A picture containing icon

Description automatically generated

|  |  |
| --- | --- |
| **Assessment Title:** | ***Transport in Ireland*** |
| **Student Full Name:** | NATALIA IOLCHIN |
| **Student Number:** | sba23303 |
| **Assessment Due Date:** | 24/11/2023 |
| **Date of Submission:** | 05/01/2024 |
| **GitHub Link:** | <https://github.com/Natidhcp/CCT_MSc_DA_CA2/blob/main/CA2.ipynb> |
| **Number of words:** | 3106 |

**Abstract**

The analysis underscores that land road transport stands out as the most unsafe mode, while showcasing a significant decrease in injury collisions and casualties from 2005 to 2012, followed by a consistent but slight decline until 2021. These trends, alongside a strong correlation between casualties, collisions, and injuries, confirm the hypotheses on declining casualties and fatalities, emphasizing a stable situation in collision trends over the past decade.

**The analysis showcases the potential effectiveness of road safety strategies in reducing fatalities, supported by a noticeable declining trend in forecasted values, aligning with governmental strategies. However, challenges persist in predicting Injury Collision numbers, emphasizing the importance of model refinement and data quality.**

**The comparison between Ireland and Denmark's average collision numbers reveals incomparability, indicating distinct differences. With the acceptance of the null hypothesis for 2 out of 3 indicators in 2021, in contrast to just 1 out of 3 for the entire 2005-2021 period, Ireland seems to align more closely with Norway's average road safety level. However, a more comprehensive and detailed dataset could provide further insights and enhance the depth of analysis.**

The public's stance on the subject is deeply polarized, with passionate supporters and strong critics. This polarization suggests the topic's significant relevance, evident from the intense emotions it evokes and the relatively low presence of neutral opinions. Moreover, this observation hints at the issue's complexity, indicating a lack of straightforward solutions.

1. Introduction

“Ireland could match the road safety record of better-performing countries” in European Union. (Eamon Ryan TD, Minister for Transport, Www.rsa.ie. (2018). RSA.ie - Home. [online] Available at: https://www.rsa.ie/.)

This study examines Ireland's transportation system safety, with a particular focus on road safety.

Road safety pertains to measures and practices implemented to reduce the risk of accidents, injuries, and fatalities on roadways. It encompasses a range of strategies, policies, and actions aimed at promoting safe behavior among road users, improving infrastructure design, and enhancing vehicle safety.

**This research aims to explore road safety in Ireland, focusing on previous years. It has two main objectives: first, to identify changing trends in road safety, and second, to assess the effectiveness of state measures in enhancing safety standards.**

**The investigation covers various aspects including accident rates, fatalities, injuries, and compliance with safety regulations. By analyzing these trends, the study aims to understand how interventions impact reducing accidents and related casualties.**

**Additionally, the study aims to compare Ireland's road safety performance with other European countries. By analyzing data from neighboring nations, it seeks to evaluate Ireland's position in terms of safety initiatives.**

**Through this analysis, the research intends to not only assess the current state of road safety in Ireland but also offer recommendations for further improvements and policy considerations.**

**The sentiment analysis will be conducted concurrently with the road safety assessment. Its objective is to acquire valuable insights into public perceptions, concerns, and feedback regarding road safety, aiming to enhance measures and effectively address pertinent issues.**

The project consists of four main parts:

1. transportation safety overview,
2. Ireland’s road safety analysis using time series,
3. Ireland’s road safety in comparison with other countries,
4. Ireland’s 'Road Safety' sentiment analysis.

Each section follows a consistent framework that combines aspects of both the KDD and SEMMA data mining processes. Some sections may combine multiple phases or involve more elaborate phases within this established framework. The data-related phases include:

* Selection
* Preprocessing
* Transformation
* Exploration
* Modeling
* Modification
* Assessment

thereby constituting the present framework.

Python Jupyter Notebook environment is used as a web-based interactive computing tool that allows to create and share documents containing live code, equations, visualizations, and explanatory text. Jupyter Notebook allows easy sharing and collaboration by exporting notebooks in various formats or hosting them on platforms like GitHub, facilitating the creation of reports or presentations, generates interactive plots, charts, and visualizations using different libraries.

