Discover necessity for Greenhouse Emission Test on Vehicle in New Zealand, based on datamining in last 15 years Fleets and Emission data

Tian Bai
tbai915
461962597
INFOSYS 722
30 Oct 2020

ABSTRACT

Human Sustainable Development is a goal for all peoples in the world that appeal by the United Nations. To restraint the global warm, the studies of control the Greenhouse Gas (GHG) emission is one of the hot topics. There are some different approaches for predict GHG emission, and different countries have their own situation. Hence, for New Zealand environment and policies publish, it is necessary to have a predictions pipeline. This study under Knowledge Discovery in Database (KDD) methodology, we base on New Zealand Transport Data to build a pipeline for predict transport Carbon Dioxide (CO2) Emission. The studies models are one of Gray Model (GM), Autoregressive Integrated Moving Average (ARIMA) or Artificial Neural Network (ANN). After discussion on this study, the most fitting model of New Zealand is Gray Model and the problem belongs to a Regression Problem. We apply Linear Model and selection best models base on Coefficient of Determination (R2) criteria to estimate emissions of different age groups of fleets in New Zealand. The models are achieved over 90% accuracy and provided decision making suggestions and implementation suggestions base on the pipeline for New Zealand Transport.

Keywords *Greenhouse gas (GHG) emissions, Carbon Dioxide (CO2), Data Mining, Knowledge Discovery in Databases (KDD), Regression, Linear Model*

Table of Content

Introduction

Nowadays, humans are facing the climate crisis, after discovering on leading international datasets which maintaining by Meteorological Organization's that year 2016 was the warmest year and year 2019 was the second warmest year [1], (World Meteorological Organization, 2020). The main reason to lead global warming and climate change is carbon dioxide (CO2) in Greenhouse gas emission (GHG) [2], (Z. Liu, et al. 2020). For dealing with climate crisis, the United Nations proposed 17 sustainable development goals and the goal 13 mentioned about reduce GHG to handle climate crisis [3] (UN, 2020). Hence many countries beginning to study on control the emission, such as introduced new emission standard and encourage using cleaning energy to replace biofuels.

According to the New Zealand's GHG emission report, 44.5 percent of CO2 were mainly contributing to the GHG emissions, and Transport produced 47 percent of CO2 emission [4], (Stats NZ, 2020). Hence, the prediction of emission has been considered as a significant topic among different countries. Unlike some countries need all the on-road vehicles need to follow the latest standard, New Zealand will apply the standard on the vehicles first come into the countries. The prediction method of totally emission will using different features among different countries which rely on its emission standard and vehicles population.

On the research problem aspect of this project is applying datamining to improved prediction on the case of New Zealand. Traditional Method is sampling to estimate the whole population. Canada try to apply deep sequence learning on road network to predict GHG emission [5], (Lama Alfaseeh, et al. 2020). Are there any specific method could best fit data set size and features like New Zealand have?

On the practical problem aspect, New Zealand applying smoky exhaust test contained in warrant of fitness (Wof), based on the colour of smoke, and it will not test on GHG Emission. However, the imported vehicles from different year have a different emission level, there are significant used vehicles on the road of New Zealand. Does the current exhaust test sufficient for reduce the emission?

This research Section Literature review will be a brief literature review that method of predicted GHG among different countries. Section Methodology illustrate the study methodology KDD process, and illustrate previous three times iterations using KDD. Section 1 describe the case study for the business and datamining objectives understanding, and Section 2 - 4 describe data collection which include data understanding, preparation and transformation. Section 5 - 7 is result and discuss that for select datamining method, algorithms, and conduct datamining. Section 8 - 9 is the final conclusion section for result interpretation and some further actions.

Literature review

From the national view, the U.S. have similar situation with New Zealand, where transportation system contributing nearly 29% of total GHG emission[8] (U.EPA, 2017). Many countries have specific prediction method for the specific national data. The Predict Models are widely used in 3 aspect, Grey Models (GM), the Auto Regressive Integrated Moving Average (ARIMA), and the Artificial Neural Networks (ANN).

GM was the most commonly used model for predict emission, it can apply on small number of data points and can reduce some unfavourable features or missing values affect [9] (A. Ö. Dengiz, et al.

2018). Although GMs able to get good results, people cannot explain the whole models in human understandable language. A study in Taiwan used GM to predict CO2 emission [10](C.-S. Lin, et al. 2011), and Brazil have a similar study[11] (H.-T. Pao and C.-M. Tsai, 2011). The difference is Brazil predict models based on people's income and energy consumption and Taiwan only using energy consumption.

ARIMA models are mainly statistical models, and they require features have linear relationship. Their predictions are basing on the historical values[12] (A. Rahman and M. M. Hasan, 2017). Example studies in Bangladesh and Bahrain predict CO2 emission by 44 years and 25 years CO2 emission time-series data to for automated forecasting prediction, respectively by [12] (A. Rahman and M. M. Hasan, 2017) and [13] (C. Tudor, 2016).

ANN models are developed with computer science, unless ARIMA need features have linear relationship. ANN able to predict with features are in non-linear relationship. One study shows that NNs have better performance than other in prediction of environment pollutants, [14] (K. P. Singh, 2012). Many European countries studies applying irrelevant features, agriculture data, various energy waste, or gross domestic product to predict GHG emission, an example study in Serbia [15] (D. Radojevi´c, 2013).

The most similar case is study about Southampton, it relies on 5 traffic features to estimate CO2 emissions by the authors developed a Linear Regression (LR) Model and ANN Model in 2018, [16] (M. Grote, et al. 2018). Another similar case is developing predict models in expressway in China, the features are the ration of volume to capacity, fuel consumption and time for petrol passenger vehicles and diesel trucks, [17] (Y.Dong, et al. 2019). This predict models are actually LR models and some of analyse works done by SPSS, However, authors test their models work on the experimental environment and have not test on real-world environment.

According to the previous studies, the prediction of CO2 emission of transport system is a **regression** problem, and it mainly rely on yearly data for all 3 type of models. GM and ARIMA models need data to have linear relationships, and ANN able to have other features such as gross domestic product, and other economic factors. The previous studies are more focus on the whole CO2 emission, and our study is employs for not only predict whole CO2 emission, but also predict each age groups CO2 emission. Another different is on the predictors, this research is using yearly data about number of vehicles, and their ages rather than some economic factors or energy consuming. Hence, the research objective is to find the most fitting model for New Zealand specific case by getting experience from other countries experience, and more like a GM model problem.

Methodology

For discover best predict models need a sequence of process rather than just build model on data sets. Hence, Data mining is an application for applying algorithms for discover patterns from data, and Methodology of Data mining is sequence of process that applying data mining to discover patterns from data. Many different Data mining methodologies was introduced in the world in 1990s, the most famous two workflows are Cross-industry standard process for data mining (CRISP-DM) and Knowledge Discovery in Databases (KDD). KDD has more specific process of data preparation than CRISP-DM thus this project selects KDD as its methodologies.

KDD process has nine steps [6] (Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996):

1. Understanding of the application domain
    1.1. Identifying the goals with relevant prior knowledge

   1.2. Access the situation
   1.3. Understanding Datamining objectives
   1.4. Set a proper project plan
2. Gathering and understanding dataset
3. Preparation of dataset
4. Data set reduction and projection
5. Select Datamining methods
6. Select Datamining algorithms
7. Conduct Datamining
8. Interpreting mined patterns
9. Acting on the discovered knowledge

This research paper will be basing on the 3 times iterations of step 1-8 among KDD process with different technologies stacks, name them ISAS for using IBM SPSS, OSAS for using Open source module and BDAS for using Spark module. The following sections will link to 9 steps of KDD process.

1 Business understanding

1.1 Business understanding

Case study reflect the first step of KDD process. The Business or Situation background understanding we already done in Section 1 and 2. The research objective is finding best model suit for NZ environment, the practical objective is applying the models to predict CO2 emissions for NZ Transport to publish new policies.

Therefore, the success goals of the study will be introduced with the following:

1. The models could accurate estimate the predict emission values against real values for match the research objective.
2. The models are able to predict partial emission value by input partial group statistics within one fleet group for match the practical objective.

Suppose one fleet group able to category by different age ranges, input all the statistics from the whole group will get the totally emission predict value. If input each category statistics will return partial emission predict value. Under accurate models, blocking other fleets and remain a specific type of fleet inputs in one year, able to get a predict value of single type against totally predict value to estimate the emission partition of this type of fleet.

1.2 Access the situation

From previous iterations then for divide fleets emission data by age groups which require the data contains number of fleets, average age of fleets and corresponding CO2 emission. New Zealand have a fully vehicle registration system, these data are findable from The Ministry of Transport.

Under previous iterations then we know some risks:

- New Zealand does not statistic age groups among motor cycles and heavy vehicles.
- SPSS are not accurate enough to build the datamining model.
- Spark have problems with data visualization.

The contingencies of these risks:

- The heavy vehicle and motorcycle data cannot split by age because The NZ Transport have not statistics the age distribution of them. However, they have taken 3.65% and 3.69% of totally vehicle, over 18 years, so that dismiss these two types of fleets.
- Combine using these 3 iterations technologies stacks for eliminate technologies problem.
- Using Git to enables other participants to join the project and this gives a better iteration of the program.

The research resource inventory:

Python 3.7+, Scikit-learn, Pandas, xlnd, yellowbrick, numpy, matplotlib, Github.

The assumption among 3 iteration will continue use in the research paper which is:

- As the Euro 5 Emission Standard published in 2009 and Euro 6 Emission published 5 years later, there will be over 50% fleets with age more than 10 years and this group will contribute over 70% CO2 emission of totally fleets.

1.3 Datamining Objectives

1. For the research objective, build a prediction model for New Zealand dataset, the model built would able to estimate the CO2 emission from the number of flees and its average ages inputs. The estimated CO2 emission data is combined for multiple fleets type groups with over 80% accuracy.
2. For practical objective, the success model able estimate parts of the CO2 emission among different age groups of vehicles with over 80% accuracy.

1.3.1 Datamining Success criteria and Benchmarks

1. The first success criterion is when input values fleets to models reflect predict values of CO2 emission and it will have a lower residual with real values of CO2 emission thus to count the accuracy, benchmark is $Real\ Value - Predict\ Value$.
2. The second success criterion is the sum of the predict values within different groups should also not have a large residual compare with the predict values, benchmark is $real\ value / \sum single\ group\ predict\ value$.

1.4 Project Plan

According to 3 times previous iterations, project plan, risks met and research paper plan illustrated as following table. Previous iterations have Gantt charts for process allocation they are included in previous reports.

Phase	Time	Risks
Iteration1 Proposal	24th July – 7th Aug (2 weeks)	• Unable to find suitable data
Iteration2 ISAS	7th Aug – 28th Aug (3 weeks)	• SPSS may not be accurate enough • SPSS might not have many model options to choose from and more customize as the

		Sci-kit Learn library in Python does.
Iteration3 OSAS	28th Aug – 2th Oct (6 weeks)	• When the result needs to be demonstrated the whole program will need to be paused while demonstrating the output figure
Iteration4 BDAS	9th Oct – 23th Oct (2 weeks)	• Require time and efforts to manage the merges or unwanted merges happens
Research paper	23th Oct – 30th Oct (1 week)	

2 Data understanding

2.1 Data Collection

According to previous iterations experiences then **Ministry of Transport issued all the data the project needs in an excel file**. The data set contains number of different fleets, age distribution between different fleets and CO2 emission for different fleets. There are some other data tables might use for mining in further insights, for example, average travel milage for different fleets. The Excel file NZ-Vehicle-Fleet-Statistics-2018_web.xlsx collect from Ministry of Transport at page[7] (https://www.transport.govt.nz/mot-resources/vehicle-fleet-statistics/).

