A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Felipe Luiz Bauer Steudel |
| *Student Number* | 2024153 |
| *Module Title* | Statistical Techniques for Data Analysis |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 21-Apr-2024 |
| *Date of Submission* | 23-Apr-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

[Task 1 – Data Analysis 1](#_Toc164777886)

[Data Source 1](#_Toc164777887)

[Analysis Variables 1](#_Toc164777888)

[DC\_POWER 1](#_Toc164777889)

[AC\_POWER 1](#_Toc164777890)

[DAILY\_YIELD 1](#_Toc164777891)

[TOTAL\_YIELD 1](#_Toc164777892)

[Conclusion 2](#_Toc164777893)

[Task 2 - Probability (Discrete) 2](#_Toc164777894)

[Two 6s in five rolls 2](#_Toc164777895)

[Probability of two accidents in a specific week 3](#_Toc164777896)

[Task 3 - Dublin Zoo 3](#_Toc164777897)

[Calculating Z-score for at most 85 minutes 4](#_Toc164777898)

[Calculating Z-score for at least 100 minutes 4](#_Toc164777899)

[Calculating Z-score for more than 100 minutes 4](#_Toc164777900)

[References 5](#_Toc164777901)

# Task 1 – Data Analysis

## Data Source

The source of the datasets is Kaggle, which provides public datasets for analysis. Two datasets generated from photovoltaic energy generation and temperature sensor data from two solar plants in India will be used. These datasets contain 34 days of generation data from the year 2020.

**Dataset names**: Plant\_1\_Generation\_Data and Plant\_2\_Generation\_Data;

## Analysis Variables

The analysis of the data below aims to understand the energy production patterns in photovoltaic systems and identify areas for improvement, as well as to assess the performance of the systems in the database.

### DC\_POWER

• Mean of DC\_POWER: 1708.541496702347

• Median of DC\_POWER: 6.05

• Mode of DC\_POWER: 0

• Std Deviation of DC\_POWER: 3222.1812733562097

• Interquartile Range (IQR) of DC\_POWER: 1155.788333333333

### AC\_POWER

• Mean of AC\_POWER: 274.80351054960653

• Median of AC\_POWER: 3.506904761904762

• Mode of AC\_POWER: 0

• Std Deviation of AC\_POWER: 380.18256929405555

• Interquartile Range (IQR) of AC\_POWER: 532.6733333333333

### DAILY\_YIELD

• Mean of DAILY\_YIELD: 3295.433783295412

• Median of DAILY\_YIELD: 2834.8035715

• Mode of DAILY\_YIELD: 0 0.0

• Std Deviation of DAILY\_YIELD: 3035.294425108002

• Interquartile Range (IQR) of DAILY\_YIELD: 5963.67857143

### TOTAL\_YIELD

• Mean of TOTAL\_YIELD: 330382090.0684921

• Median of TOTAL\_YIELD: 7269333.0

• Mode of TOTAL\_YIELD: 0

• Std Deviation of TOTAL\_YIELD: 608570527.4088567

• Interquartile Range (IQR) of TOTAL\_YIELD: 276089566.0

### Conclusion

Mean and Median: Since the means are significantly higher than the medians, it indicates that most values are concentrated around a lower value, but there are some high values that increase the mean.

Mode: Modes being 0 indicate that there is a large number of observations with zero values. This can be explained by periods when the solar panels are not generating energy, such as during nighttime periods. It could also be due to poor weather conditions or maintenance.

Standard Deviation: The values are high relative to the mean, indicating a large spread of values.

Interquartile Range (IQR): The high IQR suggests a wide range of values, similar to the standard deviation, indicating significant data dispersion.

# Task 2 - Probability (Discrete)

## Two 6s in five rolls

What is the probability of rolling exactly two 6s in five rolls of a fair die?

Binomial distribution formula:

P(X = k) = pk ⋅ (1− p)n−k

Where:

* *n* is the total number of trials.
* *k* is the number of successes (number of 6s).
* *p* is the probability of a single success.

Calculating:

P(X = 2) = ⋅ p2 ⋅ (1− p)5−2

Where:

* is the binomial coefficient.
* p is the probability of rolling a 6, which is
* 2 is the number of successes.
* 5 is the number of rolls.

Calculating the binomial coefficient:

= = = 10

Substituting in the formula:

P(X = 2) = 10 ⋅ 2 ⋅ 3

P(X = 2) = 10 ⋅ ⋅

P(X = 2) =

P(X = 2) = 0.1607

The probability of rolling exactly two 6s in five rolls of a fair die is approximately **0.1607** or **16.07%.**

## Probability of two accidents in a specific week

The number of industrial injuries on average per working week in a factory is 0.75. Assuming that the distribution of injuries follows a Poisson distribution, find the probability that in a particular week there will be no more than two accidents.

Poisson distribution formula:

Where:

* *e* is the base of the natural logarithm (approximately 2.71828).
* λ (lambda) is the mean occurrences of events in a fixed interval.
* *k* is the number of events we are interested in.

Substituting in the formula:

The probability that in a specific week there are no more than two accidents is **0.1329** or **13.29%.**

# Task 3 - Dublin Zoo

The time a person spends at Dublin Zoo is Normally distributed with a mean of 90 minutes and a standard deviation of 10 minutes.

Using this distribution, answer the following:

1. If a visitor is selected at random, find the probability that they will spend at most 85 minutes visiting the zoo.
2. If a visitor is selected at random, find the probability that they will spend at least 100 minutes visiting the zoo.
3. Given that you know that a particular visitor has spent longer than average visiting the Zoo, what is the probability that they have spent more than 100 minutes there?

Formula de Z-Score:

Where:

* *X* is the value we want.
* *μ* is the mean.
* *σ* is the standard deviation

## Calculating Z-score for at most 85 minutes

Finding the corresponding probability using a Z-table:

The probability of a visitor spending at most 85 minutes in the zoo is approximately **0.3085** or **30.85%.**

## Calculating Z-score for at least 100 minutes

Finding the corresponding probability using a Z-table, then subtracting this probability from 1:

The probability of a visitor spending at least 100 minutes in the zoo is approximately **0.1587** or **15.87%.**

## Calculating Z-score for more than 100 minutes

Finding the corresponding probability using a Z-table, then subtracting this probability from 1:

The probability of a visitor spending more than 100 minutes in the zoo is approximately 0.1562 or 15.62%.

# References

Khbuli, K., (2024) Statistical techniques for Data Analysis - Lecture 1, Study Material, CCT College Dublin, Dublin.

Khbuli, K., (2024) Statistical techniques for Data Analysis - Week 2, Study Material, CCT College Dublin, Dublin.

Khbuli, K., (2024) Probability - Lecture 3, Study Material, CCT College Dublin, Dublin.

McQuaid, D., (2024) How to Deal with Missing Data in Python. Study Material, CCT College Dublin, Dublin.