For the data visualization 2 colorblind friendly palette are used as well as the Tufte principles:

* Maximize Data-Ink Ratio
* Use Clear, Detailed Labels
* Eliminate Chartjunk
* Provide Contextual Information
* Respect the Integrity of the Data

Edward Tufte's principles are recognized as fundamental guidelines for creating effective and informative visualizations.

The hypotheses of the study are:

1. Among various modes of transport, land road transport stands out as the most unsafe, exhibiting the highest victim’s rate.
2. The number of casualties decreased in the past years.
3. The number of fatalities decreased in the past years.
4. The overall trend for collisions shows a consistent decline.
5. **The implementation of road safety strategies is expected to lead to a fatality’s reduction, as evidenced by time series data.**
6. There is no difference in the average collision and casualties’ number for Ireland and other EU countries.
7. The public's perception of road safety appears to be favorable.
8. Transportation safety overview.

Analyzing accident statistics helps to compare safety levels and identify areas for improvement across different transportation modes. By looking at accident rates, severity, contributing factors, and safety measures, we can make better decisions to improve transportation safety overall.

The objective of this section is to elaborate on statistics related to transportation accidents or incidents and perform a comparative analysis of safety records across various transportation types in Europe in general terms and for Ireland in more details.

* 1. Transportation safety in Europe

[Eurostat](https://ec.europa.eu/eurostat/web/main/data/database) was selected as a source of data as it provides most recent and updated data about victims for each transportation type for the European countries.

Four obtained files were transformed into [‘pandas’ data frames](https://pandas.pydata.org/) and then filtered by selected criteria. Resulting data frame was visualized taking in consideration [colour-blind friendly palette](https://scottplot.net/cookbook/4.1/colors/#colorblind-friendly).

A pie chart with numbers and a blue triangle

Description automatically generated

A graph showing the number of victims

Description automatically generated

**Insights:**

* land road transport stands out as the most unsafe, exhibiting the highest percentage and number of fatalities.
* The secondleast safe transport is rail.
* Avia and maritime are the safest modes of transportation.

1.2. Ireland’s transportation safety

Ireland follows similar safety trends across various transportation types.

A green circle with a yellow line

Description automatically generated A graph of a number of people

Description automatically generated

Insights:

* The number of deaths for rail transport is substantially lower compared to casualties on roads, comprising a significant proportion of the total transportation fatalities.
* **First hypothesis** **accepted**.

To explore trends related to different segments of collision over the time a [JSON file was downloaded](https://github.com/CSOIreland/PxStat/wiki/API-Cube-PxAPIv1) via API from [CSO](https://data.cso.ie/) website, with help of [**pyjstat**](https://pypi.org/project/pyjstat/) library.

The dataset was processed to ensure consistency for the period from 2005 to 2021 by aggregating date values across all statistical labels and handling missing data. It consists of 6 features, each with 204 observations.

Bar plot and one of [the palettes](https://matplotlib.org/stable/users/explain/colors/colormaps.html#colormaps) for the colour-blind people is used as the most suitable for comparisons ilustration.

A graph of different colored lines

Description automatically generated with medium confidence

**Insights:**

* Injury collisions significantly decreased during the period from 2005 to 2012, followed by a consistent decrease from 2012 to 2021.
* Similar trends were observed for the total number of casualties during identical periods.
* For the last two years, notably lower numbers were observed; however, caution is needed as these figures might be influenced by the COVID-19 pandemic.
* The total number of accidents is strongly correlate with the number of injuries.
* Number of fatalities is significantly low in comparison to total number and would need separate investigation.

The following comparison demonstrates the relationship between the number of casualties and collisions.

A graph of collision and injury collision

Description automatically generated

A graph of injury injuries

Description automatically generated with medium confidence

**Insight:** Both graphs illustrate a decline in collisions and injuries over the past decade.

A graph of falling people

Description automatically generated with medium confidence

**Insight:** The data present a clear downward trend for the years 2005-2012, with two distinct periods of sustained decrease between 2012-2016 and 2017-2021.