2.2 Data Description

After 3 iterations discovered then the data that project need 3 data tables from excel file:

- Number of Fleets: data at the content of "figure 2.1 2.2 2.3 2.4" of excel file.
  - This data table contains 34 columns and 19 rows.
  - The first column is named period for indicate the year of the data from 2000 to 2018 of following columns, the Remain 33 columns are number of fleets, average age and percentage of number fleets which generate from number of fleets. There are mainly grouped by light, motorcycle, truck and bus 4 groups, for each group have 2 categories to imported and used imported. The light fleets group has to two subgroups: passenger and commercial.
  - Columns are either int64 or float 64 type, more details illustrate in figure 4.
- Light Vehicles Age Distribution of Fleets: data at the content of "figure 2.10" of excel file.
  - This data table contains 38 columns and 7 rows.
  - This data table is transposed in excel file for human read easily, thus the age become columns and columns are the indexes.
  - Thus, Age groups information are in string type as index
  - The fleets of age groups in float type.
  - The percentages of fleets of corresponding age groups in float type between [0,1], more details illustrate in figure 4.

- CO2 Emission Data: data at the content of “figure 1.10” of excel file.
  - This data table contains 5 columns and 17 rows.
  - Columns are in float type, because of unit is the Million Tonnes CO2.and shown in figure, more details illustrate in figure 4.

	Period	Total light new	Total light used import	Total LPV new	Total LPV used	Total LCV new	Total LCV used	Total MC new	Total MC used	Total truck new	...	Truck used %	Bus used %	Light fleet average age	Light passenger average age	Light commercial average age	Motorcyc averag ag
0	2000	1527641	966687	1276498	870680	251143	96007	57728	20310	70872	...	0.257613	0.244488	11.809922	11.713870	12.404023	16.05913
1	2001	1510448	1052505	1256293	956915	254155	95590	58123	20440	70880	...	0.274291	0.269294	11.962537	11.863626	12.588450	16.44494
2	2002	1504464	1142837	1245805	1046144	258659	96693	59066	21015	71441	...	0.295766	0.313531	12.053392	11.957576	12.671381	16.66186
3	2003	1510494	1248263	1245233	1149489	265261	98774	61057	21829	72482	...	0.319948	0.339596	12.099701	12.011138	12.682297	16.66101
4	2004	1523241	1343089	1249158	1241364	274083	101725	64896	22867	74130	...	0.346510	0.354958	12.164873	12.094087	12.633977	16.30566
5	2005	1541884	1424588	1257339	1320880	284545	103708	72200	24471	75901	...	0.364654	0.369892	12.258855	12.212825	12.564520	15.51626
6	2006	1561044	1468085	1267476	1363917	293568	104168	80661	26640	76775	...	0.380547	0.387041	12.425076	12.401873	12.578584	14.84980
7	2007	1584733	1503359	1280762	1398318	303971	105041	89861	29146	77869	...	0.393303	0.413788	12.586454	12.591679	12.552226	14.19966
8	2008	1605137	1502981	1292558	1400142	312579	102839	100697	32149	79089	...	0.395025	0.426217	12.813788	12.844178	12.616798	13.59726
9	2009	1610696	1488717	1294924	1389668	315772	99049	104453	33238	78627	...	0.394553	0.416062	13.194888	13.233213	12.946859	13.87726
10	2010	1628515	1493451	1306995	1398139	321520	95312	105875	33639	77935	...	0.393464	0.409976	13.483028	13.532333	13.163049	14.31547
11	2011	1643564	1473590	1315826	1382331	327738	91259	106463	33563	77833	...	0.388384	0.399227	13.742635	13.812682	13.291567	14.82812
12	2012	1690422	1474965	1350457	1385912	339965	89053	108735	34070	78593	...	0.382446	0.394209	13.959252	14.047410	13.396964	15.28818
13	2013	1748586	1494485	1389608	1404935	358978	89550	111958	34912	80661	...	0.375790	0.387371	14.075852	14.194163	13.338720	15.65234
14	2014	1817743	1541012	1434731	1449199	383012	91813	115798	36262	83884	...	0.368115	0.379788	14.093225	14.244258	13.175905	15.97048
15	2015	1888590	1593749	1479943	1499088	408647	94661	120101	37761	86894	...	0.362077	0.370067	14.082413	14.263345	13.011491	16.18512
16	2016	1975136	1656445	1535175	1556538	439961	99907	124223	39092	89790	...	0.356496	0.348863	14.078110	14.303513	12.787277	16.43103
17	2017	2066267	1723994	1590094	1617627	476173	106367	128698	40858	93716	...	0.350588	0.332229	14.035354	14.311541	12.514544	16.67402
18	2018	2151898	1768376	1641232	1657190	510666	111186	133706	42961	97694	...	0.343582	0.310045	14.086867	14.415534	12.343552	16.96568

19 rows × 34 columns

Figure 1: Number of Fleets visualization.

	Age	2000	2001	2002	2003	2004	2005	2006	2007	2008	...	2009.1	2010.1	2011.1	2012.1	
0	0-4 years	375680.0	361278.0	371546.0	397164.0	422522.0	451885.0	475895.0	498470.0	503736.0	...	0.152161	0.143768	0.136657	0.133783	
1	5-9 years	720525.0	737685.0	730168.0	760055.0	761918.0	756408.0	661033.0	609594.0	592814.0	...	0.191736	0.202059	0.208289	0.209722	
2	10-14 years	785368.0	829982.0	887866.0	907729.0	926784.0	938797.0	1023210.0	1052695.0	1050524.0	...	0.321272	0.296736	0.259704	0.237846	
3	15-19 years	393886.0	404923.0	419316.0	446613.0	488853.0	536089.0	572331.0	616195.0	627768.0	...	0.213296	0.220129	0.246440	0.255496	
4	20+ years	218864.0	229080.0	238400.0	247192.0	266249.0	283289.0	296656.0	311134.0	333272.0	...	0.121535	0.137309	0.148910	0.163152	
5	Total	2494323.0	2562948.0	2647296.0	2758753.0	2866326.0	2966468.0	3029125.0	3088088.0	3108114.0	...	NaN	NaN	NaN	NaN	
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.334831	0.357438	0.395351	0.418649	

7 rows × 39 columns

Figure 2: Light Vehicles Age Distribution visualization.

	Year	Light passenger	Light commercial	Motorcycle	Heavy fleet
0	2001	6.765291	1.504675	0.028235	2.388743
1	2002	7.066370	1.559698	0.027965	2.572510
2	2003	7.375728	1.585027	0.028364	2.639134
3	2004	7.671046	1.582260	0.029396	2.653599
4	2005	7.549507	1.628644	0.032007	2.860993
5	2006	7.644814	1.663934	0.037437	2.919942
6	2007	7.796914	1.718954	0.040616	3.009903
7	2008	7.692528	1.764464	0.045334	3.077349
8	2009	7.626856	1.774602	0.046301	2.990650
9	2010	7.697162	1.830283	0.046434	3.109751
10	2011	7.575258	1.861729	0.045124	3.200497
11	2012	7.444775	1.879310	0.044464	3.213818
12	2013	7.415187	1.937471	0.045089	3.288810
13	2014	7.430754	1.993417	0.045306	3.345710
14	2015	7.662124	2.121840	0.046825	3.454175
15	2016	7.800260	2.237802	0.047416	3.509366
16	2017	8.077407	2.472227	0.047200	3.861072

Figure 3: CO2 Emission Data visualization.

```
Period                          int64
Total light new                 int64
Total light used import         int64
Total LPV new                   int64
 Total LPV used                 int64
Total LCV new                   int64
 Total LCV used                 int64
Total MC new                    int64
Total MC used                   int64
Total truck new                 int64
Total truck used                int64
Total bus new                   int64
Total bus used                  int64
Light passenger NZ new        float64
Light passenger used import   float64
Light commercial NZ New       float64
Light commercial used import  float64
Motorcycle NZ New             float64
Motorcycle Used Import        float64
Truck NZ New                  float64
Truck Used Import             float64
Bus NZ New                    float64
Bus Used Import               float64
Light used %                  float64
Truck used %                  float64
Bus used %                    float64
Light fleet average age       float64
Light passenger average age   float64
Light commercial average age  float64
Motorcycle average age        float64
Truck fleet average age       float64
Bus fleet average age         float64
Light used average age        float64
NZ new light average age      float64
dtype: object
```

```
Age              object
2000            float64
2001            float64
2002            float64
2003            float64
2004            float64
2005            float64
2006            float64
2007            float64
2008            float64
2009            float64
2010            float64
2011            float64
2012            float64
2013            float64
2014            float64
2015            float64
2016            float64
2017            float64
2018             object
```

```
Year                 int64
Light passenger    float64
Light commercial   float64
Motorcycle         float64
Heavy fleet        float64
dtype: object
```

Figure 4: Data Table types visualization.

2.3 Data Exploration

According to previous 3 iterations, Pandas and NumPy will be best tools for the project statistics.

- Number of Fleets:

```
column: Period, range: [2000.00, 2018.00] variance: 30.00 +/- 5.48
column: Total light new, range: [1504464.00, 2151898.00] variance: 37764660579.08 +/- 194331.32
column: Total light used import, range: [966687.00, 1768376.00] variance: 41966802550.00 +/- 204858.01
column: Total LPV new, range: [1245233.00, 1641232.00] variance: 14627376633.93 +/- 120943.69
column:  Total LPV used, range: [870680.00, 1657190.00] variance: 41214473558.83 +/- 203013.48
column: Total LCV new, range: [251143.00, 510666.00] variance: 5495977493.16 +/- 74134.86
column:  Total LCV used, range: [89053.00, 111186.00] variance: 35037479.25 +/- 5919.25
column: Light passenger NZ new, range: [12.06, 12.73] variance: 0.05 +/- 0.21
column: Light passenger used import, range: [10.90, 16.33] variance: 3.42 +/- 1.85
column: Light commercial NZ New, range: [11.16, 12.40] variance: 0.12 +/- 0.34
column: Light commercial used import, range: [12.47, 18.09] variance: 3.48 +/- 1.86
column: Light fleet average age, range: [11.81, 14.09] variance: 0.76 +/- 0.87
column: Light passenger average age, range: [11.71, 14.42] variance: 0.98 +/- 0.99
column: Light commercial average age, range: [12.34, 13.40] variance: 0.10 +/- 0.32
column: Light used average age, range: [11.05, 16.42] variance: 3.35 +/- 1.83
column: NZ new light average age, range: [11.98, 12.62] variance: 0.04 +/- 0.19
```

Figure 5: Number of Fleets Statistics visualization.

