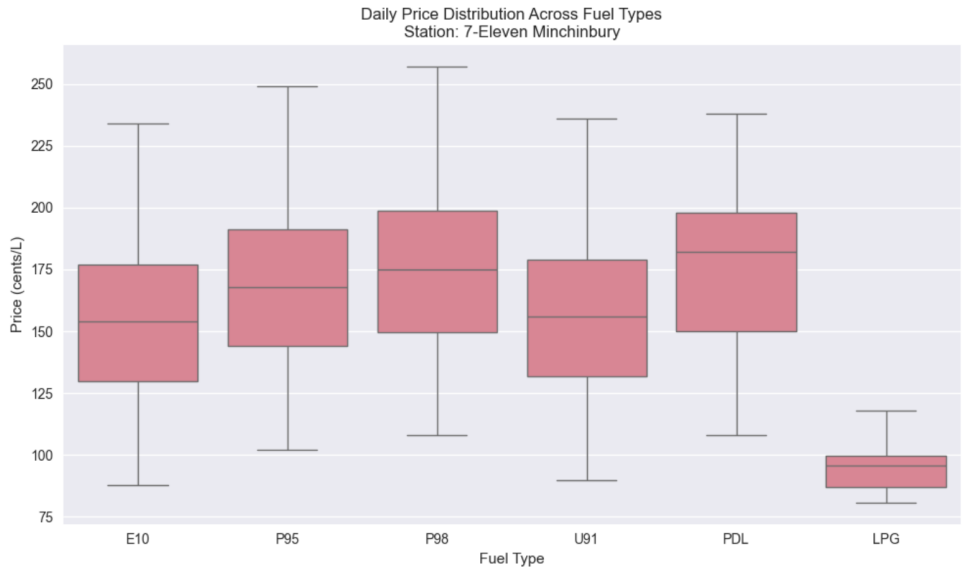

Figure 2: Box plot comparing daily price distributions across all available fuel codes. Premium fuels (PDL, P98) show higher median prices compared to regular unleaded fuels (U91, E10, P95), and higher variability compared to alternative fuel, like LPG, which exhibits low prices and very low variance.

2 1.2 Forecasting Next-Day Prices

2.1 1.2.i Problem Setup & Data Split

Forecasting Setup:

- **Inputs:** Historical price data for each (FuelCode, ServiceStation) combination
- **Target:** Next day's minimum price for each fuel type at each station
- **Time unit:** Daily frequency

Data Split Strategy:

- Training set: 70%
- Validation set: 15%
- Test set: 15% (final evaluation)

A time-series model must be trained strictly on past data, so the split follows the natural chronology to avoid leakage. I use 2016–2018 for training because it provides a long, stable window to learn seasonal and station-specific pricing patterns. 2019 serves as the validation year, giving a clean, adjacent period for hyperparameter tuning without exposing the model to future behaviour. The test set spans 2020–2025 to evaluate the model under real out-of-sample conditions, including major shifts such as COVID-era volatility. This setup keeps the learning, tuning, and final evaluation fully separated and realistic

Prevention of Information Leakage: Strict time-based splitting ensures no future information contaminates training. The model only uses data available up to day d to predict day $d + 1$.

2.2 1.2.ii Model Design & Training

Forecasting Pipeline:

- **Feature Engineering:**
  - Lag features (1-day, 7-day, 30-day price movements)
  - Station and fuel code categorical features
- **Model Family:** Random Forest Regressor
- **Key Hyperparameters:**
  - n_estimators: 300
  - max_depth: 12
- **Training Procedure:** Scikit-learn pipeline with StandardScaler for numerical features and OneHotEncoder for categorical variables

2.3 1.2.iii Evaluation & Visualisation

Metric	Validation Set	Test Set
Mean Absolute Error (MAE)	6.25 cents/L	11.16 cents/L
Root Mean Squared Error (RMSE)	7.07 cents/L	12.36 cents/L

Table 1: Model performance metrics on validation and test sets. The model performs reasonably well on the validation set, and although errors increase on the test set, the results are still within an acceptable range given the harder, more volatile 2020–2025 period.

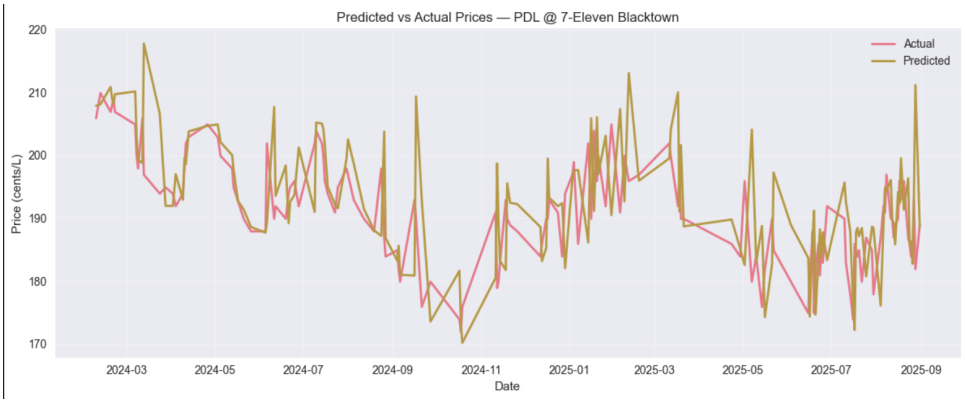

Figure 3: Predicted vs actual prices for PDL fuel at a sample station. The model captures general trends well but struggles with sudden price spikes. Errors are typically larger during rapid price change periods, suggesting the model is conservative in predicting extreme movements. The seasonal patterns and weekly cycles are reasonably well captured.

