# Group ID - MSc in Data Analytics

Group ID - MSc in Data Analytics

* Author: J. Doyle
* e-mail: sba23437@student.cct.ie
* Student ID: sba23437
* GitHub: <https://github.com/sba23437/GIT-Repository-for-CA2.git>

# 0: Title: CA2: Integrated Assessment: Transport in Ireland

## 1: Introduction

This is an investigation into Irish transport and specifically commuting times in Ireland. We have used datasets published by the Irish Central Statistics Office, the European group Eurostat and the global organisation OECD which provide us with total population commuting estimates with demographics.

Employing best practices to explore, prepare, transform and utilise large data repositories, we have programmatically developed a step-by-step python solution which describes for us key insights from the data.

Our first steps in the project were to apply descriptive and inferential statistics to understand our data as well as to help us plan how to investigate it further in our Machine Learning and Visualisation sections. We have taken the data and applied sentiment analysis with machine learning classification and clustering methodologies in python helping us to further examine these fascinating insights and to understand potential insights for our populations.

The work was accomplished using the programming language Python as an open source, quick to deploy and one of the most popular languages available for this type of undertaking (Liebowitz 2013, p261). It’s Open source and “ideal for computationally intensive applications and general-purpose systems” (McKinney, W., 2013). The Jupyter Notebook environment was employed for our development environment. It too is open source and compatible with Python.

Finally, the output from the project is an interactive and simple to use dashboard using the python Dash library. This type of dashboard has the potential to inform users as they plan their lives or develop policies. Recently a comprehensive Irish Transportation dashboard was published by the Irish CSO on their public website, indicating how infographics and interactive dashboarding can be used by scientific and statistical groups to help citizens plan, make policy and investigate by using reliable, valid and accessible data (CSO, 2023).

## 2: Project Methodology

The CRISP-DM framework is an iterative framework for data mining, modelling and deployment endorsed across many business areas. Stepping through the main phases of this planning framework became the template for this project, with specific focus on the iterative aspects needed in A diagram of data mining life cycle

Description automatically generateddeveloping this submission. In the visual (Figure 1 on left), we can see the process starts with business understanding, however, while this informs the next phases of Data Understanding, Data Preparation, and Modelling these phases still iteratively impact the business understanding and each other. The benefits of approaching data projects like this are that it provides an agile methodology that lends itself to prototyping and learning as you go while discovering new and unexpected data and environmental constrains. The downside to an agile approach like this however is that it can lead to scope creep, an unending loop with incremental improvements for diminishing gains (Chapman et al, 2000).

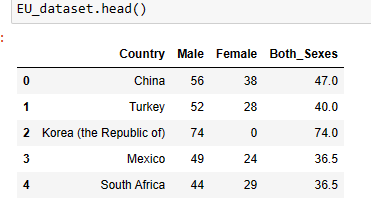
## 3: Data Preparation

The acquisition of the transport data for this project was split into two sections: Quantitative Data and Sentiment or Qualitative Data.

### 3.1: Quantitative Data

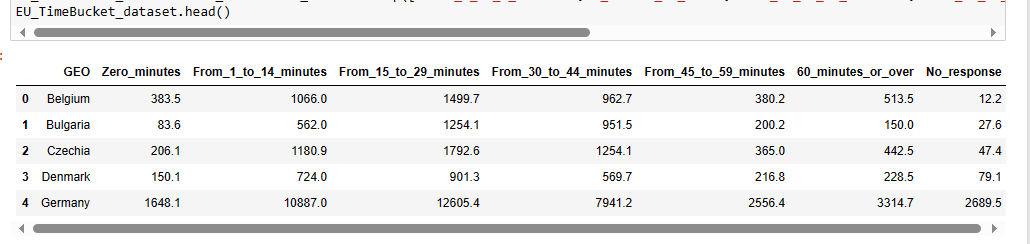
The data sources were CSO, Eurostat and OECD, which are all organisations dedicated to data and statistics for different regions (Irish, European and Global). The acquisition of this quantitative data was handled in the notebook called ‘CA2\_Data\_Acquisition&Stats’ in the section called ‘Data Acquisition’. The pd.read\_csv function was leveraged on 3 CSV sources.

**EU\_dataset**

****

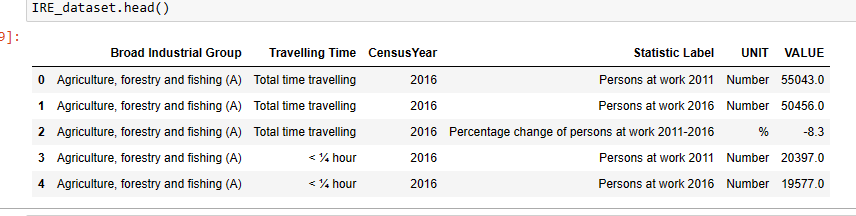
This source of data is OECD, published in 2016 and it includes average minutes per day spent travelling to paid work or study by men and women in OECD countries plus China, India and South Africa, as of 2016. It includes 29 countries from the OECD members, many of which are European, but some global countries then also such as the US, Australia, Japan and more.

**EU\_TimeBucket\_dataset**

****

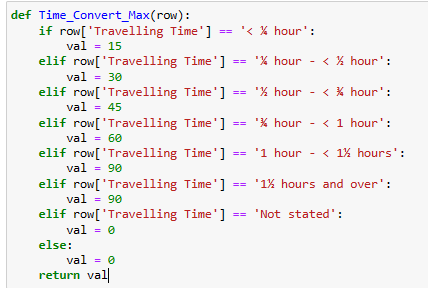
This source of data is Eurostat, published in 2019 and it describes persons in employment or education by commuting time, by country. It represents all European countries.

**IRE\_dataset**

****

Th third dataset (E6028) comes from the CSO website, published in 2016 and describes the population aged 15 years and over at work, in Ireland with their industry identified and by the commuting bracket they fall into.

