CCT College Dublin

Assessment Cover Page

|  |  |
| --- | --- |
| Module Title: | *MSc in Data Analytics*  *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| Assessment Title: | *MSC\_DA\_Integr\_CA2\_Sem1*  Scenario: Irish Construction sector |
| Lecturer Name: | *Sam Weiss*  *John O’Sullivan*  *Muhammad Iqbal*  *David McQuaid* |
| Student Full Name: | Jose Luis Alvarado Carranza |
| Student Number: | 2020092 |
| Assessment Due Date: | 26/05/2023 |
| Date of Submission: | 26/05/2023 |

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. |

Group ID - MSc in Data Analytics

Author: Jose Luis Alvarado Carranza

e-mail: [2020092@student.cct.ie](mailto:2020092@student.cct.ie)

Student ID: 2020092

Total word count: 3310 words.

**Contents**

[*Abstract and Keywords* 4](#_Toc136036071)

[Introduction 4](#_Toc136036072)

[Stage one: Sampling and exploring. 5](#_Toc136036073)

[1. State the Null Hypothesis, *H*0 and the Alternative 7](#_Toc136036074)

[2. Find the critical value in order to define the rejection region 7](#_Toc136036075)

[3. Calculate a test statistic from the data we have collected 7](#_Toc136036076)

[4. Make a decision 8](#_Toc136036077)

[5. Summarise the results 8](#_Toc136036078)

[Stage two: Exploration 9](#_Toc136036079)

[Stage three: Modifying 10](#_Toc136036080)

[Stage four: Modelling 11](#_Toc136036081)

[Stage five: Assessing 12](#_Toc136036082)

[Bibliography 12](#_Toc136036083)

[References 13](#_Toc136036084)

# *Abstract and Keywords*

*The following continuous assessment shows the results of many hours of interpretation of data and analyses done for the purpose of high marks in the topic of Construction. There are many data frames that have been chosen to surmise information. This report sets Ireland as a baseline and uses it to compare construction cost indexes. To accomplish this, the author has resorted to research and analyses of this variable that encompasses many aspects of the construction business throughout the world at the same time that utilizes methods, techniques and skills gathered in the lectures provided by this institution. Additionally, this report tries to shed some light in the existing similarities of the construction cost without having to incur in the nuances of bureaucracy.*

*Keywords: Hyperparameter, Dashboard, Statistical testing, Testing, exploratory analysis.*

# Introduction

Construction is the process of creating a plan and adding structure to real property or construction of buildings. In the modern industrialized world, construction usually involves the translation of designs into reality. However, understanding and estimating construction costs can be challenging. Luckily for us, there are a lot of data sets from all over the world to pull out data from.

The field of construction is extremely vast and there are many variables to consider when developing a project. Therefore, the focus of our project will revolve around Residential Property Price Index (RPPI) which will be simply referred as “index”. By adhering to this topic many of the other topic get considered and funnelled into a single variable, making it more comprehensible to the reader.

We begin by defining what cost index is and how it is calculated to better interpret our data.

The Residential Property Price Index (RPPI) is designed to measure the change in the average level of prices paid by households for residential properties sold in Ireland. (CSO, 2023) The RPPI specifically excludes non-household purchases, non-market purchases and self-builds (i.e. where the land is purchased separately). The index is mix-adjusted to allow for the fact that different types of property are sold in different months. (CSO, 2023)

Now that we have defined our variable, let us find out how to calculate it.

The RPPI can be used to estimate the updated value of an individual dwelling provided that a prior value is known subsequent to January 2005. (CSO, 2023)

Simply multiply the sale or valuation price by the current relevant index and divide by the index at the date of sale/valuation.

For example, consider a house sold in South Dublin in June 2010 for €220,000.

      South Dublin house price index June 2010:                           88.9

      South Dublin house price index currently:                            121.5

      Estimated current value:                                                      €220,000 x (121.5 / 88.9)   =   €300,675

This estimate provides an approximate value only, based on aggregate price movements of all dwellings. It presupposes that there has been no material change in either the dwelling or its neighbourhood in the period concerned. The CSO takes no responsibility for calculation of value based on the RPPI. (CSO, 2023).

