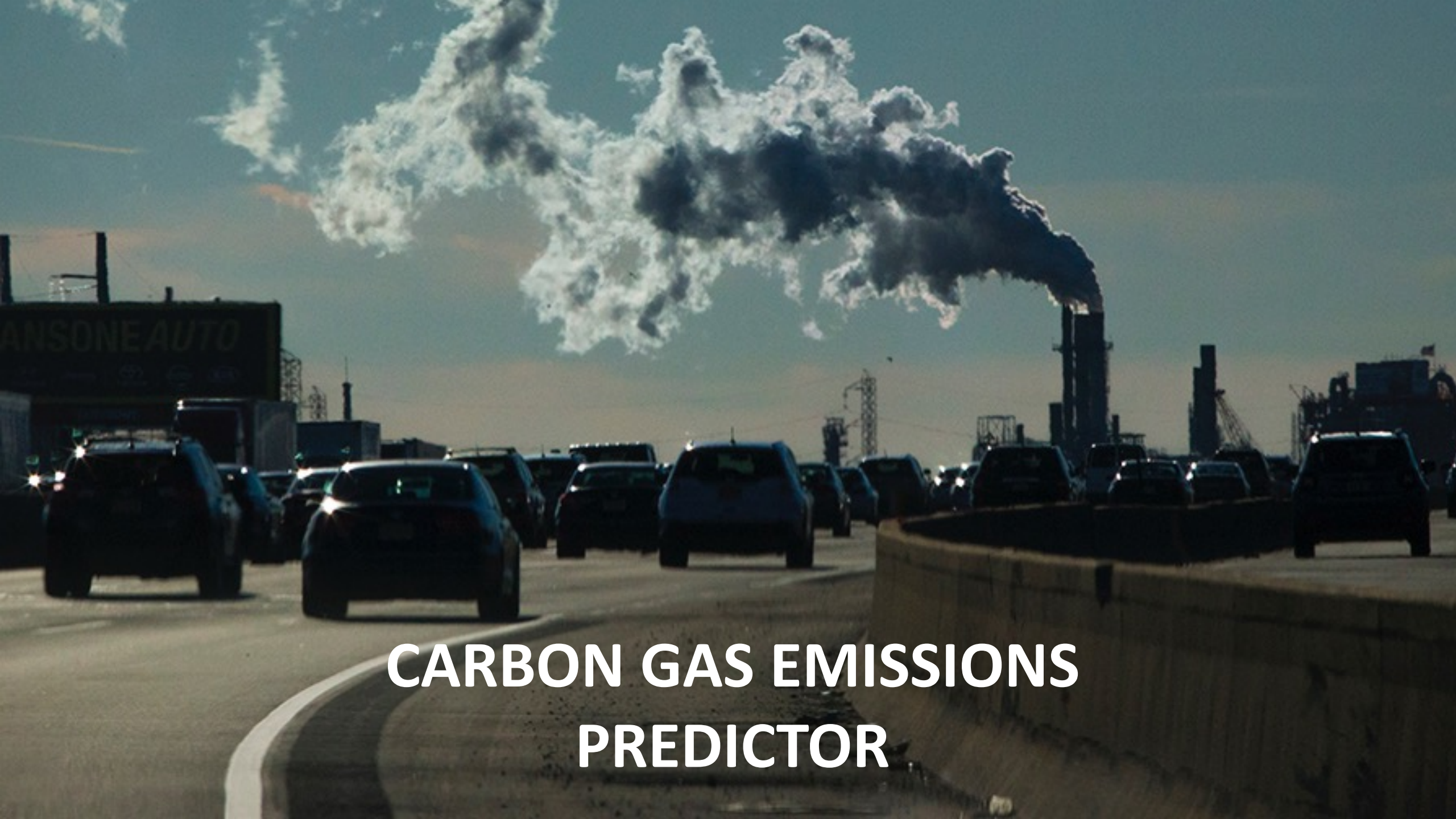


CARBON GAS EMISSIONS PREDICTOR

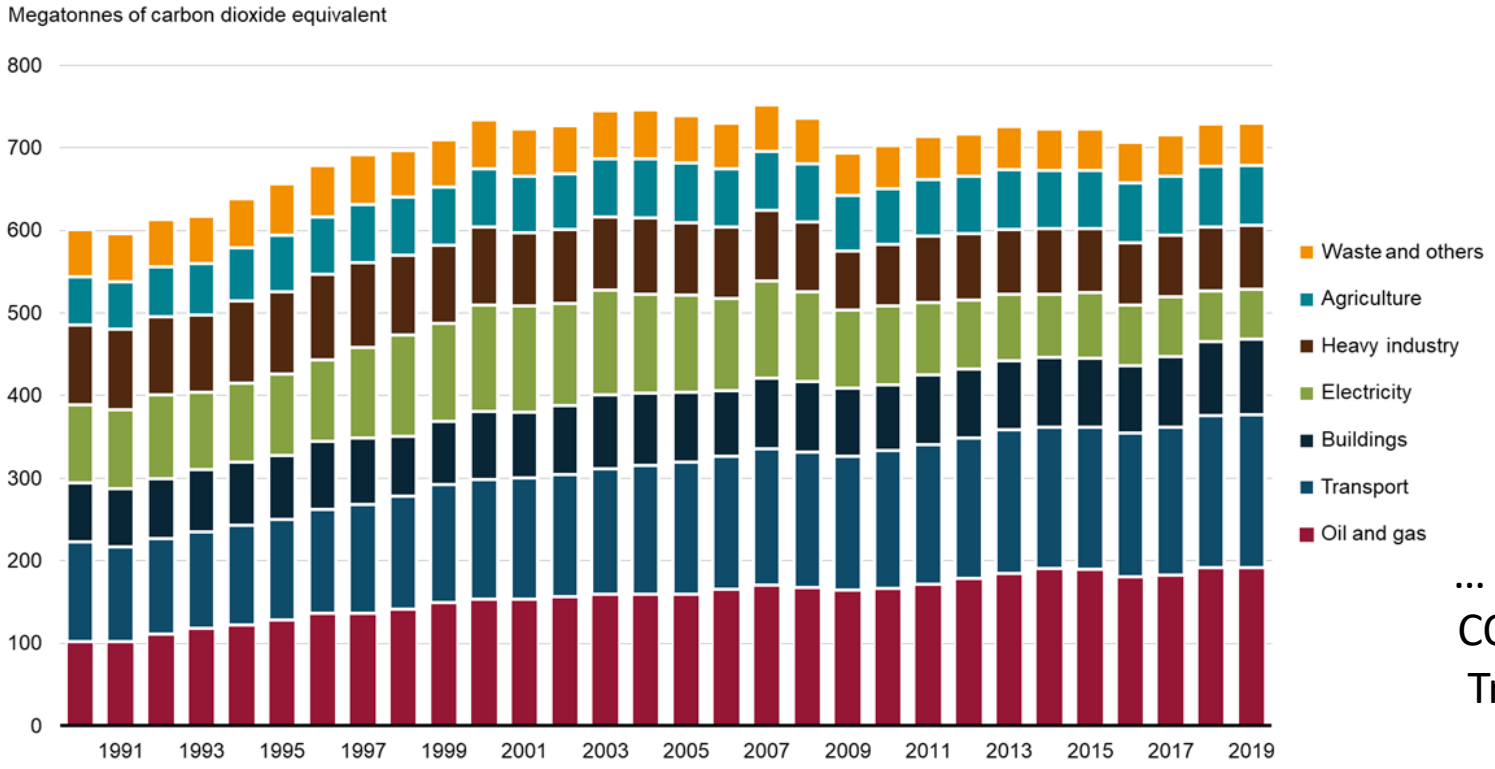
AGENDA

1. Business Problem / Data Problem
2. Stakeholder Identification
3. Process Workflow
4. Dataset Characteristics
5. Exploratory Data Analysis
6. Model Building & Selection
7. Model Deployment
8. Limitations

BUSINESS PROBLEM / DATA PROBLEM

- As carbon emissions from cars make a significant proportion of carbon emissions, policy makers need a quick method to assess and forecast carbon gas emissions emitted by automobiles in order to properly tax and disincentivize consumers from owning certain vehicles.