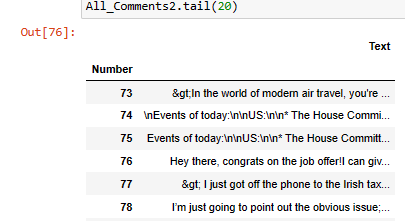
All 3 quantitative data sources are related in terms of theme, however the scope and shape of the data was a little different in each file. The Data Acquisition and EDA sections of the ‘CA2\_Data\_Acquisition&Stats’ notebook are used to understand the data types, shape and counts in each file. All are checked for nulls and duplicates too to handle before we process in later stages. Headings are relabelled to be clearer or to remove special characters making coding in later stages simpler. Finally, important updates to each dataset are conducted to remove dimensions that are not in scope for our investigation, removing categories of data that are overlapping so we are getting very clean cuts of demographics for later steps.

Having 3 data frames of clean data created for our quantitative data, made it possible throughout all the following statistics, machine learning and visualisation sections to then union data, perform a number of joins to include for example country ISO abbreviation data needed for our dashboard choropleth maps. In all sections we were able to pull together new data frames or to create series and reshape the data to fit it to the intended function. Functions were defined to convert data types too for example I needed to convert categorical labels into numerical data shown in the image to the right.

Licensing and terms of use on all 3 websites are clear that this data is appropriate for public analysis especially for research and policy development work.

### 3.2: Qualitative Data

**All\_Comments2**

****

A close up of words

Description automatically generatedFor the quantitative data or sentiment data on transport, Reddit was selected as it contains a lot of useful discussion and multitudes of comments between users on these topics. The notebook called ‘CA2\_DUBLIN\_TRANSPORT\_REDDIT\_Sentiment\_Analysis&ML Techniques’ contains this pipeline of data acquisition and EDA. A connector API is available for Reddit Community members to leverage to extract large amount of Reddit text data. The licensing for this API contains a number of terms and conditions, but for the purposes of college research assignments and where we ensure the data has been appropriately deidentified, it is appropriate for use. Deidentification is the process of removing user identification such as ids, handles, accounts, emails, labels etc and even meta data from the text commentary so it cannot be tied back to an individual user.

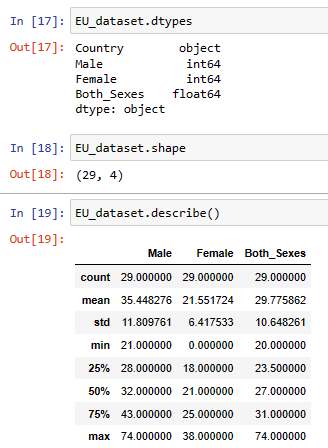
One of the first tools leveraged to understand the large amount of collected data was to create a WordCloud. This looks as follows on our Reddit data as seen in Figure2 to the left, but other steps taken to prepare the text data was to tokenise the words and remove stop words including custom stop words I identified as occurring frequently that were specific to this domain but were not valuable to analyse.

## 4: Statistics

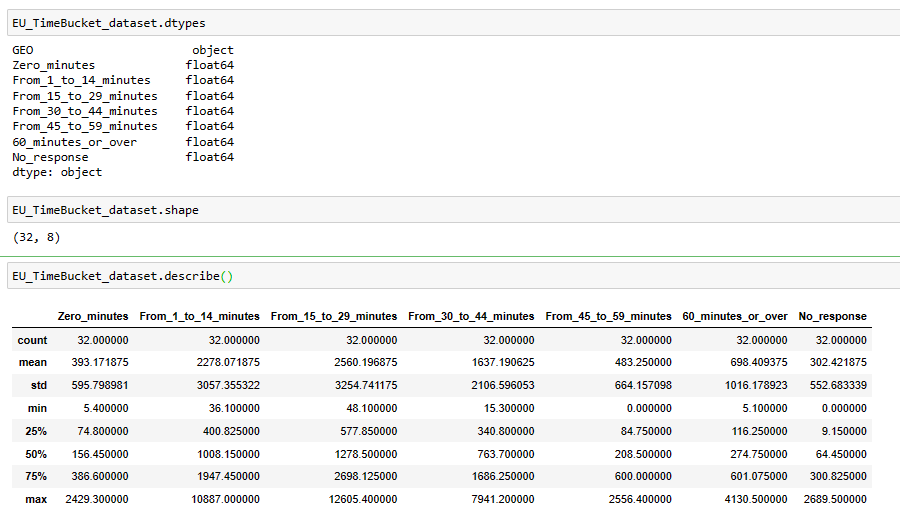
### 4.1: Descriptive Statistics

Descriptive Statistics were performed on the Quantitative Data:

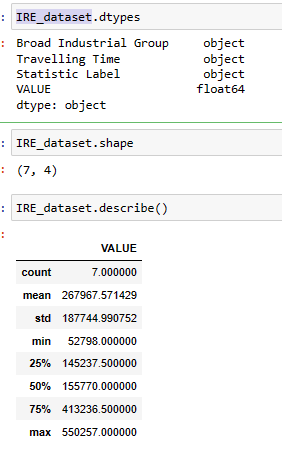
**The EU\_dataset**



**The EU\_TimeBucket\_dataset**

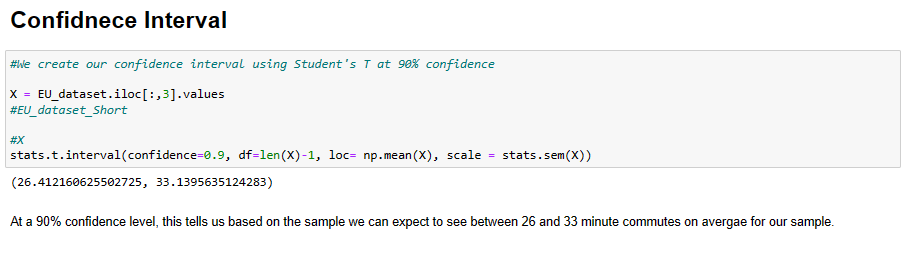


**The IRE\_dataset**



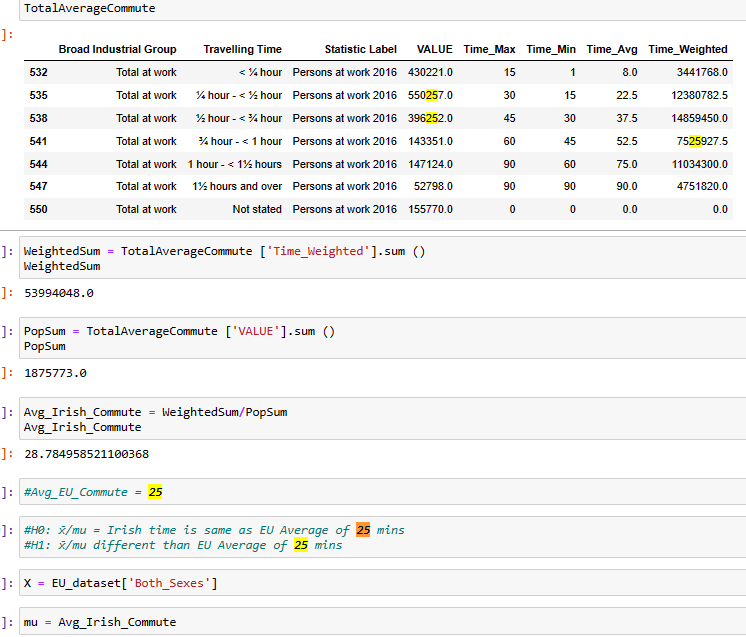
### 4.2: Confidence Interval

Looking at the mean commute time based on the EU\_dataset, we see an average of 29.8 minutes for both sexes. Applying confidence intervals to this data at a 90% confidence level, this tells us based on the sample we can expect to see between 26 and 33 minute commutes on average for our sample.



### 4.3: Inferential Statistics

Firstly, for our inferential stats, we prepared the data, creating features we could then use in our inferential tests. We started by analysing the Irish average commute times:



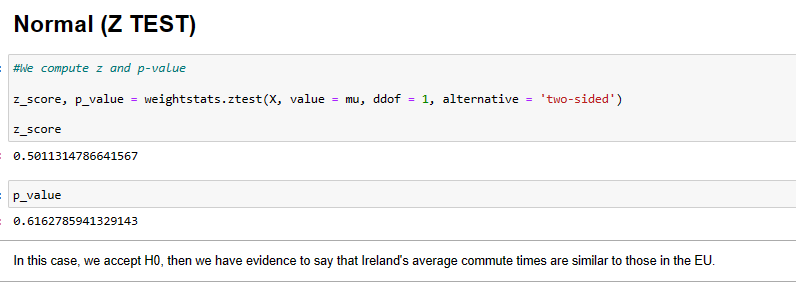
In another source from Eurostats from another time frame we saw that the EU has an average commute time of 25 minutes. (2019) as shown in figure 3 below.

**Figure 3**

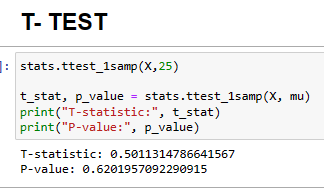
A map of europe with numbers and a number of countries/regions

Description automatically generated with medium confidence

In our first test we want to hypothesise that Ireland has a significantly different average commute time to that of the European average, but we found it is in line with Europe’s in fact.



In comparison to the Z test, we ran a T-test and generated an almost identical result where we accept the null hypothesise that Ireland’s commute times are no different to those of Europe.



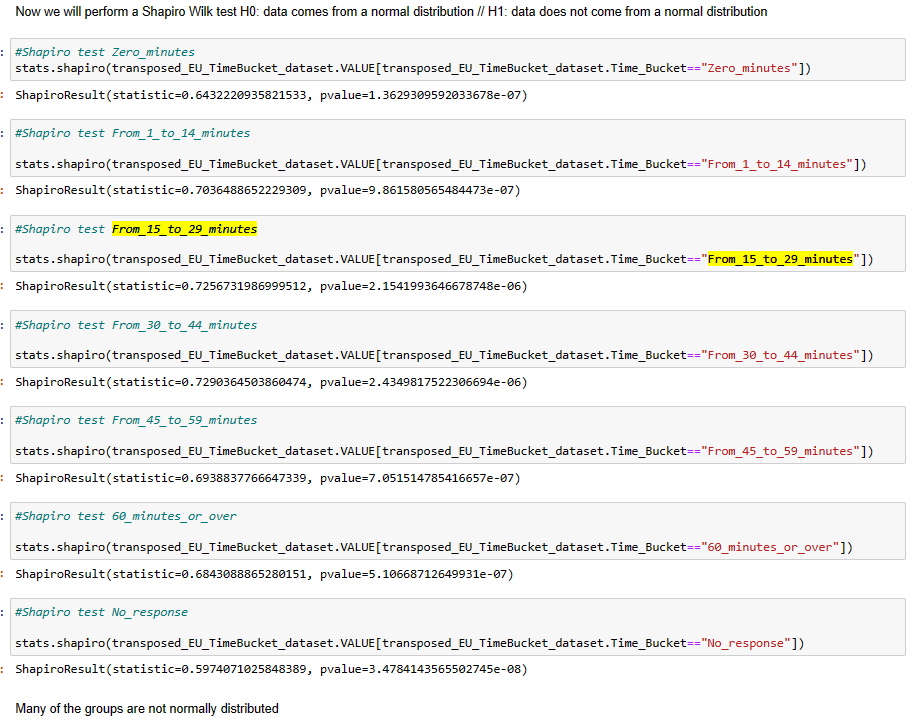
In our next round of inferential tests, we reshaped our European data into a new data frame called ’ transposed\_EU\_TimeBucket\_dataset’ which enabled the following 1 Way and 2 Way ANNOVAs.



