



.....

.....

.....

---

# Contents

**Chapter 1. Introduction to Data Preparation . . . . . 1**

**Chapter 2. Validation Rules . . . . . 3**

**Chapter 3. Validate Data . . . . . 7**

**Chapter 4. Automated Data Preparation 11**



---

## Chapter 1. Introduction to Data Preparation

1.1.1. Introduction to Data Preparation

1.1.2. Data Preparation Process

---

### Usage of Data Preparation Procedures

1.2.1. Data Cleaning

- Data cleaning is the process of identifying and correcting errors in the data.

- Data cleaning is a critical step in the data preparation process.

- Data cleaning is a time-consuming process.

1.2.2. Data Transformation



---

## Chapter 2. Validation Rules

- **Validation Rules** are used to ensure that data entered into a field meets certain criteria.
- **Validation Rules** can be used to enforce data integrity and accuracy.
- **Validation Rules** can be used to restrict the types of data that can be entered into a field.
- **Validation Rules** can be used to ensure that data is entered in a specific format.
- **Validation Rules** can be used to ensure that data is entered within a specific range.
- **Validation Rules** can be used to ensure that data is entered as a unique value.
- **Validation Rules** can be used to ensure that data is entered as a required value.
- **Validation Rules** can be used to ensure that data is entered as a specific value.
- **Validation Rules** can be used to ensure that data is entered as a specific length.
- **Validation Rules** can be used to ensure that data is entered as a specific type.





- **Wiederholungsfragen** sind Fragen, die in der Vorlesung oder in den Vorlesungsmaterialien bereits behandelt wurden. Diese Fragen sind oft einfacher zu beantworten, da sie sich auf das Gelernte beziehen.
- **Wiederholungsfragen** sind Fragen, die in der Vorlesung oder in den Vorlesungsmaterialien bereits behandelt wurden. Diese Fragen sind oft einfacher zu beantworten, da sie sich auf das Gelernte beziehen.
- **Wiederholungsfragen** sind Fragen, die in der Vorlesung oder in den Vorlesungsmaterialien bereits behandelt wurden. Diese Fragen sind oft einfacher zu beantworten, da sie sich auf das Gelernte beziehen.
- **Wiederholungsfragen** sind Fragen, die in der Vorlesung oder in den Vorlesungsmaterialien bereits behandelt wurden. Diese Fragen sind oft einfacher zu beantworten, da sie sich auf das Gelernte beziehen.



---

## Chapter 3. Validate Data

As you saw in the previous chapter, you can use the `validate` method to validate data against a schema. The `validate` method returns a `ValidationError` object if the data is invalid. The `ValidationError` object has a `message` property that contains a human-readable error message.

Let's look at an example of how to use the `validate` method. We'll create a schema for a `Person` object and then use the `validate` method to validate a `Person` object against the schema. The schema is defined in the following code block:

```
const schema = {
  type: 'object',
  properties: {
    name: { type: 'string', required: true },
    age: { type: 'number', required: true },
    email: { type: 'string', required: true }
  }
};
```

---

## Validate Data Basic Checks

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

- **Check for missing values** (e.g., NA, NaN, etc.)
- **Check for outliers** (e.g., values that are significantly higher or lower than the rest of the data)
- **Check for duplicates** (e.g., rows that are identical to each other)
- **Check for data type consistency** (e.g., all values in a column should be of the same type)
- **Check for range constraints** (e.g., values should be within a certain range)
- **Check for unique values** (e.g., all values in a column should be unique)
- **Check for data distribution** (e.g., the data should follow a normal distribution)
- **Check for data skewness** (e.g., the data should not be skewed to the left or right)
- **Check for data kurtosis** (e.g., the data should not have a heavy tail)
- **Check for data correlation** (e.g., variables should not be highly correlated)
- **Check for data multicollinearity** (e.g., variables should not be highly correlated with each other)
- **Check for data heteroscedasticity** (e.g., the variance of the data should not change with the mean)
- **Check for data non-normality** (e.g., the data should not follow a normal distribution)
- **Check for data non-independence** (e.g., the data should not be correlated with each other)
- **Check for data non-stationarity** (e.g., the data should not have a constant mean and variance over time)
- **Check for data non-ergodicity** (e.g., the data should not have a constant mean and variance over time)
- **Check for data non-ergodicity** (e.g., the data should not have a constant mean and variance over time)

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

- **Check for data non-normality** (e.g., the data should not follow a normal distribution)
- **Check for data non-independence** (e.g., the data should not be correlated with each other)
- **Check for data non-stationarity** (e.g., the data should not have a constant mean and variance over time)
- **Check for data non-ergodicity** (e.g., the data should not have a constant mean and variance over time)

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

---

## Validate Data Single-Variable Rules

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

These checks are performed on the data before it is used in the analysis. They are performed on the data before it is used in the analysis.

