A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name: John Burns* |  |
| *Student Number: SBS24019* |  |
| *Module Title: Higher Diploma Data Analytics - Strategic Thinking* |  |
| *Assessment Title: CA 2 – Capstone Project Proposal* |  |
| *Assessment Due Date: 19th May 2024 23:59* |  |
| *Date of Submission: 19th May 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.

Contents

[Introduction 3](#_Toc167009500)

[Project Management & Planning 3](#_Toc167009501)

[Data Understanding 4](#_Toc167009502)

[Data Preparation 6](#_Toc167009503)

[Machine Learning Implementation 8](#_Toc167009504)

[Summary of Findings 9](#_Toc167009505)

[Ethical Considerations 10](#_Toc167009506)

[Conclusion and Future recommendations 10](#_Toc167009507)

[Github link 11](#_Toc167009508)

[References 12](#_Toc167009509)

[Appendix 1 13](#_Toc167009510)

# Introduction

This report is a progress update on the CA1-Capstone report issued in March and is included as Appendix 1. The purpose of this assignment is to provide an update on the progress in the following respects:

* Performance against the project plan and scope
* Progress in Data Understanding phase of the project
* Any work completed in the Data Preparation stage of the project.
* Exploratory application of Machine Learning regression techniques.
* Results from the Machine Learning techniques applied.
* Conclusion including future directions for the project.

The length of this report is approximately 2,200 words.

# Project Management & Planning

The current status of the project is shown below. The following has been accomplished:



* An initial scoping of the data and some exploratory analysis has been completed and is detailed below.
* The problem definition has not changed.
* The project objectives and scope have not changed.
* The Problem Understanding phase is completed.
* Progress is in line with the high-level project plan with the Data Understanding component in progress.
* While it was anticipated that there would be some overlap in the Data Understanding and Data Preparation phases of the project, this has started to manifest itself a lot earlier than anticipated. Early elements of Modelling have also occurred.
* Initial exploration of the data has started with some notable key characteristics:
  + Key indexing variable is based on location (country, city, region etc)
  + Some of the data sets contain high volumes of time series data, some updated daily.
  + Some datasets compile location data at different level of aggregation, ranging from country level to city and county level.
  + Differences in the availability of data across locations is apparent. For instance, Covid-19 statistics are available at county level in some U.S. states while it is available only at country level for some African countries.
* To assist in getting an initial overview of the data, a data frame containing variables at a country level from a series of key datasets was compiled. The data necessary for some exploratory regression analyses were drawn from this data frame.

It is anticipated that the project will be completed within its projected timeline.

# Data Understanding

The Google Covid-19 Data Open Data Repository has a wide variety of data from around the world relating to the Covid-19 pandemic at differing levels of aggregation. A key objective was to begin to understand how the datasets related to each other and to develop an approach to provide a comprehensive view of the data. To begin this process a sample of key datasets were downloaded for exploration:

* Demographics.csv captures demographic data
* Epidemology.csv shows daily information including new cases, deaths and cumulative statistics on cases and deaths.
* Economy.csv captures economic data, mostly at a country level
* Oxford-government-response.csv chronicles government’s actions in a time series file to tackle Covid-19. It is a time series file with daily updates on changes to government actions.
* Index.csv is the index of place names used in the datasets. This includes the keys that can link the datasets together. Data can be included at many levels country, regional and city level
* The files above all have the following characteristics follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **# of Records** | **# of Attributes** | **Data Organisation** |
| Demographics.csv | 21,689 | 19 | Demographic details with regional breakdowns |
| Epidemology.csv | 12,525,825 | 10 | Time series – daily updates per country and region on key Covid-19 measures |
| Economy.csv | 404 | 4 | Point in time economic data mostly at a country level |
| Oxford-government-response.csv | 303,969 | 22 | Time series on the status of a range of government initiatives to tackle the pandemic |
| Index.csv | 22,963 | 15 | Breakdown of countries and regions within countries. The creators of the data repository sought to accommodate as much data as was available on a geographic basis. The Index.csv file shows the possible breakdowns on a geographic basis. |