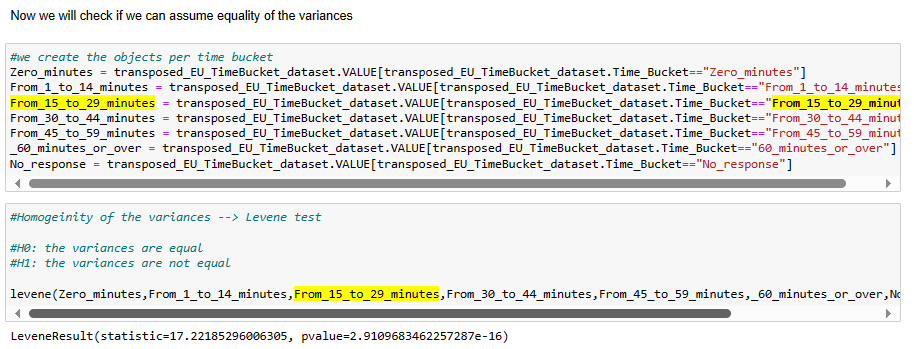
We need to understand however if our data is appropriate for parametric inferential tests such as t-tests and ANNOVAs. To understand this, we plotted a probability plot firstly and for example the following shows us the data follows a reasonably straight line indicating it may be fit for inferential tests:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| From\_1\_to\_14\_minutes | From\_15\_to\_29\_minutes | From\_30\_to\_44\_minutes | From\_45\_to\_59\_minutes | 60\_minutes\_or\_over |
|  |  |  |  |  |

We perform the Shapiro Wilk test next to identify if the data is normally distributed, but we find that most groups are not:

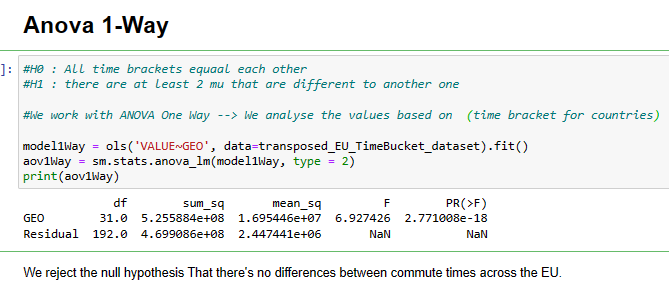


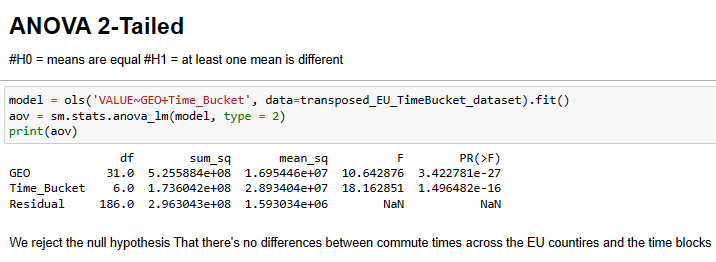
We also checked if we could assume equality of variance across our data in the following steps:



We find that the variances are not equal in our data.

We know at this stage that we should not use parametric tests on our data given the above findings. However, for the purposes of learning we proceeded with the 1 way and 2-way ANOVA tests to understand what these tests can do for us in the future.





In both tests, there is support for the hypothesis that there are differences between commute times across the EU countries in the time brackets in scope for our study. However, due to the issues with the data from the above normality tests we cannot use parametric tests and instead must use non-parametric means to describe characteristics of the data.

For this we will use the Kruskal-Wallis Test instead.



Leveraging this non-parametric test, we can show that there are significant differences between male and female commute times across the EU states.

From our statistical analysis, we know that there’s probably something worth exploring further in the commuting Time Data we’ve obtained with relation to gender differences and also to the dimension of time brackets where people are commuting in these groupings differently across countries. This gives us a good direction to explore in our visualisations and provides us with quantitative features we potentially can merge with our sentiment data in the next stage of the project with Machine Learning.

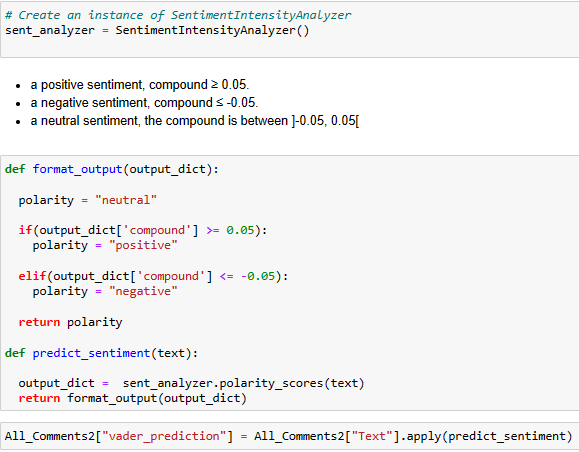
## 5: Machine Learning

### 5.1: Sentiment Analysis Section

#### 5.1.1: VADER Sentiment Analysis

After collating a large sample of comments and posts from Reddit which contained the following keywords (‘Ireland’, ‘Irish’, ‘Commute’, ‘Commuting’) we were ready to start applying different models to understand the data.

Next, we implemented VADER to analyse each text and assign a positive, neutral or negative sentiment to it:



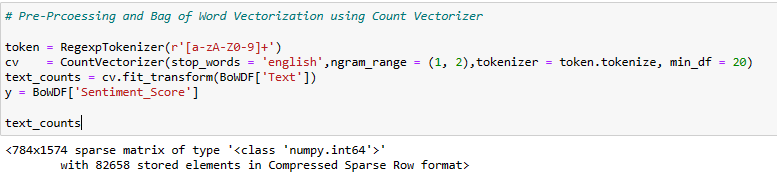
We next checked for the overall sentiment and found that the polarity of the sentiment in our reddit data is quite positive regarding Irish Transportation on average. Where 1 = positive and 0 = Not positive and the Mean is .7. This was a quick and very simple method to attempt to attribute sentiment valence to our raw data file, helping us create features we can leverage in our next steps.

Typically, in the real world we’d have a sentiment score or Likert scale with our text to guide our models, but in this learning scenario we are leveraging VADER to create these scores to leverage in our next steps.

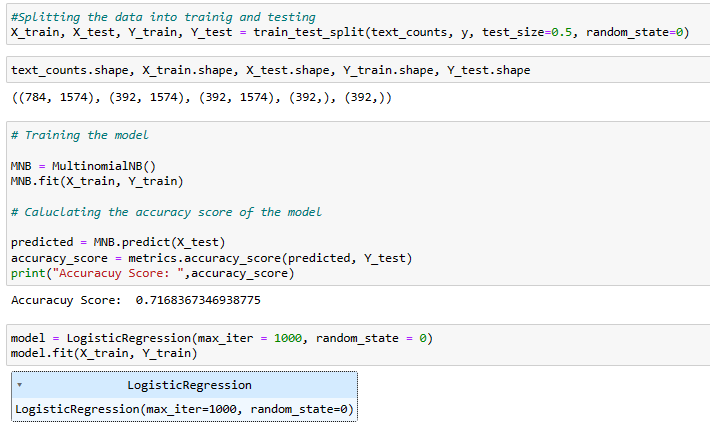
Checking the validity of these scores can be done by sampling at random the text associated with the positive or the negative scores and deciding if they truly are appropriate or not. In our sample there appears to be a consistent categorisation with negative or positive sentiment in the comments, but it appears to be quite basic and untied to context or able to detect sarcasm etc.

