# Sept 2023 FT - MSc in Data Analytics

Author: Ahmet Cihat Ceyhan

e-mail:[2023281@student.cct.ie](mailto:2023281@student.cct.ie)

Student ID: 2023281

## Abstract

*This report presents a comprehensive analysis of vehicle registrations in Ireland and Norway. The study, which includes data set research, data preprocessing, descriptive and inferential statistics, as well as machine learning models, provides a comparative evaluation of vehicle sales in the two countries. While examining the effects of global events like economic crises on the market, the study also analyses predictions of vehicle sales and public sentiment regarding vehicle tax policies.*

## Introduction

This study aims to understand trends and market dynamics in vehicle sales by analysing Irish vehicle registrations and comparing them with another country (Norway). Using both descriptive and inferential statistical methods, the research reveals similarities and differences between the two countries' vehicle purchases. It also examines the relationship between Irish vehicle registrations and population using inferential statistics. Machine learning models are used to predict future vehicle sales and analyse public perceptions of vehicle tax policies.

## Data Preprocessing

### Researching Dataset

I started to research dataset for my assignment with the idea of conducting a study on public transportation and smart card systems, but the lack of suitable datasets for objective analysis led me to change focus. Licensing issues highlighted the importance of using reliable, licensed datasets. I also considered another country's data, but its high population made unbiased analysis difficult. Shifting to vehicle registration, I found Norway's similar population and dataset licensing ideal. Both Irish and Norwegian datasets, licensed under CC BY 4.0, allowed reuse with proper attribution and complied with GDPR, ensuring usability and legal compliance.

### Loading Dataset

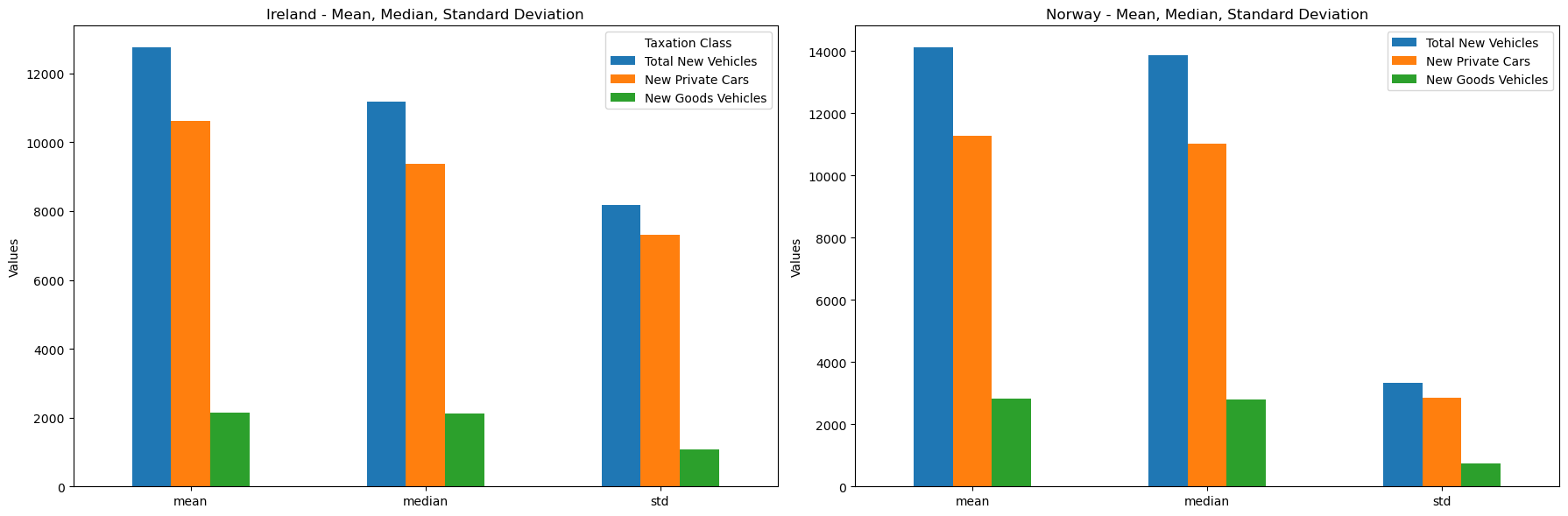
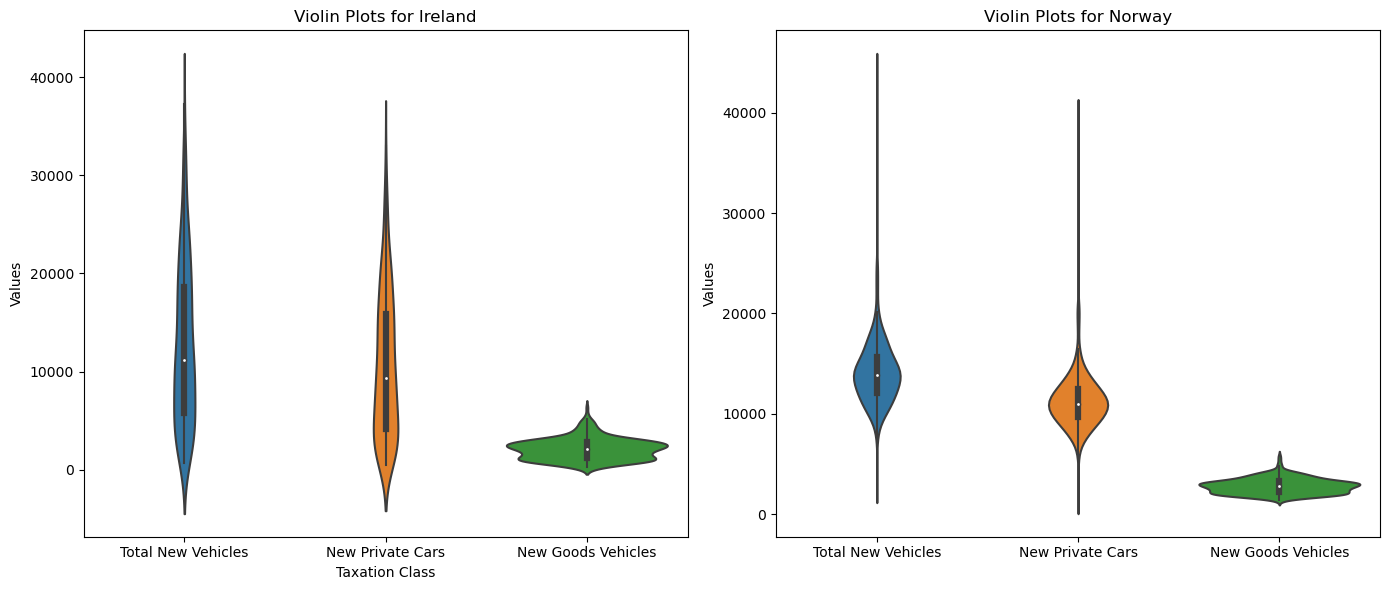
When I uploaded my Norway dataset as a CSV file, I noticed the data were not separated. Later, I realized that the data were unusually separated by semicolons. To resolve this, I added a special parameter to my code. This parameter enables the program to correctly separate the data.

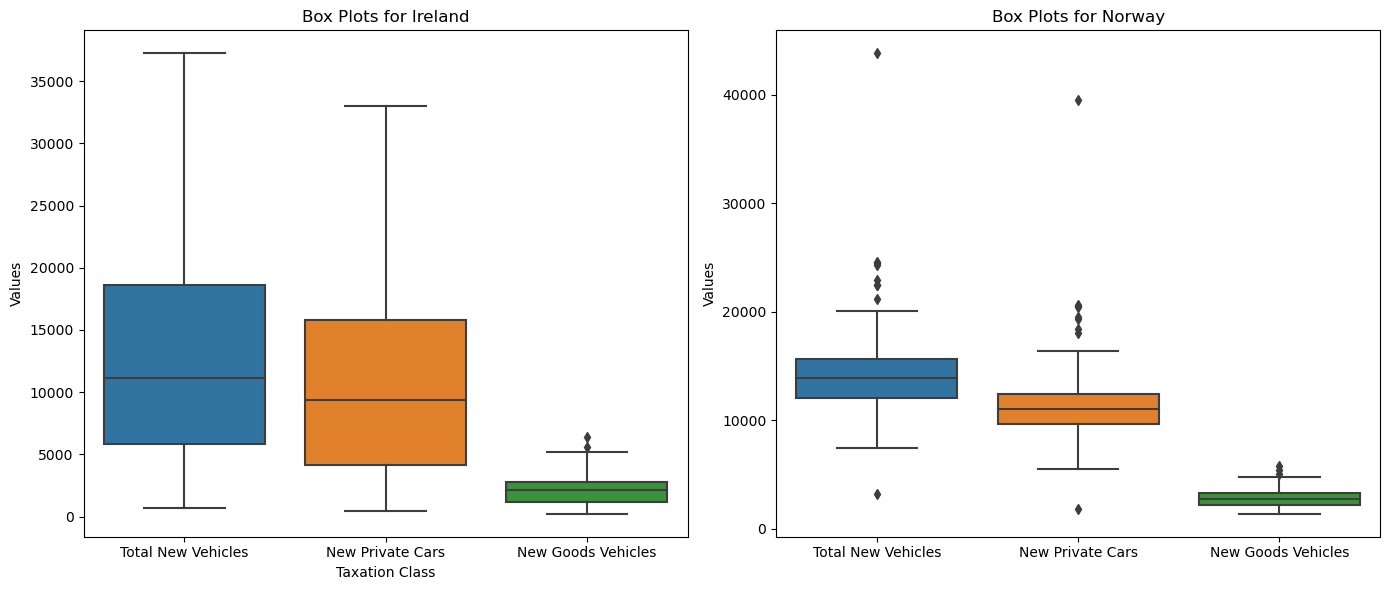
I loaded the Ireland dataset's JSON file into Python as a dictionary, focusing on the dataset key. This key included 'dimension' for data classifications and 'value' for numerical data. I used itertools product to generate dimension combinations, mapping each to its numerical value. This data was then structured into DataFrame, where each row represented a dimension combination with its corresponding numerical value, making the data interpretable for analysis.