Before delving into processing the data, a methodology has been put in to place to provide structure to the overall project. We have decided to choose SEMMA project management because we do not need to have a lot of knowledge on construction to sort the data and provide a model. After the data has been selected, we begin our assessment.

# Stage one: Sampling and exploring.

As it has been established before, the topic of construction has many aspects to it. It can be the cost of construction from the commercial, residential, industrial, a country’s infrastructure, etc. We are going to consider the residential aspect of construction.

According to the brief in the data set named “The House Building Cost Index”, it monitors labour costs in the construction industry and the cost of building materials. It does not include items such as overheads, profit, interest charges or land development. (Government, 2021) The labour costs include insurance cover and the building material costs include V.A.T. Coverage:  
The type of construction covered is a typical 3 bed-roomed, 2 level local authority house and the index is applied on a national basis”. (Government, 2021).

We first begin by gathering data from the recommended website. Since the task requires us to use Ireland as a baseline for our modelling, Our sample stage is of Ireland.

After the proper EDA has been performed, we have gathered a useful insight into our topic.

We have chosen the year 2015 as it is the latest usable information available in this data set.

A picture containing text, screenshot, diagram, plot

Description automatically generated

*Figure 1 Histogram of the Cost of construction in 2015*

This is the histogram of the index for the year 2015 in Ireland. From this figure we can gather that:

* The cost of construction in Ireland increased and reached the highest index of 207.6.
* The cost of construction in Ireland had the lowest index of 206.0.
* There is a variation in the index of 1.6 points within a year.

This is the boxplot of the index for the year 2015 in Ireland.

A picture containing screenshot, text, rectangle, square

Description automatically generated

*Figure 2 Boxplot of the Index in 2015*

From this figure we can gather that:

* There are no outliers.
* The median value is of 207.2.
* The value of 206.8 represents Q1 of the data and the value of 207.4 Q3 while Q2 is equal to 207.2.

This is the heatmap of the index throughout the years in Ireland.

A picture containing text, screenshot, line, parallel

Description automatically generated

*Figure 3 Heatmap of index per year*

From this figure we can gather that:

* 2009 and 2011 represent dips in the index.
* Even though the index has decreased in some years, it has increased in most of them.
* 2012 was the highest when compared to 2015.



*Figure 4 Statistical data for 2015*

Based on how the index is calculated and the average price of a house in Dublin in 2015 which was of €332,000, (Alliance, 2016) brings us to our first hypothesis.

## 1. State the Null Hypothesis, *H*0 and the Alternative

Hypothesis, *HA*: From our data set we know that the yearly average of the index in 2015 is of 207.05. We hypothesise that the mean index in Ireland is not 207.05. To confirm that we resort to using a Two-tailed test.

## 2. Find the critical value in order to define the rejection region

We use the conventional `significance level' of α = 0.05.

### 3. Calculate a test statistic from the data we have collected

Using python we get a p-value of 0.9917

### 4. Make a decision

Since our p-values is greater than 0.05, we decide to reject our hypothesis.

### 5. Summarise the results

This means that the values of our data are correct and that our assumption was confidently rejected.

A picture containing text, screenshot, diagram, plot

Description automatically generated

*Figure 5 Alternative hypothesis*

The result of the sampling section shows how we have gathered the necessary information to infer results on the next stage.

In order to complete this project, we have done an extensive research in the field of construction and as previously stated, we have decided to work with construction cost index. It is worth mentioning that while there are many websites where the information is available, some of them require certain permission which are obtained via subscription and or by paying a monthly fee. We have avoided them. Therefore, all the data we have gathered is through government websites and trustworthy reports from different parts of the world. This has allowed us to avoid incurring in monetary expense.

Due to the nature of the variable, we have encountered journals and handbooks in PDF files which contain useful information for our project because they discuss in detail how the information was gathered and what methods came into place when sorting the data. We have converted such information to csv files in order to process the data. The focus on indexes is, in many ways convenient, because topics like materials, workforce, location, etc have already been considered in this variable. There are journals where this information can be obtained all of them have been referenced accordingly. We have worked with 4 different data sets. One for Ireland, which was obtained from the CSO website. One for the APAC regions, one for the European block, both of which were found in journals and one for the American block, which was gathered from the Federal reserve bank website.