The interactive dashboard simplifies and confirms trends shown in the graphs above.

A graph showing the growth of the stock market

Description automatically generatedA graph showing a line

Description automatically generated with medium confidence

A graph showing the growth of a stock market

Description automatically generatedA graph showing injury injuries

Description automatically generated with medium confidence

A graph showing a line of injury collision

Description automatically generated with medium confidence A graph showing a blue line

Description automatically generated

**Insight:** The percentage of injured casualties for each collision has notably decreased since 2009, while the fatality rate has slightly declined as it follows from the below graph.

A graph of injury and injury

Description automatically generated with medium confidence

Based on the initial data exploration, a notable improvement in road safety was identified after 2012. This finding motivated the creation of two maps to further visualize and analyze the estimated fatality rate for each county.

[CSO website](https://data.cso.ie/) was utilized as a data source for the number of fatalities, while [Open Data](https://data.gov.ie/dataset/counties-national-statutory-boundaries-2019) was employed for the geographical boundaries of each county. [Geopandas](https://geopandas.org/en/stable/) package was used [to visualize geospatial data](https://prog.world/visualizing-geojson-data-using-geopandas-and-python/). Reversed [‘viridis’ color map](https://matplotlib.org/stable/gallery/color/colormap_reference.html#reversed-colormaps) was used to have a scale from darkest color for higher numbers to lighter color for lower numbers.

A map of ireland with different colors

Description automatically generated A map of ireland with different colored states

Description automatically generated

While there may be some inconsistencies in the intervals of time, there is a clear and noticeable improvement in road safety numbers post-2012 for many of the counties, particularly in the western regions.

The central region of the country has consistently demonstrated a higher rate of road safety compared to other regions.

**Insight:** Over the past decade, a stable decrease is observed in regarding to the numbers and rate of collisions and casualties, **confirming hypotheses 2, 3, and 4**.

1. Ireland's road safety analysis using time series.

This section aims to assess the effectiveness of government road safety strategies in reducing fatalities. To forecast future trends, statistical methods and Machine Learning algorithms was applied to the available historical time series data.

The analysis involves two key steps:

1. **ARIMA Analysis for Fatal Collisions:** ARIMA used to analyse the 'Fatal Collisions' data to identify trends and potential patterns.
2. **Machine Learning Approach for Injury Collisions:** General-purpose machine learning algorithms utilized to evaluate and predict trends in 'Injury Collisions' data.

The rationale to use time series analysis comes from the fact that the data is structured in time order. Time series data, arranged chronologically, offers insights into trends, patterns, and behaviours over time. This analytical approach includes examining, modelling, and predicting these patterns to gain a better understanding of future collision trends based on historical data.

* 1. Fatal collisions forecast.

The fatal collisions data for forecasting filtered from previously processed CSO dataset.

The visualization of the data pertains to a single variable; therefore, all graphs utilize a single default color.

**The next analysis employs the** Autoregressive Integrated Moving Average (ARIMA) forecasting method**. This algorithm is based on the principle that historical data can inform predictions for future values in the time series.**

**ARIMA Models** are specified by three order parameters: (p, d, q), where:

* p is the order of the AR term
* q is the order of the MA term
* d is the number of differencing required to make the time series stationary

Time series data is characterized by chronological order, displaying patterns like seasonality, trends, and cyclic behaviour. It often shows autocorrelation, with past values influencing future ones. Achieving stationarity in statistical properties is valuable, but noise or randomness may also be present. Understanding these traits is vital for accurate analysis and forecasting.

Examining the data from a time series perspective, it's crucial to visually assess the values. Spikes help identify outliers, while the visible seasonality can influence our choice of model.

The graph shows a general decrease with some fluctuations, and this pattern is consistent even when changing smoothening window. There is no apparent seasonality to the fluctuations.

A graph with blue lines

Description automatically generated

A graph showing the growth of the stock market

Description automatically generated A graph showing the growth of the company's growth

Description automatically generated with medium confidence

A graph showing the growth of a number of years

Description automatically generated A graph showing the growth of a company

Description automatically generated with medium confidence

On a histogram there is a confirmation of the decreasing trend with a left-skewing graph, followed by KDE.