### Data Cleaning and Preprocessing

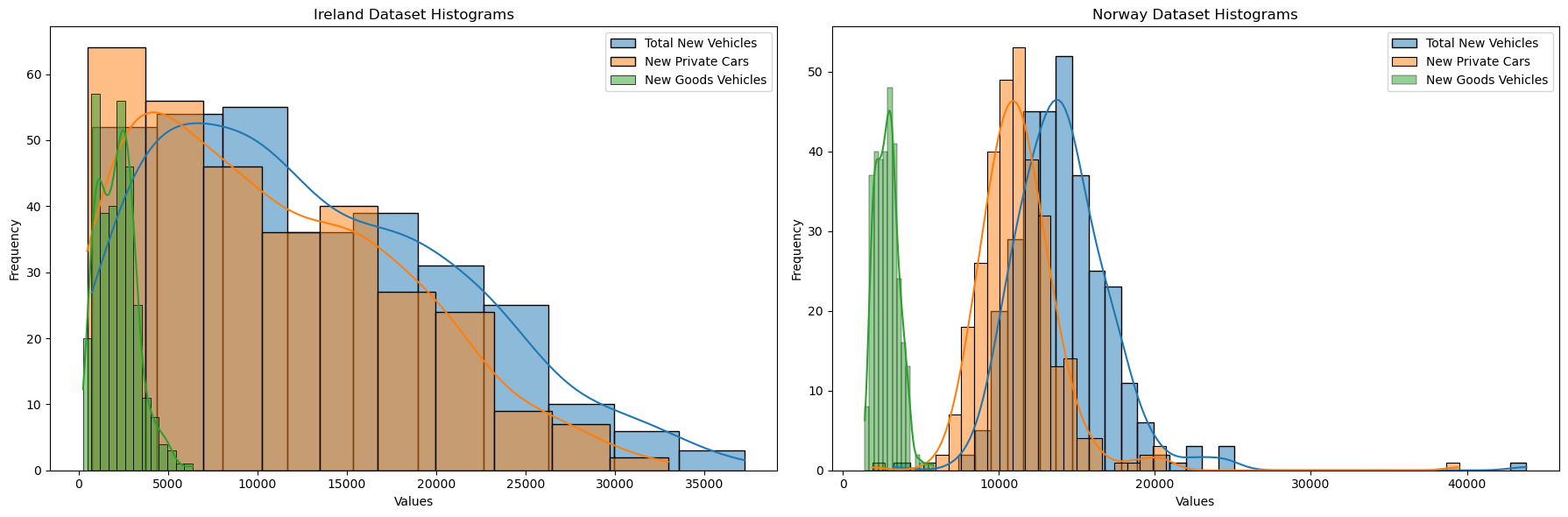
In transforming the Irish vehicle registration dataset, I focused solely on new vehicle registrations, omitting second-hand data to align with my research objectives. I used official Irish and Norwegian data sources to identify and match common vehicle types for comparative analysis, discarding irrelevant types. The datasets were then standardized in format and type for direct comparison, ensuring no missing data. This method provided a focused, clear foundation for comparing new vehicle registrations between the two countries.

## Descriptive Statistics and Comparative Analysis

Ireland's new private vehicle sales show a lower average compared to Norway, but the standard deviation, reflecting the fluctuation in sales, is quite high. This suggests that the Irish market might experience higher or lower sales than expected at times. In new commercial vehicles, Ireland also displays lower average sales than Norway, with greater variability in the number of sales. In total new vehicle registrations, Ireland's average sales lag behind Norway's, again showing higher variability.

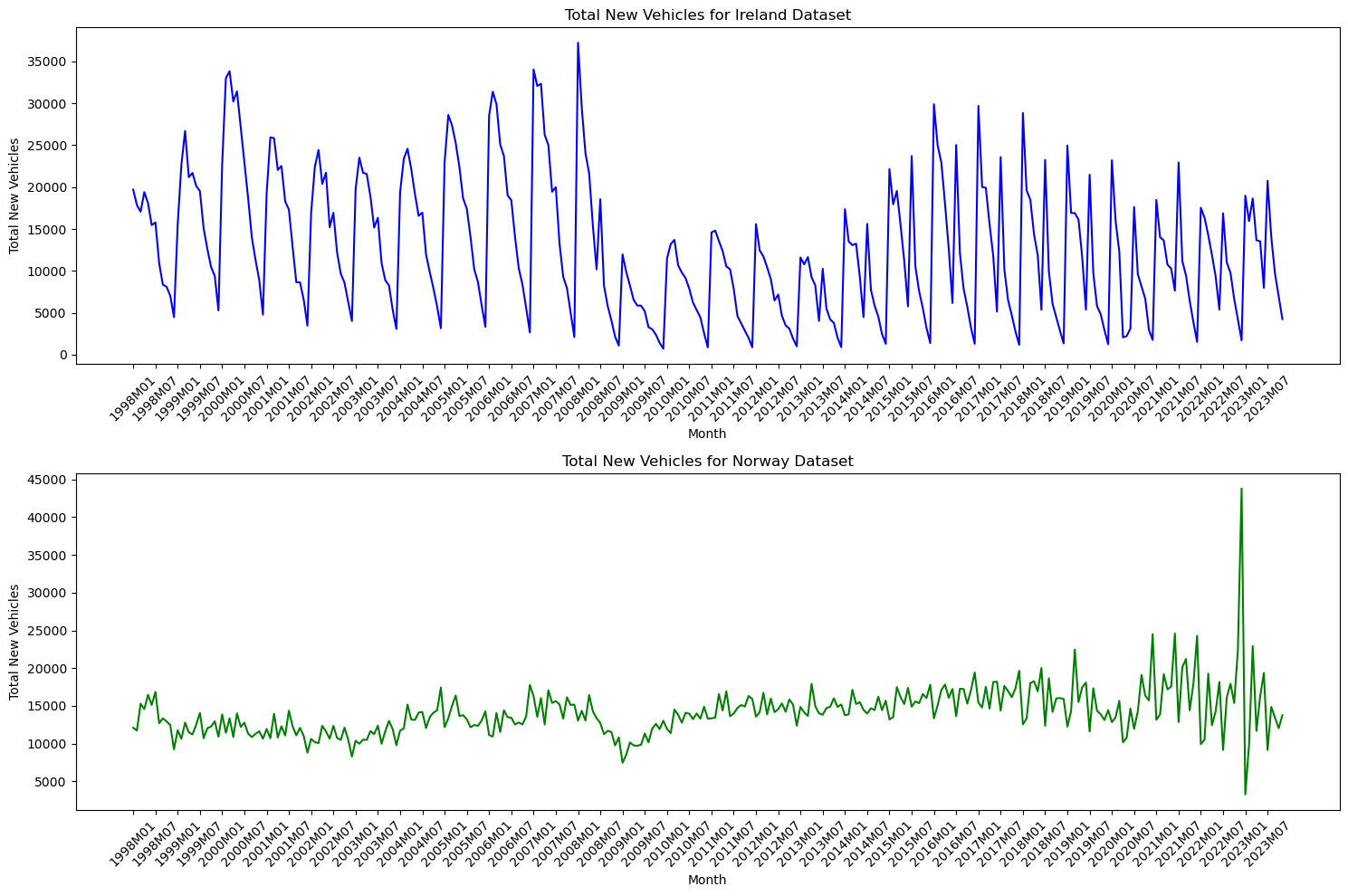
Norway's vehicle market outperforms Ireland's, with higher mean values across vehicle categories, suggesting a larger and more consistent market. The violin plots reveal that Ireland has a wide range of private car sales, with a notable concentration at lower values, hinting at a market with significant peaks. In contrast, commercial vehicle sales show a tighter and more uniform spread. Norway's plots present a balanced distribution, indicating steady sales within certain ranges.

Box plots reflect Ireland's tendency towards higher median sales in private cars with some extreme values, whereas commercial sales are more contained. Norway's data shows a broader spread in private car sales and a consistent, narrower distribution for commercial vehicles, with fewer extremes.

With these graphs, I wanted to highlight the differences in market behavior between the two countries by effectively capturing the dynamics and variability in vehicle sales.

Ireland's vehicle sales data show a high number of low-value sales skewing averages up, with new private vehicles and total registrations right-skewed and commercial vehicles more balanced. Norway's distribution is symmetric for private and total vehicles, indicating a balanced market, but narrower for commercial vehicles, suggesting fewer sales. The data outline market structure differences: Ireland has a high sales potential in specific segments, while Norway's market displays overall consistency.

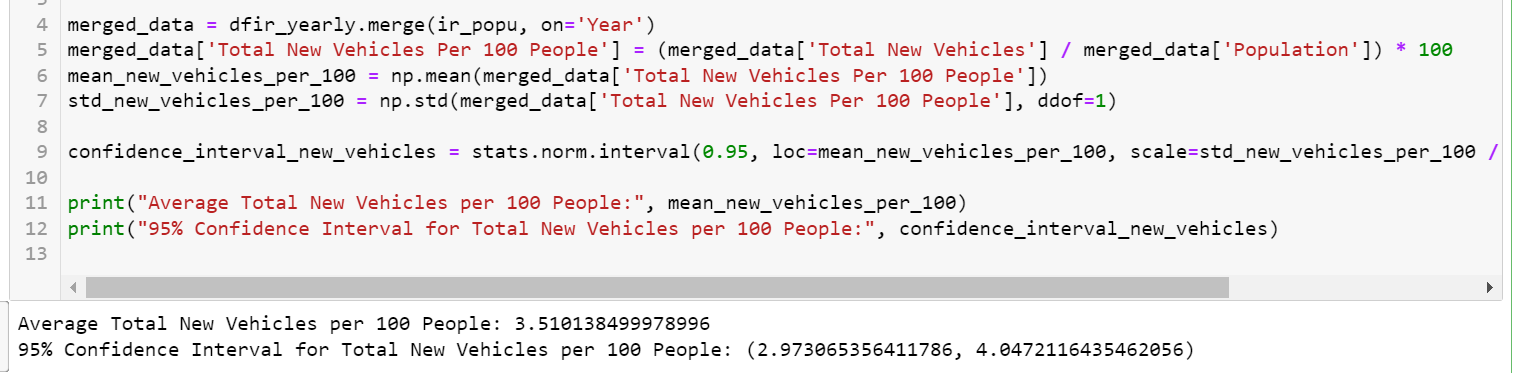
### Analysis of Monthly Data

New vehicle registrations in Ireland peak at the beginning of the year and in July, which may be related to sales campaigns coinciding with Christmas incentives and license plate change periods. Registrations decline at the end of the year and in June; consumers may be waiting for new models or new license plates. In Norway, vehicle registrations are more regular and do not show large fluctuations, suggesting a more stable market. 

Fluctuations in the Norwegian car market in 2022 may have been caused by policy changes, such as stricter emission regulations and tax increases. The removal of the VAT exemption for electric vehicles and a new weight tax affected prices, especially for non-electric vehicles. This led to increased vehicle registrations before the policy changes took effect and a decline afterwards. While these policies may cause short-term market changes, they are expected to support electric vehicles in the long run.