3 1.3 Application & Limitations

3.1 1.3.i Practical Use

To find the cheapest fuel daily using this model:

1. Generate next-day price predictions for all fuel codes at all stations
2. For each station, identify the lowest-priced fuel type available
3. Compare across stations to find the absolute minimum price
4. Recommend the station and fuel type combination with the lowest predicted price

If someone needs fuel every day, they could use the model to predict next-day prices for all station–fuel combinations. Each morning, the model outputs the forecasted price for every fuel type at every station. The user could then choose the station and fuel with the lowest predicted price for that day. This allows daily optimisation of fuel cost without manually checking all stations, effectively turning the model into a "cheapest-fuel guide" that leverages historical pricing patterns to anticipate the best purchase option.

3.2 1.3.ii Limitations & Remedy

Limitation: The current model lacks external features that significantly influence fuel prices, particularly crude oil prices and wholesale fuel costs, which are major drivers of retail price movements.The current model only uses lagged prices and categorical station/fuel identifiers. It fails to capture sudden price spikes caused by events like supply shortages, policy changes, or seasonal demand, which reduces accuracy during volatile periods.v

Proposed Remedy: Integrate external data sources including:

- Daily Brent crude oil prices
- Wholesale fuel terminal prices
- Exchange rate data (GDP/USD)
- Seasonal demand indicators

These features would be incorporated as additional inputs to the Random Forest model. Improvement would be measured by comparing MAE and RMSE against the baseline model on a held-out test period. The hypothesis is that including these fundamental price drivers would significantly reduce prediction errors, particularly during periods of market volatility driven by external factors rather than local competitive dynamics.

2.1 Multi-Linear Regression Model for Indoor Localisation

2.1.1 Model Development

A multi-linear regression model was implemented to predict X and Y coordinates based on RSSI values.

- **Input features:** RSSI values from multiple access points, encoded into a fixed-length vector (length = number of unique MAC addresses). Missing RSSI values were replaced with -100dB to indicate no signal.
- **Target variables:** X and Y coordinates (in metres).
- **Preprocessing:**
  - Converted string lists in `rssi` and `mac_addrs_idx` to Python lists.
  - Created a feature matrix where each column corresponds to a unique MAC address.
  - Applied `StandardScaler` to normalise the features.
- **Training:** Two separate linear regression models were trained for X and Y coordinates using 80% of the training data.

2.1.2 Performance

The models were evaluated on the held-out validation set (20% of the training data) using Mean Squared Error (MSE):

$$\begin{aligned} \text{MSE}_X &= 21.2061, \\ \text{MSE}_Y &= 26.6619 \end{aligned}$$

These results indicate a moderate baseline performance, with the model capturing some of the linear relationships between RSSI patterns and location coordinates.

2.2 Neural Network Model for Indoor Localisation

2.2.1 Model Architecture

A feed-forward neural network (MLPRegressor) was designed to improve upon the linear baseline.

- **Architecture:** Three hidden layers with $(128, 64, 32)$ neurons, ReLU activation.
- **Training:** Separate networks for X and Y coordinates, trained for 500 epochs with Adam optimiser.
- **Input/Output:** Same pre-processed RSSI feature matrix as in Section2.1.

2.2.2 Validation Results

The neural network achieved the following validation MSE:

$$\begin{aligned} \text{MSE}_X^{(\text{NN})} &= 21.9398, \\ \text{MSE}_Y^{(\text{NN})} &= 16.4994 \end{aligned}$$

2.2.3 Comparison with Linear Regression

The neural network outperformed the linear regression model substantially in the Y axis, however, did show decline in performance for the X axis.:

$$\begin{aligned} \text{Improvement in } X &: \frac{21.2061 - 21.9398}{21.2061} \times 100\% \approx -3.46\% \\ \text{Improvement in } Y &: \frac{26.6619 - 16.4994}{26.6619} \times 100\% \approx 38.12\% \end{aligned}$$

The non-linear transformations and higher capacity of the neural network allow it to capture more complex interactions among RSSI signals. While this led to a significantly better fit for the Y-coordinate (38% MSE reduction), the X-coordinate prediction showed slight degradation (-3.5%), suggesting that the relationship between RSSI and X may be predominantly linear or that the network requires further tuning for this dimension. The neural network achieved moderate localisation accuracy with RMSE values of approximately 4.7m for X and 4.1m for Y coordinates. While this represents a significant improvement over linear regression for the Y-coordinate (38% reduction in MSE), the absolute error of 4–5meters indicates room for refinement. Such accuracy could be sufficient for zone-based applications but falls short of fine-grained room-level localisation typically desired in indoor positioning systems. Potential factors limiting performance include signal multipath effects, environmental interference, and the inherent noisiness of RSSI measurements in indoor spaces.