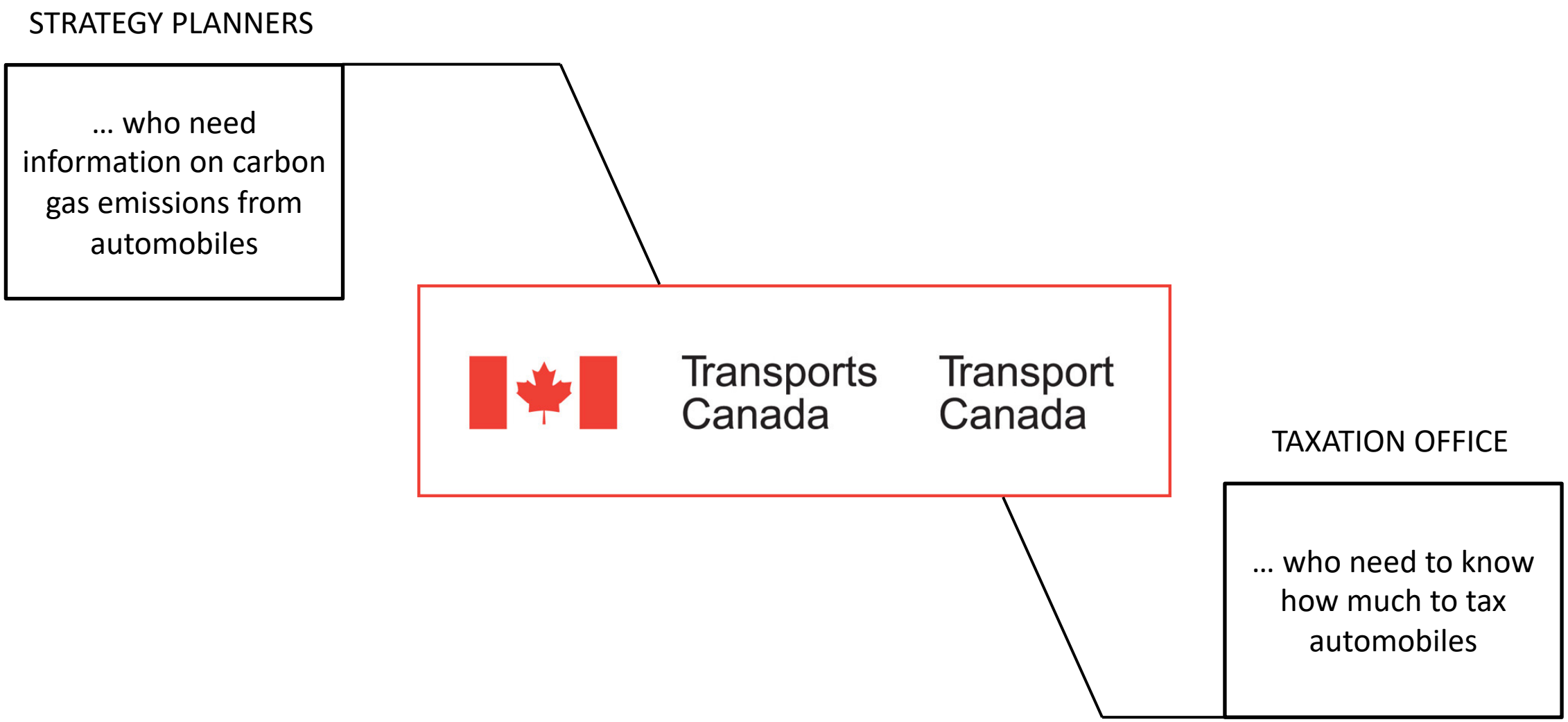
- Collecting primary historical data from a specific geographic location, as many characteristics can impact carbon emissions.



… Average of 25% of CO2 emissions from Transport Sector in recent years

Source: Government of Canada

STAKEHOLDER IDENTIFICATION



PROCESS WORKFLOW

Institute of Data

Data Collection

- Datasets sourced from Government of Canada – Open Government Data portal.

Data Cleaning

- Datasets were manipulated to fit into DataFrames, datasets were merged, columns not needed were dropped and duplicates dropped.