To get an overview and increased understanding of the data, a consolidated data frame with the following characteristics was constructed (see Figure 1):

A diagram of a computer system

Description automatically generated

* Contained all fields from each of the 5 datasets above.
* The data extracted from the source files was at a country level to facilitate an overview.
* The initial focus for a high level of analysis was on the relationships between the following:
* Stringency Index – this is a composite measure reflecting the severity of the government measures taken to combat the pandemic in each country. This field appears in the Oxford-government-response.csv file.
* Cumulative cases of Covid-19 and deaths from Covid-19 for each country are taken from the epidemiology.csv file.
* The Human-development index is a composite measure from the demographics.csv file that measures the average attainment across 3 dimensions of human development UNDP, 2024):
  + a longevity and health,
  + being educated or knowledgeable
  + standard of living.

The methodology employed at the outset was to download each of the files to:

* get the dimensions of each file as in:

A computer screen shot of a number

Description automatically generated

* Display a number of records to ensure the data corresponded to expectations as in: A screenshot of a computer

  Description automatically generated
* Get more detailed information at a field level on each file using .info() as in:

A screenshot of a computer screen

Description automatically generated

* Statistical profiles of the numeric fields in each dataset were gathered using describe():

A screenshot of a computer

Description automatically generated

* When the initial hypotheses were formulated, visualisation of the variables involved were generated before analysis began as in:

A graph of a number

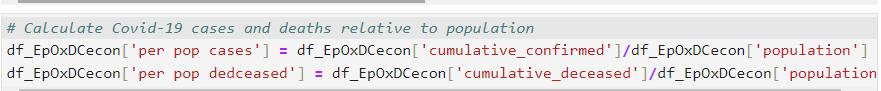
Description automatically generated

# Data Preparation

* The initial phase focused on the construction of the consolidated country-level dataset with following steps:
  + A subset of location keys from the index file was created at the “country” level of aggregation.
  + .the other files were merged to create the Consolidated file using the “country” as the key.
  + For certain cumulative values, the values were taken from the final records generated in the dataset.
  + As each file was merged, records were printed to ensure that the files had merged as expected.

As previously mentioned, exploratory analysis focused on 4 cumulative variables from the Consolidated dataset:

* + Stringency Index
  + Cumulative cases of Covid-19
  + Cumulative deaths from Covid-19
  + The Human-development index
* Cumulative cases of Covid-19 and deaths from Covid-19 are recorded as aggregates in the Epidemology.csv file. However, to be effective indicators of the seriousness of the impact of Covid-19 on a country, their incidence must be viewed in relation to the country’s population as recorded on the Demographics.csv file. 2 new variables were created in the consolidated data frame to reflect this as follows:
  + “per pop cases” is the cumulative cases of Covid-19 recorded divided by the country population.
  + “per pop dedceased” is the cumulative deaths from Covid-19 recorded divided by the country population.



* + This created a disparity in the size of the 2 new variables and the Stringency index. This was addressed by normalising the data fields (see below). The development index field was also normalised for the same reason.

|  |  |  |  |
| --- | --- | --- | --- |
| **Stringency index** | **Per Pop Dedceased** | **Per Pop cases** | **Human Development Index** |
|  |  |  |  |

A screenshot of a computer code

Description automatically generated

* It was known at the outset in constructing the Consolidated data frame that the data frames would contain null values. At least in part this reflected the differing granularity of the data. For example, for some countries there was data availability at county level but not for others.
* 3 data frames were created with subsets of the data needed for the analysis from the Consolidated data frame. Missing value analysis was performed on these subsets and small number of missing values were detected in all 3:

A screenshot of a computer program

Description automatically generated

These countries with missing values were listed to ensure that no large economies were among those without values. All were relatively small countries. There was 1 surprise in that Ireland was one of the countries without a value for its Human Development Index!:

A screenshot of a computer

Description automatically generated

# Machine Learning Implementation

The initial exploratory research questions related to whether there is any evidence of a relationship in the country level data between:

* the Stringency index (concerning the Government actions to counter the pandemic) and the number of cases of Covid-19 and deaths from the pandemic. This hypothesis is that that the stringency of government actions was a function of the severity of impact of the pandemic on their respective populations.
* the Stringency index and the Human development index. The hypothesis here is that countries with more knowledgeable citizens, higher levels of health and living standards took stronger measures to protect their populations.
* Linear regression models were used in each case with the Stringency Index as the dependent variable as shown below:
  + Hypothesis 1: Stringency Index is a function of the number of Covid-19 deaths
  + Hypothesis 2: Stringency Index is a function of the number of Covid-19 cases
  + Hypothesis 3: Stringency Index is a function of the country Development Index.
* Linear regression tries to find the line of best fit for data variables that may be related. The method endeavours to locate a line that minimises the means squared errors between points on the line fitted and the actual points (Tarek Amr, 2020). Simple Linear regression involves 2 variables, one dependent and the other independent and seeks to find the equation of the line of best fit with a slope and an intercept:

Y = β0 + B1x + error term

The effectiveness of this analysis depends on the extent to which the data points can be estimated by a line as measured by the variance of the points from the line(Oswald Campesato, 2020). If the variance is large, then it may be that a curved line will be a better fit (using say logarithmic regression). Another possible avenue is the use of multivariate regression analysis which introduces multiple independent variables.

The python “sklearn” library was used for its regression model and associated tools. The key indicators for evaluating the performance of the regression models (Https://Www.statology.org/, n.d.):

|  |  |  |
| --- | --- | --- |
| Statistic | Full title | What it tells us |
| B1 | Slope of the line of best-fit | The rate of change in the dependent variable as the independent changes. |
| R2 | Coefficient of Determination | The proportion of the variance in the dependent variable that can be explained by the independent variable |
| RMSE | Mean Squared Error | Average difference between predicted and actual values in a dataset. |
| MAE | Mean Absolute Error | The absolute difference between the predicted and actual values. |

# Summary of Findings

* The results of the analysis showed little evidence of a relationship with each of the r2 scores being close to zero. R scores for 2 of the variables were also negative indicating a poor outcome.
* These outcomes were supported by the MSE and MAE values that were relatively high given the data had been scaled.
* The coefficients showed in each case an inverse relationship between the Stringency Index and the other variables. This was unexpected for the 2 mortality and morbidity independent variables (Covid-19 cases and deaths) where it was expected that the relationship would be positive with greater stringency accompanying greater impact. However, the models performed so poorly that the inverse relationships may not be meaningful.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R2** | **MSE** | **MAE** | **Coefficient** |
| Hypothesis 1 | 0.02249 | 0.959 | 0.80 | -0.11241859 |
| Hypothesis 2 | -0.00079 | 0.711 | 0.64 | -0.35058293 |
| Hypothesis 3 | -0.15177 | 0.57 | 0.60 | -0.3457762 |

These initial findings raise more questions concerning:

* The hypothesised relationships,
* The tuning of the models used,
* The use of alternative models and methods of analysis,
* Revisiting the data to ensure that it was compiled correctly.

These issues will be tackled as the Data Exploration phase of the project continues over the summer.

# Ethical Considerations

No ethical considerations arose during the exploratory analysis to date.