- Light Vehicles Age Distribution of Fleets:

There are some **missing values** in percentage of fleets for the total groups in 5th row and 6th row, this is because the total percentage is 1 which is a trivial group in percentage columns. Instead, it using 15+ years group for the total recorded in 6th row, illustrated in figure 2. Adapting OSAS and BDAS iterations experiences, applying data table transposition which columns transpose to index and vice versa will **eliminating the missing values**. After data table transposition, it is able to statistic in figure 6.

```
column: Period, range: [2000.00, 2018.00] variance: 30.00 +/- 5.48
column: 0-4 years, range: [361278.00, 719677.00] variance: 9492113026.55 +/- 97427.48
column: 5-9 years, range: [592814.00, 761918.00] variance: 3388739045.98 +/- 58212.88
column: 10-14 years, range: [724437.00, 1174496.00] variance: 14622956139.57 +/- 120925.42
column: 15-19 years, range: [393886.00, 813648.00] variance: 17469702148.20 +/- 132173.00
column: 20+ years, range: [218864.00, 849077.00] variance: 42696096342.09 +/- 206630.34
column: Total, range: [2494323.00, 3920274.00] variance: 142570987654.51 +/- 377585.74
column: 0-4 years percentage, range: [0.13, 0.18] variance: 0.00 +/- 0.01
column: 5-9 years  percentage, range: [0.15, 0.29] variance: 0.00 +/- 0.04
column: 10-14 years percentage, range: [0.22, 0.34] variance: 0.00 +/- 0.04
column: 15-19 years percentage, range: [0.15, 0.26] variance: 0.00 +/- 0.03
column: 20+ years percentage, range: [0.09, 0.22] variance: 0.00 +/- 0.05
column: 15+ years percentage, range: [0.25, 0.42] variance: 0.00 +/- 0.06
```

Figure 6: Light Vehicles Age Distribution Statistics visualization.

- CO2 Emission Data:

There 2 rows of **missing data** founded in this file, it does not contain CO2 emission data in the year 2000 and 2018.

```
column: Year, range: [2001.00, 2017.00] variance: 24.00 +/- 4.90
column: Light passenger, range: [6.77, 8.08] variance: 0.08 +/- 0.29
column: Light commercial, range: [1.50, 2.47] variance: 0.06 +/- 0.25
```

Figure 7: CO2 Emission Data Statistics visualization.

2.4 Data Quality

2.4.1 Missing Values Imputation

According to the procedures adapted from OSAS and BDAS iterations, the miss values imputation applies **scikit-learn mean imputation** for CO2 Emission Data then won't influence the variance where pointed in Figure 7. Creating two missing rows for year 2000 and year 2018.

	Year	Light passenger	Light commercial	Motorcycle	Heavy fleet
0	2000.0	7.546587	1.830373	0.040207	3.064472
1	2001.0	6.765291	1.504675	0.028235	2.388743
2	2002.0	7.066370	1.559698	0.027965	2.572510
3	2003.0	7.375728	1.585027	0.028364	2.639134
4	2004.0	7.671046	1.582260	0.029396	2.653599
5	2005.0	7.549507	1.628644	0.032007	2.860993
6	2006.0	7.644814	1.663934	0.037437	2.919942
7	2007.0	7.796914	1.718954	0.040616	3.009903
8	2008.0	7.692528	1.764464	0.045334	3.077349
9	2009.0	7.626856	1.774602	0.046301	2.990650
10	2010.0	7.697162	1.830283	0.046434	3.109751
11	2011.0	7.575258	1.861729	0.045124	3.200497
12	2012.0	7.444775	1.879310	0.044464	3.213818
13	2013.0	7.415187	1.937471	0.045089	3.288810
14	2014.0	7.430754	1.993417	0.045306	3.345710
15	2015.0	7.662124	2.121840	0.046825	3.454175
16	2016.0	7.800260	2.237802	0.047416	3.509366
17	2017.0	8.077407	2.472227	0.047200	3.861072
18	2018.0	7.546587	1.830373	0.040207	3.064472

Figure 8: CO2 Emission Data Missing Values Imputation.

2.4.1 Check features normality

According to the CO2 emission data, check data normality within these files. If the fleet's groups follow the normal distribution or skew normal distribution then the average age of these groups must follow the same distribution, because they are from the same original fleets data. Mean and standard deviation able to get from data, then build pdf for the mean and std dev to plot the figures for check the normality.

- Number of Fleets:

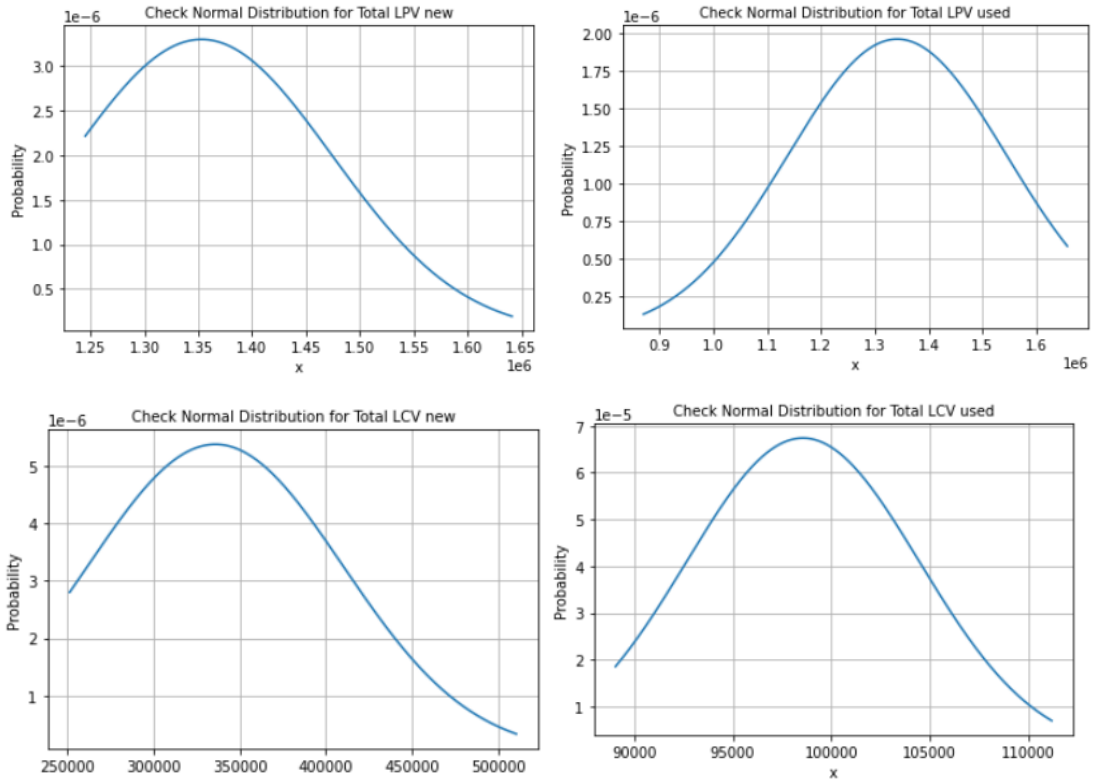
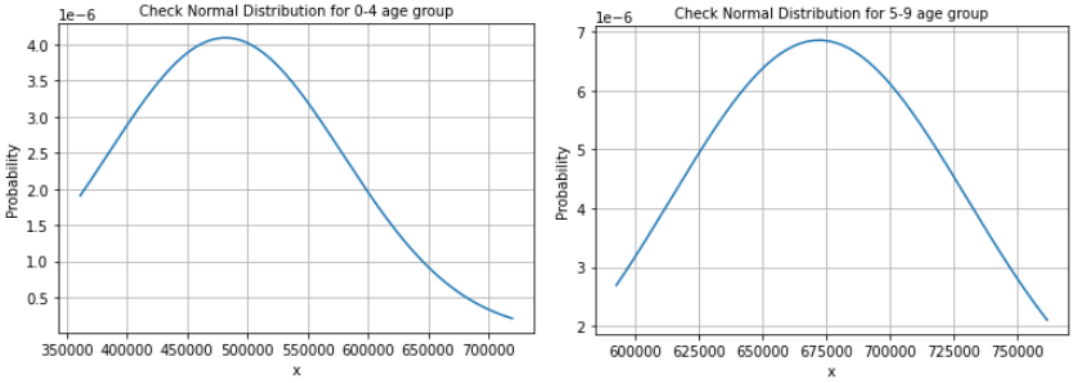


Figure 9: Number of Fleets check normality.

- Light Vehicles Age Distribution of Fleets:



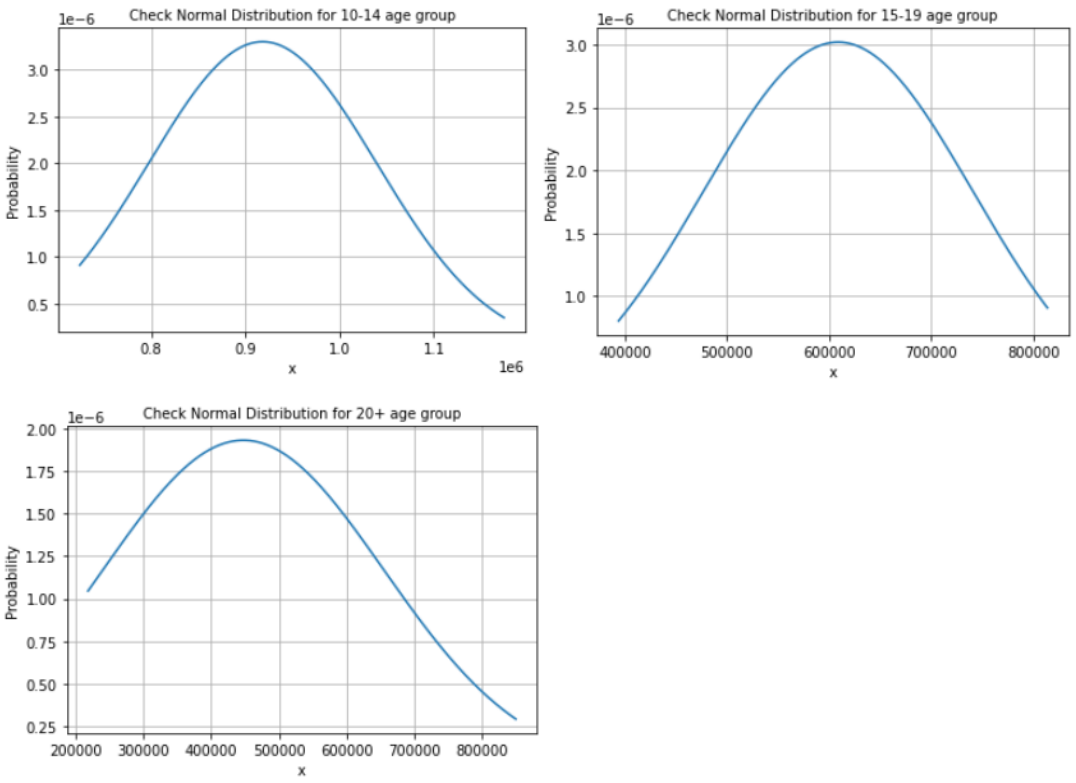


Figure 10: Light Vehicles Age Distribution check normality.