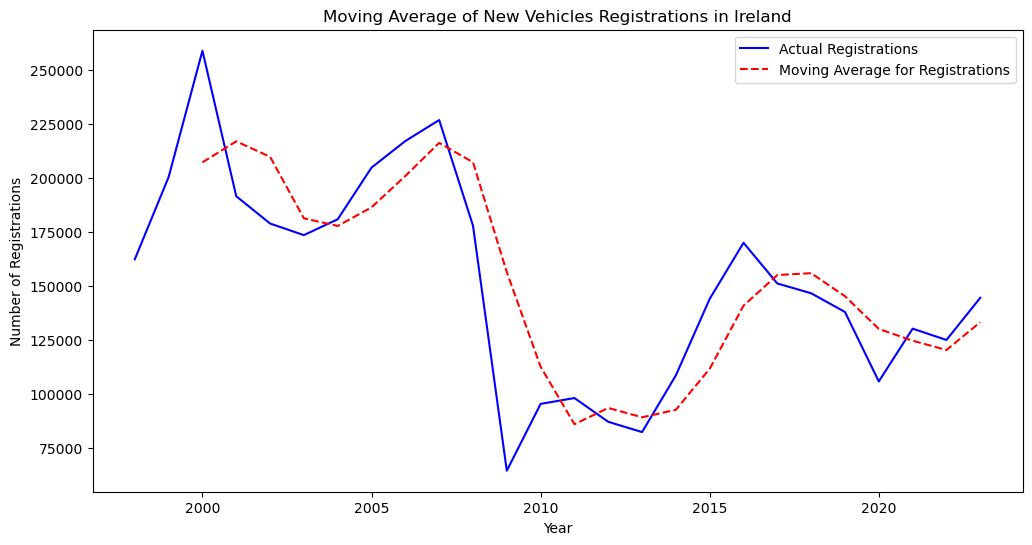
## Inferential Statistics About Vehicle Registrations and Population in Ireland

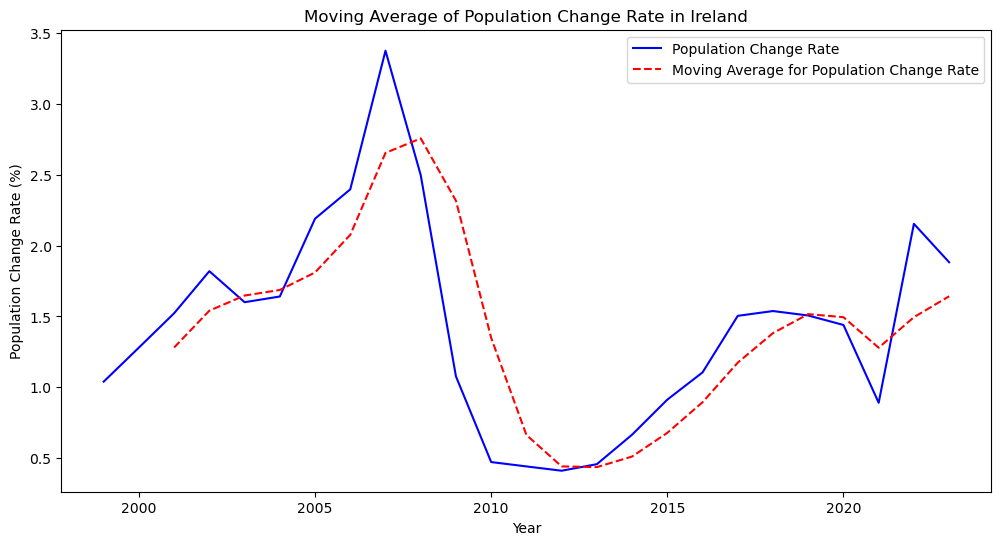
Inferential statistics provide tools that allow us to make predictions or inferences about a population based on a sample of data. In the context of understanding the dynamics of vehicle registrations in relation to population growth, inferential statistics allow us to estimate the behaviour of the entire population and make predictions by examining a subset of data. By analysing annual vehicle registration data and combining it with the Irish population, we can discover trends and draw conclusions about the relationship between the number of new vehicle registrations and population growth over time.



The 95% level is that it contains those values of α that are rejected by a significance test at the 5% level. (G. Berry, 1995) I merged Ireland's vehicle registration data with population figures to determine new registrations per hundred people. Calculating the average and standard deviation led to a 95% confidence interval, estimating the average rate between 2.97 and 4.05 per hundred people.

### Moving Average of Registrations and Population Change

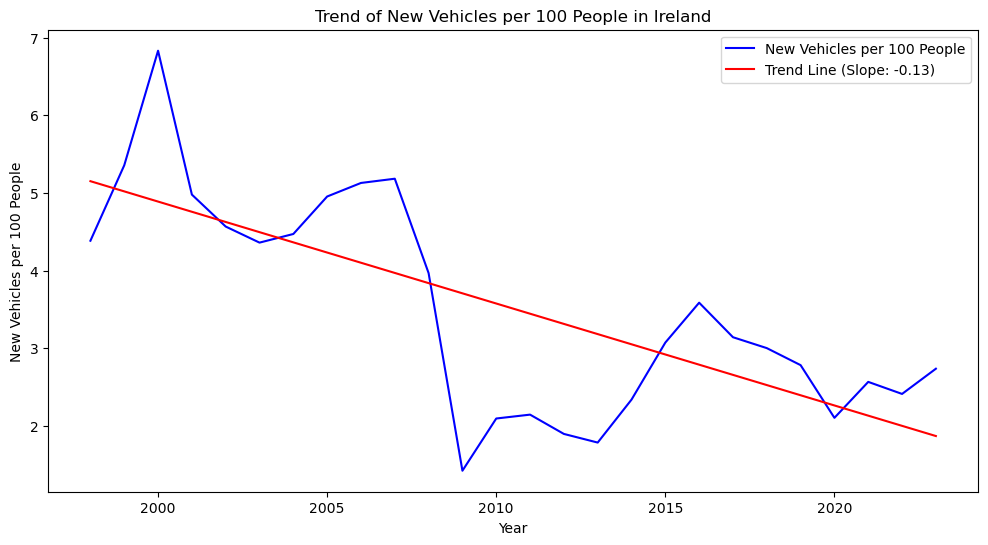




I drew this graph to observe the relationship between the annual new vehicle registrations in Ireland and the rate of population change. The first graph shows the number of new vehicle registrations over the years along with their three-year moving average, while the second graph visualizes the moving average with the rate of population change. The purpose of using a moving average is to smooth out short-term fluctuations and highlight longer-term trends.

Examining the graphs, a significant decline around 2008 is noticeable in both indicators. This drop could be a reflection of the global economic crisis experienced during that period. Specifically, the decrease in consumers' purchasing power and confidence during economic crises could be connected to the observed decline in new vehicle registrations.

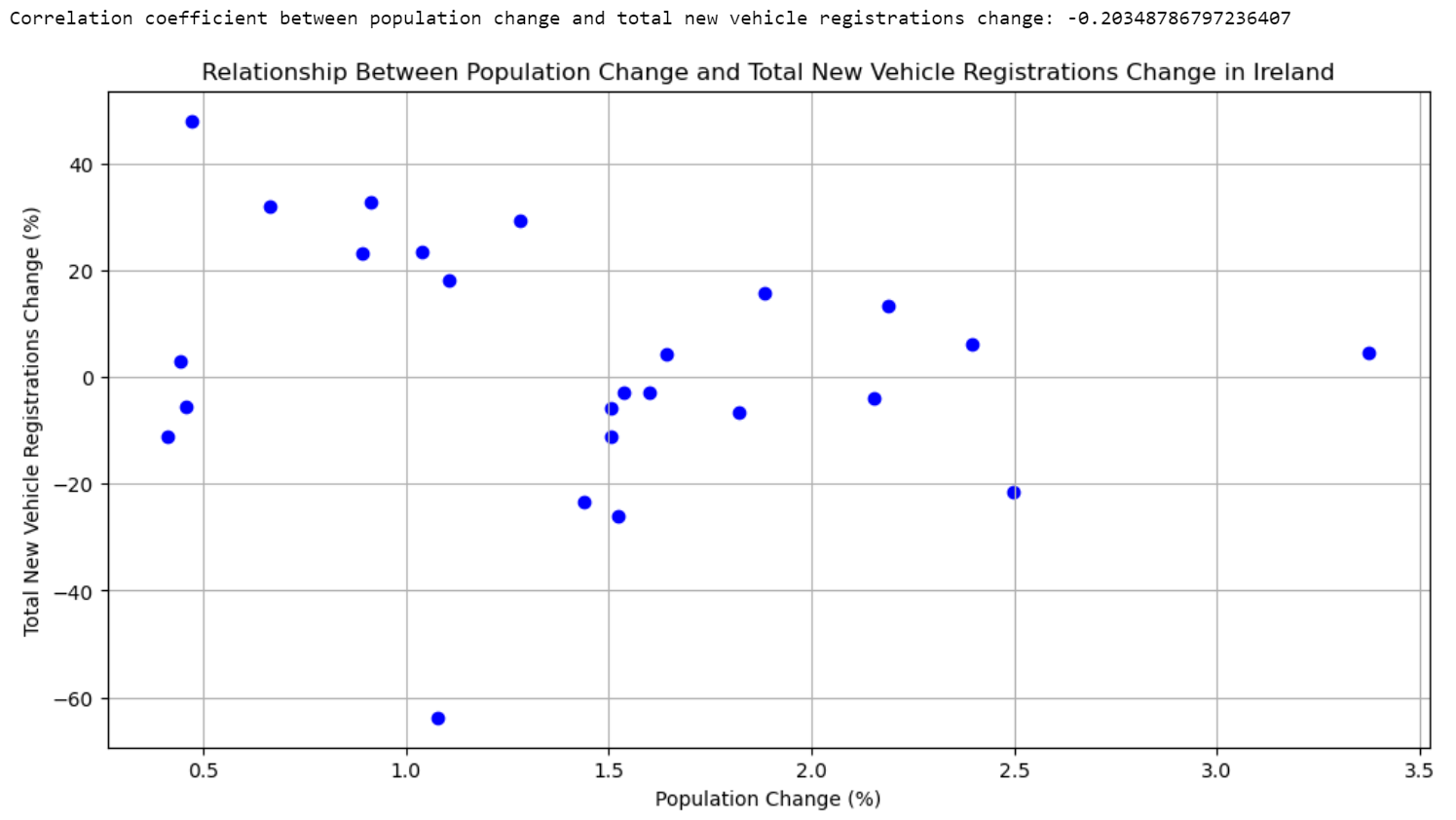
### Linear Regression Analysis



I used linear regression to analyze the trend of new vehicle registrations per person in Ireland. The slope (-0.131) indicates a decrease in registrations per hundred people, with a strong negative correlation (-0.718) showing a significant decrease relative to population growth. The p-value (3.543e-05) and standard error (0.02593) confirm the statistical significance and reliability of this trend. Declines around 2008 and 2020 likely reflect the global financial crisis and COVID-19 pandemic impacts, respectively.