A graph of falling graph

Description automatically generated with medium confidence

Ensuring stationarity in time series data is essential for building robust models and making accurate predictions. Stationarity in a time series context refers to statistical properties remaining constant over time and can be checked by using the **Augmented Dickey Fuller test** (ADF Test), from the stats package. In case the series is non-stationary, differencing is needed.

The null hypothesis (H0) of the ADF test is that the time series is non-stationary.

A screenshot of a computer code

Description automatically generated

Since p-value (0.3) is greater than the significance level (0.05), our H0 is accepted and thus continuing with differences.

Differencing aims to achieve stationarity in a time series, yet caution is required to avoid excessive differencing. An over-differenced series can impact the model parameters. Optimal differencing involves attaining a near-stationary series hovering around a specific mean, with the Autocorrelation Function (ACF) plot approaching zero.

Determining the correct order of differencing involves several considerations:

* If autocorrelations remain positive across multiple lags (e.g., 10 or more), further differencing is necessary.
* Conversely, an excessively negative autocorrelation at lag 1 suggests potential over-differencing.
* When uncertainty arises between two differing orders, opting for the order resulting in the least standard deviation in the differenced series is advisable.

A group of graphs showing different types of data

Description automatically generated

It appears that the time series achieves stationarity after undergoing one order of differencing (d=1).

The next step involves determining the number of AR terms. Any autocorrelation observed in a stationarized series can be corrected by incorporating an adequate number of AR terms. Initially, the order of AR terms is set to match the number of lags that surpass the significance threshold in the Partial Autocorrelation Function (PACF) plot.

A graph of a graph

Description automatically generated with medium confidence

PACF reveals significance in the initial three lags, consistently below the threshold, suggesting 'p' can be fixed at 4. (**p=4**)

The ACF plot identifies the required number of MA terms ('q'), which represent lagged forecast errors, to eliminate autocorrelation in the stationary series.

A graph of a graph

Description automatically generated with medium confidence

Given the second lag falling below the significance line, 'q' will be set to 1 (q=1).

* 1. Auto Arima Forecasting

ARIMA model utilized for a time series dataset that has been differenced at least once to achieve stationarity and combines AR and MA terms to analyze and forecast the series.

The [pmdarima package](https://pypi.org/project/pmdarima/) provides auto\_arima() function which is used to automate the process of ARIMA forecasting, it uses a stepwise approach to search multiple combinations of p,d,q parameters and chooses the best model that has the least AIC.

A screenshot of a computer

Description automatically generated

The residual plots review:

A group of graphs and diagrams

Description automatically generated

**Standardized residual**: The residual errors seem to fluctuate around a mean of zero and have a uniform variance.

**Histogram**: The density plot suggests normal distribution.

**Theoretical Quantiles**: Mostly the dots fall perfectly in line with the red line. Any significant deviations would imply the distribution is skewed.

**Correlogram**: The Correlogram shows the residual errors are autocorrelated.

A graph showing a blue line

Description automatically generated

**Insights:**

* In the final forecast graph, a noticeable declining trend in the predicted values for the future is evident.
* This observation suggests that the **fifth hypothesis is accepted for Fatalities**.
  1. Injury Collisions prediction

In this section, machine learning algorithms utilized to analyze and forecast another crucial indicator of road safety: **injury collisions number**.

Various methods like decision trees, random forests, and others are crafted to manage temporal patterns, trends, and dependencies inherent in sequential data. These techniques are tailored to interpret time-based information, enabling the identification of patterns, and making predictions in time series datasets.

Filtering applied on the previously explored data frame as a first step.

