# Working With Missing And Duplicate Data: Takeaways



## Syntax

### IDENTIFYING MISSING VALUES

- Identify rows with missing values in a specific column:

```
missing = df[col_name].isnull()
df[missing]
```

- Calculate the number of missing values in each column:

```
df.isnull().sum()
```

### REMOVING MISSING VALUES

- Drop rows with any missing values:

```
df.dropna()
```

- Drop specific columns:

```
df.drop(columns_to_drop, axis=1)
```

- Drop columns with less than a certain number of non-null values:

```
df.dropna(thresh = min_nonnull, axis=1)
```

### REPLACING MISSING VALUES

- Replace missing values in a column with another value:

```
df[col_name].fillna(replacement_value)
```

### VISUALIZING MISSING DATA

- Use a heatmap to visualize missing data:

```
import seaborn as sns
sns.heatmap(df.isnull(), cbar=False)
```

### CORRECTING DUPLICATE VALUES

- Identify duplicates values:

```
dups = df.duplicated()
df[dups]
```

- Identify rows with duplicate values in only certain columns:

```
dups = df.duplicated([col_1, col_2])
df[dups]
```

- Drop duplicate values. Keep the first duplicate row:

```
df.drop_duplicates()
```

- Drop rows with duplicate values in only certain columns. Keep the last duplicate row:

```
combined.drop_duplicates([col_1, col_2], keep='last')
```

## Concepts

- Missing or duplicate data may exist in a data set for many reasons. Sometimes, they may exist because of user input errors or data conversion issues; other times, they may be introduced while performing data cleaning tasks. In the case of missing values, they may also exist in the original data set to purposely indicate that data is unavailable.
- In pandas, missing values are generally represented by the `NaN` value or the `None` value.
- To handle missing values, first check for errors made while performing data cleaning tasks. Then, try to use available data from other sources (if it exists) to fill them in. Otherwise, consider dropping them or replacing them with other values.

## Resources

- Working with Missing Data