### Correlation Coefficient Analysis

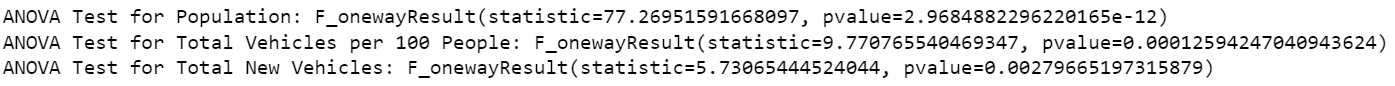
In my graph analysis, I'm looking at how changes in Ireland's population relate to the number of new cars registered each year. I used a method called correlation analysis to see if there's a link between these two things. The result I got, -0.203, shows that there's a small and negative connection between them. This means there isn't a strong link between how fast the population is growing and the changes in new car registrations.



My analysis points out that big events like economic crises and the COVID-19 pandemic affect both population growth and new car registrations. However, just because the population is growing doesn't mean more cars are being sold. Things like pandemics and economic crises might have a bigger impact on car sales than the number of people increasing. This tells us that the link between population growth and new car sales isn't straightforward and is more complicated.

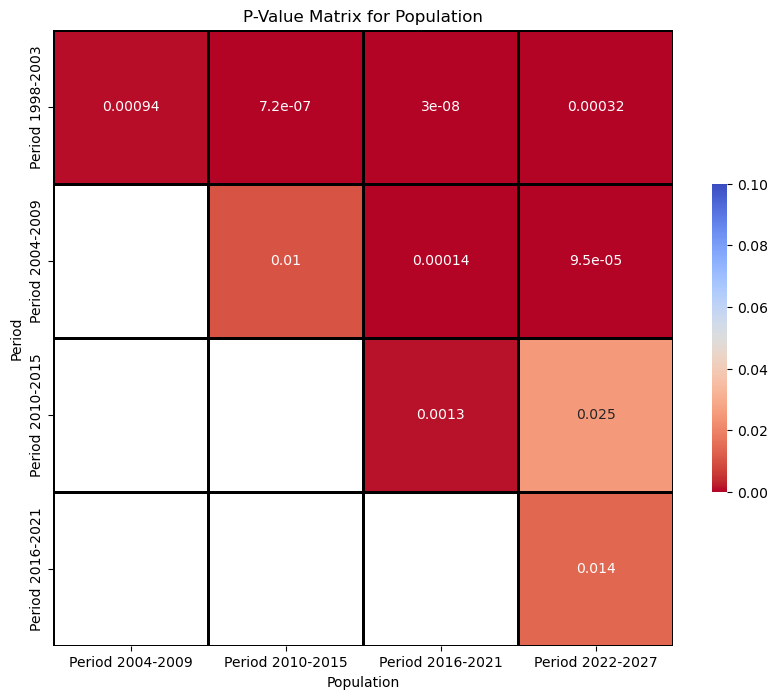
### Six-Year Period ANOVA Analysis

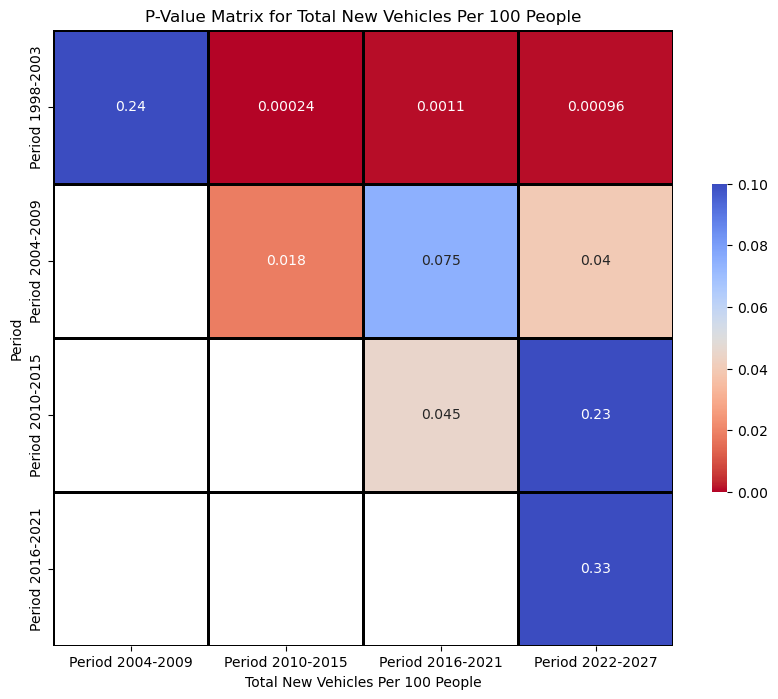
The ANOVA test is similar to the t-test in that it compares the differences between group means. (Pompeii, 1998) I did this test to compare the variations within groups against the variations between groups. For the Irish population, the number of new vehicles per hundred people, and the total number of new vehicles, I performed the ANOVA test for six-year periods. I chose 6 because I wanted to compare the years with the most fluctuations.

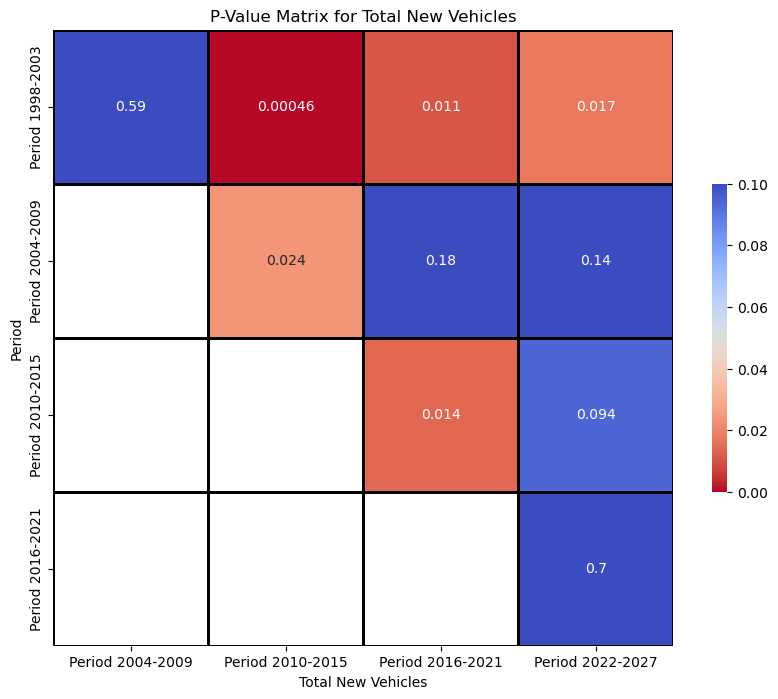
The p-values from the test results indicate that all three measures show statistically significant differences between different six-year periods. We usually consider there to be significant differences between groups when the p-values are less than 0.05. In this case, the p-value for the population (2.96e-13), for the new vehicles per hundred people (0.000125), and for the total new vehicles (0.002796) are all quite low. This shows that there are significant differences between the six-year periods.

### Six-Year Period T-Test Analysis

To explore differences highlighted by the ANOVA test, I conducted t-tests over six-year intervals to see if there is a difference between the mean weight for two groups (Pompeii, 1998). Where p-values were below 0.05, marked in red, it showed significant differences, while blue indicated higher p-values and no significant variance. The population change matrix did not show p-values over 0.05, suggesting notable differences in population size over time. However, the p-value of 0.59 between 1998-2003 and 2004-2009 in the total new vehicles data suggests no significant change, possibly reflecting steady economic and market conditions during those periods.

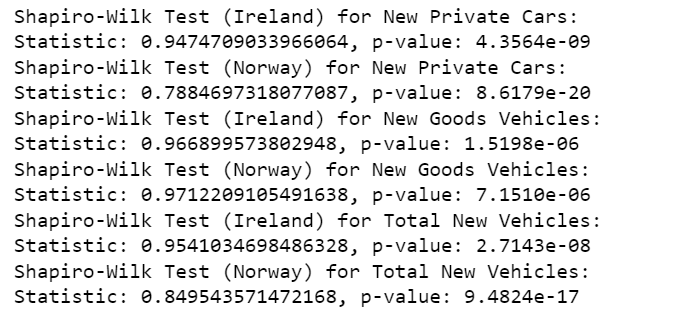






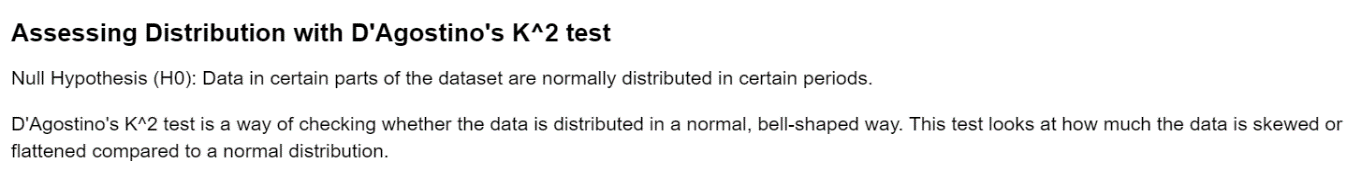
## Comparative Inferential Statistics Between 2 Countries