#### 5.1.2: Logistic-regression Model Sentiment Analysis

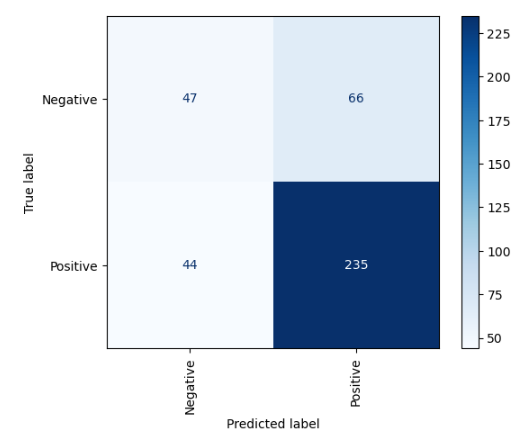
Using this model, we were able to call a few additional functions relatively easily including Count Vectorizer, removal of Stop Words and tokenisation of the data.



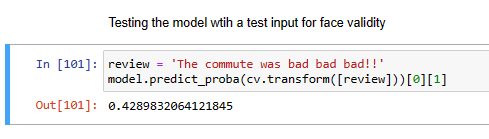
Next the data was split into training and test data, before the model was trained. Upon evaluation it had an accuracy score of 72% which was quite good based on the condition of this experiment and the reliability of the sentiment scores applied by VADER in the previous steps.



Using a confusion matrix, we see that our trained model positively characterised the data 235 plus 47 times aligning with the true label and only 66 plus 44 times incorrectly labelling the data. This is a good outcome.



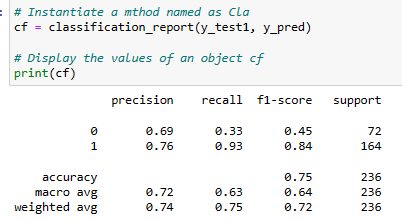
How does the model compare in unit testing? I faked a bad comment and entered it as a test as follows:



It passed the test with the model providing amore negative score in line with what I’d expect for the text I’d used.

#### 5.1.3: MultinomialNB Model Sentiment Analysis

In comparison, we tried leveraging a multinomialNB model. This resulted in a slightly better accuracy score:



### 5.2: Classification Section

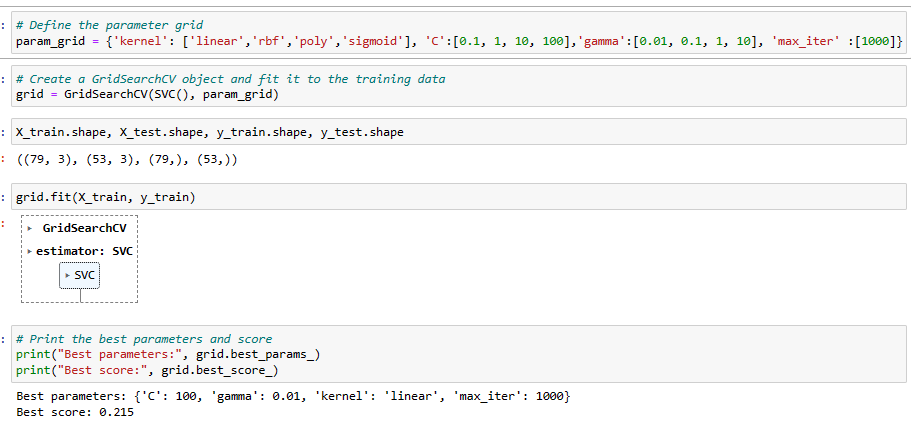
In our classification section, we are exploring the commuter data for Ireland to understand if we have ‘types’ or ‘groups’ that can be described. Imagine creating a way of discussing commuters in Ireland, like how large supermarkets classify their customers for targeted advertisements. In the following steps, we use GridSerachCV to evaluate models that will help us analyse the data and then we apply the model to discover potential insights.

Based on the scikit-learn library and the needs of this study, we are interested in clustering or classification models. These give us methods to analyse the data and create clusters that could be used then to understand types of commuters. For classification, when dealing with <100K samples SVC is a good option to start with. For clustering KMeans is a good model to leverage (SicitLearn, 2023).

#### 5.2.1: GridSearchCV

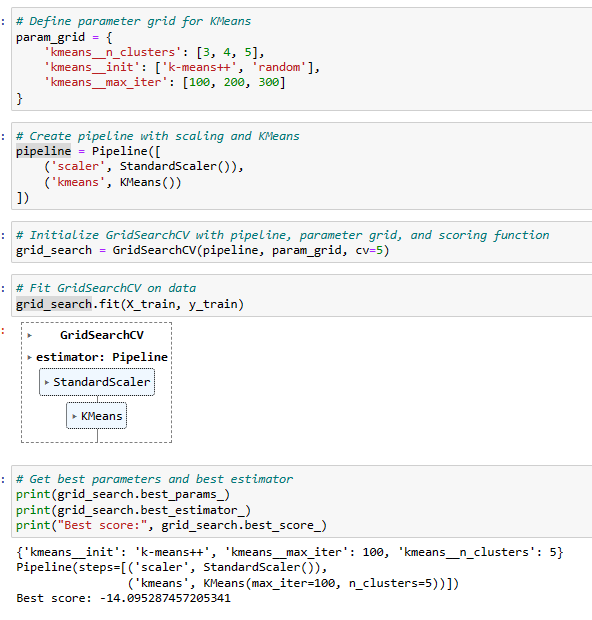
**Support Vector Machines**

We leveraged GridSearchCV to understand the best parameters to use and were recommended to use Linear SVC with a C of 100, gamma of 0.01 and max iterations of 1000. However, the accuracy was really low at 22%.



