# Introduction

# Data

# Exploratory Data Analysis

# Statistical Analysis

## Descriptive Statistics

Descriptive statistics summarizes the characteristics of a dataset (Bhandari, 2020) mainly uses central tendency, distribution, measure of variability and correlation. Central tendencies are based of the central limit theorem and include mean, median and mode. The central limit theorem works off the principle that a sample population has a mean of μ and standard deviation σ. Measure of variability are based on variance which include deviation, variance, kurtosis and skewness. Kurtosis represents the tailedness of the data, higher values mean a narrow peak and lower values mean a wider peak on data distribution plots. Skewness represents the symmetry of the data. Negative skew means the tail is to the left said of the distribution, positive values mean the tail is to the right side of the distribution. The closer the skewness value is to 0, the more symmetrical the data is said to be.

Boxplots are useful visualizations for checking for outliers and visualizing the spread of data between the different quartiles. Figure 1 below shows the boxplot for average weekly hours worked and average hourly earnings in the construction industry. From the plot there are a small number of outliers present for the hourly earnings and no outliers present for the average hours worked. The mean of the hourly earnings is located towards the first quartile more than the centre of the box. The whiskers of the box are also significantly longer for the hourly earnings than the hours worked indicating a larger spread of data which is supported by the presence of outliers. The displot shows us the distribution of the data from which we can conclude the data is approximately normal. From the plot it can be said the average earnings per hour data is slightly positively skewed. This is confirmed by the fact the mean is greater than median. The average hours worked is negatively skewed which is supported by the median being greater than the mean.

Figure Boxplot of hours worked/ hourly earnings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Median | Skew | Kurtosis |
| Average earnings per hour | 8.83 | 7.57 | 0.69 | -0.20 |
| Average hours worked | 43.82 | 44.70 | -1.00 | 0.04 |

For the production index in construction dataset the median is greater than the mean. There are a significant number of outliers present in the data, mainly at the upper end. The mean line in the boxplot is located toward the upper quartile. From the histplot it can be said the data is not normally distributed. The standard deviation is very large at 65.57. This is an indication of XX.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Median | Skew | Kurtosis |
| Production Index | 83.89 | 93.3 |  |  |

## Inferential Statistics

Inferential statistics uses analytical tools. It is useful for drawing conclusions about the population. In order for inferential statistics to be meaningful the sample taken should be representative of the entire population.wh

A number of inferential statistics tests were used to compare Ireland with other countries.

* T-test is used to compare the means of two groups and is frequently used in hypothesis testing to gauge whether two groups are similar (Bevans, 2022). T-test is limited as it is only useful when comparing two groups at most. T-test is a parametric test which makes assumptions about the data. These assumptions include both groups are independent and normally distributed.
* Kruskal-Wallis H-test
* Mann Whitney U-test
* Wilcoxon Signed-Rank Test
* ANOVA

When performing these tests, it was decided to consider the number of construction workers in Ireland against that of the number of construction workers in New Zealand. This comparison was chosen as both countries have similar populations.

|  |  |  |
| --- | --- | --- |
| **Inferential Statistic Test** | **Statistic Value** | **P value** |
| Paired T-test | -14.311 | 0.0001 |
| Kruskal-Wallis H-test | 12.500 | 0.002 |
| Mann Whitney U test | 0.000 | 0.008 |
| Wilcoxon Signed-Rank Test | 0.000 | 0.062 |
| ANOVA | 121.000 | 4.18e-06 |

# Machine Learning Models

## Classification

## Regression

## Clustering and PCA

## Sentiment Analysis

# Programming

## Data structures

## Data Manipulation