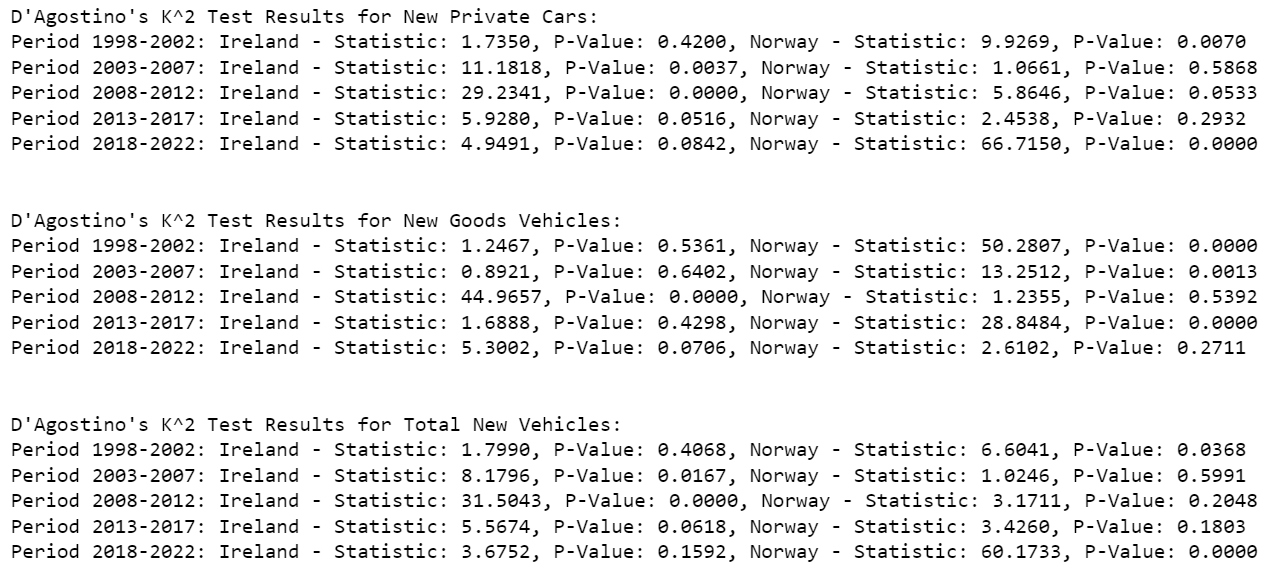
### Checking Datasets Distribution

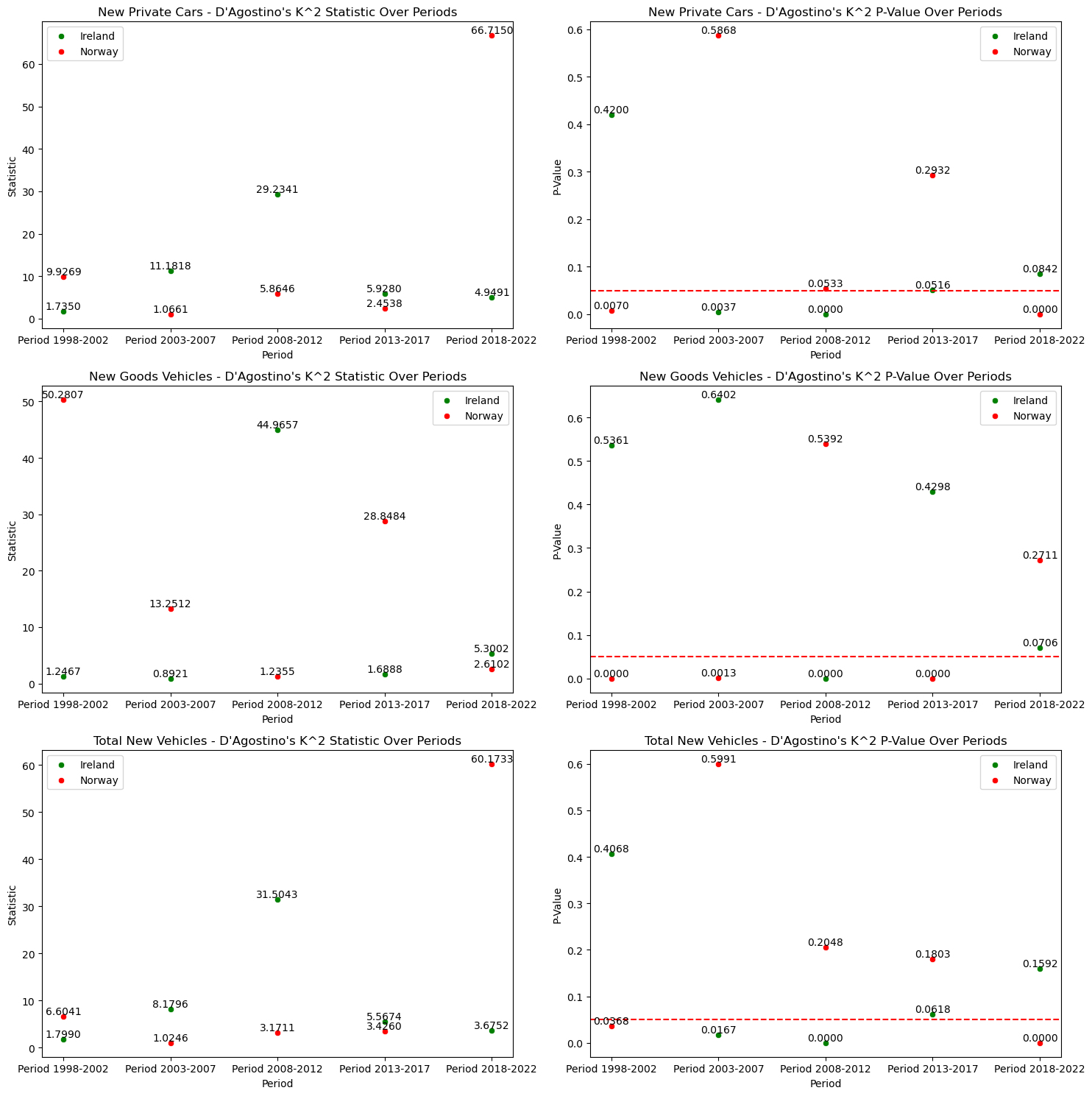


I used the Shapiro-Wilk test to check if the samples in my dataset follow a normal distribution. This statistical method is helpful to test the normality of data (Shapiro & Wilk, 1965). I conducted the Shapiro-Wilk tests for the new private vehicles, new commercial vehicles, and total new vehicle numbers in Ireland and Norway. The tests showed that in all categories, the p-values came out as 0.0000. This result indicates that the examined data do not follow a normal distribution, meaning the distributions are skewed or have extreme values. This could suggest that in both countries, vehicle registrations in a particular year may show abnormal variations due to specific factors or events, instead of the expected smooth distribution.

### Assessing Distribution with D'Agostino's K^2 test

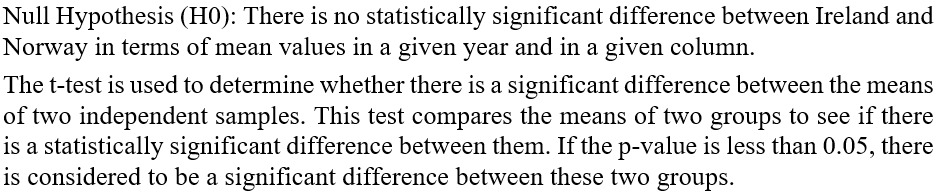


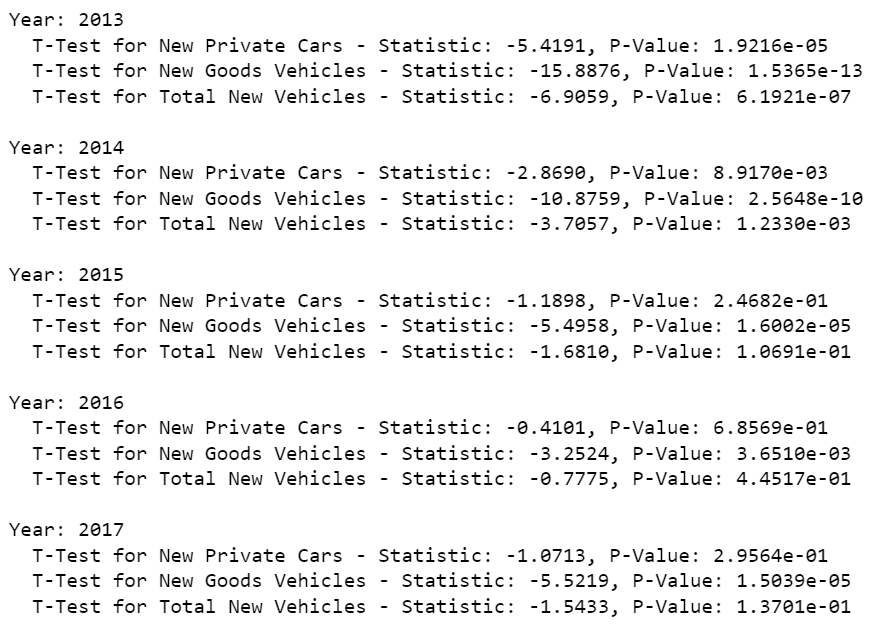




When I conducted D'Agostino's tests for 2013-2017, the p-values for the Irish and Norwegian vehicle registration data in the two columns were above 0.05 (RB., 1970). This implies that the data for this period may be normally distributed and generally close to what we would expect in terms of how they are skewed or flattened. However, this does not fully confirm that the data are normally distributed. Nevertheless, with these initial results I can assume that the data are normally distributed and accept hypothesis H0 and use more detailed tests. A 95% confidence interval is the estimated range of values within which it is 95% possible or likely that the precise or true population effect lies

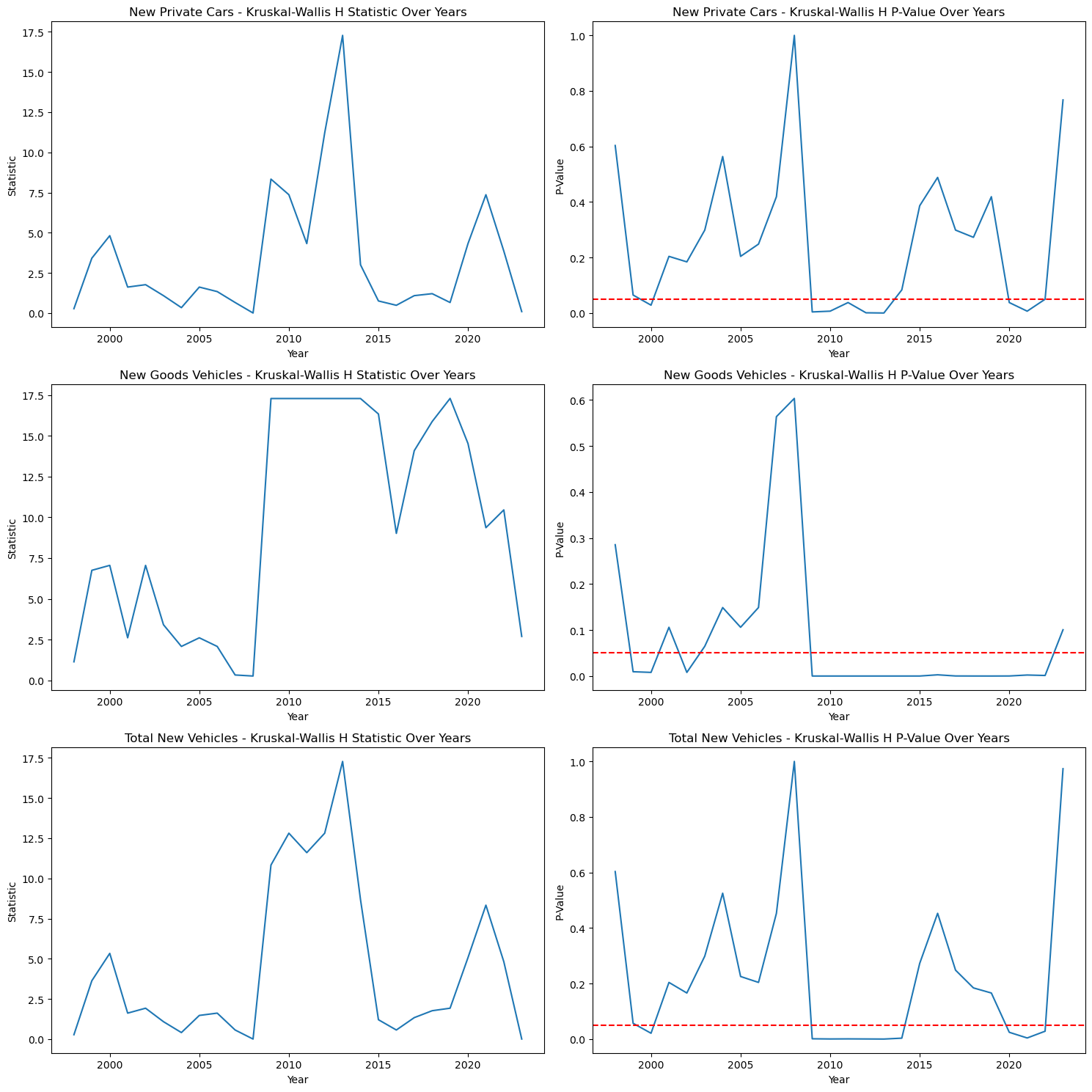
### T-Test on Normally Distributed Periods