**KMeans**

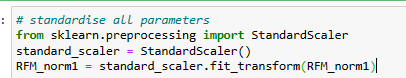
With KMeans, I had to rerun the model with different clusters to understand the best model to leverage. For the purposes of creating a commuter taxonomy, I would prefer a simple 3 to 5 clustering and the model recommended using 3 to 5 (small return on increasing from 3 to 5) on 300 max iterations. However, there is very low accuracy on versions of the model I reran.



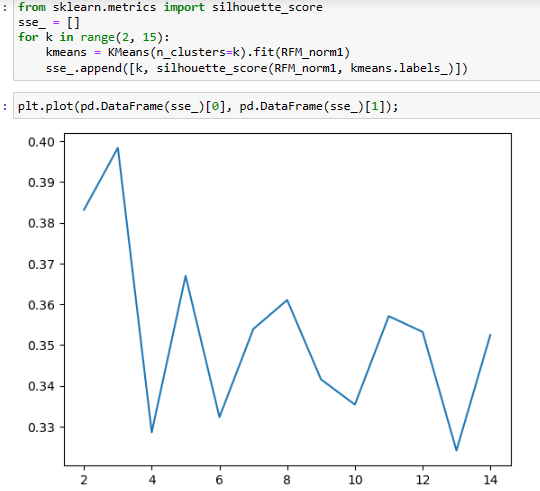
#### 5.2.2: K-Means with some K

I’ve decided to leverage a clustering model to identify types of commuters in the data. We know the accuracy is unfortunately low, but in a real-world scenario this would not be the correct pat to follow and we’d hope for a much larger data set. It’s possible there are no clusters naturally in our dataset also. We are at this point demonstrating how we could find these clusters if they did exist in the data.

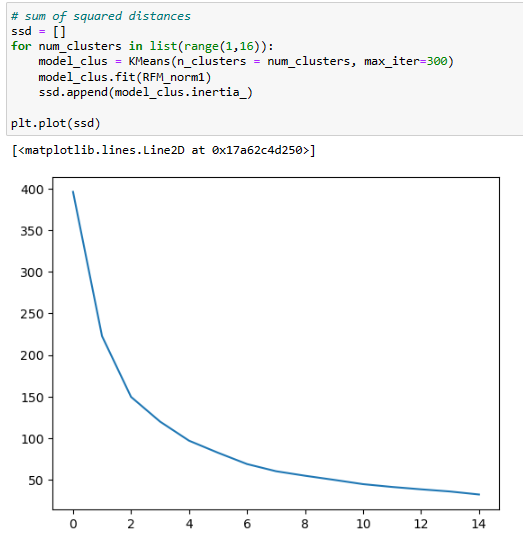
Step1 in this process was to standardise the data for the model. This increases the model’s accuracy and increases the speed it can compute or train on the training data set:



Next up, we employed a silhouette analysis and looked at the Sum of Squared Distances to understand what the best number of clusters was in our data. We see that that 3 or perhaps 5 are good options form the silhouette analysis:



And leveraging the Sum of Squared Distances, 3 seems to me the turning point for clusters providing the most distinctive explanatory power:



What do the results tell us? We have been suggested 3 clusters by our model. The first commuter group is typically related to small industries (Sample Size is small) with typically the longest commute times but also with the largest increase in commuting time between 2011 and 2016. I would consider this group to be the ‘**Worst Impacted Commuters’** who see to get hit hardest by increases in commuting times, but because the population is small, they probably are not considered in many planning processes due to a low return on investment potentially. This type of grouping and naming is based on domain expertise and research or might help with marketing strategies.

Group 2 and 3 are quite similar with large populations and low changes in commuting times over time, but group 2 having longer commute times and group 3 having shorter commute times. We can call group 2 ‘**Majority Long Distance Commuters’** and group 3 ‘**Majority Short Distance Commuters’.**

* Group 1: Worst Impacted Commuters
* Group 2: Majority Long Distance Commuters
* Group 3: Majority Short Distance Commuters

|  |  |  |
| --- | --- | --- |
| Sample\_Size | Commute\_Duration | Sample\_Volatility |
|  |  |  |

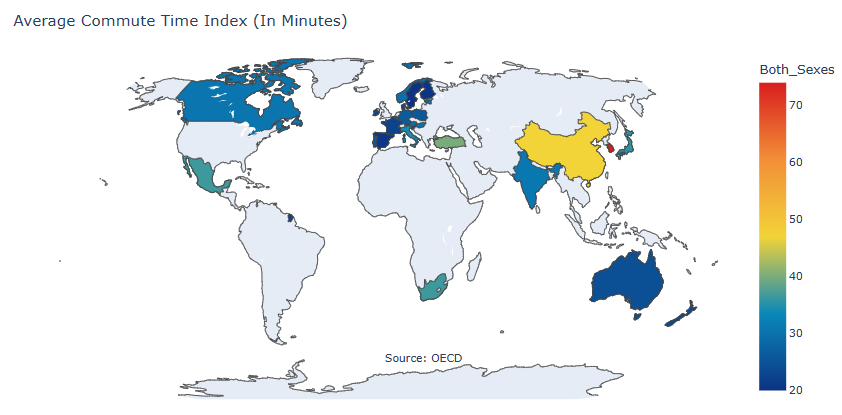
## 6: Visualisations

### 6.1: Prototyping

From our previous sections in the project, we have acquired the data, cleaned the data, transformed the data including the creation of features and prepared data sets that we can build visualisations on. Our statistics section helped us identify that country and gender and commute time are the most impactful aspects of our data with which we can explore in our interactive application in the next step.