As part of this assignment, we have been tasked with processing data in different formats. This proved to be a challenge for the same reasons mentioned above. Furthermore, data found in API in JSON format is tricky to manipulate because of the limited number of requests and the time it takes to establish connections. Our approach to circumvent this obstacle was to work with csv files and then store them in our own database using MONGODB. However, it even that gave us something to struggle with.

It is worth mentioning that the data for the sentiment analyses were particularly troublesome to find because it is extremely abundant, and so extremely valuable. We mean that in more than one sense. Thus, we didn’t include it in our project. We understand the relevance of a sentiment analysis. It is the voice of the people in important matters for our society. Scrutinizing people’s opinions has turned into a form of currency for big corporations.

We were fortunate enough to have found relevant information for our investigation. Nevertheless, in some instances, this information was incomplete. It became apparent that they had been published before all the data was gathered. In such cases, we opted to discard that data as it would have created problems throughout the development process of our project. This decision was followed in a case by case basis as it would vary depending on the type of analysis and the method that was followed.



*Figure 6: Incomplete data*

# Stage two: Exploration

As previously mentioned, the subject of construction is vastly different. It contains in itself various approaches. We decided to include representative data to cover most of it, nonetheless, most of it has been overlooked for this report.

The construction indexes can be separated in branches. These are Infrastructure for any nation such as roads, bridges, government facilities, etc., Residential construction i.e., houses and apartment buildings, commercial construction like shopping malls, department stores, supermarkets, etc. All these branches are represented in the different data sets we used in our project.

We have performed a test to compare the average index in Europe with the index in Ireland. It is worth noting that for the European block, the data is averaged from 2011 – 2016. To match it we have chosen again the year 2015 for Ireland.

We hypothesised that in Ireland the index is higher than the rest of Europe.

A screenshot of a computer code

Description automatically generated with medium confidence

Figure 7 Comparing Ireland to Europe

The result of our p-value tells us that it was not the case.

We hypothesised that there was a difference in the construction of Asian pacific block index. To prove that we ran an ANOVA test.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 8 ANOVA test Ireland v APAC

Our p-value indicate that it is the case that they are different.

The variables IreM and EUM represent values for means for both of the different populations. In this case we hypothesised that the means were different.

A screenshot of a computer code

Description automatically generated with low confidence

Figure 9 Mean values for Ireland and the EU

The result of our p-value in this case indicates that there is no difference in the means for both groups. I presume there must have been an error in the data as it is highly unlikely the result turned out to be one. We can manipulate the data set as needed. However, we are not allowed to modify the data and assign values in order to get the results we hoped for.

It is important to mention that our data sets do not include categorical data. Therefore, categorical test cannot be done in any of the data sets.

# Stage three: Modifying

As we have gathered data from different sources, it is very important to make sure we have the respective data types to be able to process the data and use the supervised learning model in the Machine learning portion of our assignment. Since we have gathered the data from handbooks and journals, the data type is as an object. We used several techniques learnt throughout our course to be able to manipulate it and obtain results in the data type needed to fulfil the task. In other words, the data has been parsed and cleaned, then passed onto the modelling stage, and explored if the data requires refinement and transformation.

Regarding the data visualization, we have created a dashboard depicting a map of Europe. It shows the index data per country. To accomplish this, we relied on several libraries. We used resources online and the libraries that were used in the lectures e.g., plotly.express. We have adapted them for the purposes of this continuous assessment. Additionally, we have considered alternative libraries like “folium”. While both are very powerful libraries, there are drawbacks to both of them. While the library plotly.express achieves the objective of this continuous assessment, it is not the best looking result and so, we tried to use “folium”. The downside of folium is that we could not find the proper file to adapt it to our needs. Its capabilities seem to be more user-friendly.

# Stage four: Modelling

When talking about modelling, there are certain characteristics that must be incorporated into the process. This means that the model must be independent and repeatable. In our case, the code was written following the lectures throughout the semester. Therefore, by adhering to the conditions set in class we can ensure that they are repeatable. It is independent because each of the variables, analyses and interpretations have been rigorously determined. And so, any potential outcomes have been considered. This is confirmed by the fact that each single line of code has a single purpose.

