MAIS 202 – Deliverable 1: Project Proposal

Team: Leping Wang, Chengyu Wang, Brandon Ku

1) Choice of Dataset

The goal of our team project is to forecast the average retail price of regular gasoline in Canada, especially in metropolitan regions such as Montreal, Toronto and Vancouver.

Type	Dataset Source	Reasons
Target	Statistics Canada regular gasoline price series (Canada-wide / regional averages)	- Exact outcome we want to predict (retail pump prices) - Official + reliable source - Model national trends and compare provinces.
Drivers	- WTI crude oil (FRED: MCOILWTICO) - CAD-USD exchange rate (FRED: EXCAUS).	- Strong correlation between crude oil costs and gasoline prices - Foreign exchange rate due to WTI crude oil priced in USD, leading to potential fuel fluctuation costs.
Variation	- Extreme Weather Event flags - Regional fuel taxes	- Temporary spike due to extreme weather transport cost - Tax implication differ by Canadian region.

2) Methodology

a. Preprocessing:

- Pull WTI and exchange rate series from FRED
- Pull Canadian gas price series form Statistics Canada
- Merge with tax and weather/event tables
- Normalize units (CAD/L) and timestamps (align timeseries by months)
- Clean data (handle missing values and outliers)
- Create Lag features (1-3 months): start small, then slowly increase it

b. Model:

We want to predict the monthly average retail price of regular gasoline in Canada (by region) for a chosen forecast horizon using both target and driver datasets.

Proposed machine learning algorithm: SARIMAX

Reason: Gas prices are a time series with strong dependence on recent months, and seasonal patterns (summer/winter demand). SARIMAX is built specificially for this, since it lets us include outsider drivers (Oil + FX) as exogenous variables.

Pros	Cons

- Designed for time-series forecasting - Model seasonality directly - Take exogenous inputs (oil x FX)	- Assumes fairly structured relationship (mostly in this case) - Requires checking parameter choices

Alternative algorithms considered:

Algorithms	**Pros**	**Cons**
Linear Regression	- Very simple, fast - Good Benchmark	- Relationship may not be linear - Underfit if effects are nonlinear
XGBoost & Random Forest	- Captures nonlinear patterns and interactions - Strong accuracy	- not "Time-aware" - Overfitting risk with small datasets
Recurrent Neural Networks	- Learns time dependence automatically from sequences	- Requires larger datasets (daily) - More tuning complexity - Prone to time-series preprocessing mistakes.

c. Evaluation Metrics
Because we are trying to forecast monthly retail gasoline prices, we will evaluate SARIMAX using time-based set – a walk-forward evaluation to simulate real forecasting: train on earlier months, and test on the next months, repeating over the test period.

Primary metrics:

- MAE (Mean Absolute Error): average absolute difference between predicted and true price
- RMSE (Root Mean Squared Error): penalizes large mistakes more.

Baseline for comparison

- Baseline: "Next month's price = this month's price"
- SARIMAX is considered successful if it predicts the MAE/RMSE lower than these baselines.

3) Application

Model Integration

An interactive dashboard (webapp) that helps users see a visual forecast of monthly regular gasoline prices in Canada using our SARIMAX model.

(Optional) If time allows it, we hope to train our model offline and save it, and create a small REST API endpoint, where the frontend calls the API and renders results.

User inputs (and how they provide them)

- Region selector: dropdown (Canada-wide, or a province/region).
- Forecast horizon: dropdown or buttons (e.g., 1 month, 2 months, 3 months ahead).

Outputs

- Predicted gas price for the selected horizon.
- Trend indicator (up/down/flat compared to the latest month).
- Line chart showing recent historical prices plus the forecasted points.
- A short "What's driving the forecast?" text showing the main drivers used (WTI oil + CAD–USD) for transparency.