Exploratory Data Analysis

- Statical and visual exploratory data analysis was conducted

Pre processing & Feature Selection

- Categorical features were label encoded
- Data was scaled using StandardScaler for some models
- Feature selection by RandomForest used for some models

Model Building & Selection

- Assessment of model accuracy metrics
- Selection of model to be deployed

Deployment for Stakeholder Use

- Model deployed via a webapp for easy use by stakeholders
- Webapp was stored in a docker container for easy reproduction

DATASET CHARACTERISTICS

Institute of Data

Dataset statistics	
Number of variables	16
Number of observations	2891

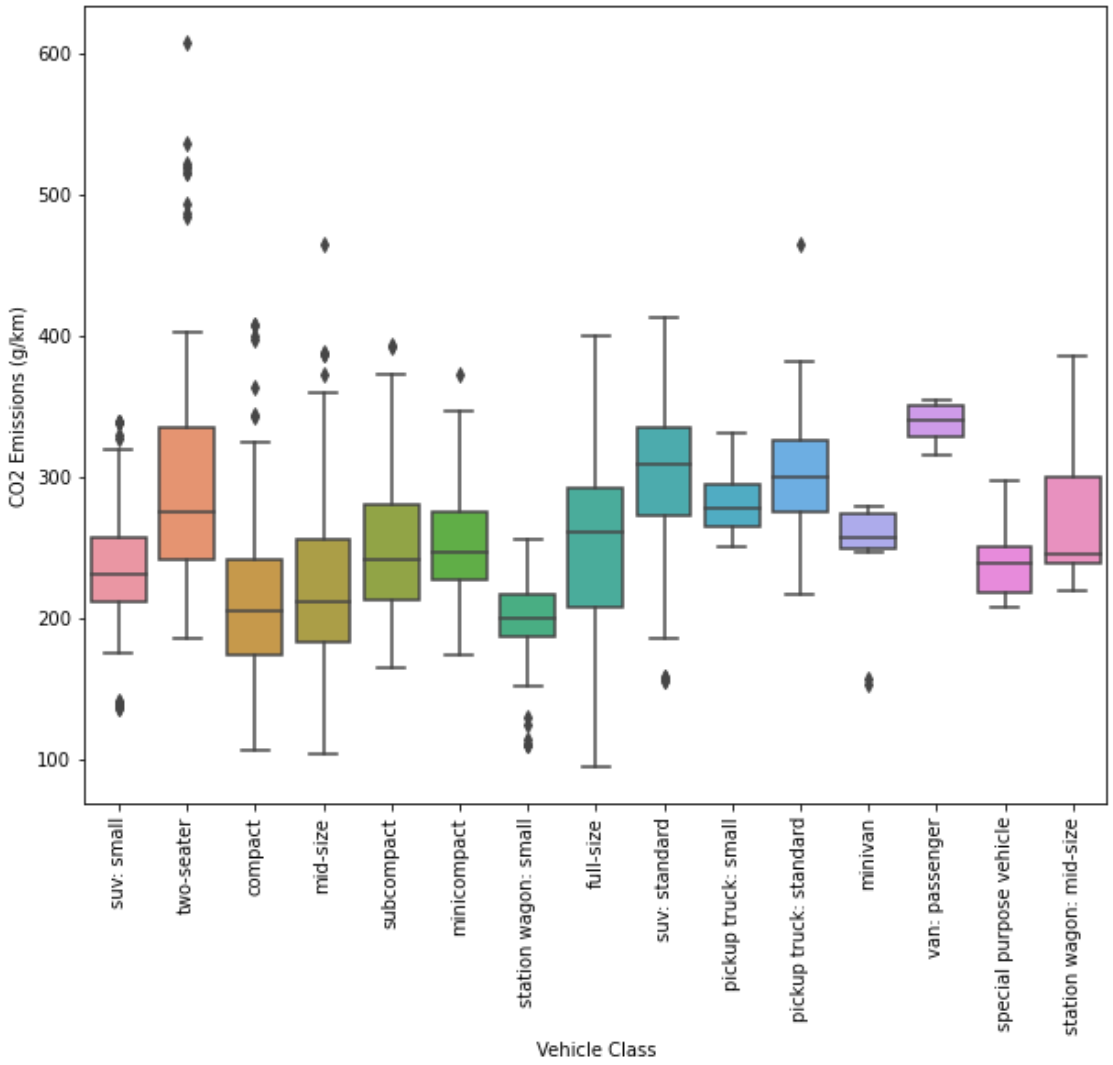
Variable types	
Numeric	9
Categorical	7

- Data downloaded from: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

Source: Canada Government – Open Data - Carbon Fuel Datasets

EXPLORATORY DATA ANALYSIS

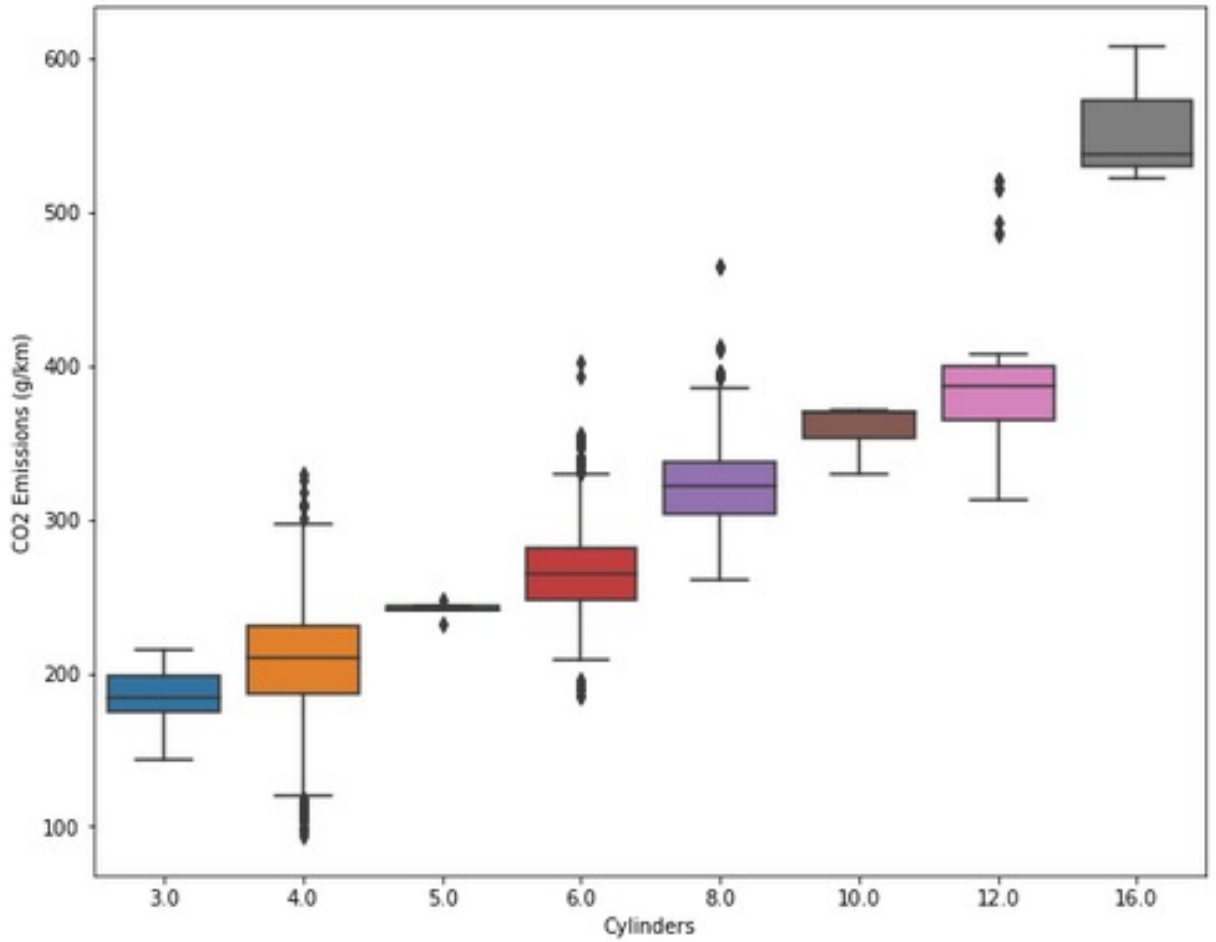
- Two-seater vehicles released highest amount of carbon gas emissions.
- Passenger vans had the highest mean carbon emissions released.



Source: Canada Government – Open Data - Carbon Fuel Datasets

EXPLORATORY DATA ANALYSIS

- Cars with higher number of cylinders released the most carbon emissions.