- **Check for data non-normality** (e.g., the data should not follow a normal distribution)
- **Check for data non-independence** (e.g., the data should not be correlated with each other)
- **Check for data non-stationarity** (e.g., the data should not have a constant mean and variance over time)
- **Check for data non-ergodicity** (e.g., the data should not have a constant mean and variance over time)

•  $n$  is the number of observations in the dataset

---

## Validate Data Cross-Variable Rules

•  $n$  is the number of observations in the dataset

---

## Validate Data Output

- $n$  is the number of observations in the dataset
- $n$  is the number of observations in the dataset
- $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

- $n$  is the number of observations in the dataset
- $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

---

## Validate Data Save

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

•  $n$  is the number of observations in the dataset

1. The first part of the document is a list of names and titles, including "The Hon. Mr. Justice G. D. C. O'Connell, Chief Justice of the Supreme Court of the State of New South Wales" and "The Hon. Mr. Justice G. D. C. O'Connell, Chief Justice of the Supreme Court of the State of New South Wales".

---

## Chapter 4. Automated Data Preparation

Automated data preparation is a process that uses machine learning algorithms to identify and correct errors in data. This process is often used in conjunction with data mining and data analysis. Automated data preparation can help to improve the accuracy and reliability of data, and can also help to reduce the time and effort required to prepare data for analysis.

Automated data preparation can be used in a variety of ways, including:

- Identifying and correcting errors in data.
- Identifying and removing outliers.
- Identifying and removing duplicate records.
- Identifying and removing missing data.

**Note:** Automated data preparation is a process that uses machine learning algorithms to identify and correct errors in data. This process is often used in conjunction with data mining and data analysis. Automated data preparation can help to improve the accuracy and reliability of data, and can also help to reduce the time and effort required to prepare data for analysis.

Automated data preparation can be used in a variety of ways, including:







... ..

## Adjust Measurement

... ..



1. The first step in the process of identifying field names is to determine the scope of the data. This involves understanding the geographical area covered by the data and the time period over which it was collected. Once the scope is defined, the next step is to identify the specific fields or variables that are being measured.

## Field Names

2. The second step is to identify the specific fields or variables that are being measured. This involves looking at the data and determining what each field represents. For example, if the data is about crop yields, the fields might include the type of crop, the yield per acre, and the total yield for each field.

3. The third step is to identify the units of measurement for each field. This is important because different fields may be measured in different units, and it is necessary to know these units in order to compare and analyze the data. For example, crop yields might be measured in bushels per acre, and total yield might be measured in bushels.

4. The fourth step is to identify the sources of the data. This involves determining where the data was collected and how it was collected. This information is important for understanding the reliability and accuracy of the data. For example, data collected from a government agency might be more reliable than data collected from a private company.

IMPORT

TMS MERGE

TMS

---

## Analysis Tab

*Note:*

1.

...

...

- ...
- ...
- ...

- ...
- ...
- ...
- ...

...

- **Field** is a collection of related data items called **records**.
- **Field** is a collection of related data items called **records**. Each record is a collection of related data items called **fields**.
- **Field** is a collection of related data items called **records**. Each record is a collection of related data items called **fields**. Each field is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**. Each field is a collection of related data items called **records**.

## Fields

Each record is a collection of related data items called **fields**. Each field is a collection of related data items called **records**.

- **Field** is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**.

Each field is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**.

- **Field** is a collection of related data items called **records**.
- **Field** is a collection of related data items called **records**.

*Note*

Each record is a collection of related data items called **fields**.

Each field is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**.

- **Field** is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**.

Each field is a collection of related data items called **records**.

Each record is a collection of related data items called **fields**.

Each field is a collection of related data items called **records**.

- $\frac{1}{2} \int_{-\infty}^{\infty} \delta(x) dx = 1$  (normalization)
- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)
- $\int_{-\infty}^{\infty} \delta(x) dx = 1$  (normalization)
- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)
- $\int_{-\infty}^{\infty} \delta(x) dx = 1$  (normalization)
- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)

**Note**

- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)
- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)
- $\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$  (sifting property)

- **Verfahren zur Ermittlung**

Die Ermittlung der **Wahlberechtigung** erfolgt durch die **Eintragung** in das **Wahlverzeichnis**.  
Die **Eintragung** erfolgt durch die **Eintragung** in das **Wahlverzeichnis**.  
Die **Eintragung** erfolgt durch die **Eintragung** in das **Wahlverzeichnis**.