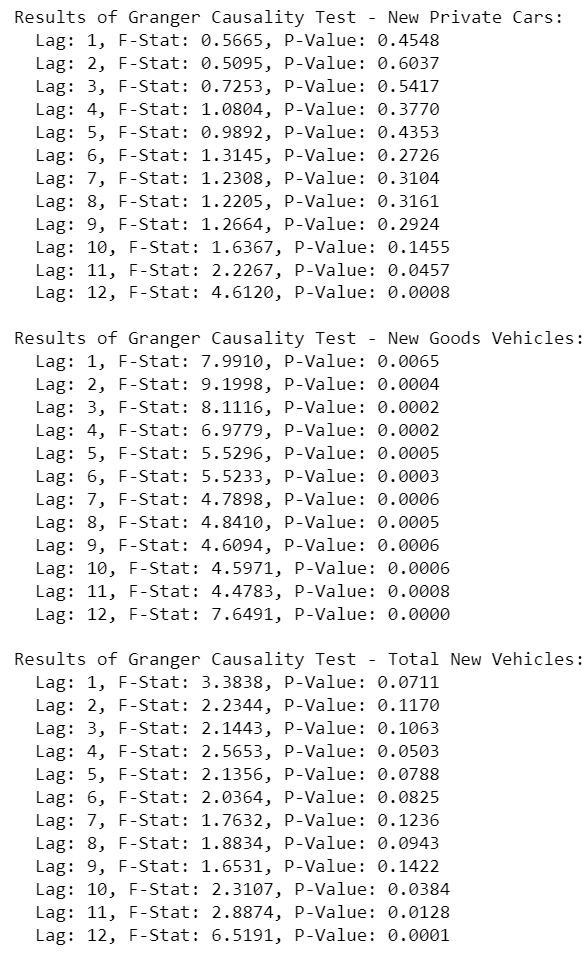
My t-tests show that the comparison of the all columns in Ireland and Norway between 2013 and 2017 reveals statistically significant differences in some years. In particular, the p-values for new private and commercial vehicles in 2013 were extremely low and the H0 hypothesis was rejected. This indicates that there are significant differences in the number of sales between the two countries for this year. Similar results were observed in other years, but in some years (e.g. 2015 and 2016) the p-values were higher and the H0 hypothesis was accepted, suggesting that the differences were not as pronounced.

### Compare The Distributions



In the graphs, I have visualized the annual Kruskal-Wallis H statistics and p-values for all columns and the fluctuations in these values. Years with p-values below 0.05 indicate a significant difference between Ireland and Norway for that year. For example, if the p-values for new private cars are low, this indicates that there are different distributions of new private car sales in Ireland and Norway in that year. In 2008, I observe that the p-values are higher in all columns. To better understand the reasons for these changes, I will apply causality test.

### Causality Test

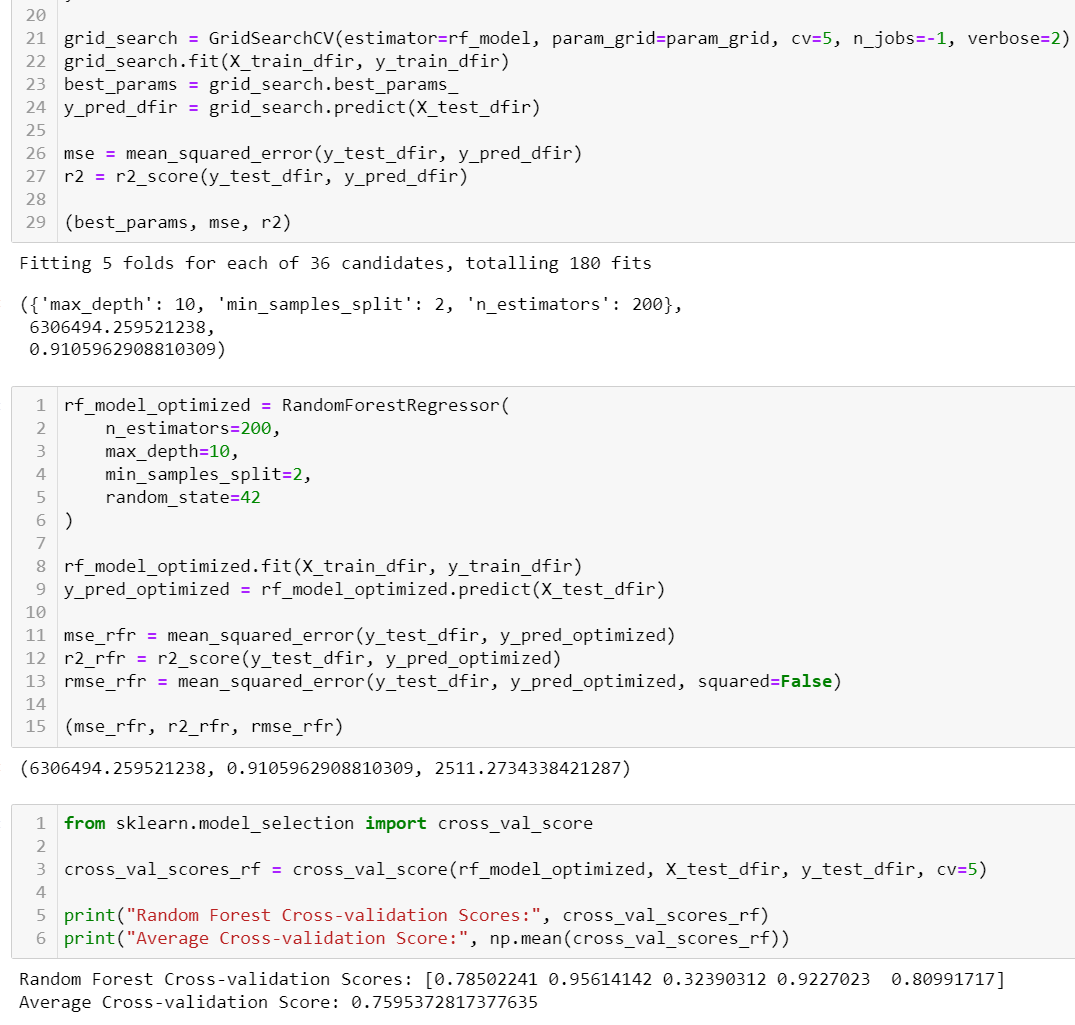


Null Hypothesis (H0): There is no significant causal difference between the number of vehicles sold annually in Ireland and Norway.

I conducted the Granger causality test to analyse the potential causal relationship between the two datasets. In the new private cars data, lag 12 is significant with a low p-value, while in the new goods vehicles series, significant results are observed at all lags. This indicates a potential causal relationship on vehicle sales. However, for total new vehicles, significant results were obtained at lags 10, 11 and 12. In the years around the 2008 economic crisis, this test can be a guide to a deeper understanding of the effects of the crisis on vehicle sales.

