**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Data preparation and visualization, Machine Learning, Statistics and Programming |
| **Assessment Title:** | Irish and british transportation |
| **Lecturer Name:** | David McQuaid, Muhammad Iqbal, Taufique Ahmed, Sam Weiss |
| **Student Full Name:** | Fernando Tupa Leniz |
| **Student Number:** | 2023341 |
| **Assessment Due Date:** | 05/01/2024 |
| **Date of Submission:** | 07/01/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

Transportation is a vital component of any thriving society, playing a pivotal role in fostering economic development, enhancing social connectivity, and promoting overall well-being. In the context of Ireland, understanding and optimizing the efficiency of transportation systems is crucial for the nation's growth. This research project delves into the intricacies of Irish transportation, with a specific focus on the analysis and prediction of bus types using the National Public Transport Access Nodes (NAPTAN) dataset.

The NAPTAN dataset serves as a comprehensive repository of information pertaining to public transportation nodes, offering insights into the network and infrastructure that facilitate the movement of people across Ireland. By leveraging this dataset, the research aims to employ machine learning models to predict the types of buses operating within the Irish transportation system. This predictive analysis involves the utilization of various models, including Naive Bayes and Logistic Regression, to discern patterns and relationships within the data.

To ensure the robustness and applicability of the developed models, a comparative analysis will be conducted by applying the same machine learning methodologies to both Irish and UK transportation datasets. By examining the similarities and differences between the two regions, this research seeks to identify factors that contribute to the variance in bus types and transportation dynamics. Such a comparative approach can enhance our understanding of the unique characteristics of Irish transportation and contribute to the development of targeted strategies for improvement.

Furthermore, beyond the quantitative analysis of transportation data, this research project also incorporates a qualitative dimension through sentiment analysis. By exploring public sentiments related to Irish transportation, the study aims to gauge the overall satisfaction and perception of the transportation system among the populace. This sentiment analysis will draw on textual data from various sources, such as reddit, providing a holistic view of the public's experiences and opinions regarding transportation in Ireland.

In conclusion, this research project aspires to contribute to the optimization of Irish transportation by combining NAPTAN analysis, machine learning prediction models, and sentiment analysis. Through a comprehensive examination of the transportation landscape, this study seeks to provide valuable insights that can inform policy-making, infrastructure development, and the enhancement of the overall transportation experience for the Irish populace.

**Data preparation and visualisation part:**

**1)**

Obtaining the NaPTAN dataset for both Ireland and the UK presented a challenging yet crucial task for the research due to the requirement of acquiring the data in two different formats. This dual-format necessity added a layer of complexity to the data acquisition process, as compatibility issues and differing standards had to be addressed to ensure a seamless integration of information from both regions.

The challenge stemmed from the fact that the NaPTAN dataset for the UK was available under the Open Government Licence (OGL), as outlined in the official documentation provided by the National Archives at <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> . This license, while fostering openness and collaboration, required careful adherence to its terms and conditions, including proper attribution and compliance with the specified usage guidelines.

Simultaneously, acquiring the NaPTAN dataset for Ireland posed a different challenge, as the dataset was made available under the Creative Commons Attribution 4.0 International License. This license, was found at <https://creativecommons.org/licenses/by/4.0> allowed for the sharing and adaptation of the data with the requirement of providing appropriate attribution to the data source.

Despite the challenges associated with navigating two distinct licensing frameworks, our research successfully obtained the necessary permissions to use those datasets for both Ireland and the UK. This achievement was vital in ensuring the ethical and legal use of the data, as well as facilitating a comprehensive and unified analysis that encompassed both regions.

The dual-licensing approach not only underscores the complexities of dealing with geographically diverse datasets but also highlights the commitment to upholding the principles of open data and responsible research practices. By navigating these challenges and securing the appropriate licenses, our research endeavors to contribute meaningfully to the understanding and enhancement of public transportation systems in countries.

**2)**

Understanding the NaPTAN datasets for Ireland and the UK is crucial for meaningful analysis and decision-making in data-related tasks. The dimensions of the datasets, (16491, 24) for Ireland and (437106, 31) for the UK, immediately convey the scale and complexity of the information at hand.

To begin the exploration, employing the `head()` method allows a quick glimpse of the initial rows in each dataset. This aids in understanding the structure of the data and the types of information contained in the columns. The concise summary provided by the `info()` method is invaluable at this stage, offering insights into data types, non-null counts, and the potential presence of missing values.