Source: Canada Government – Open Data - Carbon Fuel Datasets

EXPLORATORY DATA ANALYSIS – FEATURE CORRELATION

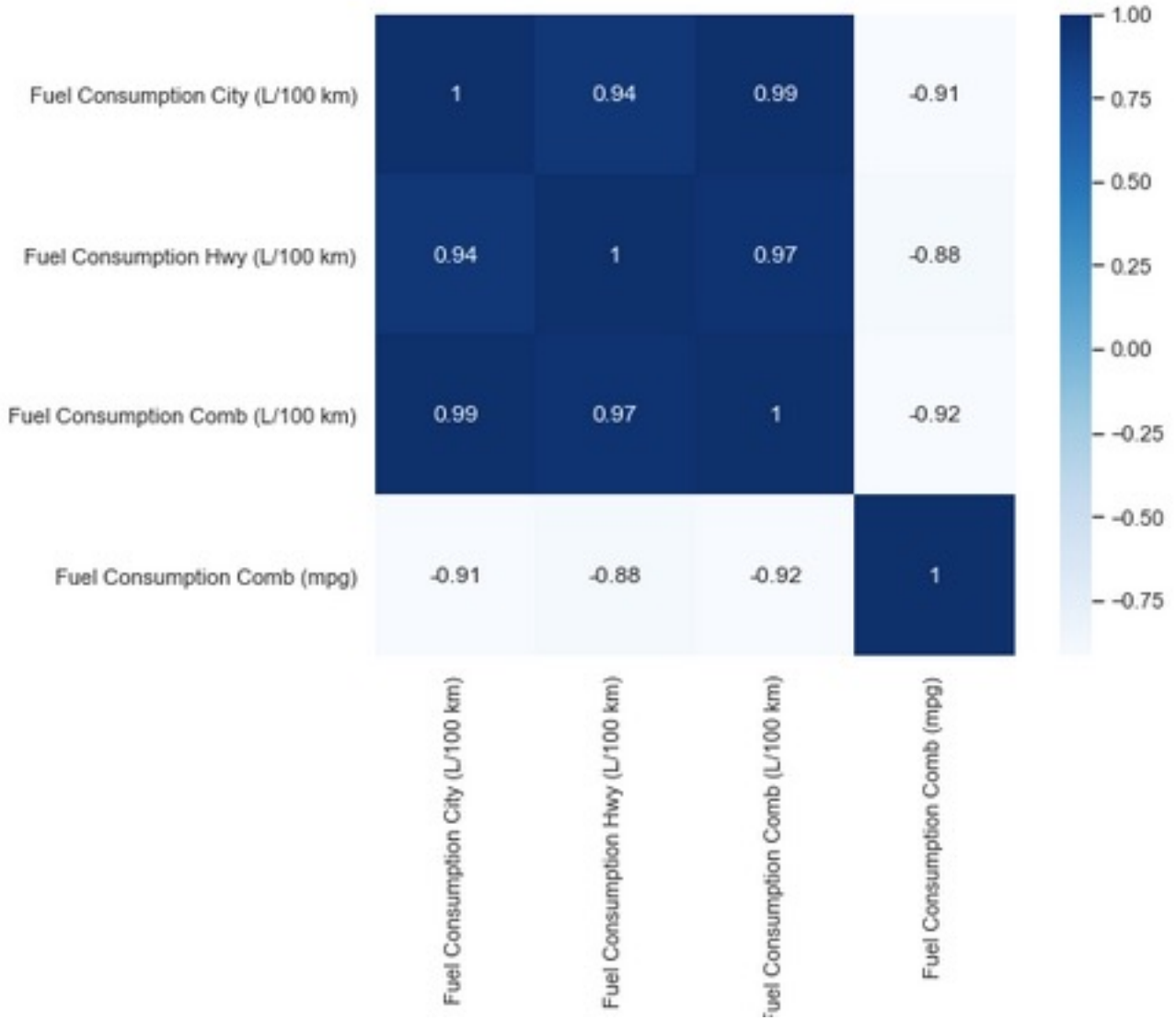
- Fuel consumption features found to be most highly correlated with target.