# Conclusion and Future recommendations

* The project is still at the data exploration stage. Good progress has been made in exploring the data and in particular how the indexing works to link the files.
* Initial exploration has concentrated on data at a country level at a point in time when real time updates ceased (September 2022)
* Some initial data analysis explored whether the stringency index (in terms of Government actions) was related to the impact on their respective populations in terms of deaths recorded as caused by Covid-19 or number of cases recorded with Covid-19. In both cases, based on a linear regression analysis, there was no evidence of any relationship between stringency index and the number of deaths and Covid-19 cases.
* Further data analysis explored whether the stringency index (in terms of Government actions) was related to the Development Index of the country. Again, based on a linear regression analysis, there was no evidence of any relationship between Stringency index and the Development Index.
* The conclusion from this analysis is preliminary as further exploratory work on the data and the development of the models needs to be completed.
* In terms of future development of the project, the following will be prioritised:
  + Initial analysis using Linear regression has offered no evidence of the relationships hypothesised. Other models will be trialled to see if they perform better.
  + Much of the data is time series data, with in some cases daily data updates. This offers the possibility of using artificial intelligence techniques suit to this type of data. In a continuation of the themes explored so far, a possible avenue for exploration is whether daily variations in the impact of Covid-19 (cases and deaths) are related to daily changes in the stringency of government counter measures. If such a relationship exists, it is possible that government action lagged the progress of cases and deaths. This will be a focus going forward.
  + Due to the volume of data, it may be more manageable to initially download the data into MYSQL for ease of data exploration and manipulation and then allowing Python access to the data through an API.

## Github link

The Github link for this report and associated material is at: https://github.com/CCT-Dublin/capstone-project-feb-2024-pt-burnsjohn1

# References

Tarek Amr. (2020). *Hands-on machine learning with scikit-learn and scientific Python toolkits : a practical guide to implementing supervised and unsupervised machine learning algorithms in Python / Tarek Amr.* Packt Publishing Ltd., July,2020, London.

‌Oswald Campesato. (2020). *Python 3 for machine learning*. Mercury Learning & Information, 2020, Dulles,Virginia.

‌*https://www.statology.org/*. (n.d.). Statology. https://www.statology.org/

‌

# Appendix 1

A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name: John Burns* |  |
| *Student Number: SBS24019* |  |
| *Module Title: Higher Diploma Data Analytics - Strategic Thinking* |  |
| *Assessment Title: CA 1 – Capstone Project Proposal* |  |
| *Assessment Due Date: 29th March 2024 23:59* |  |
| *Date of Submission: 26-03-2024* |  |

**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.

Contents

[Introduction 1](#_Toc162371520)

[Objectives 1](#_Toc162371521)

[Problem Definition 2](#_Toc162371522)

[Scope & Methodology 2](#_Toc162371523)

[Business (Problem) Understanding 2](#_Toc162371524)

[Data Understanding 2](#_Toc162371525)

[Data preparation 2](#_Toc162371526)

[Modeling 3](#_Toc162371527)

[Evaluation 3](#_Toc162371528)

[Deployment 3](#_Toc162371529)

[Data Sources 3](#_Toc162371530)

[Timeline 3](#_Toc162371531)

[Ethical Considerations 4](#_Toc162371532)

[Conclusion 5](#_Toc162371533)

[References 6](#_Toc162371534)

# Introduction

The recent Covid-19 pandemic disrupted the lives of billions throughout the world. In Ireland it led to rapid changes across the economic and social lives of citizens with enormous impacts on numerous areas such as physical and mental health, social care, job losses associated with business closures, a forced shift to remote working, major constraints on the activities and movements of citizens. Governments across the world acted to mitigate the impact on their populations. A variety of measures, proactive and reactive, were introduced in efforts to curb infections. The effects of these measures varied from hugely successful (as with vaccines developed in record time) to more questionable (many older, more vulnerable patients were moved to higher risk social care homes from hospital and subsequently died alone). For many countries, a top priority became to gauge the effects of the pandemic on their populations to provide input to decisions as to the measures that could be implemented to limit the damage to their societies. Its impacts were closely followed as the top stories in the nightly news and agencies charged with data collection data found that many of their outputs became key indicators on the state of the nation. As a result, this became perhaps the most tracked and documented health crisis in history.

# Objectives

While the Covid-19 emergency has receded, the disease and many of its human casualties are still with us (Christakis, 2024) and will be for some time with 920 deaths still being reported as being due to the disease in 2023 (Report on COVID-19 deaths reported in Ireland, 2024). However, it is timely to look at the various initiatives taken when the Covid-19 crisis was at its peak and assess what worked well and what did not. This is the overall objective of this project with a focus on the actions of governments and their impact on the crisis.