Moving forward, the application of `value\_counts()` becomes essential, particularly for categorical variables. This method enables the examination of the distribution of unique values, shedding light on the diversity and prevalence of categories within specific columns.

The determination of missing values is a critical aspect of data quality assessment. Utilizing the `.isnull().sum()` method allows for a systematic count of missing values in each column. Recognizing the extent and locations of missing data is imperative for decision-making regarding imputation, removal, or other handling strategies.

In addition to these methods, the visual exploration through a heatmap is valuable for intuitively grasping the distribution of missing values. Empty cells in the heatmap serve as visual cues, indicating the presence and patterns of missing data. A heatmap is particularly effective in revealing if certain columns or rows exhibit higher concentrations of missing values, guiding further investigation.

Furthermore, employing a box plot aids in understanding the distribution of numerical data, identifying potential outliers, and showcasing statistical measures. In the context of missing values, the box plot can reveal the impact of these gaps on the overall data distribution. Empty values may manifest as gaps or outliers, influencing the interpretation of central tendencies and variability.

**3)**

In the Irish dataset, a preliminary examination revealed the presence of empty values across several columns, indicating potential data gaps that required careful consideration and handling. The report provided details on the count of missing values for each relevant column, offering insights into the nature and scope of the missing data. Addressing these empty values was crucial for ensuring data completeness and reliability in subsequent analyses.

Among the columns with missing values, several were notable for the extent of their data gaps. For instance, columns like "CommonNameGA," "ShortCommonName," "ShortCommonNameGA," "BusStopType," "TimingStatus," "CompassPoint," "Street," "PlateCode," and "StopAreaRef" exhibited varying degrees of missing information. The challenge presented by these gaps necessitated a thoughtful approach using appropriate methods within the Jupyter notebook environment.

Similar to the procedures applied to the UK dataset, the `fillna()` method was employed to handle missing values in the Irish dataset. Columns such as "CommonNameGA," "ShortCommonName," "ShortCommonNameGA," "BusStopType," "TimingStatus," and "CompassPoint" benefited from imputation techniques, filling empty cells with statistically meaningful substitutes like the median, mode, or other relevant values. This process aimed to maintain the integrity of the existing data distribution while mitigating biases in subsequent analyses.

In cases where entire columns were deemed irrelevant or the missing values were extensive, the `drop()` method was utilized to remove those columns directly. For example, the columns "DefaultWaitTime," "Notes," "NotesLang," and "GrandParentLocalityName" in the UK dataset were directly dropped, streamlining the dataset and facilitating a more focused analysis.

Visualizations within the Jupyter notebook, such as heatmaps or bar charts, complemented these methods by providing a visual understanding of the distribution of missing values. These visual representations aided in identifying patterns, clusters, or correlations among missing values, informing the choice of imputation strategies and ensuring a comprehensive approach to data cleansing.

**4)**

In developing an interactive dashboard tailored to modern transport planning, the choice of Dash, Plotly Express, and associated components reflects a commitment to Tuft's principles of clarity, simplicity, and effective communication. Dash, as a versatile Python framework, enables the seamless integration of Python code with HTML and CSS, simplifying the development process while maintaining a clear and cohesive user interface.

In terms of model comparison, the inclusion of accuracy as a metric serves as a common and straightforward measure to evaluate the overall performance of different models. However, the decision to incorporate a confusion matrix provides a more detailed breakdown of model performance, aligning with Tuft's principle of effective communication by offering insights into model behaviour across specific categories (true positives, true negatives, false positives, and false negatives).

The specific comparison between logistic regression and a naive classifier is motivated by the need to assess the effectiveness of logistic regression against a baseline. This evidence-based approach aligns with Tuft's principles, emphasizing the importance of informed decision-making in the context of transport planning. By providing a detailed view of model performance through the confusion matrix, the dashboard aims to empower users with valuable insights for making strategic decisions in modern transport planning scenarios.

**Machine Learning part:**

**1)**

For the Irish dataset, the choice of logistic regression and a naive classifier was influenced by the nature of the classification task. Logistic regression is a versatile algorithm known for its simplicity and interpretability, making it suitable for predicting categorical outcomes such as bus stop types. The naive classifier, often used as a baseline model, provides a simplistic approach based on class frequencies, allowing for a comparative analysis of model performance.