- CO2 Emission Data:

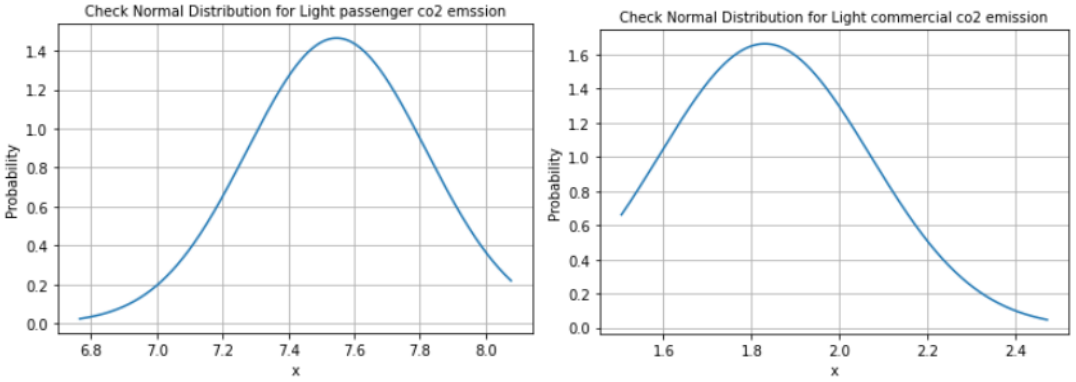


Figure 11: CO2 Emission check normality.

3 Data Preparation

3.1 Data Selection

- Number of Fleets:

Considering the business and datamining objectives is around real emission values and model predict emission values. In the Data Selection part will start from CO2 emission data, and **there are no age distributions data for the motor cycle fleets and heavy fleets**, and we mentioned above in step 2.3.1 that we **discard these two groups data**.

- Light Vehicles Age Distribution of Fleets:

All the light age distribution columns are needed, the project will construct new features using these data in next step 3.3.

- CO2 Emission Data:

CO2 emission data are using for target, mentioned in step2.2, thus only Light passenger and Light commercial columns are select for corresponding output.

3.2 Data Cleaning

According to the KDD process, the Data Cleaning step is clean the missing values, redundant values extreme values and outliers. There are no outliers and extreme values basing on step 2.3 data sets statistics. Hence, we clean the mining values and redundant values.

- Number of Fleets:

Excluding the columns which are redundant columns. There are some fields are redundant for Light fleets groups. For example, the “Total Light New” field are including the number of new fleets both for Light passenger and Light commercial, thus we can clean some cross-hold information groups between Light Passenger and Light Commercial. The project needs columns direct have relationship with LPV and LCV, because they are the target output values columns

- Light Vehicles Age Distribution of Fleets:

Cleaning for the null values, in step 2.3, Data Transposition have cleaned the nan (not a number) values.

- CO2 Emission Data:

In step 2.4, we imputation the missing rows.

3.3 Data Construct

During 3 previous iterations find that applying Number of Fleets and Light Vehicles Age Distribution of Fleets able to construct new features that increase the model accuracy.

It can estimate the number of vehicles in different age ranges for different groups by construct new features, applying the percentage of age groups to the number of Light Passenger and Light Commercial to get the number of age groups in LPV new, LPV used, LCV new or LCV used. There are 24 columns which generated by 6 age groups * 4 number fleets group and 1 column for the Period.

	0-4 years of LPV new	5-9 years of LPV new	10-14 years of LPV new	15-19 years of LPV new	20+ years of LPV new	15+ years of LPV new	0-4 years of LPV used	5-9 years of LPV used	10-14 years of LPV used	15-19 years of LPV used	...	10-14 years of LCV new	15-19 years of LCV new	20+ years of LCV new	15+ years of LCV new	0-4 years of LCV used	5-9 years of LCV used	10-14 years of LCV used
0	192258	368737	401921	201576	112006	313582	131137	251510	274144	137492	...	79075.4	39658.7	22036.5	61695.2	14460	27733.2	30229
1	177089	361595	406836	198483	112289	310772	134889	275426	309886	151184	...	82305.2	40154.2	22716.7	62871	13474.5	27513.4	30955.8
2	174848	343614	417826	197328	112190	309518	146826	288544	350862	165703	...	86750.6	40970.1	23293.3	64263.4	13570.8	26669.5	32429.5
3	179270	343070	409726	201590	111576	313166	165486	316692	378223	186090	...	87280.4	42943	23768.1	66711.1	14220	27212.9	32500.2
4	184137	332047	403897	213044	116033	329077	182988	329976	401377	211715	...	88620.7	46745	25459.2	72204.1	14995.2	27040.2	32891.3
5	191532	320604	397910	227222	120072	347294	201211	336806	418018	238704	...	90049.8	51421.9	27173.2	78595.1	15797.9	26444.1	32820.4
6	199129	276596	428141	239480	124130	363610	214280	297642	460718	257702	...	99164.5	55467.5	28750.5	84218	16365.5	22732.1	35187
7	206737	252825	436598	255562	129041	384603	225712	276030	476671	279019	...	103620	60654.2	30626	91280.1	16955.4	20735.3	35807.3
8	209487	246531	436877	261067	138596	399664	226923	267051	473240	282797	...	105650	63133.8	33516.7	96650.5	16667.2	19614.6	34759
9	197037	248284	416022	276203	157378	433581	211453	266450	446461	296411	...	101449	67353.1	38377.3	105730	15071.4	18991.3	31821.6
10	187904	264089	387832	287708	179462	467169	201008	282506	414878	307771	...	95406.5	70775.9	44147.4	114923	13702.8	19258.6	28282.5
11	179817	274072	341725	324272	195940	520213	188905	287924	358996	340662	...	85114.7	80767.8	48803.6	129571	12471.2	19008.2	23700.3
12	180668	283220	321201	345037	220330	565367	185412	290656	329634	354096	...	80859.5	86859.8	55466	142326	11913.8	18676.4	21180.9
13	189838	301834	310411	348637	238889	587526	191931	305163	313835	352482	...	80188.6	90063.4	61712.3	151776	12233.6	19451	20003.7
14	214795	298521	327128	330623	263665	594288	216961	301531	330427	333957	...	87329.2	88262.1	70387.3	158649	13745.4	19103.3	20934
15	237588	283004	363130	309776	286445	596221	240662	286665	367827	313784	...	100269	85536.5	79094.2	164631	15196.8	18101.7	23226.7
16	262337	266038	406590	272947	327263	600209	265988	269740	412248	276745	...	116523	78222.9	93789.2	172012	17072.5	17313.4	26460.3
17	284523	254698	453647	251079	346148	597227	289449	259108	461502	255426	...	135850	75188.6	103658	178847	19032.7	17037.6	30346
18	301294	251947	491706	240817	355468	596285	304224	254396	496486	243159	...	152993	74929.8	110603	185533	20411.3	17068.2	33310.8

19 rows × 24 columns

Figure 12: New Constructed Age Distribution.

3.4 Data Integration

After imputation on CO2 emission data, all the data sets are from 2000 to 2018, from previous experiences, pandas are able to integrate all the dataset into one. Hence, we got one data frame for input values with 30 columns and 1 period column, last 2 columns for CO2 emission target values. There are **no problems** find when integrate data frames, since pandas provided concat dataframe object method and able to select columns with their column names.

Total LCV new	Total LCV used	Light passenger average age	Light commercial average age	0-4 years of LPV new	5-9 years of LPV new	10-14 years of LPV new	...	20+ years of LCV new	15+ years of LCV new	0-4 years of LCV used	5-9 years of LCV used	10-14 years of LCV used	15-19 years of LCV used	20+ years of LCV used	15+ years of LCV used	Light passenger	Light commercial
251143	96007	11.713870	12.404023	192258	368736	401920	...	22036	61695	14459	27733	30228	15160	8424	23584	7.546587	1.830373
254155	95590	11.863626	12.588450	177089	361594	406836	...	22716	62870	13474	27513	30955	15102	8543	23646	6.765291	1.504675
258659	96693	11.957576	12.671381	174847	343613	417825	...	23293	64263	13570	26669	32429	15315	8707	24023	7.066370	1.559698
265261	98774	12.011138	12.682297	179270	343070	409726	...	23768	66711	14220	27212	32500	15990	8850	24840	7.375728	1.585027
274083	101725	12.094087	12.633977	184137	332047	403896	...	25459	72204	14995	27040	32891	17349	9449	26798	7.671046	1.582260
284545	103708	12.212825	12.564520	191531	320603	397909	...	27173	78595	15797	26444	32820	18741	9903	28645	7.549507	1.628644
293568	104168	12.401873	12.578584	199128	276595	428141	...	28750	84217	16365	22732	35186	19681	10201	29883	7.644814	1.663934
303971	105041	12.591679	12.552226	206736	252824	436597	...	30625	91280	16955	20735	35807	20959	10583	31543	7.796914	1.718954
312579	102839	12.844178	12.616798	209486	246531	436876	...	33516	96650	16667	19614	34758	20771	11027	31798	7.692528	1.764464
315772	99049	13.233213	12.946859	197036	248283	416022	...	38377	105730	15071	18991	31821	21126	12037	33164	7.626856	1.774602
321520	95312	13.532333	13.163049	187903	264089	387832	...	44147	114923	13702	19258	28282	20980	13087	34068	7.697162	1.830283
327738	91259	13.812682	13.291567	179816	274072	341724	...	48803	129571	12471	19008	23700	22489	13589	36079	7.575258	1.861729
339965	89053	14.047410	13.396964	180668	283220	321201	...	55466	142325	11913	18676	21180	22752	14529	37281	7.444775	1.879310
358978	89550	14.194163	13.338720	189837	301833	310410	...	61712	151775	12233	19450	20003	22467	15394	37861	7.415187	1.937471
383012	91813	14.244258	13.175905	214794	298520	327127	...	70387	158649	13745	19103	20933	21157	16872	38030	7.430754	1.993417
408647	94661	14.263345	13.011491	237588	283003	363129	...	79094	164630	15196	18101	23226	19814	18321	38135	7.662124	2.121840
439961	99907	14.303513	12.787277	262336	266038	406590	...	93789	172012	17072	17313	26460	17762	21297	39060	7.800260	2.237802
476173	106367	14.311541	12.514544	284522	254697	453646	...	103658	178846	19032	17037	30346	16795	23155	39950	8.077407	2.472227
510666	111186	14.415534	12.343552	301294	251946	491705	...	110603	185532	20411	17068	33310	16314	24081	40395	7.546587	1.830373

Figure 12: Integrated data set.

3.5 Data Formatting

Since the new construct data is object type shown as figure below which actually is float, in this step we need to convert to integer, because there are not exist that parts of vehicles could run on the road.

