**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Statistics for Data Analytics |
| **Assessment Title:** | MSC\_DA\_Integr\_Repeat\_Sem1 |
| **Lecturer Name:** | John O’Sullivan |
| **Student Full Name:** | Sule Beste Kapci |
| **Student Number:** | 2022327 |
| **Assessment Due Date:** | 26th May 2023 |
| **Date of Submission:** | 26th May 2023 |
| **Github** | https://github.com/besteli/Summer2023RepeatAssessment |

**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. |

In this study, CRISP-DM stages were followed, and only statistical operations were discussed within the scope of this report. Within the scope of this project, the average value generation data by years of the construction sector in Ireland and Spain were examined.

Since the construction sector dataset of Ireland contains more data, statistically more information was obtained in the exploration phase compared to the Spanish dataset.

The information obtained from the Irish dataset is as follows:

A picture containing text, diagram, plot, line

Description automatically generated

Figure 1. Average Value Production in Construction Sector by Year for Ireland

The average value of the sector has increased in the process starting from the beginning of the 2000s and continuing until 2006. It is observed that there was a decrease in the period from 2006 to 2012. Although there is a surplus between 2012 and 2019, the average value of the sector in the 2000s could not be caught in this time period.

A graph with blue and orange lines

Description automatically generated with low confidence

Figure 2. Value Production in Construction Sector by Year and Type for Ireland

Although the Residental Building has shown up until 2006, there is a rapid decline after 2006. Despite the acceleration in the ongoing process, the values obtained in the 2000s could not be reached.

When the Non-Residental Building data were examined, it was observed that the process continued in a more stable manner and there were increases or decreases with small accelerations.

. A picture containing text, screenshot, font, diagram

Description automatically generated

Figure 3. Total Value Production Distribution by Type of Building and Construction

Approximately 70% of the buildings that have been constructed in total are Residental Buildings, and there is a big difference between them and the Non-Residental Building.

Although there are more graphics in this study, the main subject of these graphics is related to the situations mentioned above.

The information obtained from the Spain Dataset is as follows;

A picture containing text, plot, diagram, line

Description automatically generated

Unlike the situation in Ireland in this graph, a net increase is observed in the sector until 2006. Although the rate of increase decreased between 2006 and 2007, a net decrease was observed after 2007. This decline continues until 2013.

Although an increase can be observed in the sector after 2013, it is far from the value obtained in the 2000s.

During the data exploration phase, it was seen that the Irish dataset contains data between 2000-2022, while the Spain dataset contains data between 2005-2020. During the process, the Irish data set will be reduced to 2005-2020. This process was carried out to make the data set comparable.

STATISTICAL ANALYZES

Before starting the statistical analyzes, the data for the years other than the years 2005-2020 belonging to the Irish data set were excluded from the data set so that the two data sets could have the same time interval.

The hypothesis of this study is the idea that the growth trends in the construction sector between Ireland and Spain are different. It is thought that construction expenditures in Ireland are increasing and decreasing more steadily. In addition, it can be thought that construction expenditures in Spain follow a more volatile course or change under the influence of different factors.

1. Checking the Distribution of Values (Billion Euros) Column

It is this variable that is important in this study. For this reason, the distribution of the variable was controlled primarily for both data sets. The Shapiro-Wilk test was used at this point after seeing both columns.The data set was examined using this approach to check whether it had a normal distribution. The data is considered to have a normal distribution if the p-value from the test is less than 0.05. The p-values discovered are listed below.

|  |  |
| --- | --- |
| Dataset Name | P-Val |
| Ireland Dataset | 0.0006685575353913009 |
| Spain Dataset | 0.0037719435058534145 |

As can be seen, the two data sets do not show a normal distribution.

1. Confidence Level

Confidence interval shows how accurate the statistical result obtained as a result of the tests performed is within a certain range. The Confidence Level range obtained from the Irish data set at the 95% confidence level is (108.58111207216182, 252.65246796502964). The Confidence Level range obtained for Spain is: (151.76234660122756, 241.86265339877244). The width of the confidence interval gives the estimation precision. If the confidence interval is wide, the estimate is less accurate, while if the confidence interval is narrow, it indicates that the predictive value is more precise.

1. . Skewness

The resulting distributions structured are given below.

A picture containing text, plot, diagram, line

Description automatically generated

A picture containing text, plot, diagram, line

Description automatically generated

The skewness values ​​obtained from the data sets are as follows, respectively;

ireland skewness 1.28, spain skewness 0.80. Positive skewness values ​​for the two countries indicate that the distribution is skewed to the right. Here, in the distribution, the tail extends to the right. This means: Among the data in the data set, data with high degrees are more common than data with low degrees. This is clearly seen in the charts above.

1. Hypothesis Tests

Comparison will be made in the two data sets in the study. The hypothesis tests to be applied here will be carried out between two independent groups. In this case, it is theoretically appropriate to perform an independent two-sample T-Test or Mann-Whitney U test. However, the independent two-sample test is a test used when assumptions are met.

The data is assumed to be normally distributed and have equal variances between groups. We will be using the nonparametric Mann-Whitney U test to analyze the data. The objective of conducting this hypothesis test is to ascertain whether there exists a noteworthy difference between the two sample groups. It appears that there are no established principles to adhere to.It is important to note that these methods can be utilized in scenarios where the data does not conform to a standard pattern or exhibit similar levels of variability as in the previous example.To conduct this test, we assign ratings to the values that were randomly selected from each group. We then calculate the overall ranking scores based on these ratings. Following this step, a test statistic is derived by determining the gap that exists between the middle values of each category. During the process of putting the hypothesis to the test, the outcome is determined by comparing the acquired test statistic to a relevant critical value and using the p-value.

When Mann-Whitney U Test is applied; The null hypothesis cannot be rejected. When this situation is interpreted in line with the determined hypothesis, it means that the growth trends in the construction sector between Ireland and Spain are not different.

For individuals working in the field of data engineering, it is crucial to acknowledge that in instances where the distribution of a given data set is not known, it becomes imperative to employ alternative statistical tests. In this instance, alongside the Mann-Whitney U Test, the Permutation test was also employed. The findings of the random test were discovered to be The Wilcoxon Signed-Rank Test is a non-parametric hypothesis test used in statistical analysis to determine the equality of two independent datasets based on their individual properties. Because it makes no assumptions about the distribution's normality, the test is classified as non-parametric. This evaluation's conclusions are consistent with those of previous evaluations.

The Wilcoxon Rank-Sum Test is a commonly employed statistical test for comparing a specific characteristic between two distinct datasets. The normal distribution assumption is not required for this test, given its non-parametric nature. The Wilcoxon Rank-Sum Test is a statistical procedure employed to assess the degree of similarity or dissimilarity between the primary patterns of two distinct datasets. This test ranks the values in the data sets and calculates the test statistic using these ranks. The same results are obtained when this test is applied to the data sheet.

Finally, two sample Median Test was applied. It is a nonparametric test used to compare the medians of two independent data sets. This test does not require the assumption of normal distribution of data sets. The results obtained in this post-test applied did not change.

Although different non-parametric tests are applied, all tests show the same result. In this case, it can be said that the result of the data set is determinant rather than the difference of statistical methods. This result emphasizes the importance of the analyzed data set, along with the correct application of statistical methods.