IT3212 Høst 2025

# Øving 1: Data Preprocessing

Gruppenummer: 37

|  |  |  |
| --- | --- | --- |
| NTNU epost | Etternavn | Fornavn |
|  |  |  |
| oddvarhf@stud.ntnu.no | Oddvar | Hanevik Folkestad |
| @stud.ntnu.no |  |  |
| @stud.ntnu.no |  |  |
| @stud.ntnu.no |  |  |

[Øving 1: Data Preprocessing 1](#_Toc208334031)

[Task 1 2](#_Toc208334032)

[Task 2) 5](#_Toc208334033)

[Task 3) 6](#_Toc208334034)

[Task 4) 8](#_Toc208334035)

# Task 1

A)

Et bilde som inneholder tekst, skjermbilde

KI-generert innhold kan være feil.

**Figure 1. First rows, summary statistics, and data types for each column.**

B)

Et bilde som inneholder tekst, diagram, Plottdiagram, skjermbilde

KI-generert innhold kan være feil.

**Figure 2. Original versus log-transformed data.**

To identify outliers as requested in the assignment, I chose to examine the data distribution. As we can see in Figure 2, the histogram is highly right-skewed, so I chose to log transform it to make it normally distributed. I did this so I could use Z-score (Z>3) to find outliers. This works perfectly since all harvest data etc. is positive.

Et bilde som inneholder tekst, skjermbilde, Font

KI-generert innhold kan være feil.

**Figure 3. Unique values in categorical columns.**

****



**Figure 4. Outliers.**

** **



**Figure 5. Missing values.**

** **

# Task 2)

Et bilde som inneholder tekst, Font, skjermbilde, hvit

KI-generert innhold kan være feil.

**Figure 6. Missing values removed.**

** **

Justification for Handling Missing Data

For this agricultural dataset, I chose deletion as the method for handling missing values, which involves removing all rows that contain missing values in the Value column. This decision is based on key factors that make this approach the most appropriate for the specific characteristics of this dataset.

The missing data pattern in this dataset shows that only 6.8% of the total records contain missing values, all of which occur in the *Value* column containing numerical agricultural measurements. Removing these rows will retain over 93% of the dataset, leaving a sufficiently large sample for robust analysis.

Et bilde som inneholder tekst, skjermbilde, meny, Font

KI-generert innhold kan være feil.

**Figure 7. Missing values per country (top 20 countries with missing values).**

** **

While a few countries such as Palestine (47.6% missing) exhibit very high proportions of missingness, the issue is not concentrated in just a handful of cases. Instead, it affects 237 countries, suggesting that missingness is broadly distributed across the dataset. This means that, at a global level, removing rows with missing values will not distort the overall representativeness of the dataset.

This systematic pattern suggests that the missing data reflects real-world limitations in data collection infrastructure rather than random measurement errors or data loss. Since the missing values are not randomly distributed, imputation methods that assume Missing At Random would be inappropriate and could introduce bias into the analysis. For example, in Afghanistan large gaps in the data — particularly between 1961 and 2014 — coincide with periods of prolonged conflict and political instability, which likely disrupted the country’s capacity to collect and report agricultural statistics. This illustrates that the missingness is driven by context. We also cant know

Et bilde som inneholder tekst, skjermbilde

KI-generert innhold kan være feil.

**Figure 8. Missing data across years.**

** **

Agricultural data is characterized by high natural variance due to factors such as weather conditions, economic policies, technological changes, and environmental shifts that vary significantly both across countries and over time. A single year may see exceptional yields, while the next year might suffer from droughts, floods, or pest outbreaks. This inherent volatility means that temporal imputation methods like forward or backward filling are inappropriate, as they assume continuity and smooth trends that do not reflect the true dynamics of agricultural systems.

# Task 3)

As mentioned in Task 2, I chose to handle outliers by log-transforming the values. This provides a more uniform distribution, reduces skewness in the data, and makes extreme values less influential on the analyses. An alternative could have been to combine capping and log-transformation. However, as shown in the figures (figure reference), and given the nature of the dataset, agricultural data can naturally contain real extreme values (for example, exceptionally good or very poor harvest years). By capping at the 99% level, there is a risk of removing this important variation.

I addition to the log-transformation, I chose to remove observations with a Z-score greater than 3. Such values are more than three standard deviations away from the mean, which makes them highly extreme and unlikely to represent the general distribution of agricultural data (maybe wrong reporting so on so on). Removing these 0.40% of the records ensures that the analysis is not distorted by extremely rare outliers.

Et bilde som inneholder tekst, diagram, Plottdiagram, skjermbilde

KI-generert innhold kan være feil.

**Figure 9. Comparison of distributions.**

** **

Et bilde som inneholder tekst, Font, skjermbilde, sort

KI-generert innhold kan være feil.

**Figure 10. Removal of outliers.**

** **

# Task 4)

Et bilde som inneholder skjermbilde, tekst

KI-generert innhold kan være feil.

Et bilde som inneholder mønster, stoff

KI-generert innhold kan være feil.