Convert object type shown in rectangle below to int64 type in yellow below:
age and CO2 emission columns already in float64 type, others should be in int64 type, only new constructed columns are in object type thus it will only change type for these columns.

```
In  [11]: cleaned_data_df.dtypes

Out[11]: Period                          int64
         Total LPV new                   int64
          Total LPV used                 int64
         Total LCV new                   int64
          Total LCV used                 int64
         Light passenger average age     float64
         Light commercial average age    float64
         0-4 years of LPV new            object
         5-9 years of LPV new            object
         10-14 years of LPV new          object
         15-19 years of LPV new          object
         20+ years of LPV new            object
         15+ years of LPV new            object
         0-4 years of LPV used           object
         5-9 years of LPV used           object
         10-14 years of LPV used         object
         15-19 years of LPV used         object
         20+ years of LPV used           object
         15+ years of LPV used           object
         0-4 years of LCV new            object
         5-9 years of LCV new            object
         10-14 years of LCV new          object
         15-19 years of LCV new          object
         20+ years of LCV new            object
         15+ years of LCV new            object
         0-4 years of LCV used           object
         5-9 years of LCV used           object
         10-14 years of LCV used         object
         15-19 years of LCV used         object
         20+ years of LCV used           object
         15+ years of LCV used           object
         Light passenger                 float64
         Light commercial                float64
         dtype: object
```

```
Period                          int64
Total LPV new                   int64
 Total LPV used                 int64
Total LCV new                   int64
 Total LCV used                 int64
Light passenger average age     float64
Light commercial average age    float64
0-4 years of LPV new            int64
5-9 years of LPV new            int64
10-14 years of LPV new          int64
15-19 years of LPV new          int64
20+ years of LPV new            int64
15+ years of LPV new            int64
0-4 years of LPV used           int64
5-9 years of LPV used           int64
10-14 years of LPV used         int64
15-19 years of LPV used         int64
20+ years of LPV used           int64
15+ years of LPV used           int64
0-4 years of LCV new            int64
5-9 years of LCV new            int64
10-14 years of LCV new          int64
15-19 years of LCV new          int64
20+ years of LCV new            int64
15+ years of LCV new            int64
0-4 years of LCV used           int64
5-9 years of LCV used           int64
10-14 years of LCV used         int64
15-19 years of LCV used         int64
20+ years of LCV used           int64
15+ years of LCV used           int64
Light passenger                 float64
Light commercial                float64
dtype: object
```

Figure 13: Converted Data Types.

4 Data Transformation

4.1 Data Reduction

Basing on BDAS and OSAS experience, using Principle Component Analysis (PCA) for find are there any features could be reduced. PCA will return n ratios to show how many variances that the n features under this PCA component can explained. Aiming for the components can explained over 90% variances.

- For LPV:

The first 2 components explained 0.9095881686967918 variance, ranking the features **unimportance** for first 2 components in ascending.

```
[8.15390324e-01 9.41978448e-02 8.17119244e-02 7.50059559e-03
 8.88775169e-04 2.98629476e-04 7.93352645e-06 2.14514515e-06
 1.68783289e-06 1.40183256e-07 6.51477294e-13 4.36104833e-13
 1.36339300e-13 1.06180843e-13 2.28458149e-14 2.16830936e-16]
```

Figure 14: LPV PCA ratios.

```
[-6.17384026e-01 -4.22388682e-01 -3.48123730e-01 -3.32328525e-01
 -2.73625879e-01 -2.45265666e-01 -1.48761857e-01 -1.29053217e-01
 -1.02857621e-01 -8.72487626e-02 -8.38773119e-02 -1.72400829e-05
 -2.98352232e-06  6.11808248e-03  1.79344876e-02  9.69261430e-02]
[2, 15, 9, 1, 14, 8, 13, 10, 7, 4, 12, 0, 3, 6, 11, 5]
[-5.33210681e-01 -4.48752542e-01 -3.62034556e-01 -1.77394300e-01
 -9.84292324e-02  2.65330341e-06  5.14555414e-06  9.80370954e-03
  5.77236642e-02  9.07354028e-02  1.35142869e-01  1.39055422e-01
  1.61312178e-01  2.43986527e-01  2.52049435e-01  3.79131164e-01]
[12, 2, 6, 10, 4, 3, 0, 11, 1, 14, 8, 5, 13, 7, 15, 9]
```

Figure 15: LPV top two components ranked features.

- For LCV:

The first component explained 0.94477867 variance, ranking the features **unimportance** for first components in ascending.

```
[9.43506488e-01 4.29729645e-02 1.27173317e-02 6.00974134e-04
 1.27722380e-04 5.70790317e-05 1.47682085e-05 2.01237408e-06
 4.94915320e-07 1.33894605e-07 2.23233991e-08 8.46538499e-09
 8.90326264e-13 1.46161184e-13 2.27102080e-14 2.42373903e-15]
```

Figure 16: LCV PCA ratios.

```
[-6.99980964e-01 -5.97771524e-01 -3.46025781e-01 -1.32969219e-01
 -8.39788622e-02 -6.55031413e-02 -5.08754388e-02 -2.31958257e-02
 -1.26990705e-02 -6.32865822e-03 -2.81013492e-03 -2.79350672e-05
 -7.54374729e-07  2.08409574e-03  1.55001342e-02  1.89187709e-02]
[15, 9, 1, 8, 7, 4, 6, 14, 5, 13, 10, 0, 3, 2, 12, 11]
```

Figure 17: LCV top component ranked features.

From above information **reduce features related to 15+ year groups** both for LPV and LCV inputs, which is 15 and 9 in figure 15 and 16. Since these groups have potential influence on the model accuracy, their high variance are made from other groups. Period we will use for indicate the result and it won't attend the datamining model.

```
In [ ]: # ---- step 4.1 ----
reduced_LPV_cols = ['Total LPV new', ' Total LPV used', 'Light passenger average age','0-4 years of LPV new', '5-9 years of LPV new',
                    '10-14 years of LPV new', '15-19 years of LPV new', '20+ years of LPV new', '0-4 years of LPV used',
                    '5-9 years of LPV used', '10-14 years of LPV used', '15-19 years of LPV used', '20+ years of LPV used',
                    'Light passenger']

reduced_LCV_cols = ['Total LCV new', ' Total LCV used', 'Light commercial average age', '0-4 years of LCV new', '5-9 years of LCV new',
                    '10-14 years of LCV new', '15-19 years of LCV new', '20+ years of LCV new', '0-4 years of LCV used',
                    '5-9 years of LCV used', '10-14 years of LCV used', '15-19 years of LCV used', '20+ years of LCV used',
                    'Light commercial']
```

Figure 18: Reduced LPV and LCV columns for next steps

4.2 Data Transformation

From Step 2.3 Data statistics, different columns have different ranges, for avoid overfitting in next steps of machine learning models this step we will do normalization under L2 normalization. The L2 normalization will consider all the feature ranges, it will prevent weights from going to large, and transform the values for all data in range (0,1), but not equal to 0 or 1.

From OSAS and BDAS experience, Scikit machine learning models provide **more powerful** L2 normalization and force L2 normalization on the models, it could change the parameter lambda and omega. If the model input values already done l2 normalization then the model will force do l2 normalization again which will influence the accuracy. Hence here we represent we are **able to apply** l2 normalization on data before the training models and applying powerful normalization method on spark models from step 6.

$$L' = L + \frac{\lambda}{2n}\sum_{w} w^2 \quad w \leftarrow \left(1 - \frac{\eta\lambda}{n}\right) w - \eta\frac{\partial L}{\partial w}$$

Example of LPV data under L2 normalization will transform to range 0 to 1 which make the input smoother for the model:

```
In [30]: # ---- step 4.2 ----
LPV_df, LCV_df = step_4_2_normalization(cleaned_data_df,reduced_LPV_cols,reduced_LCV_cols)
LPV_df
```

Out[30]:

	Total LPV new	Total LPV used	Light passenger average age	0-4 years of LPV new	5-9 years of LPV new	10-14 years of LPV new	15-19 years of LPV new	20+ years of LPV new	0-4 years of LPV used	5-
0	0.742510709923	0.506455337114	0.000006813699	0.111832234808	0.214485591935	0.233787992251	0.117251728050	0.065151417844	0.076278916580	0.146
1	0.714235993429	0.544031635655	0.000006744787	0.100679807848	0.205575808994	0.231297089630	0.112842866022	0.063839283882	0.076687416615	0.156
2	0.687310716569	0.577157727152	0.000006596996	0.096463103664	0.189571319149	0.230514085391	0.108865873134	0.061894680131	0.081003364058	0.159
3	0.660092946235	0.609339441433	0.000006367055	0.095030297520	0.181860010990	0.217194085355	0.106862038696	0.059145983578	0.087723455209	0.167
4	0.638363673527	0.634380665396	0.000006180504	0.094100483488	0.169687695795	0.206405061876	0.108872977208	0.059296433091	0.093513304075	0.168
5	0.621636080669	0.653051139139	0.000006038095	0.094694096156	0.158508081250	0.196728639788	0.112339450129	0.059364330127	0.099479452869	0.166
6	0.613374294298	0.660045339996	0.000006001684	0.096364741009	0.133853629522	0.207191839321	0.115892431887	0.060070200759	0.103697303761	0.144
7	0.608718792218	0.664590645332	0.000005984556	0.098257200189	0.120161841094	0.207505218398	0.121463153948	0.061329952753	0.107276087227	0.131
8	0.611819784687	0.662743626956	0.000006079667	0.099158164984	0.116693056202	0.206791014604	0.123573530727	0.065603071490	0.107411326363	0.126
9	0.616349577400	0.661445215724	0.000006298659	0.093783925028	0.118176141708	0.198015469548	0.131464847340	0.074907765855	0.100646035744	0.126
10	0.619061376415	0.662231954797	0.000006409623	0.089000715238	0.125086400358	0.183697574771	0.136273123787	0.085002141303	0.095207456868	0.133
11	0.626165715044	0.657813631166	0.000006573079	0.085569531394	0.130423391735	0.162617133882	0.154312202942	0.093242503344	0.089894738666	0.137
12	0.634350720120	0.651005011802	0.000006598495	0.084865253690	0.133037046683	0.150877877380	0.162073901699	0.103495231476	0.087093185024	0.136
13	0.639754966500	0.646811290565	0.000006534782	0.087398146510	0.138959448135	0.142908172054	0.160506856970	0.109980501290	0.088362193133	0.140
14	0.641032021600	0.647496265621	0.000006364277	0.095969092497	0.133377531459	0.146159023629	0.147720575526	0.117804011305	0.096936852557	0.134
15	0.640317289155	0.648600631487	0.000006171229	0.102795650978	0.122445062940	0.157112656970	0.134028762300	0.123933857976	0.104125225853	0.124
16	0.639310558085	0.648206997547	0.000005956576	0.109247593640	0.110789260020	0.169320943744	0.113666037805	0.136285473552	0.110768021505	0.112
17	0.636917702684	0.647946141951	0.000005732538	0.113966280360	0.102019772492	0.181709488968	0.100570169408	0.138650789795	0.115939807410	0.103
18	0.638267727303	0.644473721576	0.000005606136	0.117171878583	0.097980663808	0.191221858247	0.093652645870	0.138239903005	0.118311342376	0.098

Figure 19: LPV after L2 normalization

5 Datamining Objectives

5.1 Understanding Datamining Objectives

For the research objective predict: Given **whole** New Zealand one year's vehicles fleets statistics then able to predict the **whole** year CO2 emission values.

For the practical objective predict: Given **partial** New Zealand one year's **different age** vehicles fleets statistics on the model then could predict **partial** year CO2 emission values

5.2 Datamining Method Selection

There are three main categories of Machine Learning:

- Supervised Learning: indicate to problem that the model input data **have** labels, two problems are derived, they are:
    - Classification: predict a **label** for given features vector, accepted numeric, String, and Boolean type of data.
    - Regression: predict a **value** for given features vector, accepted numeric type of data, but the data should have some variance, better not equal to 0 or 1 which is same as Boolean values. If some values are equal to 0 it might make some features useless, because whatever 0 relation to other features will be useless. If some values are equal to 2 then it might cause this feature become useless, because everything relation to1 will be themselves.
- Unsupervised Learning: indicate to problem that the model input data **don't have** labels; one problem belongs to this is:
    - Clustering: Partitioning data into **n groups**.
- Reinforcement Learning: build software agents in an environment to optimized the notion of cumulative reward.