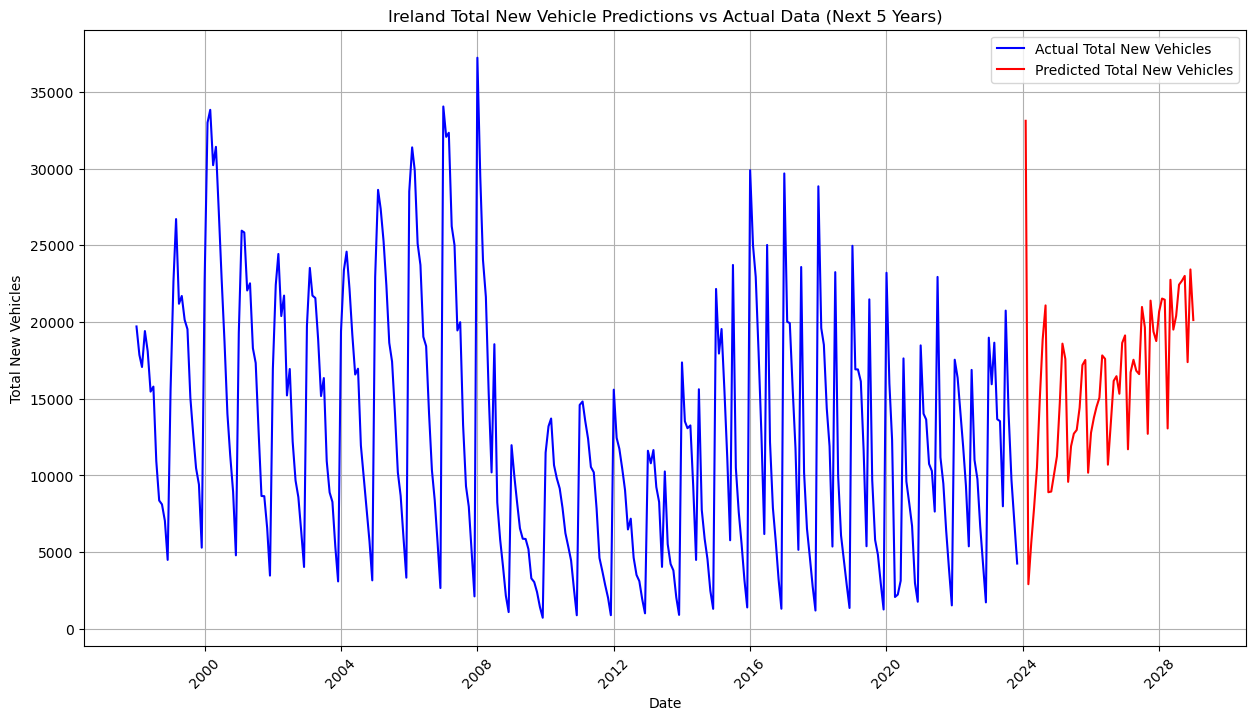
## Machine Learning

### Predict Future of Ireland Data with Random Forest Regression

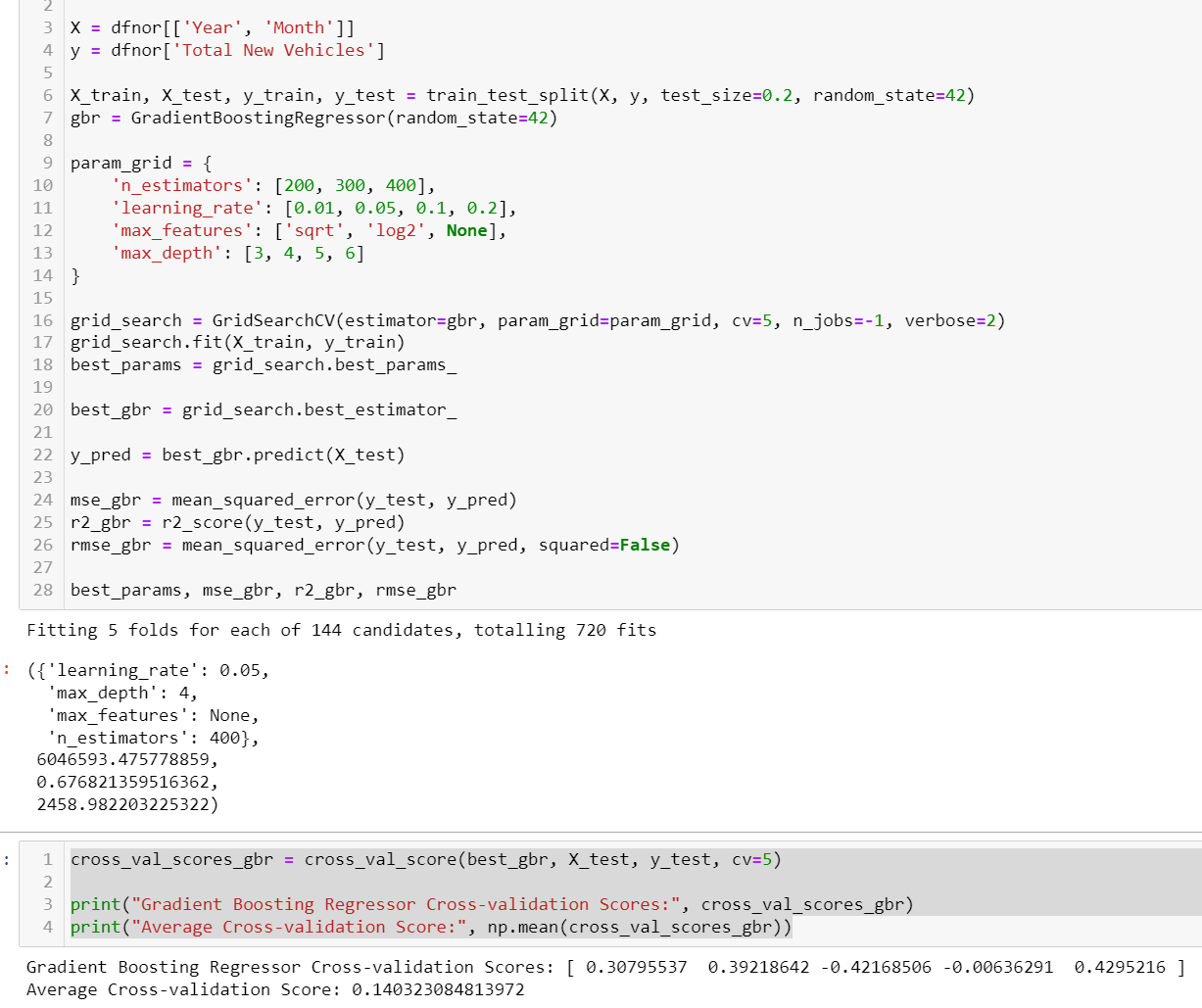


In this machine learning model I created, I start by dividing the data into years and months. Then, I add lagged features, lag features are values from previous time steps used as inputs for future predictions. In time series forecasting, they capture patterns like trends and seasonality, allowing the model to learn and identify recurring patterns over time for accurate future predictions.

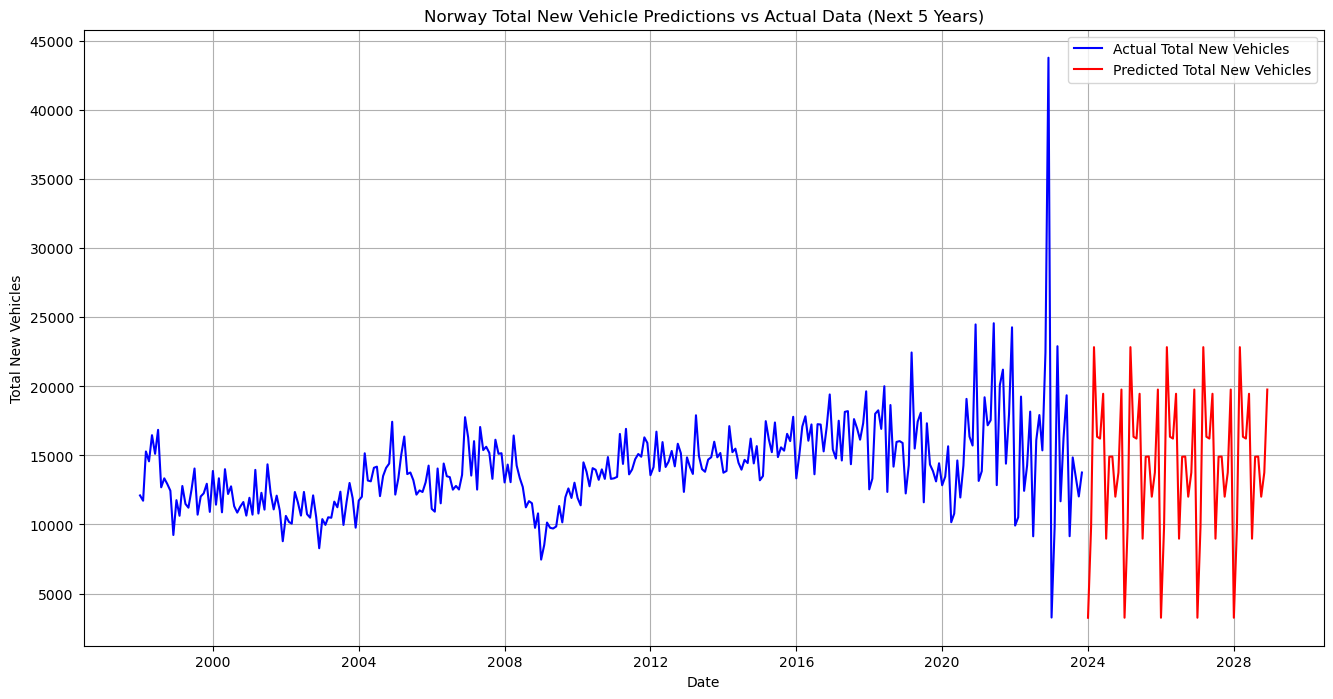
To make sure the model works well, I use a method called cross-validation. This involves splitting the data into several parts, using some parts to train the model, and testing it on the other parts. This step is really important to avoid overfitting, where the model works well only on the data it was trained on but not on new data. My model's average score from cross-validation is about 0.7595, which means it can predict new vehicle registrations with around 75.95% accuracy.



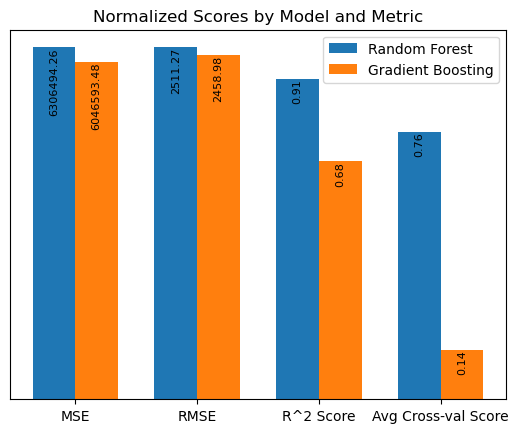
### Predict Future Data with Gradient Boosting Regressor



I used Gradient Boosting Regressor model, known for handling complex datasets, was used for time series analysis of vehicle registrations. I use the grid search technique in hyperparameter optimization to methodically find the most effective hyperparameter combination for my model, impacting its learning and prediction accuracy. However, its average cross-validation score of 14.03% suggests limited predictive power, potentially due to not capturing underlying patterns or the presence of noisy data. The around 2020 spike and decline, likely related to external events, may have disrupted the model's learning. I tried to improve my model by adding new parameters and by taking the sinus and cosinus of the monthly value for a better time series analysis. The method of changing the parameters didn't work well. Changing the monthly values improved the model by 20%. However, that time I had problems with graphing and dashboard.