	CO2 Emissions (g/km)
CO2 Emissions (g/km)	1.000000
Fuel Consumption Comb (L/100 km)	0.951656
Fuel Consumption City (L/100 km)	0.948930
Fuel Consumption Hwy (L/100 km)	0.917122
Engine Size_L	0.835887
Cylinders	0.834904
Smog Rating	-0.514353
Fuel Consumption Comb (mpg)	-0.910652
CO2 Rating	-0.954422

Source: Canada Government – Open Data - Carbon Fuel Datasets

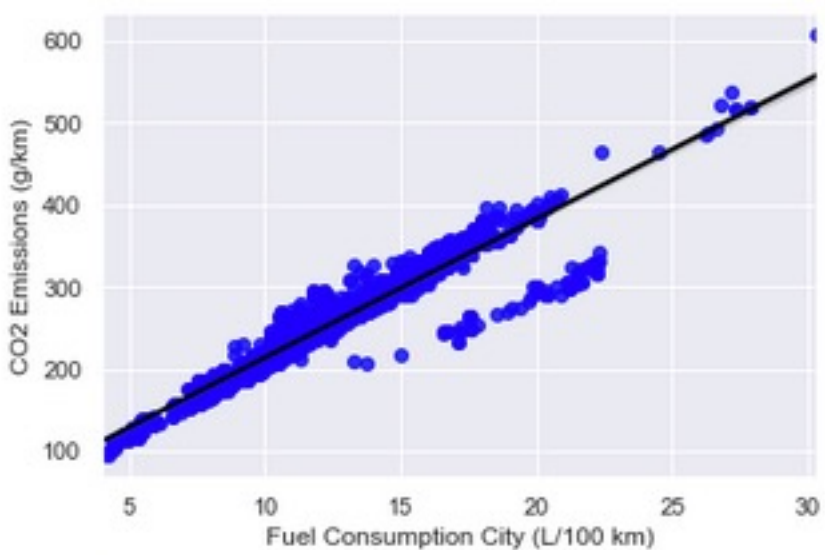
EXPLORATORY DATA ANALYSIS – FEATURE CORRELATION

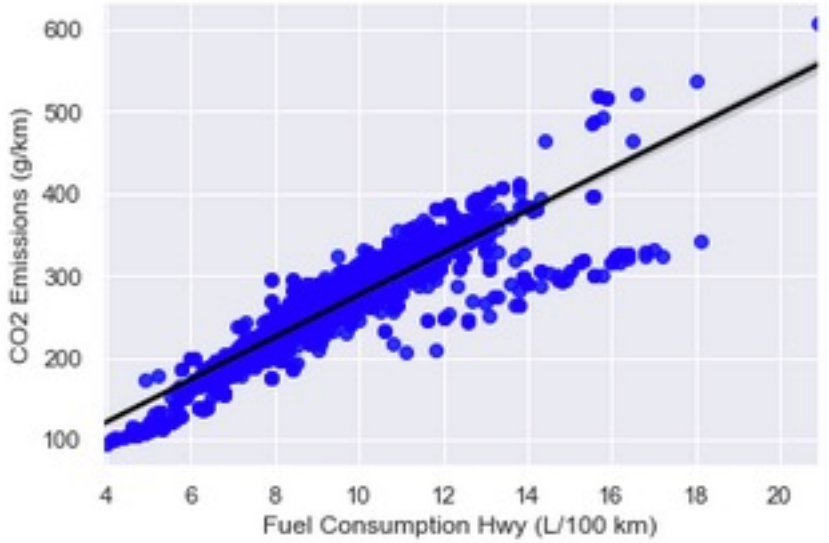
- All 4 feature columns labelled “Fuel Consumption” are highly correlated.