Explanation: After match Machine Learning Algorithms and Objectives, this project is a supervised learning problem its data have labels, otherwise it will be an un-supervised learning problem. It will be either a classification problem or regression problem. The Objective is for predict values, which is means the target predict value have relation with all feature values thus it is a regression problem.

6 Datamining Algorithm Selection

6.1 Conduct Datamining Exploratory analysis and discuss

From literature review, BDAS iteration and OSAS iteration, either Linear Model or Neural Network are able to match research objectives. From Literature review, the Southampton study, Serbia study and many European studies use Neural Network implement prediction and the China, Bangladesh and Bahrain studies are applying Linear Model, both of them have good results for prediction. Nearly none of studies using Case-Based Learning Models such as Decision Tree, SVM and KNN. These conclusions coincide with previous iterations, thus this study select models from Linear Model and Neural Network.

- Linear Model

The most traditional way is model working on **Linear Model**, here is its equation denote equation 1:

$$equation\ 1: \ y_i = \beta_0 + \ \beta_1 x_{i1} + \cdots + \ \beta_p x_{ip} + \varepsilon_i \ \ i = 1, \ldots, n.$$

where y_i is our target values, where $(\beta_0\, \beta_1 \ldots \beta_p)$ are the model coefficients, $(x_{i1} \ldots x_{ip})$ are the features. Objective for the linear model is for find best coefficients based on our data that could predict accuracy target values. We will discuss how to choose the best method of optimization in step 5.2.

Advantages: Simple to use, able to train data fast, predicting values not rely on other data.
Disadvantages: Easily overfitting or underfitting.

There are several types of Linear Regression:
Ordinary least squares Regression: Using original data for training.

Ridge Regression: Applying L2 regularization on data before training.
Lasso Regression: Applying L1 regularization on data before training.
Elastic Net Regression: Combine Applying L1 + L2 regularization on data before training

- Artificial Neural Network

Artificial Neural Network is inspired from animal's brain to simulated biological neural networks. It is a power technique cross computer science, statistics, and optimization. Example in figure 20, similar with equation 1, x indicates input features, then follow some arrows to the next node in hidden layer, the arrows indicate some activation function with coefficients like equation 1. After feed-forward propagation from input nodes through hidden nodes to the output nodes.

Advantages: Accuracy, after training model, feedforward run the model able to get the predicts, predicting values not rely on other data.
Disadvantages: Hardly to train, need much time to train the model, gradient vanish and gradient exploding problems.

The Artificial Neural Network models are always different when they have different architectures, optimization algorithms, loss functions, or activity functions.

- To measure model performance.

Definition: Residuals means the difference values between real values with predict values.

$$Residual(y, \hat{y}\,) = (yi - \hat{y}\,) \ \text{where}\ y\ is\ real\ value,\ \hat{y}\ is\ the\ predict\ value$$

Definition: RMSE is root of MSE.

$$\mathrm{RMSE}(y, \hat{y}\,) = \sqrt{\sum i (yi - \hat{y}\imath\,) 2}$$

Definition: R2 is combine use sum of squares and sum of residuals.

$$\mathrm{R2}(y, \hat{y}\,) = 1 - \sum ni = 1 (yi - \hat{y}\,) 2 \sum ni = 1 (yi - \overline{y}) 2$$

Some previous studies using either RMSE or R2 for measure the performance, R2 will return a value from 0 – 1, good models will more and more close to 1. In this iteration using R2 measurement for comparing models in a Regression problem.

6.2 Select datamining algorithm based on discussion

This study will select **Linear Regression model**, the reasons are:

1. This study data has strong linear relationship.
2. The Dataset have small number of features.
3. Lack of economic data like gross domestic product data.

From the above reasons, this study belongs to a GM problem. Linear relationship data are more suit for Linear Models, conclude by previous studies from literature review. Hence, the algorithm selection become select from **Ordinary least squares Regression, Ridge Regression, Lasso Regression, Elastic Net Regression.**

Select model Criteria One: The success models will have R2 with greater than 0.8. For Matching Datamining Objective One.

Select model Criteria Two: The success models will have lower residuals, which means avoid model overfitting in the case of optimized with R2 to 1 but have residuals away from 0. For Matching Datamining Objective Two

6.3 Build Appropriate Models

From OSAS iteration, the Ridge regressor best solvers are either sag or saga. From BDAS iteration, the study find Lasso and Elastic Net regressor are also good solutions for this problem. Hence, the study builds appropriate from these regressors and selection base on R2 score and residuals.

The result shows that under same environment all the regressor performance similar, and **Ridge Regressor** performance best according to these four regressors. After change lambda and omega of the L2 normalization, which indicate Lasso Regressor and Elastic Net Regressor, then the r2 score decreased.

```
Select algorithm for LPV data set:
-------------------------------------------------
Ordinary Least Square Regression: r2 score = 0.380950,
Ridge with sag solver: r2 score = 0.998467,
Ridge with saga solver: r2 score = 0.998483,
Lasso Regression: r2 score = 0.995819,
Elastic Net Regression: r2 score = 0.995817,
-------------------------------------------------

Select algorithm for LCV data set:
-------------------------------------------------
Ordinary Least Square Regression: r2 score = 0.836357,
Ridge with sag solver: r2 score = 0.986505,
Ridge with saga solver: r2 score = 0.990198,
Lasso Regression: r2 score = 0.963290,
Elastic Net Regression: r2 score = 0.964135,
-------------------------------------------------
```

Figure 21: Regressors Results

```
Check Residuals for LPV
Ridge with sag solver: r2 score = 0.998465,
[ 0.00162934 -0.05485702 -0.11219658 -0.18137276  0.02533252  0.16037346
  0.01919133]
Ridge with saga solver: r2 score = 0.998454,
[ 0.02907332 -0.09667338 -0.15894218 -0.15322426  0.01611881  0.12073345
  0.04476539]

Check Residuals for LCV
Ridge with sag solver: r2 score = 0.986527,
[ 0.11150293 -0.07894497  0.02694912  0.10151121 -0.08202497 -0.0295616
  0.02576853]
Ridge with saga solver: r2 score = 0.990211,
[ 0.10286461 -0.05174725  0.03237584  0.09467447 -0.05756324 -0.01399293
  0.02833125]
```

Figure 22: Residuals Results

Comparing residuals results, Ridge with sag solver have lower range residuals on large terms and it also have a better R2 score for LPV data set. For LCV data set, residuals are similar. All the residuals are in small range. Hence, basing on the criteria, select **Ridge with sag solver for** models.

7 Datamining

7.1 Create and justify test design

- **For Datamining Objective One:**

Splitting data set to 70% as a training set and 30% as a testing set to avoid overfitting and underfitting the model. After splitting the training set and testing set, adding noisy data into these files respectively, for avoid overfitting and also make the model learn that no vehicles on road indicate to no CO2 emissions.

- **For Datamining Objective Two**:

Prepare splitting age groups set base on the data in 2017, One reason is the real emission value for 2018 is missing which we imputed in Data Pre-processing. Comparing using mean values, data in 2017 are more fit current situation. This study creates files with 10 instances of 2017 and trim it only contains one group data.

	Total LPV new	Total LPV used	Light passenger average age	0-4 years of LPV new	5-9 years of LPV new	10-14 years of LPV new	15-19 years of LPV new	20+ years of LPV new	0-4 years of LPV used	5-9 years of LPV used	10-14 years of LPV used	15-19 years of LPV used	20+ years of LPV used
0	1590094.0	1617627.0	14.311541	284522.0	254697.0	453646.0	251078.0	346148.0	289449.0	259108.0	461501.0	255426.0	352141.0
1	284522.0	0.0	2.975102	284522.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	254697.0	0.0	7.975102	0.0	254697.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	453646.0	0.0	12.975102	0.0	0.0	453646.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	251078.0	0.0	17.975102	0.0	0.0	0.0	251078.0	0.0	0.0	0.0	0.0	0.0	0.0
5	346148.0	0.0	22.975102	0.0	0.0	0.0	0.0	346148.0	0.0	0.0	0.0	0.0	0.0
6	0.0	289449.0	9.130000	0.0	0.0	0.0	0.0	0.0	289449.0	0.0	0.0	0.0	0.0
7	0.0	259108.0	14.130000	0.0	0.0	0.0	0.0	0.0	0.0	259108.0	0.0	0.0	0.0
8	0.0	461501.0	19.130000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	461501.0	0.0	0.0
9	0.0	255426.0	24.130000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	255426.0	0.0
10	0.0	352141.0	29.130000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	352141.0

Figure 23: LPV 2017 splitting group dataset

	Total LCV new	Total LCV used	Light commercial average age	0-4 years of LCV new	5-9 years of LCV new	10-14 years of LCV new	15-19 years of LCV new	20+ years of LCV new	0-4 years of LCV used	5-9 years of LCV used	10-14 years of LCV used	15-19 years of LCV used	20+ years of LCV used
0	476173.0	106367.0	12.514544	85203.0	76272.0	135849.0	75188.0	103658.0	19032.0	17037.0	30346.0	16795.0	23155.0
1	85203.0	0.0	2.868357	85203.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	76272.0	0.0	7.868357	0.0	76272.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	135849.0	0.0	12.868357	0.0	0.0	135849.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	75188.0	0.0	17.868357	0.0	0.0	0.0	75188.0	0.0	0.0	0.0	0.0	0.0	0.0
5	103658.0	0.0	22.868357	0.0	0.0	0.0	0.0	103658.0	0.0	0.0	0.0	0.0	0.0
6	0.0	19032.0	7.980000	0.0	0.0	0.0	0.0	0.0	19032.0	0.0	0.0	0.0	0.0
7	0.0	17037.0	12.980000	0.0	0.0	0.0	0.0	0.0	0.0	17037.0	0.0	0.0	0.0
8	0.0	30346.0	17.980000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30346.0	0.0	0.0
9	0.0	16795.0	22.980000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16795.0	0.0
10	0.0	23155.0	27.980000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23155.0

Figure 24: LCV 2017 splitting group dataset

7.2 Conduct Datamining

Run the pipeline for LPV and LCV models, setting Random Seed with 722, split the training set and test set with 70% and 30% respectively for the LPV and LCV data sets.

