# Group ID - MSc in Data Analytics

Author: S. Clifford

e-mail: [sbs23100@student.cct.ie](mailto:sbs23100@student.cct.ie)

Student ID: 23100

Contents

[Group ID - MSc in Data Analytics 1](#_Toc135085795)

[1 Abstract 3](#_Toc135085796)

[2 Introduction 3](#_Toc135085797)

[3 Data Preparation and Visualization 3](#_Toc135085798)

[3.1 Acquiring Raw Data 3](#_Toc135085799)

[3.2 Exploratory Data Analysis 3](#_Toc135085800)

[3.3 structure and enrich your data 3](#_Toc135085801)

[3.4 interactive dashboard 3](#_Toc135085802)

[4 Statistics 4](#_Toc135085803)

[4.1 summarise the dataset 4](#_Toc135085804)

[4.2 inferential statistics 4](#_Toc135085805)

[4.3 parametric and non-parametric inferential statistical techniques 4](#_Toc135085806)

[4.4 challenges 4](#_Toc135085807)

[5 Machine learning for data analytics 4](#_Toc135085808)

[5.1 Describe the rationale and justification for the choice of machine learning models 4](#_Toc135085809)

[5.2 Perform a sentimental analysis for an appropriate construction topic 5](#_Toc135085810)

[5.3 learning models 5](#_Toc135085811)

[5.4 similarities and contrast of the Machine Learning modelling 5](#_Toc135085812)

[6 Programming 5](#_Toc135085813)

[6.1 Programming 5](#_Toc135085814)

[6.2 Data structures 6](#_Toc135085815)

[6.3 Testing & Optimisation 6](#_Toc135085816)

[6.4 Data manipulation 6](#_Toc135085817)

[7 Conclusion 6](#_Toc135085818)

[8 Next Steps 6](#_Toc135085819)

[9 References 7](#_Toc135085820)

# Abstract

# Introduction

# Data Preparation and Visualization

## Acquiring Raw Data

Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data. [0-15]

## Exploratory Data Analysis

Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.[0-25]

## structure and enrich your data

Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.[0-30]

## interactive dashboard

Modern construction has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern farmers, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development. Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).[0-30]

# Statistics

## summarise the dataset

Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. [0-20 ]

## inferential statistics

Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international home building, you could find a confidence interval for the population proportion of yearly apartment builds out of all home builds). [0-20 ]

## parametric and non-parametric inferential statistical techniques

Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. [0-40]

## challenges

Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. [0-20]

# Machine learning for data analytics

## Describe the rationale and justification for the choice of machine learning models

Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

[0 - 30]

## Perform a sentimental analysis for an appropriate construction topic

Collect and develop a dataset based on the construction topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate construction topic (e.g., house price, availability of labour etc…) for producers and consumers point of view in Ireland.

[0 - 25]

## learning models

You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

[0 - 30]

## similarities and contrast of the Machine Learning modelling

A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

[0 - 15]

# Programming

## Programming

Programming: The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. [0-20]

## Data structures

Data structures: You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format. [0-20]

## Testing & Optimisation

Testing & Optimisation: You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas. [0-20]

## Data manipulation

Data manipulation: For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques. [0-20]

# Conclusion

# Next Steps

# References

Birchwood University, 2023. *birchwoodu.org.* [Online]   
Available at: https://www.birchwoodu.org/top-programming-languages-for-data-scientists/  
[Accessed 24 March 2023].

Clifford, S., 2023. *CPI.ipynb,* Killarney: s.n.

Clifford, S., 2023. *house\_registrations\_by\_area.ipynb,* Killarney: s.n.

Clifford, S., 2023. *MIR.ipynb,* Killarney: s.n.

Clifford, S., 2023. *population.ipynb,* Killarney: s.n.

Coursera , 2023. *Python or R for Data Analysis: Which Should I Learn?.* [Online]   
Available at: https://www.coursera.org/articles/python-or-r-for-data-analysis  
[Accessed 14 April 2023].

Data Science Process Alliance, n.d. *What is CRISP DM.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-2/  
[Accessed 13 April 2023].

Data Science Process Alliance, n.d. *www.datascience-pm.com.* [Online]   
Available at: https://www.datascience-pm.com/semma/  
[Accessed 23 April 2023].

DATA.GOV.IE, 2021. *HSA10 - New House Registrations.* [Online]   
Available at: https://data.gov.ie/dataset/hsa10-new-house-registrations?package\_type=dataset  
[Accessed 02 March 2023].

datacamp.com, 2022. *datacamp.com.* [Online]   
Available at: https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022  
[Accessed 24 March 2023].

datadriveninvestor.com, n.d. *datadriveninvestor.com.* [Online]   
Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb  
[Accessed 13 April 2023].

Dhingra, D., n.d. *How to transform features into Normal/Gaussian Distribution.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/05/how-to-transform-features-into-normal-gaussian-distribution/  
[Accessed 1 April 2023].

Dissanayake, A., 2021. *Python vs R for Data Science.* [Online]   
Available at: https://medium.com/octave-john-keells-group/python-vs-r-for-data-science-320e167ffe90  
[Accessed 14 April 2023].

Iantorno, M., 2023. *Lecturer.* Dublin: s.n.

IBM.com, n.d. *CRISP-DM Help Overview.* [Online]   
Available at: https://www.ibm.com/docs/it/spss-modeler/saas?topic=dm-crisp-help-overview  
[Accessed 13 April 2023].

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. An introduction to statistical learning. In: *An introduction to statistical learning.* New York: Springer.

jetbrains.com, 2023. *The State of Developer Ecosystem 2022.* [Online]   
Available at: https://www.jetbrains.com/lp/devecosystem-2022/?source=google&medium=cpc&campaign=19867226536&term=best%20programming%20languages&content=651996376797&gclid=Cj0KCQjwlPWgBhDHARIsAH2xdNdnHHpx-c5mgllSOCY5\_IjYN6Mnu6j8Q4UmWJPl38747STLRfDTPQsaAq5rEALw\_wcB  
[Accessed 24 March 2023].

Khete, T., 2022. *R vs Python for Data Science and visualization: The language debate.* [Online]   
Available at: https://medium.com/@tusharkhete118/r-vs-python-for-data-science-and-visualization-the-language-debate-1aac453e7e29  
[Accessed 14 April 2023].

Korstanje, J., 2020. *Is Python faster than R?.* [Online]   
Available at: https://towardsdatascience.com/is-python-faster-than-r-db06c5be5ce8  
[Accessed 14 April 2023].

numpy.org, n.d. *numpy.org.* [Online]   
Available at: https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html  
[Accessed 1 April 2023].

Reichenbach , M., 2021. *R vs Python: Which Is Best for Data?.* [Online]   
Available at: https://blog.boot.dev/python/r-vs-python/  
[Accessed 14 April 2023].

sas.com, n.d. *SAS Analytics Software & Solutions.* [Online]   
Available at: http://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm  
[Accessed 13 April 2023].

scikit-learn.org, n.d. *https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html  
[Accessed 8 04 2023].

scipy.org, n.d. *docs.scipy.org.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html  
[Accessed 1 April 2023].

scripy.org, n.d. *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html  
[Accessed 8 April 2023].

seaborn.pydata.org, n.d. *www.seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/generated/seaborn.pairplot.html  
[Accessed 31 March 2023].

seaborn.pydata.org, n.d. *www.seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/generated/seaborn.regplot.html  
[Accessed 31 March 2023].