In the pandemic government bodies took center stage in tackling not only the direct medical implications of Covid-19 but in preventing its spread and supporting the economic and social infrastructure. These initiatives included:

* School closings
* Cancellation of public events
* International travel controls
* Income support.
* Workplace closings
* Limits on mobility.

The broad objectives of this project are:

* To explore the impact of various government initiatives on the incidence of Covid-19
* Determine if there were any such measures that were particularly effective.
* Compare the effectiveness of government actions across the EU and UK.
* Track the impact of vaccines on the pandemic from their introduction.

# Problem Definition

During the Covid-19 pandemic governments took various actions with the intention of reducing the impact on their populations. Some of these worked well, some had limited impact while others are likely to have had little effect. To be able to deal more effectively with future pandemics, it is essential to learn from this recent experience the value of these initiatives. This will increase the effectiveness of future responses to pandemics.

# Scope & Methodology

The scope of the project is as follows:

* It will focus on Government actions to contain the pandemic and maintain public health.
* Geographically it will extend to EU countries and the UK. If time allows it may be extended further.
* The time for the analysis will extend from the 11th of March 2020 when the World Health Organization declared the pandemic to the 5th May 2023 when the global health emergency crisis caused by the disease was over.
* Further exploration of the Google Health COVID-19 Open Data Repository may reveal additional opportunities for insightful analysis of the effectiveness of measures to mitigate the effects of Covid-19.

Process models provide a structure in progressing a project. There are several in the data science arena of which CRISP-DM (Cross Industry Standard Process for Data Mining) is the best known. Schroder et al characterize it as the de facto standard in their systemic literature review of the CRISP-DM (Schröer, Kruse and Gómez, 2021). It was developed by a consortium of private companies, based on their collective real-world experience (Shearer, 2000) and seeks to put structure on knowledge discovery endeavors (Wirth and Hipp, 2000). It will be used for this project. It is a sequential model with the following phases:

Business (Problem) Understanding

This concentrates on defining the problem to be addressed and forms the basis for project objectives. These in turn drive the formulation of analysis and machine learning objectives that combine to define key project deliverables.

## Data Understanding

This phase includes gathering of data and familiarization with the data. This includes getting to know the nature and extent of the data available, its provenance and identification of any shortcomings or quality issues with the data.

## Data preparation

This phase encompasses all the activities to develop the final dataset(s) that will be used as input to the modelling tools. This includes feature engineering, data consolidation and cleaning. It has the potential to be the step with the longest duration. The principles encompassing tidy data as outlined by Hadley Wickham (Wickham, 2014) will be followed during this phase.

## Modeling

This involves the selection and application of the modeling methods chosen to meet project objectives. It includes model selection, training, and validation of models, tuning hyperparameters, developing visualisations and performance metrics.

## Evaluation

This phase sees the model’s deployment evaluated in the light of the project objectives. Particular emphasis will be placed on fine-tuning the models and ensuring that all important variables have been considered.

Deployment

This phase is the final step that sees the model used to further the original project objectives be they reporting, forecasting, etc. For this project the final deliverables will include a report and working models developed during the project. While this may be the final step in the project, it may set the stage for the development of further iterations and enhancements to the models in the future.

While the description of the process may read like a typical sequential waterfall type process, the developers of the methodology recognized that the process is creative and often iterative and that there may be several iterations before a satisfactory conclusion is reached (Wirth and Hipp, 2000).

# Data Sources

The primary data source will be the Google Health COVID-19 Open Data Repository at <https://health.google.com/covid-19/open-data/> (health.google.com, n.d.). This repository began to collect data from across the globe during the Covid-19 pandemic to assist in epidemiologic studies and also to track numerous associated variables such as government interventions, weather, population movements etc. to assist in understanding and isolating patterns (Wahltinez et al., 2022). Other local data sources (such as the Irish Central Statistics Office) may emerge during the exploratory stages of the project that may be useful in meeting project objectives.