Verification R2 score and residuals with the test sets using regression models for predict single age groups values.

```
LPV_regressor = Ridge(solver='sag')
LPV_regressor.fit(LPV_X_train, LPV_y_train)
print('LPV: r2 score = %f,' %(LPV_regressor.score(LPV_X_train, LPV_y_train)) )
print('LPV residuals:', np.subtract(LPV_regressor.predict(LPV_X_test),LPV_y_test.values) )
```

```
LPV: r2 score = 0.984496,
LPV residuals: [ 0.00135812 -0.05456789 -0.1117188  -0.18174419  0.02543127  0.16081591
  0.01907663]
```

Figure 25: LPV model R2 score and residuals.

```
LCV_regressor = Ridge(solver='sag')
LCV_regressor.fit(LCV_X_train, LCV_y_train)
print('LCV: r2 score = %f,' %(LCV_regressor.score(LCV_X_train, LCV_y_train)) )
print('LCV residuals:', np.subtract(LCV_regressor.predict(LCV_X_test),LCV_y_test.values) )
```

```
LCV: r2 score = 0.910515,
LCV residuals: [ 0.11150094 -0.0788993   0.02700045  0.10149281 -0.0819709  -0.02953021
  0.02580158]
```

Figure 26: LCV model R2 score and residuals.

```
LPV 2017 predictions of groups:
0-4 years new: 0.976542 percentage:  10.802376 %
5-9 years new: 1.389575 percentage:  15.371284 %
10-14 years new: 2.269714 percentage:  25.107268 %
15-19 years new: 1.246504 percentage:  13.788661 %
20+ years new: 0.358798 percentage:  3.968978 %
0-4 years used: 0.194920 percentage:  2.156174 %
5-9 years used: 0.572353 percentage:  6.331289 %
10-14 years used: 0.918900 percentage:  10.164746 %
15-19 years used: 0.599088 percentage:  6.627030 %
20+ years used: 0.513674 percentage:  5.682193 %
```

```
LCV 2017 predictions of groups:
0-4 years new: 0.273341 percentage:  10.596666 %
5-9 years new: 0.659302 percentage:  25.559291 %
10-14 years new: 0.051102 percentage:  1.981088 %
15-19 years new: 0.442954 percentage:  17.172090 %
20+ years new: 0.348348 percentage:  13.504451 %
0-4 years used: 0.143159 percentage:  5.549877 %
5-9 years used: 0.158620 percentage:  6.149264 %
10-14 years used: 0.205294 percentage:  7.958655 %
15-19 years used: 0.133985 percentage:  5.194233 %
20+ years used: 0.163396 percentage:  6.334385 %
```

Figure 27: LPV and LCV models for 2017 groups predicitions.

7.3 Search for patterns

	LPV	Groups Predict Emission		LCV	Groups Predict Emission	
0-4 years NEW	284522	0.974931	10.802376 %	85203	0.384250	10.482671 %
5-9 years NEW	254697	1.390561	15.371284 %	76272	0.528610	14.420937 %
10-14 years NEW	453646	2.265125	25.107268 %	135849	0.602434	16.434932 %
15-19 years NEW	251078	1.242139	13.788661 %	75188	0.337329	9.202630 %
20+ years NEW	346148	0.357665	3.968978 %	103658	0.530958	14.485000 %
0-4 years USED	289449	0.197123	2.189383 %	19032	0.240762	6.568200 %
5-9 years USED	259108	0.577938	6.351886 %	17037	0.226516	6.179552 %
10-14 years USED	461501	0.921495	10.167598 %	30346	0.340285	9.283260 %
15-19 years USED	255426	0.598005	6.625376 %	16795	0.198131	5.405174 %
20+ years USED	352141	0.510036	5.672847 %	23155	0.276298	7.537645 %

The relaitionships that we discover for datamining are:

1. Figures 25 and 26 show the model are accuracy enough to predict CO2 emission by the fleet's information, proved by all residuals are smaller and LPV predictions are greater than LCV predictions.
2. Figure 27 compare with 28, new import fleets have more CO2 emissions than Used import fleets
3. Figure 27 in LPV categories, 10-14 age groups contribute the most CO2 emissions among new import groups or used groups.
4. Figure 27 in LCV categories, 10-14 age groups contribute the most CO2 emissions among new import groups, and 20+ age groups for used groups.

8 Result Analysis

8.1 Study and discuss the mined patterns

The Datamining goal is a regression problem, the first need is the model should be accuracy and the mined **pattern 1** proved that the residuals are quite small in range [0.01 to 0.2] compare with target values range [6.90 to 8.0]. Under an accuracy model then study can discuss the next project needed which is finding CO2 distribution of age groups.

The Pattern 2, 3 and 4 results caused by the fleets numbers and age trend, **it means the models explained age trend in correct way.** Becase in the pattern 4, 20+ year groups of LCV have samiliar number of fleets among different groups, but the used fleets contribute more CO2 emission, it is same as common sense that is 20+ year is counting since they come into New Zealand. They are import as used vehicles that they must have higher used age than new import group. The pattern 2

and the pattern 3 proved the result follow the normal distribution, the median position groups have the biggest proportion. For example, there are 453646 vehicles and 461601 vehicles for the median position group (10-14 age group) and they contribute the most CO2 emission.

However, this study do not know that LPV model explaind more on new vehicles? Since new vehichle contribute more CO2 emission when the similar number of vehicles in a same age range. Becase this study models are **Gray Models**.

8.2 Visualize the data, results, models and patterns

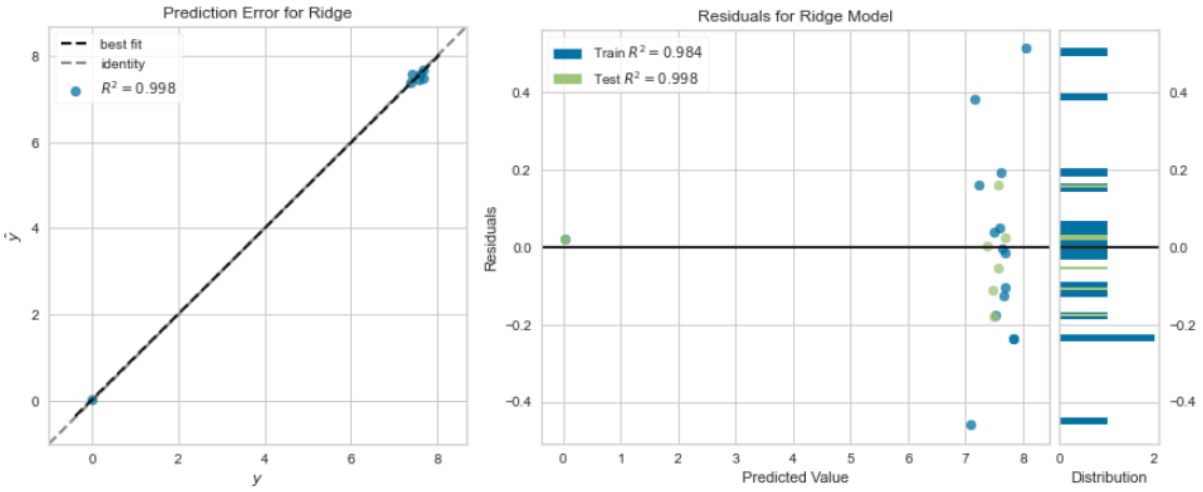


Figure 28: LPV model for 2017 groups predicitions.

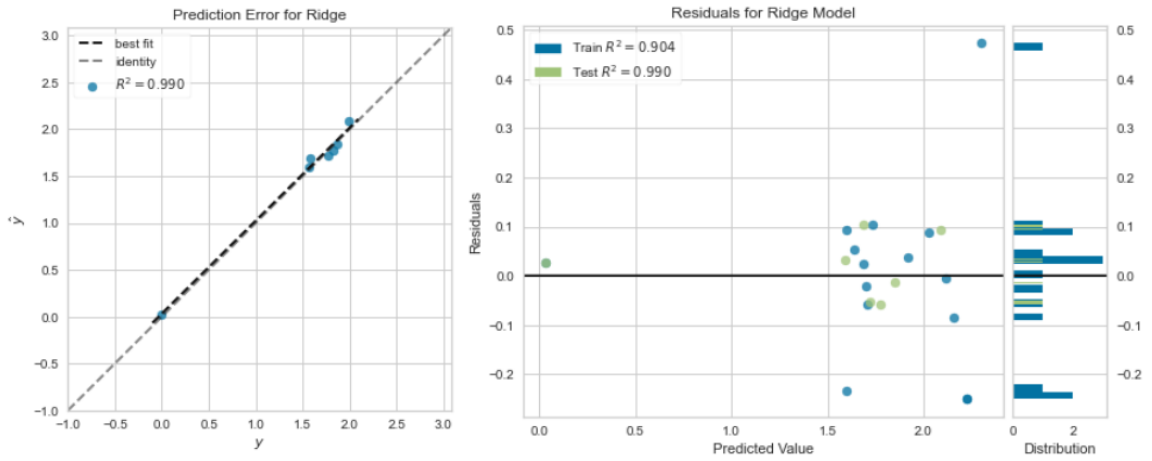


Figure 29: LCV model for 2017 groups predicitions.

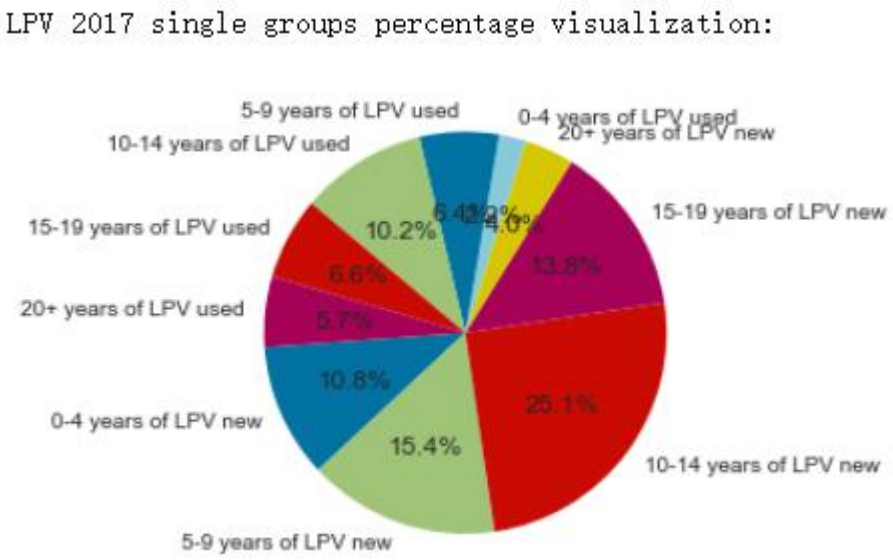


Figure 30: LPV model for 2017 single groups percentage.

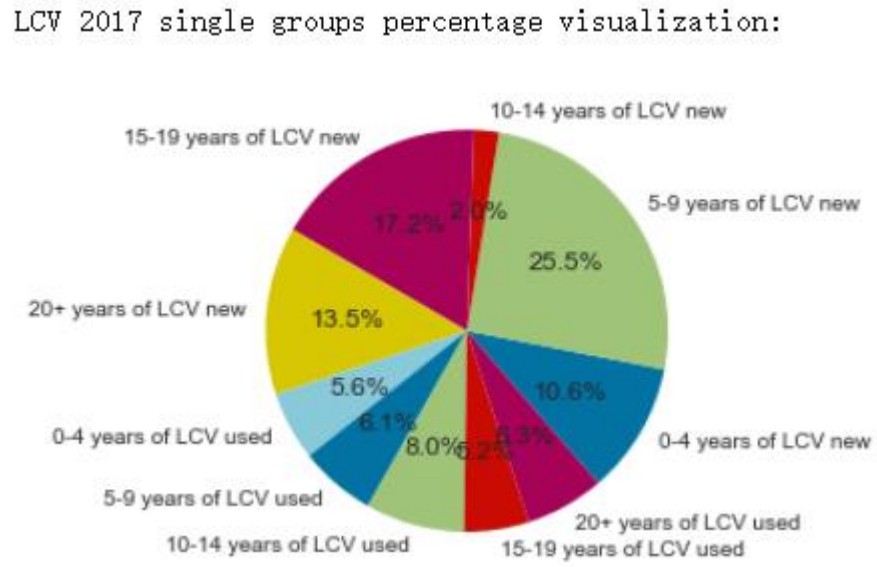


Figure 31: LCV model for 2017 single groups percentage.

