

**MSc data analytics ca1**



ireoluwa olajitan

SBA23029@STUDENT.CCT.IE

Contents

[ABSTRACT 2](#_Toc148809910)

[1. INTRODUCTION 3](#_Toc148809911)

# ABSTRACT

# INTRODUCTION

Over the course of this report, there were 4 focus areas in which the selected data was analysed- Data Preparation and Visualization, Machine Learning, Statistics and Programming. The data used in this report was extracted from the Central Statistics Office (CSO) and is titled ‘Estimated Population (Persons in April)’.

# DATA PREPARATION AND VISUALIZATION

You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained [0-20]

You must also rationalise justify and detail all the methods used to prepare the data for ML. **[0-30]**

Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-20]**

All design and implementation of your visualizations must be justified and detailed in full., making reference to Tufts Principles **[0-30]**

# MACHINE LEARNING

Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. **[0 - 20]**

Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**

Show the results of two or more ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**

Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. **[0 - 20]**

# STATISTICS

You need to analyse the chosen dataset using statistical logic and statistical techniques. Note: ALL Statistical work MUST be carried out using Python.

You are required to:

Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]

Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]

Use Normal distribution to explain or identify some information about your dataset. [0-20]

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

# 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. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

# RESULTS AND CONCLUSION

# REFERENCES