1. The first step is to identify the problem or goal.

2. Next, you need to gather information.

3. Then, you should analyze the information and develop a plan.

*Note:* The following steps are part of the process.

4. The next step is to implement the plan.

- Identify the problem or goal.
- Gather information.
- Analyze the information and develop a plan.
- Implement the plan.

## Action Details

5. The first action detail is to identify the problem or goal.

6. The second action detail is to gather information.

- Identify the problem or goal.

*Note:* The following steps are part of the process.

7. The next step is to implement the plan.

8. Then, you should analyze the information.

9. The next step is to develop a plan.

- Identify the problem or goal.

10. The next step is to implement the plan.

11. Then, you should analyze the information.

- Identify the problem or goal.
- Gather information.
- Analyze the information and develop a plan.
- Implement the plan.

12. The next step is to implement the plan.

13. Then, you should analyze the information.

14. The next step is to develop a plan.

- Identify the problem or goal.
- Gather information.

- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\cos(x)$  when  $x = 0$
- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\sin(x)$  when  $x = \frac{\pi}{4}$
- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\sin(x)$  when  $x = \frac{3\pi}{4}$

$\frac{1}{\sqrt{2}}$  is the maximum value of  $\cos(x)$  when  $x = 0$

$\frac{1}{\sqrt{2}}$  is the maximum value of  $\sin(x)$  when  $x = \frac{\pi}{4}$  and  $x = \frac{3\pi}{4}$

- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\cos(x)$  when  $x = 0$
- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\sin(x)$  when  $x = \frac{\pi}{4}$  and  $x = \frac{3\pi}{4}$
- $\frac{1}{\sqrt{2}}$  is the maximum value of  $\sin(x)$  when  $x = \frac{3\pi}{4}$

$\frac{1}{\sqrt{2}}$  is the maximum value of  $\cos(x)$  when  $x = 0$  and  $x = \pi$







...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

---

## Identify Unusual Cases Output

...the ... of ...

- ...
- ...
- ...
- ...

...the ... of ...

...the ... of ...

- ...
- ...
- ...
- ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...

...the ... of ...



MLWEIGHT CRITERIA

*Command Syntax Reference*



---

## Chapter 6. Optimal Binning

...the optimal binning problem is NP-hard. This means that there is no known algorithm that can solve the problem in polynomial time. In other words, the time required to solve the problem grows exponentially with the number of data points.

- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.

...the optimal binning problem is NP-hard. This means that there is no known algorithm that can solve the problem in polynomial time. In other words, the time required to solve the problem grows exponentially with the number of data points.

...the optimal binning problem is NP-hard. This means that there is no known algorithm that can solve the problem in polynomial time. In other words, the time required to solve the problem grows exponentially with the number of data points.

...the optimal binning problem is NP-hard.

- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.

...the optimal binning problem is NP-hard.

1. The optimal binning problem is NP-hard.
2. The optimal binning problem is NP-hard.
3. The optimal binning problem is NP-hard.

...the optimal binning problem is NP-hard. This means that there is no known algorithm that can solve the problem in polynomial time. In other words, the time required to solve the problem grows exponentially with the number of data points.

---

### Optimal Binning Output

...the optimal binning problem is NP-hard.

- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.
- The optimal binning problem is NP-hard.

...the optimal binning problem is NP-hard.



---

## OPTIMAL BINNING Command Additional Features

2020-01-28 10:15:00 AM CST

- [OPTIMAL BINNING Command Additional Features](#) CRITERIA

[OPTIMAL BINNING Command Syntax Reference](#)



---

## Notices

1. The first notice is a general notice of the meeting of the Board of Directors of the Corporation, to be held on the 15th day of May, 1910, at the office of the Secretary, at New York, New York, for the purpose of electing directors and officers for the year ending on the 31st day of December, 1910.

2. The second notice is a notice of the meeting of the Board of Directors of the Corporation, to be held on the 15th day of May, 1910, at the office of the Secretary, at New York, New York, for the purpose of electing directors and officers for the year ending on the 31st day of December, 1910.

3. The third notice is a notice of the meeting of the Board of Directors of the Corporation, to be held on the 15th day of May, 1910, at the office of the Secretary, at New York, New York, for the purpose of electing directors and officers for the year ending on the 31st day of December, 1910.

.....

.....  
.....  
.....  
..... 666

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....















Printed in USA