8.3 Interpret the results, models and patterns

8.3.1 Interpretation of regression model, indicate Pattern One:

Visualization interpretation: The regression figures 79 and 80, the solid diagonal lines are for the fitting regression lines. it shows LCV model have better result than LPV model because LCV scatters are closer the line and LPV scatters seem a small away from line.

The residual figures 80 and 81 could find test data distributed on both side of distribution graph, fit for the normal distribution performance. It shows that there are no over fitting and under fitting problem because LPV have r2 score 0.9845 test sets and 0.9987 for LCV All the accuracy is exceeded 80%. For the Decision maker of New Zealand, they are able to control the number of fleets registrate or work off from road then they can base on these data to estimate trend of CO2 emission.

8.3.2 Interpretation of single groups predictions, indicate Pattern Two to Four:

Visualization interpretation: It has shown that the percentage of ages could be divided from whole groups, from Figure 83, and 84. There percentage are giving the common sense for the real business decision making, we will assess these percentage in real business aspect in step 8.4 for decision making for the project business objectives.

- The business objective is for find the different age group of fleets take the occupation of the CO2 emission, and notice that the assumption of this project in Step 1.2 mentioned if the fleets with age greater or equal to 10 take 70% of CO2 Emission then the NZ Transport needs to consider apply extra Greenhouse Exhaust Test. The fleets with over 10 years are 5 - 20+ groups in used import vehicles and 10 – 20+ groups in new import vehicles.
- Counting the occupation:

Compute the weight average of occupation among LPV and LCV:

$$\left(\frac{\#\ LPV}{\#\text{LPV}+\#\text{LCV}} * 71.602\% + \frac{\#\ LCV}{\#\text{LPV}+\#\text{LCV}} * 68.528\%\right) * 100\% = \left(\frac{3207721}{3207721 + 582540} * 71.602\% + \frac{582540}{3207721 + 582540} * 68.528\%\right) * 100\% = 71.130\%$$

In conclusion, rely on the project assumption and the weight average of the occupation of CO2 emissions reflect that it is **necessary** to suggest the NZ Transport apply extra Greenhouse Exhaust Test.

8.4 Assess and evaluate results, models and patterns

8.4.1 Assessment of the Prediction:

After above interpretation, all of our prediction is basing on vehicle's **age**, and **number of vehicles**. Compare with same size of groups, Higher age groups will contribute more CO2 emission than Lower age groups. People always have common sense of this situation. We compare with 0-4 average groups with 5-9 average groups where 0-4 groups contribute significantly smaller than 5-9 groups. This could prove **Harder Emission Standard** reduce the CO2 emission in other side. This model able to help New Zealand Transport to estimate future emission combine with Emission Standard to make decision about New Zealand Fleets control.

8.4.2 Assessment of the Percentage:

Follow the results percentage, we can find in the pie char graph that used vehicle used take near over 40% CO2 Emission Proportion in Commercial vehicles, However the used vehicles are just nearly 30% of the new vehicles. Connect to the real-world situation, government are more encouraging reducing CO2 emission and the taxies you can see on the road are always hybrid engine system which give some evidence that hybrid or electrical vehicles will significantly improve the CO2 emission. New Zealand Transport could use this to enlarge investment on replacing old commercial vehicles to more eco selections.

8.5 Multiple Iteration.

8.5.1 Test the Pipeline Robust:

Test the robust through change the random seed to 123, while the random seed has changed then the training and test data set will be different.

```
LPV: r2 score = 0.991516,
LPV residuals: [ 0.78578614 -0.00145467 -0.22034035  0.01036245 -0.05993759 -0.04397604
  0.00786783]

LCV: r2 score = 0.909101,
LCV residuals: [ 0.03667286 -0.05810854  0.10088362 -0.00381486  0.12035398  0.00406225
  0.02545228]
```

Figure 32: Under random seed 123.

The R2 score increase a little for LPV from 0.984495 to 0.991516, and decrease a bit from 0.910515 to 0.909101. The pipeline has good predictions with lower residuals generated from another iteration that the pipeline has a good robustly and stability for Data Mining.

8.5.2 Get a better result

Replacing the Ridge Regressor with **saga** solver which have similar r2 result under step 6.

```
LPV: r2 score = 0.983666,
LPV residuals: [ 0.02959306 -0.09478853 -0.15837284 -0.15246599  0.01502861  0.12016893
  0.04292094]

LCV: r2 score = 0.904099,
LCV residuals: [ 0.10286656 -0.05174695  0.03240236  0.09466663 -0.05755595 -0.01399769
  0.02830274]
```

Figure 33: Under saga solver.

Using saga solver will still satisfy the data mining objective. The R2 score decrease a little, from 0.984495, 0.910515 under sag to 0.983666, 0.904099. For the LPV test set, saga solver has similar results with sag solver and for LCV data set it performance poor than sag solver

Finally, we could confirm this iteration results above is the optimal and the pipeline is robust and stable for various data.

9 Action

9.1 Apply the knowledge and deploy the implementation

Every policy publish will be like a butterfly effect for sociality. Hence, every decision making need to be thinking into consideration. Although the study conclude that it is necessary to suggest NZ Transport check GHG emission on WOF. Not all the private garages have the instrument for test the GHG emission, thus this study recommend NZ Transport on reasonable samples to estimate the whole population. From previous paper, China using this kind of measurement on partial volume vehicles, [17] (Y.Dong, et al. 2019). To make the result more accuracy, NZ Transport could use clustering sampling follow the normal distribution, each sample size able to count from 95% Confidence Intervals Estimates, example on the following equations.

$$\overline{X} \pm Z\frac{\sigma}{\sqrt{n}} \rightarrow n = p(1-p)\left(\frac{Z}{E}\right)^2 = 0.5 * 0.5 * \left(\frac{1.96}{0.05}\right)^2 = 384.2$$

Assume under 95% confidence interval estimate, then the size of samples could be 385. If NZ Transport select 385 vehicles in every area of NZ then might able to make the decision to publish the GHG emission test policy.

9.2 Monitor the implementation

NZ Transport collect the data from samples then to estimate the population, after data gathering, they still need the models establish by this study to predict the new emission. Moreover, From BDAS experience, the AWS EC2 and Jupyter are great tools use for monitor the implementation. Combining with the NZ Transport data statistics and monitoring tools for the implementation is necessary. For avoid human's bias, build the models on the AWS EC2 instance and provided interface for users to input new data, treat the models as a black box and run AB test for different samples would progress the decision making. Provide Jupyter interface for the experienced studies to get specific results they want.

9.3 Maintain the implementation

As mentioned in step 1 and experience from BDAS iterations, using Github to make the project opensource then allow other studies to join the project to provide their opinions.

The Github link:https://github.com/tbai915/New-Zealand-Transport-CO2-Emission-Predict-Study

To let the models, keep training, it is necessary to keep NZ transport dataset growing, the public version only have 18 years data and there might have more data at inner of NZ transport, but their might storage by paper. Hence, one maintaining work is make these data to become digital data, which need specialist to working on. Furthermore, other maintaining works for models, if there are new studies publish to have better modelling strategies for the population similar like NZ then specialist could tune this study's models base on the result.

9.4 Enhance the solution in future

On one way, the models base on vehicle distribution data provided by NZ Transport. The study using normal distribution to count the distribution of fleets, hence, if NZ Transport provide more reliable data then the models are able to be more accuracy.

On another way, from previous studies from other countries, gross domestic product and energy consumption are also good predictors for GHG prediction. Since there are not studies for GHG prediction in New Zealand, the following enhancement of the solution are able to start from these aspects, have trials on ANN models using these predictions. Controlled by the data gathering, this study cannot get GDP or energy consumption data related to the transport from the total data. Strongly appeal NZ government could build team for the GHG prediction and from this study experiences then combing with inner data to get a better result.

Reference:

[1] World Meteorological Organization. (15 January 2020). “WMO confirms 2019 as second hottest year on record”. Retrieved from https://public.wmo.int/en/media/press-release/wmo-confirms-2019-second-hottest-year-record

[2] Z. Liu, F. Wang, Z. Tang, and J. Tang. (2020). Predictions and driving factors of production-based co2 emissions in beijing, china. Sustainable Cities and Society, 53:101909.

[3] The United Nations. (2020). “The 17 Goals”. Retrieved from https://sdgs.un.org/goals

[4] Stats NZ. (15 October 2020). “New Zealand’s greenhouse gas emission” Retrieved from https://www.stats.govt.nz/indicators/new-zealands-greenhouse-gas-emissions

[5] L. Alfaseeh, R. Tu, B. Farooq, M. Hatzopoulou. (2020). Greenhouse Gas Emission Prediction on Road Network using Deep Sequence Learning. arXiv:2014.08286.

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37.

[7] Ministry of Transport. (05 August 2020).” Vehicle Fleet Statistics”. Retrieved from https://www.transport.govt.nz/mot-resources/vehicle-fleet-statistics/

[8] U.EPA. Source of greenhouse gas emission. Retrieved December, 2017.

[9] A. Ö. Dengiz, K. D. Atalay, and O. Dengiz. Grey forecasting model for co 2 emissions of developed countries. In The International Symposium for Production Research, pages 604–611. Springer, 2018.

[10] C.-S. Lin, F.-M. Liou, and C.-P. Huang. Grey forecasting model for co2 emissions: A taiwan study. Applied Energy, 88(11):3816–3820, 2011.

[11] H.-T. Pao and C.-M. Tsai. Modeling and forecasting the co2 emissions, energy consumption, and economic growth in brazil. Energy, 36(5):2450–2458, 2011.

[12] A. Rahman and M. M. Hasan. Modeling and forecasting of carbon dioxide emissions in bangladesh using autoregressive integrated moving average (arima) models. Open Journal of Statistics, 7(4):560–566, 2017

[13] C. Tudor. Predicting the evolution of co2 emissions in bahrain with automated forecasting methods. Sustainability, 8(9):923, 2016.

[14] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla. Linear and nonlinear modeling approaches for urban air quality prediction. Science of the Total Environment, 426:244–255, 2012.

[15] D. Radojevi´c, V. Pocajt, I. Popovi´c, A. Peri´c-Gruji´c, and M. Risti´c. Forecasting of greenhouse gas emissions in serbia using artificial neural networks. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 35(8):733–740, 2013.

[16] M. Grote, I. Williams, J. Preston, and S. Kemp. A practical model for predicting road traffic carbon dioxide emissions using inductive loop detector data. Transportation Research Part D: Transport and Environment, 63:809–825, 2018.

[17] Y. Dong, J. Xu, X. Liu, C. Gao, H. Ru, and Z. Duan. Carbon emissions and expressway traffic flow patterns in China. Sustainability, 11(10):2824, 2019.