# Timeline

The high-level timeline below assumes the following:

* The project continues over the summer break.
* The first phase of the project – Business (Problem) Understanding begins in week 6 of Semester 1
* The broad project plan outlined below will act as a baseline against future changes to the project plan as the project progresses.
* There is limited contingency if deadlines slip.

A screenshot of a computer screen

Description automatically generated **Semester One**:

* Business (Problem) Understanding – 8 weeks.
* Data Understanding – 15 weeks.
* Data preparation including preprocessing, cleaning and feature engineering - 5 weeks.

**Semester Two**:

* Modelling includes model selection, training and validation - 6 weeks.
* Modelling part 2 including hyperparameter tuning and production of visualisations - 4 weeks.
* Evaluation and assessment of the model - 5 weeks.
* Documentation and reporting - 4 weeks.

# Ethical Considerations

The focus of this project concerns government actions to control the impact of a pandemic focusing on the recent Covid-19 outbreak. The World Health Organization emphasises the imperative to learn what worked and what did not so as to be able to increase resilience in preparation to meet future pandemics (World Health Organisation, 2021). The aims of this project are aligned with this objective. No major ethical issues are anticipated during the project.

* The report will use the data from the Google Health COVID-19 Open Data Repository in line with best standards of transparency, attribution, and proper dataset usage,
* The Harvard Reference system will be used to acknowledge and attribute the work of others cited in the project.
* Much of the data comes from authoritative and reliable sources such as public health agencies or government agencies charged with collecting statistics. A subset of these will be the focus of this project.
* There is also data from locations that do not collect reliable data and may introduce bias. These have been collated from press sources or through crowd sourcing. These may also introduce bias. The locations for which this type of data is available are outside the scope of the project (Wahltinez et al., 2022).
* There is no personal information included in the database.

# Conclusion

At the conclusion of the project a project report and presentation will be produced along with associated Jupyter notebooks showing the analysis performed and the results. These will also document the sequence and content of the analysis performed.

# References

Christakis, N. (2024). “Four Years Later, Covid Isn’t Done With Us”. *Wall Street Journal*, 10 Mar., p.C3.

‌ Report on COVID-19 deaths reported in Ireland -. (2024). [online] Health Protection Surveillance Centre, Ireland: Health Protection Surveillance Centre, pp.1–8. Available at: https://www.hpsc.ie/a-z/respiratory/coronavirus/novelcoronavirus/surveillance/covid-19deathsreportedinireland/2023/COVID-19\_Death\_Report\_Website\_v1.8\_09012024.pdf [Accessed 13 Mar. 2024].

health.google.com. (n.d.). *COVID-19 Open Data — Google Health*. [online] Available at: <https://health.google.com/covid-19/open-data/>.

Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, [online] 181(1), pp.526–534. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921002416.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mini*. Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mini. Great Britain: Practical Application Company, pp.29–40.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10). doi:https://doi.org/10.18637/jss.v059.i10.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), pp.13–22.

‌Wahltinez, O., Cheung, A., Alcantara, R., Cheung, D., Daswani, M., Erlinger, A., Lee, M., Yawalkar, P., Lê, P., Navarro, O.P., Brenner, M.P. and Murphy, K. (2022). COVID-19 Open-Data a global-scale spatially granular meta-dataset for coronavirus disease. *Scientific Data*, 9(1). doi:https://doi.org/10.1038/s41597-022-01263-z.

‌ World Health Organisation (2021). *Response to the COVID-19 pandemic: lessons learned to date from the WHO European Region*. [online] *World Health Organisation - IRIS (Institutional Repository for Information Sharing)*, Geneva: World Health Organisation, pp.1–13. Available at: https://iris.who.int/bitstream/handle/10665/343157/71wd06e-rev1-PR-Response-LessonsLearned-210693.pdf?sequence=1&isAllowed=y [Accessed 24 Mar. 2024].