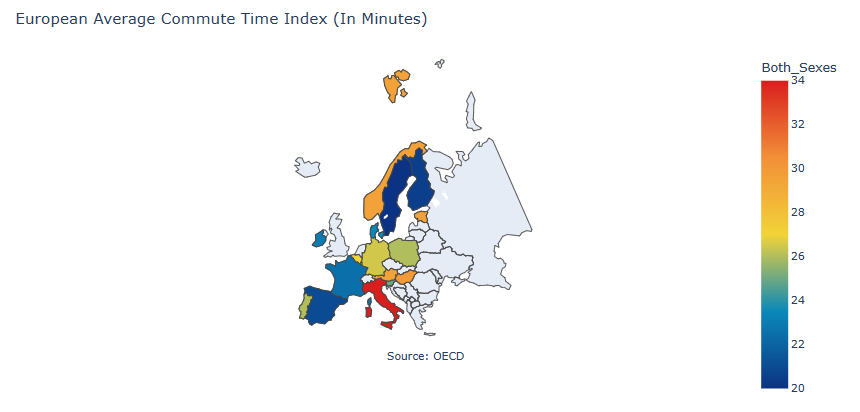
Firstly a few standalone visuals were created as prototypes for potential features in a larger dashboard. Standalone Visualisation 1 displays the world with colours that have a heat map effect drawing our attention to those countries with the longest commute times.

*Standalone Visualisation 1*



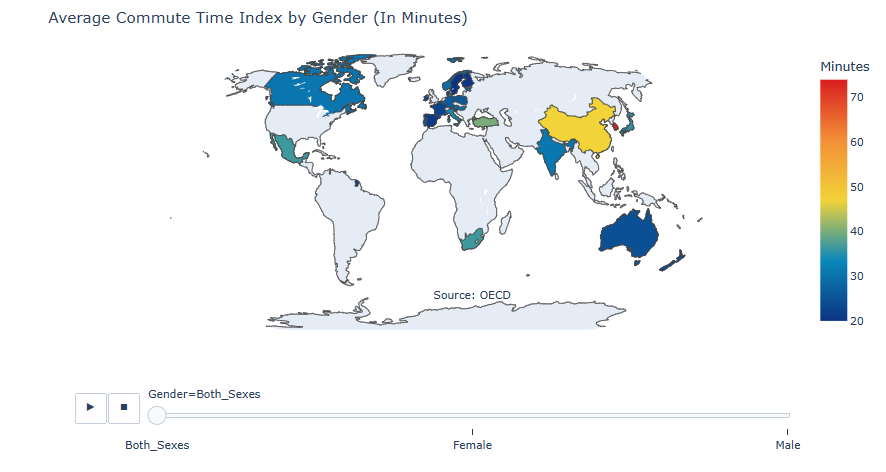
Standalone Visualisation 2 is related but where we could zoom into the European space to understand commuting across Europe. This is useful because globally we can see Korea and China being extreme outliers. This is good information in and of itself, however it means we don’t really see an intra-European visual analysis. In the updated EU view, we now see where Ireland compares for example to the extremes of Italy (Long Commutes) and Finland (Short Commutes).

*Standalone Visualisation 2*



Finally, we prototyped one last image called Standalone Visualisation 3 which provided an interactive control for the end user along the bottom of the map where they can specify if they want to see commute times for ‘Both Sexes’ or just for Males or Females. This is the direction we want to go further to enable analysis and deeper understanding of the data.

*Standalone Visualisation 3*



### 6.2: The Dashboard

All charts and graphs for the dashboard were developed with best practice concepts where simply put ‘less is more’. The key concepts employed here were (Tufte 2001, p105):

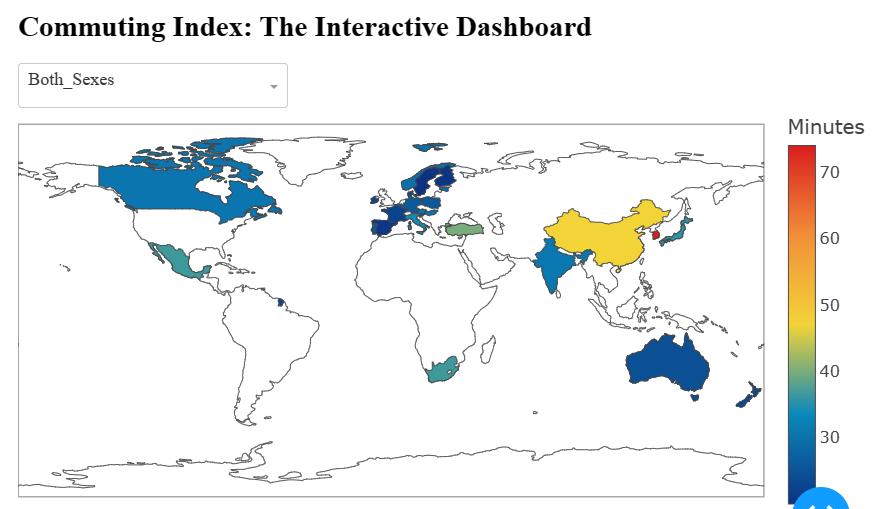
* Show the data first and foremost
* Maximize the data to ink ratio
* Remove non-data
* Erase redundant data
* Reiterate

The Dash library was utilised for this because of its user interface and design flexibility. Often, we need complex relationships between charts and tables, especially where end users are leveraging the tool for analysis ad-hoc and often unplanned work often-times. This requires a design that seems simple and intuitive, but underneath has a deeply thought out and well-constructed data structure.

In our use case, we are looking at average commute times in Ireland across industries, across genders and comparing it to the rest of the world. We were able to achieve this using 3 prompts/ user inputs and 4 graphs/ maps as outputs.

The first visual gives the user the full picture of the world, but immediately gives them the prompt to investigate the data by sex if they wish or otherwise, they get the default value for both sexes. The heat map colours will update to larger or smaller ranges depending on the selection made.

*Dashboard Visualisation 1*



Immediately following after this are 2 more optional user inputs/ prompts for Country and for Industry. This means you can select for example ‘France’ and look at ‘Total at work’ or look at the ‘Farming’ industry for example.