This is not to say that there may not be any errors. It is to say that the errors are a result of a lack of complete proficiency in the tools we have used in the development of this assignment. Foreseeable issues have been identified and avoided so as to come up with results that satisfy the needs of this assignment. Additionally, there are many aspects of the project that were not explored. There have been many instances in which typical values were tested resulting in errors. Therefore, it is up to the analyst to determine if a result is acceptable within the boundaries of the assignment.

We have summarized the results of the machine learning portion in this table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | LINEAR REGRESSION | RIDGE | LASSO |
| R-Squared of  the model in  training set | 0.9794719301272705 | 0.9552862318393665 | -0.7624516730123474 |
| R-squared of  the model in test | 0.7510572889077689 | 0.4105282510653957 | -0.5394646788565747 |
| Root mean squared error of the  prediction | 0.7341915121186402 | 1.1297728917702028 | 0.49687094584517905 |
| Mean absolute  percentage error of the prediction | 46.40428305159508 | 83.17654693226828 | 0.19392890591092773 |

Table 1 Statistical values for supervised machine learning models

# Stage five: Assessing

There are many reasons to compare one nation with the rest of the world. Our research indicates that even when compared to more powerful (in terms of infrastructure) nations, Ireland measures up to the standards of the first world. Our different tests, and analyses support this claim.

With regards to our approach, we presume that it covers the most relevant details. We found that sorting out the information and its data were as time consuming as the coding part.

In terms of the efficacy of our research, we must admit that our variable index, while concise is very limiting. The possible outcomes that could come if the results of this report were to be deployed would be less than desirable.

However, different aspects of the world of construction that were out of the scope of this assignment could show different results. Which brings us to the areas that could be developed further.

The data that was processed in this report for Ireland, is outdated. The fact that there are handbooks and journals with data from 2021 in other parts of the world, suggest that there is a lot of information that has been left out. This is not even accounting for other factors like the Pandemic, Brexit or even the war in Ukraine i.e., workforce, material, and its transport. An important aspect to be considered in a future project would be in terms of the technology that is implemented in the field, like machinery or spatial techniques to maximize the number of households in taller apartment buildings.

# Bibliography

Alliance, R. E., 2016. *https://www.realestatealliance.ie/q4-2015---average-house-price-index-p6726.* [Online]   
Available at: https://www.realestatealliance.ie/q4-2015---average-house-price-index-p6726  
[Accessed 23 05 2023].

CBRE, 2022. *https://www.cbre.com.* [Online]   
Available at: https://www.cbre.com/-/media/project/cbre/shared-site/insights/reports/global-cost-trends-guide-2020-2021/global-cost-trends-2020-2021-apac.pdf  
[Accessed 16 05 2023].

CSO, I., 2023. *https://www.cso.ie/.* [Online]   
Available at: https://www.cso.ie/en/methods/surveybackgroundnotes/residentialpropertypriceindex/  
[Accessed 21 05 2023].

datascience-pm, 2021. *https://www.datascience-pm.com/.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-2/  
[Accessed 10 April 2023].

Eurostat, 2021. *https://ec.europa.eu/eurostat/statistics-explained/index.php.* [Online]   
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Construction\_producer\_price\_and\_construction\_cost\_indices\_overview#Construction\_costs\_-\_development\_since\_2005  
[Accessed 5 5 2023].

geeksforgeeks, 2020. *https://www.geeksforgeeks.org/kdd-process-in-data-mining/.* [Online]   
Available at: https://www.geeksforgeeks.org/kdd-process-in-data-mining/  
[Accessed 10 April 2023].

Government, I., 2021. *https://data.gov.ie.* [Online]   
Available at: https://data.gov.ie/dataset/national-house-construction-cost-index?package\_type=dataset  
[Accessed 05 05 2023].

linode, 2022. *https://www.linode.com.* [Online]   
Available at: https://www.linode.com/docs/guides/pros-and-cons-of-python/  
[Accessed 11 April 2023].

Matplotlib, 2012. *https://matplotlib.org.* [Online]   
Available at: https://matplotlib.org/3.5.3/api/\_as\_gen/matplotlib.pyplot.html  
[Accessed 7 April 2023].

Review, W. P., 2023. *https://worldpopulationreview.com/.* [Online]   
Available at: https://worldpopulationreview.com/world-cities/dublin-population  
[Accessed 8 April 2023].

