Outliers

UNIVERSITETET
I OSLO

Introduction

Outliers are data points that are significantly different from other data points in a dataset. They are observations that stand out from the rest of the data, and can have a significant impact on statistical analyses.

Definitions of Outliers

Outliers can be defined in various ways depending on the context of the data. One common way to define outliers is using the interquartile range (IQR) and the concept of a “fence”. The IQR is the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset, and the fence is a boundary to identify outliers.

- **Mild outliers**: Data points that fall outside the range of 1.5 times the IQR from the upper or lower quartile.
- **Extreme outliers**: Data points that fall outside the range of 3 times the IQR from the upper or lower quartile.

Why Outliers Are a Problem

Outliers can cause significant problems because they can distort the results of statistical analyses. For example, the mean and standard deviation of a dataset can be affected by outliers, making it difficult to draw meaningful conclusions from the data. Additionally, machine learning models can be sensitive to outliers and produce suboptimal predictions.

What Can Be Done About Outliers

There are various techniques that can be used to deal with outliers in data science. Some of the techniques are:

- **Remove outliers**: Remove the data points that are identified as outliers from the dataset. This can be done by using the fences discussed earlier.
- **Replace outliers**: Replace the data points that are identified as outliers with another value, e.g. a max or min value.
- **Transform data**: Apply a transformation to the data to make it more normally distributed, reducing the impact of outliers. For example, a log transformation can be applied to data that has a skewed distribution.
- **Use robust statistical methods**: Use statistical methods that are less sensitive to outliers, such as the median and weighted mean.

Missing Data

UNIVERSITETET
I OSLO

Introduction

- **Definition of missing data:** Missing data refers to the absence of values for one or more variables in a dataset.
- **Importance of addressing missing data in data science:** Missing data can lead to biased or inaccurate results, making it important to address in data science.
- **Business problem behind missing data:** The reasons why data is missing can provide insight into the problem being studied.

Business Problem Behind Missing Data

- **Understanding why data is missing:**
  - **Missing Completely at Random (MCAR):** Data is MCAR when the probability of a value being missing is the same for all observations, regardless of the values of other variables in the dataset. In other words, missingness is completely random and unrelated to any aspect of the data.
  - **Missing at Random (MAR):** Data is MAR when the probability of a value being missing is related to other variables in the dataset, but not to the missing value itself. In other words, the missingness is related to other variables in the dataset, but is independent of the missing value.
  - **Missing Not at Random (MNAR):** Data is MNAR when the probability of a value being missing is related to the missing value itself. In other words, the missingness is related to the unobserved value, which can introduce bias into the data if not accounted for.
- **Types of missing data and their implications on business decisions:** The type of missing data can impact the conclusions drawn from data analysis and the decisions made based on those conclusions.

Methods to Address Missing Data

- **Dropping data:**
  - **Complete Case Analysis (CCA):** CCA involves dropping all observations that have missing values. This method only works when the probability of missingness is MCAR.
- **Imputing data:**
  - **Mean imputation:** Mean imputation involves replacing missing values with the mean of the observed values for that variable. This method only works when the probability of missingness is MCAR or MAR.
  - **Regression imputation:**
    - **Simple Linear Regression:** Simple linear regression involves using a linear model to predict missing values based on the observed values of a related variable. This method works when the probability of missingness is MAR.
    - **Multiple Linear Regression:** Multiple linear regression involves using a linear model to predict missing values based on the observed values of multiple related variables. This method works when the probability of missingness is MAR.

Complete Case Analysis (CCA)

    - **Multiple imputation:** Multiple imputation involves creating multiple plausible imputations for missing values based on a statistical model that accounts for the uncertainty in the imputed values. This method works when the probability of missingness is MAR.
- **Definition of CCA:** CCA involves dropping all observations that have missing values.
- **Advantages and disadvantages of CCA:**
    - **Advantages:** CCA is simple and easy to implement.
    - **Disadvantages:** CCA can result in loss of data and reduced statistical power. It only works when the probability of missingness is MCAR.
- **Assumptions of CCA:**
    - **MCAR:** CCA assumes that the probability of missingness is MCAR.
- **When to use CCA:** CCA should only be used when the probability of missingness is MCAR.

Mean Imputation

- **Definition of mean imputation:** Mean imputation involves replacing missing values with the mean of the observed values for that variable.
- **Advantages and disadvantages of mean imputation:**
  - **Advantages:** Mean imputation is simple and easy to implement.
  - **Disadvantages:** Mean imputation can introduce bias and underestimate the standard error of the estimate. It only works when the probability of missingness is MCAR or MAR.
- **Assumptions of mean imputation:**
  - **MCAR or MAR:** Mean imputation assumes that the probability of missingness is MCAR or MAR.