STA 3180 Statistical Modelling: Data Wrangling

STA 3180 Statistical Modelling: Data Wrangling Lecture Notes

Data wrangling is the process of transforming and mapping data from one “raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. It is a key part of the data science process, and involves gathering data from multiple sources, cleaning and validating it, and transforming it into a format that is suitable for analysis.

Key Concepts

* **Data Gathering**: The process of collecting data from multiple sources, such as databases, spreadsheets, text files, web services, etc.

* **Data Cleaning**: The process of removing or correcting inaccurate, incomplete, or irrelevant data. This includes identifying and correcting errors, filling in missing values, and removing duplicate data.

* **Data Validation**: The process of ensuring that the data is accurate and complete. This includes verifying that the data is consistent with the source, checking for outliers, and ensuring that the data is in the correct format.

* **Data Transformation**: The process of transforming data from one form to another. This includes converting data from one format to another, normalizing data, and creating new variables.

Definitions

* **Raw Data**: Data that has not been processed or transformed in any way. It is usually in its original form, as collected from the source.

* **Clean Data**: Data that has been processed and transformed in order to make it more suitable for analysis. This includes removing errors, filling in missing values, and transforming the data into a format that is more suitable for analysis.

* **Outliers**: Data points that are significantly different from the rest of the data. Outliers can be caused by errors in the data, or they can indicate interesting patterns in the data.

* **Normalization**: The process of transforming data so that it has a mean of 0 and a standard deviation of 1. This is often done to make it easier to compare different datasets.

Coding Examples

```
Start of Code
# Import necessary packages
import pandas as pd
import numpy as np
```

```
# Read in the data
data = pd.read_csv('data.csv')
# Remove outliers
data = data[(np.abs(data - data.mean()) < (3 * data.std()))]
# Fill in missing values
data = data.fillna(data.mean())
# Normalize the data
data = (data - data.mean()) / data.std()
End of Code
```

Practice Multiple Choice Questions

Q1. What is the process of transforming data from one form to another called?

A. Data Gathering

B. Data Cleaning

C. Data Validation

D. Data Transformation

Answer: D. Data Transformation