### Comparing Two Machine Learning Models



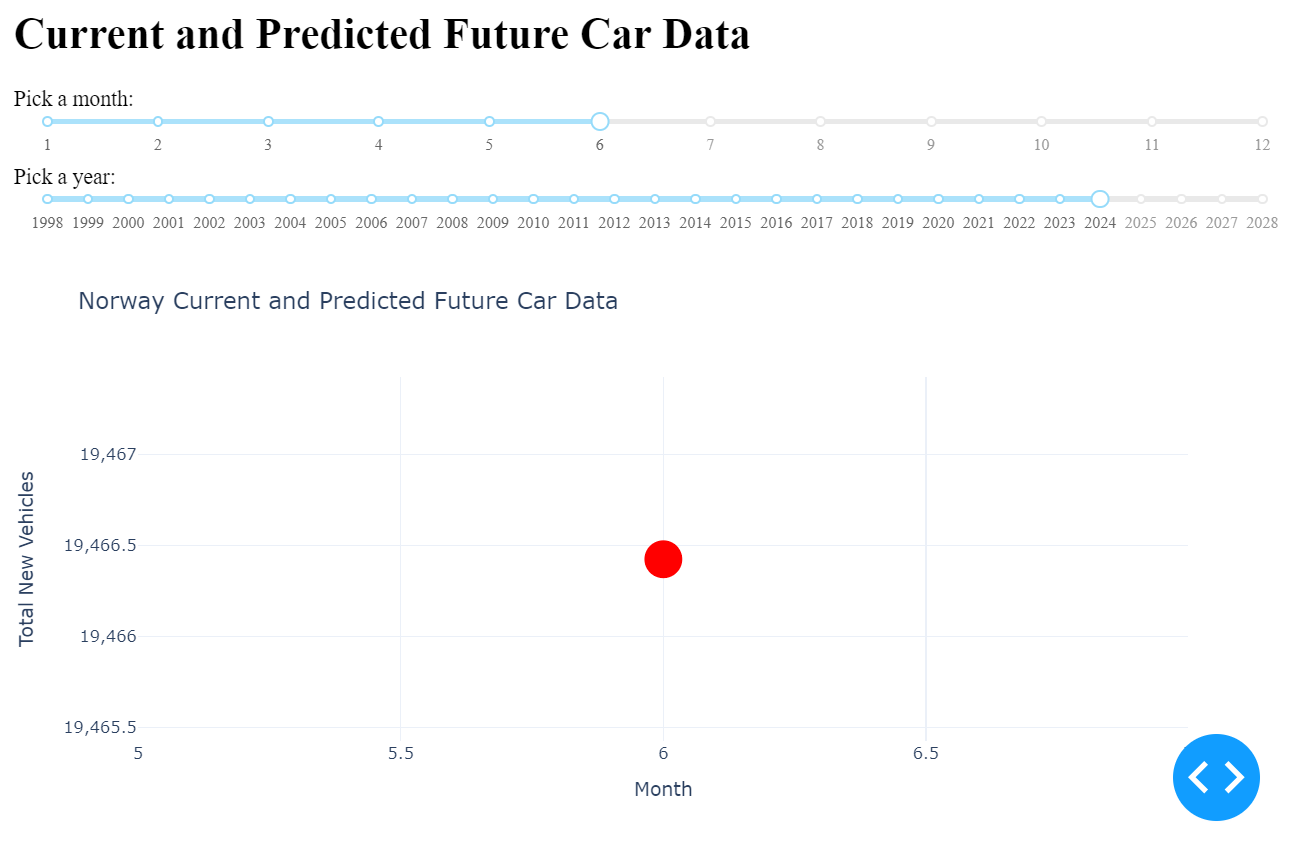
This chart I created compares the normalized MSE, RMSE, R^2, and average cross-validation scores of the Random Forest and Gradient Boosting models. Normalization makes the MSE and RMSE values of both models comparable as ratios, while the actual scores written on top of the bars allow us to see the full scale of the values. The bar graph clearly displays the performance of each model across the metrics, providing a quick assessment of which model performs better or worse under certain conditions.

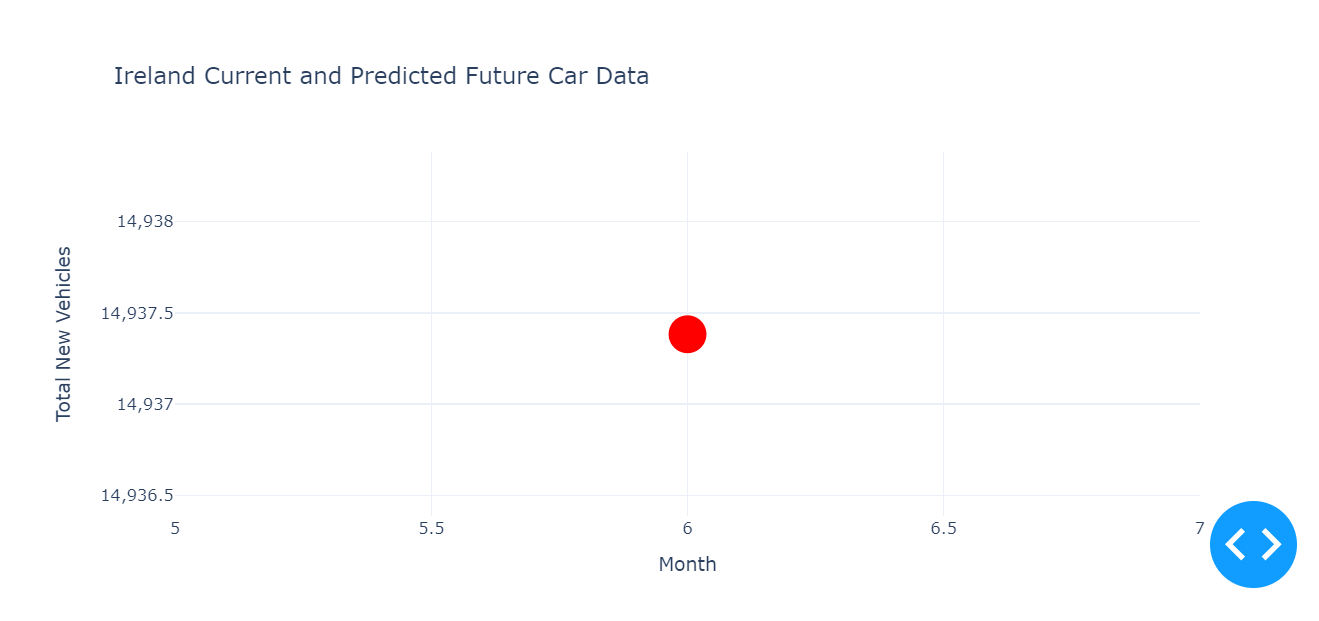
It's noticeable that the Gradient Boosting model has lower R^2 and average cross-validation scores compared to the Random Forest model. This could indicate that incorporating lagged analysis while constructing the Random Forest model may have allowed for better results. It may also suggest that the dataset used by the GBR model could be noisy, leading to some inconsistencies.

### Dataset Preparation for Dashboard

To prepare the data for the dashboard, I worked on two separate datasets containing current and forecasted vehicle sales data for Norway and Ireland. First, I created a time frame by creating combinations for each year and each month between 2024 and 2028. Then, I duplicated the Norwegian dataset and removed some columns, and did the same for the Irish dataset. For both countries, I used the predicted future vehicle sales data from my previous machine learning model and combined it with the year and month information to create the final datasets.

### Dashboard





While creating the dashboard, I added two sliders for users to select year and month, which allow users to filter the data according to the time period they want. I then defined a callback function that will generate two charts that will show current and projected vehicle sales data for Norway and Ireland based on the users' selections. I set it up so that when users select year and month, two separate scatter plots will be updated and displayed with the data filtered according to these selections. At first I wanted to create a line plot. but for some reason the visibility of the graph was problematic. so I drew a dot plot and made some customizations for the dot.

### Dataset Preparation for Sentiment Analysis

I wanted to analyse Reddit comments responding to news about upcoming taxes on newly purchased vehicles to encourage public transport and reduce car usage. I began the analysis by simplifying the text. Stop words, which are commonly used words that carry little meaning in text and language processing tasks, were removed first. Then, I defined a function to calculate metrics like the number of characters, average word length, stop word count, special character count, and most frequent words for each comment. This function tokenizes the words in the text and counts the frequency of words made up of alphabet characters and excluding stop words. As a result, I created a list with these metrics for each comment.

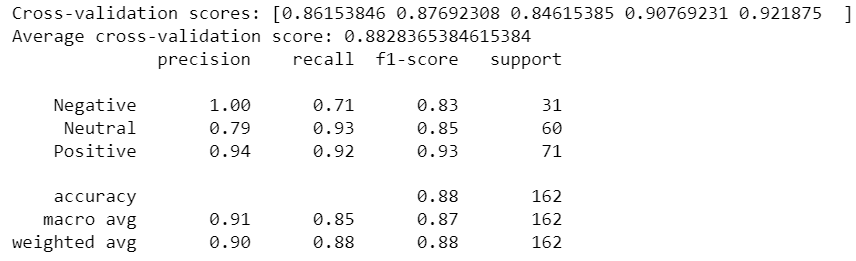
Next, I turned my analytical findings into a dataset. Using a new function, I calculated the number of characters, average word length, stop word count, special character count, and most frequent words for each comment. I filtered the comments with an average word length greater than zero and that were not deleted, and reset their indexes. Finally, to further process the texts, I reduced words to their base forms (lemmatized) and removed stop words. These steps cleaned the comments, and I stored the lemmatized versions in a separate column.

### Sentiment Analysis

I performed sentiment analysis on the lemmatized comment texts. By calculating the sentiment score of each comment, I obtained polarity and subjectivity values to determine the positivity or negativity of the texts and whether they are subjective. Polarity measures emotional charge, while subjectivity indicates whether the text contains personal opinions or emotional expressions.

I then defined function based on the polarity values to label each comment as 'Positive', 'Negative' or 'Neutral'. This categorization made the results from the sentiment analysis clearer. The categorized sentiment categories served as target variables for my machine learning model.

To train the machine learning model, I used polarity and subjectivity values as features and trained a Logistic Regression model with these features. To evaluate the performance of the model, I split the dataset into training and test sets and calculated the classification report, accuracy and F1 score of the model. To improve the performance of the model, I conducted a parameter search and used GridSearchCV to find the best hyperparameters. The cross-validation scores of the best model were consistent and showed high overall performance.



Since I chose half of them as the test size, I was able to classify 162 comments, of which 71 were positive, 60 were neutral and 31 were negative. The average cross-validation score was 88.28% and the accuracy score on the test set was 88%.

## Results

I used machine learning models, descriptive, and inferential statistics to compare and predict the impact of global events like economic crises and pandemics on vehicle registrations in two countries. It reveals key findings: Ireland's vehicle registrations fluctuate more compared to Norway's stable records; significant global events notably affect vehicle registrations; and the relationship between vehicle sales and population growth is complex, influenced by external factors such as economic conditions and policies. Norway showed steady growth until 2020 but was significantly affected by a change in government policy that year, indicating how a single policy shift can have a large impact.

# References

G. Berry, P. A., 1995. Mid-P Confidence Intervals: A Brief Review. *The Statistician,* Issue 44, pp. 417-423.

Pompeii, L. A., 1998. Inferential and Advanced Analysis of Research Data. *AAOHN Journal,* Issue 46, pp. 266-267.

RB., D., 1970. Transformation to normality of the null distribution. *Biometrika,* Issue 57, pp. 679-681.

Shapiro, S. & Wilk, M. B., 1965. An analysis of variance test for normality (complete samples). *Biometrika,* Volume 52, pp. 591-611.