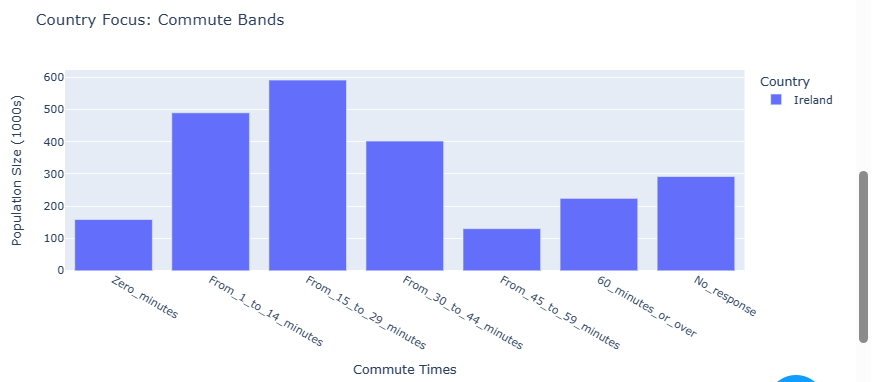
This section has 3 charts then with Dashboard Visualisation 2 showing us gender differences for the specified country. Dashboard Visualisation 3 shows us the break-down of commuting bands for the population of that country and lastly Dashboard Visualisation 3 gives us the commuting bands again but for the specified Industry selected.

Each chart is labelled clearly and the flow from top to bottom helps create a story of meaning for the user as they step through the dashboard step by step. Everything is defaulted on arrival, but the user can choose then to do their own analysis or ad hoc data exploration as needed.

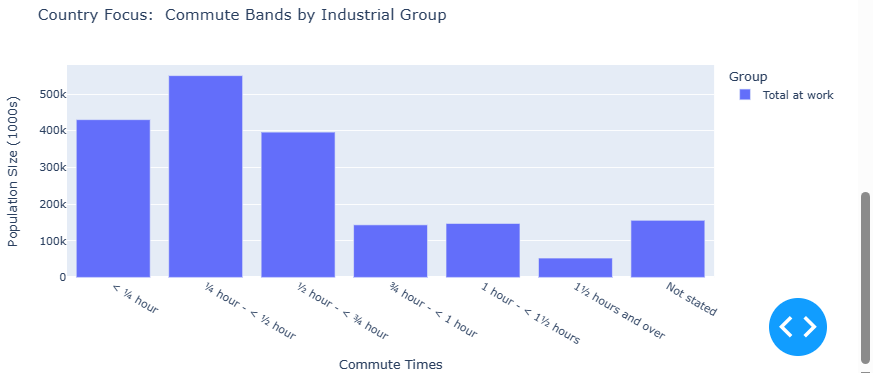
Dashboard Visualisation 2



*Dashboard Visualisation 3*



*Dashboard Visualisation 4*



## 7: Programming Discussion

For all programming tasks, please refer to the 3 Jupyter notebooks attached to this submission.

For task 2, please note the use of the use of CSV formats in the main data acquisition notebook called ‘CA2\_Data\_Acquisition&Stats’ whereas APIs in JSON format are used in the notebook called ‘CA2\_DUBLIN\_TRANSPORT\_REDDIT\_Sentiment\_Analysis&ML Techniques’.

For tasks 4 & 5, please note the use of pandas otherwise known as pd in the notebook called ‘CA2\_Data\_Acquisition&Stats’ and the use of ‘requests’ and ‘Jason libraries in conjunction with pandas in the notebook called ‘CA2\_DUBLIN\_TRANSPORT\_REDDIT\_Sentiment\_Analysis&ML Techniques’. In comparison, these libraries are all serving very effective methods for connecting to and reading data from different sources. In this instance they read CSVs and Jason format API data. In contrast to each other however, they are all independent and serve specific utility with the key to success being our ability to pass our data seamlessly in Jupyter between the functions. This requires an understanding of the conditions needed within each function and to have the data pipeline in the correct shape and format as it is passed/ processed from function to function. There is added bonus to leveraging JASON formatted data and passing the data into parquet files where our data becomes large and pushes the processing time for our data too long. With the current size of the data set in this use case we didn’t meet the threshold to transform the data any further than the data frames and Jason APIs we leveraged.

## 8: Concluding Remarks

Through this work, we have been able to acquire data from multiple sources, in different formats and transform this data into useable tables and features. We leveraged an array of parametric and non-parametric tests to understand the data and the significance of specific features within it. We were able to enrich our quantitative data with sentiment data scraped from social media sites and apply positive and negative polarity to it. We were able to generate classifications or types of commuter groups in Ireland too which was really intriguing:

1. Group 1: Worst Impacted Commuters
2. Group 2: Majority Long Distance Commuters
3. Group 3: Majority Short Distance Commuters

Finally, we pulled it all together into an intuitive and easy to use dashboard. Interestingly in the real world at this same time, we saw the CSO launce the Irish Transport Hub on their web site which is an interactive dashboard on Irish Transport Data. This shows us how this type of work could be translated into a real-world job following our course completion.

## 9: References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide,* SPSS Inc

CSO (2023) The Transport Statistics Hub. Available at: <https://lnkd.in/eHJNzJky> (Accessed: 11 Dec 2023)

CSO (2020) Population Aged 15 Years and Over at Work, Usually Resident and Present in the State 2011 to 2016. Available at: [E6028 - Population Aged 15 Years and Over at Work, Usually Resident and Present in the State 2011 to 2016 (cso.ie)](https://data.cso.ie/) (Accessed: 1 Dec 2023)

Eurostat (2016) Average minutes per day spent travelling to paid work or study by men and women in OECD countries plus China, India and South Africa, as of 2016 | Statista. Available at: [Statistics | Eurostat (europa.eu)](https://ec.europa.eu/eurostat/databrowser/product/page/lfso_19plwk28$defaultview) (Accessed: 1 Dec 2023)

Eurostat (2019) Average European Commute Times. Available at: [Majority commuted less than 30 minutes in 2019 - Products Eurostat News - Eurostat (europa.eu)](https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20201021-2) (Accessed: 5 Dec 2023)

Liebowitz, J. (2013) *Big Data and Business Analytics,* CRC Press

McKinney, W., (2013) Python for Data Analysis. Sebastopol: O’Reilly Media, Inc. pp. 1-3

Scikit-learn (2023) Choosing the Right Estimator. Available at: [Choosing the right estimator — scikit-learn 1.3.2 documentation](https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html) (Accessed: 5 Dec 2023)

Tufte, E. (2001) *The Visual Display of Quantitative Information, (2nd ed.)*, Graphics Press