In the case of the UK dataset NaPTAN and the prediction of LocalityCentre, the decision to create a class for machine learning models was driven by the need for a systematic and efficient approach to model development. The class incorporated classification prediction methods, providing a standardized framework for building, training, and evaluating models. The inclusion of GridSearchCV and random search for hyperparameters ensured an exhaustive search for the best model configurations, optimizing performance and adhering to best practices in model tuning.

In the domain of sentiment analysis, the choice of MultinomialNB, a variant of Naive Bayes, was driven by its effectiveness in handling text data. The algorithm is well-suited for classification tasks involving textual features, making it an appropriate choice for sentiment analysis. The decision to reuse the logistic regression class highlights the modular and reusable nature of machine learning code. By adapting the logistic regression class for sentiment analysis, it streamlines the development process and promotes code efficiency.

In summary, the choices of logistic regression, naive classifier, MultinomialNB, and the creation of a machine learning class with GridSearchCV and random search were guided by the specific requirements of each task. These decisions reflect a thoughtful consideration of algorithm suitability, model development practices, and the need for efficient and reusable code in the context of predicting bus stop types and LocalityCentre, as well as sentiment analysis.

**3)**

The decision to use Support Vector Machine (SVM) models, cross-validation, and dimensionality reduction techniques in the UK and Irish datasets reflects thoughtful considerations for optimizing model performance and handling high-dimensional data.

In the UK dataset, the use of SVM models is likely due to their effectiveness in handling complex and non-linear relationships in data. SVMs are powerful classifiers suitable for scenarios where the decision boundary is not easily linearly separable. Cross-validation is employed to assess the generalization performance of the SVM model by splitting the dataset into training and testing sets multiple times. This helps mitigate the risk of overfitting and ensures a robust evaluation of the model's performance on different data subsets.

Linear Discriminant Analysis (LDA) is applied for dimensionality reduction in the UK dataset. LDA is particularly useful when dealing with classification tasks, as it seeks to maximize the separation between classes. By reducing the dimensionality of the data while preserving class-related information, LDA can enhance the SVM model's ability to capture relevant patterns, potentially leading to improved classification performance.

In the Irish dataset, the choice of SVM models and cross-validation suggests a similar motivation for robust model evaluation and handling non-linear relationships. SVMs can capture intricate decision boundaries, making them suitable for tasks with complex data patterns. Cross-validation aids in estimating how well the SVM model will generalize to unseen data, providing a more reliable performance evaluation.

Principal Component Analysis (PCA) is applied for dimensionality reduction in the Irish dataset. PCA is a widely used technique to reduce the number of features while retaining most of the original data's variance. By applying PCA before training the SVM model, the computational burden is reduced, and the risk of overfitting is mitigated. This allows the SVM model to focus on the most informative features, potentially improving its generalization performance.

**4)**

Comparing metrics such as Accuracy, Precision, Recall, F1 Score, Random Accuracy, and Grid Accuracy between logistic regression and a naive classifier offers a thorough evaluation of their performance in classification tasks. Accuracy, as a foundational metric, provides an overarching view of correctness, though it may fall short in scenarios with imbalanced datasets. Precision and Recall offer insights into the accuracy of positive predictions and the model's ability to capture all relevant positive instances, respectively. The F1 Score, a harmonic mean of precision and recall, strikes a balance between false positives and false negatives.

Comparing these metrics between logistic regression and a naive classifier is essential for discerning their respective strengths and weaknesses. Logistic regression, with its capacity to capture intricate relationships, is expected to outperform the naive classifier. However, the naive classifier serves as a baseline for comparison, revealing how well the logistic regression model performs beyond random chance. This comprehensive assessment facilitates informed decision-making, guiding the selection of the most suitable model based on the specific requirements of the classification task at hand.

In general it worked better in the irish dataset using the classes about the machine learning model.

**Statistics part:**

**1)**

Developing a function like "custom\_descriptive\_statistics" that encompasses key descriptive statistics such as "Count," "Mean," "Variance," "Standard Deviation," "Median," "Min," "Max," and "Mode" holds paramount importance in the realm of data analysis. These statistics collectively offer a nuanced and holistic understanding of the dataset's characteristics, contributing to a comprehensive summary that aids in informed decision-making.