Seaborn, 2022. *https://seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/tutorial/introduction  
[Accessed 7 April 2023].

stxnext, 2022. *https://www.stxnext.com/.* [Online]   
Available at: https://www.stxnext.com/python-vs-other-programming-languages/#:~:text=dynamically%20typed%20vs.-,compiled%20and%20statically%20typed,code%20readable%20by%20the%20machine.  
[Accessed 11 April 2023].

w3schools, 2023. *https://www.w3schools.com/.* [Online]   
Available at: https://www.w3schools.com/python/numpy/numpy\_intro.asp#:~:text=NumPy%20is%20a%20Python%20library,you%20can%20use%20it%20freely.  
[Accessed 7 April 2023].

W3Schools, 2023. *https://www.w3schools.com/.* [Online]   
Available at: https://www.w3schools.com/python/pandas/pandas\_intro.asp#:~:text=Pandas%20is%20a%20Python%20library,by%20Wes%20McKinney%20in%202008  
[Accessed 7 April 2023].

# References

Alliance, R. E., 2016. *https://www.realestatealliance.ie/q4-2015---average-house-price-index-p6726.* [Online]   
Available at: https://www.realestatealliance.ie/q4-2015---average-house-price-index-p6726  
[Accessed 23 05 2023].

CBRE, 2022. *https://www.cbre.com.* [Online]   
Available at: https://www.cbre.com/-/media/project/cbre/shared-site/insights/reports/global-cost-trends-guide-2020-2021/global-cost-trends-2020-2021-apac.pdf  
[Accessed 16 05 2023].

CSO, I., 2023. *https://www.cso.ie/.* [Online]   
Available at: https://www.cso.ie/en/methods/surveybackgroundnotes/residentialpropertypriceindex/  
[Accessed 21 05 2023].

datascience-pm, 2021. *https://www.datascience-pm.com/.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-2/  
[Accessed 10 April 2023].

Eurostat, 2021. *https://ec.europa.eu/eurostat/statistics-explained/index.php.* [Online]   
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Construction\_producer\_price\_and\_construction\_cost\_indices\_overview#Construction\_costs\_-\_development\_since\_2005  
[Accessed 5 5 2023].

geeksforgeeks, 2020. *https://www.geeksforgeeks.org/kdd-process-in-data-mining/.* [Online]   
Available at: https://www.geeksforgeeks.org/kdd-process-in-data-mining/  
[Accessed 10 April 2023].

Government, I., 2021. *https://data.gov.ie.* [Online]   
Available at: https://data.gov.ie/dataset/national-house-construction-cost-index?package\_type=dataset  
[Accessed 05 05 2023].

linode, 2022. *https://www.linode.com.* [Online]   
Available at: https://www.linode.com/docs/guides/pros-and-cons-of-python/  
[Accessed 11 April 2023].

Matplotlib, 2012. *https://matplotlib.org.* [Online]   
Available at: https://matplotlib.org/3.5.3/api/\_as\_gen/matplotlib.pyplot.html  
[Accessed 7 April 2023].

Review, W. P., 2023. *https://worldpopulationreview.com/.* [Online]   
Available at: https://worldpopulationreview.com/world-cities/dublin-population  
[Accessed 8 April 2023].

Seaborn, 2022. *https://seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/tutorial/introduction  
[Accessed 7 April 2023].

stxnext, 2022. *https://www.stxnext.com/.* [Online]   
Available at: https://www.stxnext.com/python-vs-other-programming-languages/#:~:text=dynamically%20typed%20vs.-,compiled%20and%20statically%20typed,code%20readable%20by%20the%20machine.  
[Accessed 11 April 2023].

w3schools, 2023. *https://www.w3schools.com/.* [Online]   
Available at: https://www.w3schools.com/python/numpy/numpy\_intro.asp#:~:text=NumPy%20is%20a%20Python%20library,you%20can%20use%20it%20freely.  
[Accessed 7 April 2023].

W3Schools, 2023. *https://www.w3schools.com/.* [Online]   
Available at: https://www.w3schools.com/python/pandas/pandas\_intro.asp#:~:text=Pandas%20is%20a%20Python%20library,by%20Wes%20McKinney%20in%202008  
[Accessed 7 April 2023].