2.3 Dimensionality Reduction for Feature Visualisation

2.3.1 Feature Extraction and Clustering

Feature representations were extracted from the first hidden layer of the neural network (128-dimensional activations). These embeddings capture the model's internal representation of RSSI patterns. A k-means clustering ($k = 3$) was applied to the extracted features without using any positional labels.

- **Cluster distribution:**
  - Cluster 0: 32.9% of samples
  - Cluster 1: 38.3% of samples
  - Cluster 2: 28.8% of samples
- **Spatial correspondence (for analysis only):**
  - Cluster 0: average position ($X = 64.64,\ Y = -29.01$)
  - Cluster 1: average position ($X = 35.69,\ Y = -24.65$)
  - Cluster 2: average position ($X = 48.57,\ Y = -29.38$)

The clusters map to distinct spatial regions of the environment, showing that the neural network internally groups RSSI patterns according to their geometric structure—even though the clustering was performed without using the coordinate labels.

2.3.2 Dimensionality Reduction Visualisation

A 2D visualisation of the extracted 128-dimensional features was produced using t-SNE (perplexity = 30). The plot shows clear grouping patterns that align with both the k-means clusters and the underlying spatial layout, confirming that the network learns a meaningful low-dimensional structure from raw RSSI signals.

- **Left plot (coloured by cluster):** Shows three well-separated groups, confirming the k-means clustering results. Each cluster corresponds to a distinct spatial region within the environment.
- **Right plot (coloured by true Y coordinate):** Reveals a gradual colour gradient along the t-SNE dimensions, with warmer colours (higher Y values) concentrated in certain areas. This indicates that the learned features preserve spatial ordering, particularly along the Y-axis.

These visualisations demonstrate that the neural network has learned a representation space where geometric proximity correlates with signal-pattern similarity, a desirable property for the localisation task.

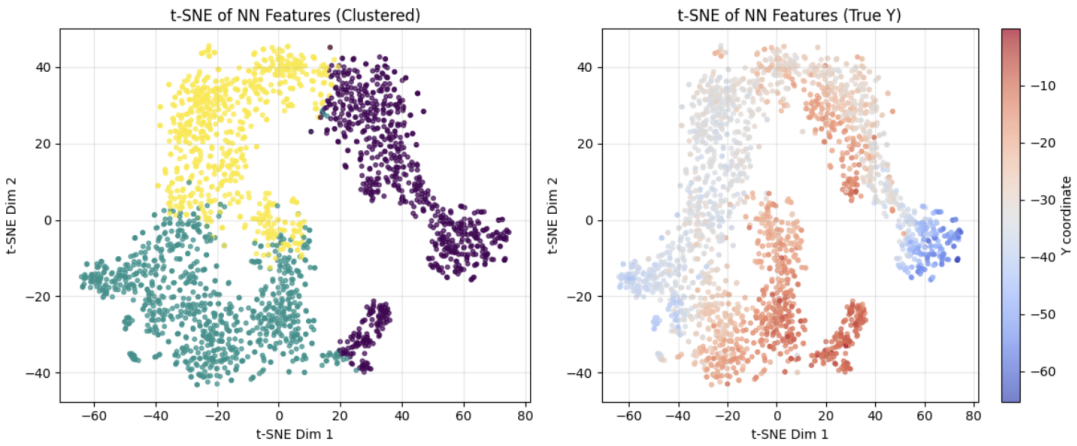

Figure 4: t-SNE visualisation of neural network features. **Left:** Coloured by k-means clusters (k=3). **Right:** Coloured by true Y-coordinate. The clear separation and spatial gradient demonstrate that the network has learned a representation that preserves geometric relationships.

2.4 Prediction and Analysis on Test Data

2.4.1 Model Selection and Prediction

The neural network model was chosen for final testing due to its superior validation performance. The held-out test set was processed identically to the training data:

- Same MAC-address mapping and missing-value imputation (-100dB).
- Features scaled with the same `StandardScaler` fitted on the training set.
- Predictions generated for both X and Y coordinates.

Prediction ranges:

$$X : [22.00, 84.62]$$
$$Y : [-67.84, -4.76]$$

Predicted X and Y values fall fully within the expected spatial range and show smooth, non-erratic behaviour, indicating that the model generalises appropriately to unseen test data.

2.4.2 Submission File

A text file named `24111778.txt` was created containing two columns (X,Y) with comma-separated predicted coordinates. The file follows the required format and is ready for evaluation.