STATISTICS FOR DATA SCIENCE

Data Cleaning

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Data Cleaning

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

STATISTICS FOR DATA SCIENCE

What is Data Cleaning?

Data cleaning or cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Source: www.towardsdatascience.com

Remove Unwanted Observation

Duplicate observations

Duplicate observations most frequently arise during **data collection**, such as when you:

- Combine datasets from multiple places
- Scrape data
- Receive data from clients/other departments

Source : https://elitedatascience.com/data-cleaning

Remove Unwanted Observation

Irrelevant observations

Irrelevant observations are those that don't actually fit the **specific problem** that you're trying to solve.

For example, if you were building a model for Single-Family homes only, you wouldn't want observations for Apartments in there.

This is also a great time to review your charts from Exploratory Analysis. You can look at the distribution charts for categorical features to see if there are any classes that shouldn't be there.

Checking for irrelevant observations **before engineering features** can save you many headaches down the road.

Source: https://elitedatascience.com/data-cleaning

Uninformative / Repetitive

Sometimes one feature is uninformative because it has too many rows being the same value.

We can create a list of features with a high percentage of the same value.

We need to understand the reasons behind the repetitive feature. When they are genuinely uninformative, we can toss them out.

Source: www.towardsdatascience.com

Solutions to Clean

Irrelevant

Data needs to provide valuable information.

We need to skim through the features to identify irrelevant ones.

For example, a feature recording the temperature in Toronto doesn't provide any useful insights to predict Russian housing prices.

When the features are not serving the goal, we can remove them.

Source: www.towardsdatascience.com

Solutions to Clean

Duplicates

The duplicate data is when copies of the same observation exist.

Sometimes it is better to remove duplicate data based on a set of unique identifiers.

For example, the chances of two transactions happening at the same time, with the same square footage, the same price, and the same build year are close to zero.

Source: www.towardsdatascience.com

STATISTICS FOR DATA SCIENCE

Fix Structural Errors

The next bucket under data cleaning involves fixing structural errors.

Structural errors are those that arise during measurement, data transfer, or other types of **"poor housekeeping.“**

For instance, you can check for **typos** or **inconsistent capitalization.**

This is mostly a concern for categorical features, and bar plots can be used to check.

Source : https://elitedatascience.com/data-cleaning

Solutions to Clean

It is also crucial to have the dataset follow specific standards to fit a model.

We need to explore the data in different ways to find out the inconsistent data.

Capitalization

Inconsistent usage of upper and lower cases in categorical values is a common mistake.

To avoid this, we can put all letters to lower cases (or upper cases).

Source: www.towardsdatascience.com

Solutions to Clean

Formats

Another standardization we need to perform is the data formats.

One example is to convert the feature from string to DateTime format.

Categorical Values

A categorical feature has a limited number of values.

Sometimes there may be other values due to reasons such as typos. We can set criteria to convert these typos to the correct values.

Source: www.towardsdatascience.com

STATISTICS FOR DATA SCIENCE

Filter Unwanted Outliers

Outliers can cause problems with certain types of models.

In general, if you have a **legitimate** reason to remove an outlier, it will help your model's performance.

However, outliers are **innocent until proven guilty.**

You should never remove an outlier just because it's a "big number." That big number could be very informative for your model.

We can't stress this enough: you must have a good reason for removing an outlier, such as suspicious measurements that are unlikely to be real data.

Source : https://elitedatascience.com/data-cleaning

Solutions to Clean

Depending on whether the feature is numeric or categorical, we can use different techniques to study its distribution to detect outliers.

Histogram / Boxplot

When the feature is numeric, we can use a histogram and box plot to detect outliers.

Descriptive Statistics

Also, for numeric features, the outliers could be too distinct that the box plot can’t visualize them. Instead, we can look at their descriptive statistics.

Source: www.towardsdatascience.com

Solution to Clean

Bar Chart

When the feature is categorical. We can use a bar chart to learn about its categories and distribution.

While outliers are not hard to detect, we have to determine the right solutions to handle them.

The methods of handling outliers are somewhat similar to missing data. We either drop or adjust or keep them.

Source: www.towardsdatascience.com

Handle Missing Data

Missing data is a deceptively tricky.

First, just to be clear, **cannot simply ignore missing values in your dataset.**

The two most commonly recommended ways of dealing with missing data are,

Dropping observations that have missing values.

Imputing the missing values based on other observations.

Source : https://elitedatascience.com/data-cleaning

Solution to clean

Missing Data Heatmap

When there is a smaller number of features, we can visualize the missing data via heatmap.

Missing Data Percentage List

When there are many features in the dataset, we can make a list of missing data % for each feature.

Missing Data Histogram

Missing data histogram is also a technique.

Source: www.towardsdatascience.com

Solution to clean

Impute the Missing

When the feature is a numeric variable, we can conduct missing data imputation.

We replace the missing values with the average or median value from the data of the same feature that is not missing.

When the feature is a categorical variable, we may impute the missing data by the mode (the most frequent value).

Source: www.towardsdatascience.com

Solution to clean

Replace the Missing

For categorical features**,** we can add a new category with a value such as "_MISSING_".

For missing numeric data, you should **flag and fill** the values

Then, fill the original missing value with 0 just to meet the technical requirement of no missing values.

Source: www.towardsdatascience.com

STATISTICS FOR DATA SCIENCE

Demonstration on Data Cleaning

Click here for Demonstration of Data Cleaning

THANK YOU

Prof. Uma D
Prof. Silviya Nancy J
Prof. Suganthi S

Department of Computer Science and Engineering