Resource

Checklist 1: Data preparation

Introduction

As you will learn in this program, the fundamental goal of data science is to answer questions. Although this question component is critical and defines the scope of any data analysis, it is also useful to have a strategy in place to prepare and evaluate data for this application.

To assist you with developing a strategy, several resources will be provided to you in this program to supplement the IDE activities. These resources will guide you on what information to record in the IDE and provide a checklist to ensure that all steps have been followed. This is the first of these resources, which specifically describes the steps involved in preparing and evaluating data.

Process overview

This framework contains five steps, which are briefly described below. These steps are not straightforward, and often require multiple iterations to obtain a structured, tidy, and complete data set.

Step 1: Prepare the coding environment

The first step in the framework is to record important details, such as:

- The author and date of evaluation.
- The data source and unit of observation.
- Any evaluation criteria.

This will ensure that third parties and future users (including yourself) understand the decisions that were made, the reasoning behind those decisions, and the findings from the analysis.

Recording the details is followed by loading the required statistical packages along with the data being evaluated.

Step 2: Summarize the data

This step defines the dimensions of the data set; in other words, the total number of rows (observations) and the total number of columns (variables or attributes). It also defines which variables are numerical and which are categorical in nature, along with the total number of variables in each case. This is important for gaining a better sense of the types of data included in the data set.

- **Numerical data:** The distribution of numerical data can be examined by extracting measures of centrality – such as the mean, median, minimum value, and maximum value – as well as measures of skewness and kurtosis – such as the standard deviation and range. This information can assist in identifying anomalies within the data set that may need to be removed, and in providing insight into whether the data follows a normal distribution – which has implications for the types of statistical tests that can be used.

- **Categorical data:** Categorical data is evaluated by examining the types of categories within each variable. For example, are the categories nominal, ordinal, or binary in nature? Additionally, information can be extracted on the frequency of observing each category within each variable. This process will help to identify whether there are any categories with missing features and whether these categories need to be recoded or removed.

- **Dimensional data:** It is rare to observe a single moment in time or space; however, there must be enough observations at each point in order to provide a meaningful story. In real estate, it is particularly important to understand both the granularity of an event and where it is occurring.

Step 3: Clean and wrangle the data

Wide and disparate data – such as real estate data – needs to be pulled or wrangled together in order to tell an accurate story. This step of the framework provides an opportunity to convert (wrangle) data into different formats and to identify anomalies – including missing (NA) terms. These anomalies need to be resolved, either by:

- Recoding the data (completing the data).

- Or – as a last resort – by removing variables or observations (deleting the data).

Step 4: Evaluate the data set

The goal of all prior steps in this framework is to arrive at a tidy and complete data set.

- A tidy data set has observational units in every row or every column.

- A complete data set is one where each observational unit is either a numerical value or an encoded categorical value.

A data set can be evaluated by ensuring that each data record in the set has observational units – whether numerical or categorical – in each row or column.

This tidy and complete data set – typically formatted as a matrix of numbers – is then ready to be used in a regression analysis and to answer any real estate questions.

Step 5: Widen the data set using joins (if required)

Important variables can be added to the data set by using a common ID or key to join data. Spatial joins are particularly useful in real estate data and typically use geospatial coordinates – specifically latitude and longitude. This joining process may generate additional anomalies, which will need to be resolved before the data is used for downstream analysis.

Note: These steps have been structured into a checklist shown on the following page.

Checklist 1: Data preparation

1. Prepare the coding environment:

    a. Provide a title, date, and information on the author and data – such as the data source and the unit of observation. €

    b. Provide any important details on criteria for data preparation and evaluation; for example, criteria for resolving anomalies. €

    c. Check the working directory to ensure that these details are correct. €

    d. Load the relevant statistical packages. €

    e. Load the relevant data. €

2. Summarize the data to understand its characteristics:

    a. Define the dimensions of the data set – specifically the number of observations (rows) and variables or attributes (columns). €

    b. Note which variables contain missing (NA) values. €

    c. Note which variables contain numerical data (continuous or discrete), categorical data (nominal, ordinal, or binary), or other types of data. €

    d. Examine the summary statistics:

        i. Continuous numerical data: mean, median, minimum value, maximum value, standard deviation, and quartiles €

        ii. Discrete numerical and categorical data: category names and counts (frequency of each category within the variable) €

3. Clean and wrangle the data to create a tidy and complete data set:

    a. Recode the missing (NA) terms. €

    b. Convert data types where required. €

    c. Identify and resolve anomalies. €

    d. Remove data, but only if absolutely necessary. €

4. Evaluate the data set:

a. Check whether the data set is tidy. €

b. Check whether the data set is complete; in other words, whether all missing values (NAs) have been removed. €

If the data set is not yet tidy and complete, repeat Step 3.

5. Join data (if required):

a. View the data frames. €

b. Identify the join keys and common values. €

In cases where there are common values with incompatible data types or formats, repeat Step 3.

c. Join data frames using inner joins, left joins, right joins, and full joins. €

d. Confirm the join by examining the new data set. €

e. Create new variables using spatial joins (if required). €

f. Re-evaluate the new data set by repeating Step 4. €

If the new data set is not yet tidy and complete, repeat Step 3.