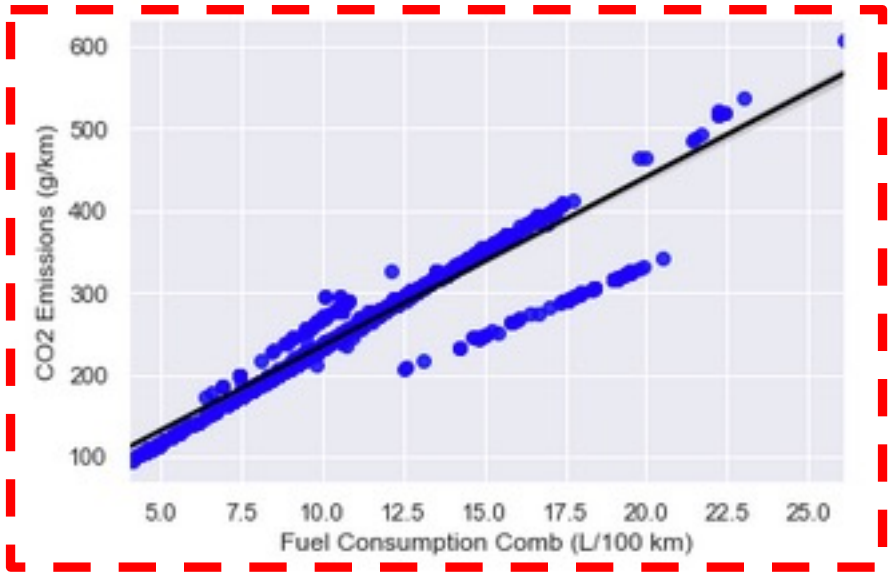


Source: Canada Government – Open Data - Carbon Fuel Datasets

EXPLORATORY DATA ANALYSIS – FEATURE CORRELATION







Source: Canada Government – Open Data - Carbon Fuel Datasets

MODEL SELECTION

- Selected model can can predict price with an **upper/lower bound of 16.71 CO2 g/km (7% of mean CO2 g/km of dataset)**

Model Type	Features	R2	RMSE (CO2 Emissions (g/km))
Simple LR (Not Scaled)	1	0.87	21.08
Simple LR (Scaled)	1	0.87	21.08
Multi LR (Features Selected by RandomForest)	3	0.90	17.08
Multi LR (All features – not scaled)	7	0.92	16.71

RECOMMENDATIONS

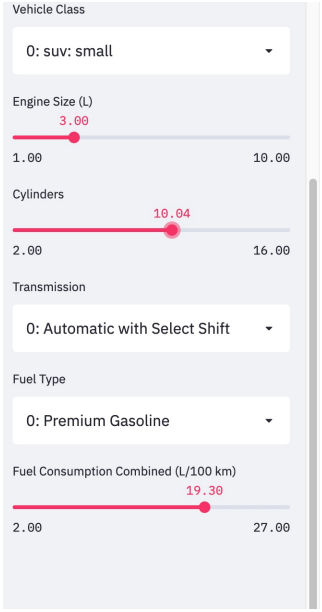
Institute of Data

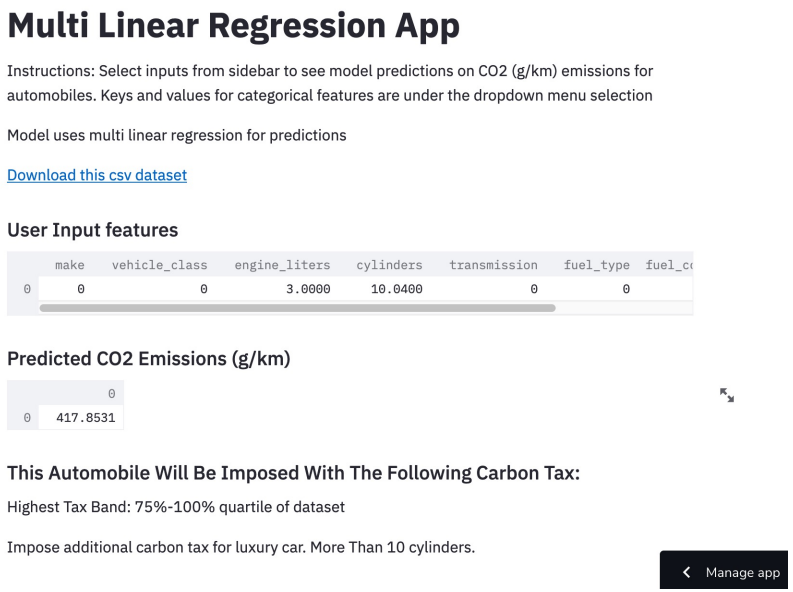
- We propose a 4-tier carbon tax pricing system based on the quartiles of the dataset.
- An additional carbon tax for luxury cars with more than 10 cylinders.
- An additional carbon tax for passenger vans.

MODEL DEPLOYMENT

- Our model was deployed onto a webapp that makes predictions and recommendations: https://share.streamlit.io/adireksa/streamlit-linregapp-project2/main/app.py

- The app was also stored into a docker container for easy reproducibility.





IMPROVEMENTS FOR FUTURE

- Expand dataset to include hybrid vehicles and electric vehicles.
- Compare other more advanced machine learning algorithms that potentially outperform linear regression.

Thank you