A graph showing the number of injuries

Description automatically generated

**The presence of high standard deviation justifies the use of reshaping from the beginning of the analysis. 70/30 proportion is used for model splitting to train and test.**

A graph with numbers and lines

Description automatically generated

A graph of different colored lines

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

A graph of different colored lines

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generated with medium confidence

**The bar plot accumulates accuracy scores for different models. A zero-value inserted for cases where accuracy results in a negative number.**

A graph with blue and white bars

Description automatically generated

**Insight: Most of the applied models and data manipulation approaches yield poor performance. This could be due to their simplicity, poor training data, or the inherently unpredictable nature of the outcome. Simple Ridge and Random Forest give a negative R-squared prediction score. This means that the model is doing worse than simply predicting the average for every data point.**

* 1. **Laggs on the time series**

The time series data consistently displays a downward trend, indicating the potential benefit of incorporating lagged values to enhance the performance of machine learning algorithms. Therefore, 12 months of lagged values were constructed.

A screenshot of a computer

Description automatically generated

A graph with text on it

Description automatically generated

“Gradient Boosted Tree” model has the highest score: 0.32, however it’s still too low to be accepted for future prediction.

**Insights:**

* Various models applied to predict Injury Collision numbers in time series data showed low performance level.
* Incorporating lagged variables improved prediction scores across models, highlighting the significance of temporal dependencies in enhancing model performance.
* The presence of negative scores suggests data issues, emphasizing the need for careful investigation into feature engineering and data quality to address these challenges.
* The ARIMA model forecasted a downtrend for Fatalities in the upcoming years, aligning with the government's strategy for the next decade.
* The ability of statistical models like ARIMA to align with governmental strategies and forecast trends consistent with long-term policy objectives highlights their potential to provide valuable insights for decision-makers.
* **The fifth hypothesis is confirmed solely for Fatalities. Prediction of Injury Collision numbers remains elusive.Top of Form**

1. Ireland’s road safety in comparison with other countries.

In this section, a comparative analysis conducted between Ireland and other European countries concerning key road safety indicators.

Three crucial factors were compared to justify the selection of the countries: the number of fatalities, population density, and accessibility of public transport, all within the most recent data available on [Eurostat](https://ec.europa.eu/eurostat/web/main/data/database).

The countries that appear most often on the comparison list are Denmark, Iceland, Norway, and Greece.

A graph of different sizes and colors

Description automatically generated with medium confidence

The use of data visualization techniques helped during the final decision phase in excluding Greece due to a significant variance in fatalities number and Iceland due to disparities across all three parameters considered.

Statistical road safety data, officially accessible to the public, has been obtained and downloaded for [Ireland](https://data.cso.ie/), [Denmark](https://www.dst.dk/en/Statistik/emner/transport) and [Norway](https://www.ssb.no/en/statbank/list/vtu).

The availability and consistency of data across official statistical websites are vary, making it challenging to compare countries on key indicators. However, it’s possible to compare average collision numbers between Ireland and Denmark, and average casualty numbers between Ireland and Norway for time frames: 2005-2021 and 2021 only.

1. Average collision numbers comparison.

**The aggregated data set was evaluated to ensure its suitability for the ANOVA analysis with already available independence in the collected data.**

A graph of a line with a red line

Description automatically generated A graph with a red line

Description automatically generated

A screenshot of a computer code

Description automatically generated

**The datasets for both countries were checked to see if they were normally distributed, which is a requirement for using the ANOVA test. Although they looked like they were normally distributed at first (the points on the plot closely follow a straight line), the Shapiro test showed that they were not. This means that the ANOVA test cannot be used, and a different approach needs to be applied.**

A non-parametric version of ANOVA called **U-Mann Whitman used as it doesn’t** have to assume homogeneity and normal distribution for the original variable.

A screenshot of a computer

Description automatically generated

**Insight: Both test results indicate that Ireland and Denmark can’t be compared in the average number of collisions.**

1. Average Casualties Number comparison

**The same approach applies for average number of casualties comparison.**

A graph of a number of people

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

Null hypothesis for data normality is rejected based on the Shapiro-test results.

As per performed U-test:

* The average number of total casualties is the same for the whole period as well as for 2021.
* The average number of fatalities for 2021 is also comparable.
* All other means are different.

**Insights:**

* Considering that our null hypothesis is accepted for 2 out of 3 indicators in 2021, as opposed to only 1 out of 3 for the entire period from 2005 to 2021, it leads to the conclusion that Ireland is increasingly aligning with the average level of Norway regarding road safety.