By incorporating "Count" into the function, it ensures a fundamental grasp of the dataset's size, providing insights into potential missing or incomplete data. This foundational statistic is crucial for establishing the dataset's integrity and completeness, forming the basis for robust analyses.

The inclusion of "Mean" serves as a measure of central tendency, offering an average value that represents the dataset's central theme. This statistic is invaluable for understanding the typical magnitude of the data, providing a reference point for further analysis and interpretation.

"Variance" and "Standard Deviation" contribute to the function by quantifying the dispersion or spread of the data points. These statistics provide insights into the variability of values around the mean, aiding in the identification of datasets with more tightly clustered or widely dispersed values.

The incorporation of "Median" is essential for capturing the central point of the dataset without being influenced by extreme values. It complements the mean by offering an alternative measure of central tendency, particularly effective in the presence of skewed distributions.

"Min" and "Max" offer a range perspective, delineating the minimum and maximum values in the dataset. These statistics are instrumental in identifying outliers or extreme data points that could significantly impact the analysis, providing a basis for outlier detection and understanding the data's boundaries.

Lastly, the inclusion of "Mode" provides insights into the most frequently occurring value(s) in the dataset. Understanding the mode enhances the comprehension of data distribution, particularly in cases where certain values appear more frequently than others.

The importance of this function lies not only in its ability to calculate these descriptive statistics comprehensively but also in its potential for reuse across different datasets or specific columns within datasets. The function's versatility facilitates efficient and consistent data analysis, contributing to code maintainability, readability, and adaptability as analytical requirements evolve over time.

**2)**

When applying the chi-square test for independence, the decision-making process involves comparing the obtained p-value to a predetermined significance level, commonly denoted as alpha (α). In hypothesis testing, alpha is the threshold used to determine the level of significance, typically set at 0.05. The significance level represents the probability of rejecting the null hypothesis when it is actually true.

In the context of comparing AtcoCode and BusStopType, the null hypothesis (H0) posits that there is no effect or association between the two categorical variables. The alternative hypothesis (H1), on the other hand, suggests the presence of an effect or association. The decision to accept or reject the null hypothesis is contingent upon the p-value resulting from the chi-square test.

If the calculated p-value is greater than alpha, it indicates that there is insufficient evidence to reject the null hypothesis. In this scenario, where the p-value exceeds the predetermined significance level (α > p-value), the analyst accepts H0. This decision implies that the observed distribution of AtcoCode and BusStopType is consistent with what would be expected under the assumption of independence. The deviation observed is not statistically significant, and any apparent association can be attributed to random chance.

The justification for accepting H0 based on a higher p-value than alpha underscores the principle of caution in making claims about associations. By adhering to a predetermined significance level, analysts mitigate the risk of making spurious claims of association when the observed data patterns may occur by chance. It ensures a rigorous and conservative approach to hypothesis testing, enhancing the reliability and validity of statistical conclusions.

In summary, accepting H0 when the p-value is greater than alpha signifies a cautious interpretation, acknowledging that the observed data does not provide sufficient evidence to reject the null hypothesis. This decision-making process aligns with the principles of statistical hypothesis testing, promoting robust and reliable conclusions in the assessment of the association between AtcoCode and BusStopType.

**3)**

**ANOVA:**

ANOVA is suitable when it is necessary to compare means among multiple groups and assess whether there are significant differences.

The F-statistic is a measure of the difference in variability between groups relative to the variability within groups. A high F-statistic suggests that the means of the groups are significantly different. In your case, the F-statistic of 53589 is very high, indicating substantial differences among the groups.

The p-value is the probability of observing a test statistic as extreme as the one computed from the sample data, assuming that the null hypothesis is true. A p-value of 0 means that the observed differences are highly unlikely to have occurred by random chance alone. In practice, a p-value of 0 is often rounded due to computational limitations.

**t-test:**

The t-test is a statistical method used to compare the means of two groups and assess whether the observed differences are statistically significant. In this case it is better to use “Independent Samples t-Test”, because it is used when comparing the means of two independent groups.

The t-statistic measures the difference between the means of the two groups relative to the variability within the groups. In this case, a t-statistic of 231.49458572732723 indicates a substantial difference in the means of the groups being compared.

The p-value is the probability of observing a t-statistic as extreme as the one computed from the sample data, assuming that the null hypothesis is true. A p-value of 0 suggests that the observed difference is so large that it is extremely unlikely to have occurred by random chance alone.

**Mann-Whitney U test:**

Since the p-value is much smaller than the typical significance level. It has to reject the null hypothesis.

The Mann-Whitney U test is used in situations where the assumptions of parametric tests (such as the t-test) cannot be met or when it is necessary to deal with non-normally distributed data.

The Mann-Whitney U statistic is a measure of the rank-based difference between two groups. In this case, the large U statistic indicates a substantial difference in the distributions of the two groups (NaPTAN Ireland and the UK).

The p-value is extremely small (close to zero), indicating strong evidence against the null hypothesis. In other words, it suggests that there is a significant difference in the distribution of active and inactive bus stops.

**Two-Proportion Z-test:**

The Two-Proportion Z-test is used to compare proportions between two independent groups. There is strong evidence to suggest that there is a significant difference in the proportion of active bus stops.

The Z-statistic measures how many standard deviations an observed proportion is from the expected proportion under the null hypothesis. In this case, a Z-statistic of 23.97 is quite large, indicating a substantial difference between the proportions .

The P-value is extremely small (close to zero), indicating strong evidence against the null hypothesis. In the context of this test, the P-value represents the probability of observing a difference as extreme as the one in the sample, assuming the null hypothesis is true. A very small P-value suggests that the observed difference in proportions is unlikely to have occurred by random chance alone.

**Sign test:**

This test is important because it provides a robust and distribution-free method for assessing the existence of systematic differences in paired ordinal data. It's a versatile tool, especially when dealing with small samples or data that may not conform to parametric assumptions.

Since the p-value is significantly below the chosen significance level (alpha), the decision is reject the null hypothesis. This implies that there is a significant and systematic difference in bus stop status.

In practical terms, it means that the observed differences in bus stop status are not likely due to random chance, and there is evidence to suggest a systematic difference between the two datasets.

**4)**

The substantial differences observed between the NaPTAN datasets in Ireland and the UK are multifaceted and require a comprehensive analytical approach encompassing both statistical metrics and machine learning models. By examining statistical metrics such as mean, variance, standard deviation, median, and mode, insights into the central tendencies and distributions of the datasets can be gained. Disparities in these metrics may highlight variations in data patterns, potential outliers, and the overall structural differences between the Irish and UK datasets.

The application of hypothesis tests, such as the chi-square test for independence, further contributes to understanding the relationships between categorical variables within each dataset. Testing hypotheses about associations provides valuable insights into specific factors that contribute to dissimilarities between the two datasets.

In the realm of machine learning, logistic regression serves as a powerful tool for modeling the relationship between predictor variables and a binary target variable. When applied to predict different targets in each dataset, logistic regression captures dataset-specific patterns, unveiling unique associations and relationships specific to the Irish or UK NaPTAN dataset.

Interpreting these differences requires a contextual understanding of the transportation systems, data collection methodologies, and regional variations. Geographical distinctions, cultural nuances, and regulatory frameworks may contribute to the observed dissimilarities, influencing the effectiveness of predictive models.

**Programming part:**

Executing the line `custom\_descriptive\_statistics('test function')` proves to be a valuable testing strategy, particularly when the function is designed to accommodate either a list or a DataFrame and incorporates a try-except block to manage potential errors. This testing approach serves multiple purposes that contribute to the overall reliability and robustness of the function.

By supplying a non-list or non-DataFrame input, in this case, the string 'test function', the test scenario simulates unexpected or unsupported data types, allowing for the examination of how the function responds to such cases. The try-except block within the function is put to the test, verifying its effectiveness in catching and gracefully handling errors. This proactive error handling is crucial in preventing the function from breaking or producing undesirable outcomes when faced with unanticipated inputs.

Testing with unconventional inputs, such as strings, contributes to the overall robustness of the function. It ensures that the function can gracefully handle unexpected scenarios, maintaining stability and preventing crashes, which is particularly important when dealing with diverse or dynamic datasets in real-world applications.

The try-except block not only enhances the reliability of the function by gracefully handling errors but also supports a user-friendly experience. Instead of allowing the function to crash or generating obscure error messages, the try-except block enables the function to respond with clear and informative messages, guiding users on correct usage and helping to troubleshoot issues.

Testing with intentional errors also aids in the debugging process, allowing developers to trace and identify specific issues related to unexpected data types. This iterative debugging approach facilitates refining the function and addressing any unforeseen challenges in its implementation.