1. Sentiment analysis of road safety.

Sentiment analysis is a valuable tool across various domains, providing insights into opinions, emotions, and attitudes expressed in text data, thereby aiding decision-making processes, and improving customer experiences.

* The data collection phase involves extracting comments about road safety in Ireland from the Reddit social network, widely used within Ireland. [Reddit API Wrapper](https://praw.readthedocs.io/en/latest/tutorials/comments.html) used to extract posts for Subreddit 'Ireland' and then comments with the words 'road safety' to ensure that the results are relevant.
* Data cleaning step involves removing errors, inconsistencies, and irrelevant information from the data, to ensure its reliability.
* Polarity scoring involves [assigning a numerical value to the sentiment](https://pypi.org/project/vaderSentiment/) of each piece of text, typically ranging from -1 (very negative) to 1 (very positive), to identify overall trends and patterns.

A graph of different colored squares

Description automatically generated

The bar chart illustrates sentiment polarity, revealing a slight negative bias.

**Insights:**

* The topic sparks strong emotions and low neutrality, indicating its importance to the public.
* The **hypothesis suggesting a positive favorability toward road safety has been rejected.**

The Bag of Words Vectorization method used for sentiment analysis shows a low accuracy rate of 77%, suggesting that the model is unlikely to yield accurate predictions.

A [word cloud](https://pypi.org/project/wordcloud/) was generated as a straightforward and intuitive method to [visually represent textual information](https://python-charts.com/ranking/wordcloud-matplotlib/). Word Cloud insights are limited, but the analysis suggests mixed emotions.

A close-up of words

Description automatically generated

1. Challenges:

The main challenge of this work is a data collection phase as:

1. **Some of the discovered data lacks sufficient detail.**
2. **Various countries store data with varying levels of detail, posing challenges in finding comparable indicators for later analysis.**
3. **Acquiring essential data may necessitate using supplementary tools like APIs or complicate the collection process, often entailing additional steps and potential monetary expenses.**
4. ****References:****
5. pip.pypa.io. (n.d.). *pip freeze - pip documentation v23.3.2*. [online] Available at: https://pip.pypa.io/en/stable/cli/pip\_freeze/ [Accessed 6 Jan. 2024].
6. www.packtpub.com. (n.d.). *Time Series Analysis with Python 3.x [Video] | Packt*. [online] Available at: https://www.packtpub.com/product/time-series-analysis-with-python-3x-video/9781838640590 [Accessed 6 Jan. 2024].
7. ‌Pandas (2018). *Python Data Analysis Library — pandas: Python Data Analysis Library*. [online] Pydata.org. Available at: https://pandas.pydata.org/.
8. ‌Numpy (2009). *NumPy*. [online] Numpy.org. Available at: https://numpy.org/.
9. ‌Matplotlib (2012). *Matplotlib: Python plotting — Matplotlib 3.1.1 documentation*. [online] Matplotlib.org. Available at: <https://matplotlib.org/>.
10. matplotlib.org. (n.d.). *Choosing Colormaps — Matplotlib 3.8.2 documentation*. [online] Available at: https://matplotlib.org/stable/users/explain/colors/colormaps.html#colormaps [Accessed 6 Jan. 2024].
11. scottplot.net. (n.d.). *Colors - ScottPlot 4.1 Cookbook*. [online] Available at: https://scottplot.net/cookbook/4.1/colors/#colorblind-friendly [Accessed 6 Jan. 2024].
12. seaborn (2012). *seaborn: statistical data visualization — seaborn 0.9.0 documentation*. [online] Pydata.org. Available at: https://seaborn.pydata.org/.
13. plotly.com. (n.d.). *Plotly Express*. [online] Available at: https://plotly.com/python/plotly-express/.
14. ‌Zygomatic (2019). *Free online word cloud generator and tag cloud creator*. [online] wordclouds.com. Available at: https://www.wordclouds.com/.
15. Scipy.org. (2019). *Statistical functions (scipy.stats) — SciPy v1.3.3 Reference Guide*. [online] Available at: https://docs.scipy.org/doc/scipy/reference/stats.html.
16. scikit-learn.org. (n.d.). *3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.22.1 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/model\_evaluation.html.
17. PyPI. (n.d.). *pmdarima: Python’s forecast::auto.arima equivalent*. [online] Available at: https://pypi.org/project/pmdarima/.
18. Statsmodels.org. (2023). Available at: https://www.statsmodels.org.
19. docs.python.org. (n.d.). *string — Common string operations — Python 3.9.1 documentation*. [online] Available at: https://docs.python.org/3/library/string.html.
20. NLTK (2009). *Natural Language Toolkit — NLTK 3.4.4 documentation*. [online] Nltk.org. Available at: https://www.nltk.org/.
21. scikit-learn (2019). *scikit-learn: machine learning in Python*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/.
22. Martín, M.E. (n.d.). *pyjstat: Library to handle JSON-stat data in python using pandas DataFrames.* [online] PyPI. Available at: https://pypi.org/project/pyjstat/ [Accessed 6 Jan. 2024].
23. praw.readthedocs.io. (n.d.). *PRAW: The Python Reddit API Wrapper — PRAW 7.4.0 documentation*. [online] Available at: https://praw.readthedocs.io/en/stable/.
24. geopandas.org. (n.d.). *GeoPandas 0.10.2+0.g04d377f.dirty — GeoPandas 0.10.2+0.g04d377f.dirty documentation*. [online] Available at: https://geopandas.org/en/stable/.
25. Tufte, E.R. (1983). *The Visual display of quantitative information*. Cheshire, Conn.: Graphics Press.
26. Eurostat (n.d.). *Database - Eurostat*. [online] ec.europa.eu. Available at: https://ec.europa.eu/eurostat/web/main/data/database.
27. Data.cso.ie. (2020). Available at: https://data.cso.ie/.
28. GitHub. (n.d.). *API Cube PxAPIv1*. [online] Available at: https://github.com/CSOIreland/PxStat/wiki/API-Cube-PxAPIv1 [Accessed 6 Jan. 2024].
29. proger (2022). *Visualizing GeoJSON data using GeoPandas and Python*. [online] Prog.World. Available at: https://prog.world/visualizing-geojson-data-using-geopandas-and-python/ [Accessed 6 Jan. 2024].
30. data.gov.ie. (n.d.). *Counties - National Statutory Boundaries - 2019 - data.gov.ie*. [online] Available at: https://data.gov.ie/dataset/counties-national-statutory-boundaries-2019 [Accessed 6 Jan. 2024].
31. SSB. (n.d.). *Road traffic accidents involving personal injury. Statbank Norway*. [online] Available at: https://www.ssb.no/en/statbank/list/vtu [Accessed 6 Jan. 2024].
32. www.dst.dk. (n.d.). *Transport*. [online] Available at: https://www.dst.dk/en/Statistik/emner/transport [Accessed 6 Jan. 2024].
33. www.youtube.com. (n.d.). *PRAW - Using Python to Scrape Reddit Data!* [online] Available at: https://www.youtube.com/watch?v=Y7BSe7EiBTs [Accessed 6 Jan. 2024].
34. www.youtube.com. (n.d.). *How-to Use The Reddit API in Python*. [online] Available at: https://www.youtube.com/watch?v=FdjVoOf9HN4 [Accessed 14 Dec. 2022].
35. docs.python.org. (n.d.). *string — Common string operations — Python 3.9.1 documentation*. [online] Available at: https://docs.python.org/3/library/string.html.
36. Python (2009). *re — Regular expression operations — Python 3.7.2 documentation*. [online] Python.org. Available at: https://docs.python.org/3/library/re.html.
37. Wong, B. (2011). Points of view: Color blindness. *Nature Methods*, 8(6), pp.441–441. doi:https://doi.org/10.1038/nmeth.1618.
38. PYTHON CHARTS | The definitive Python data visualization site. (2022). *Wordclouds in Python*. [online] Available at: https://python-charts.com/ranking/wordcloud-matplotlib/.