Data Cleaning Walkthrough: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2022

Syntax

- Combine dataframes:

```
z = pd.concat([x,y])
```

- Copy or add columns:

```
survey["new_column"] = survey["old_column"]
```

- Filter to keep columns:

```
survey_fields = ["DBN", "rr_s", "rr_t"]
survey = survey.loc[:,survey_fields]
```

- Add 0s to the front of the string until the string has desired length:

```
zfill(5)
```

- Apply function to Series:

```
data["class_size"]["padded_csd"] = data["class_size"]["CSD"].apply(pad_csd)
```

- Convert a column to numeric data type:

```
data["sat_results"]["SAT Math Avg. Score"] = pd.to_numeric(data["sat_results"]["SAT Math Avg. Score"])
```

Concepts

- A data science project usually consists of either an exploration and analysis of a set of data or an operational system that generates predictions based on data that updates continually.
- When deciding on a topic for a project, it's best to go with something you're interested in.
- In real-world data science, you may not find an ideal dataset to work with.

Resources

- Data.gov
- /r/datasets
- Awesome datasets
- rs.io

Data Cleaning Walkthrough: Combining the Data: Takeaways

Syntax

- Reset the index:

  ```
  class_size.reset_index(inplace=True)
  ```

- Group a dataframe by column:

  ```
  class_size=class_size.groupby("DBN")
  ```

- Aggregate a grouped Dataframe:

  ```
  class_size = class_size.agg(numpy.mean)
  ```

- Display column types:

  ```
  data["ap_2010"].dtypes
  ```

- Perform a left join:

  ```
  combined.merge(data["ap_2010"], on="DBN", how="left")
  ```

- Display the shape of the dataframe (row, column):

  ```
  combined.shape
  ```

- Performing an inner join:

  ```
  combined = combined.merge(data[class_size], on="DBN", how="inner")
  ```

- Fill in missing values:

  ```
  combined.fillna(0)
  ```

Concepts

- Merging data in Pandas supports four types of joins -- `left` , `right` , `inner` , and `outer` .
- Each of the join types dictates how pandas combines the rows.
- The strategy for merging affects the number of rows we end up with.
- We can use one or multiple aggregate functions on a grouped dataframe.

Resources

- Data Cleaning with Python
- Dataframe.groupby()
- agg() documentation

Analyzing and Visualizing the Data

Syntax

- Find correlations between columns in a dataframe:

```
In [ ]: combined.corr()
```

- Specify a plot type using Dataframe.plot():

```
In [ ]: combined.plot.scatter(x='total_enrollment', y='sat_score')
```

- Convert a Pandas series to list:

```
In [ ]: longitudes = combined["lon"].tolist()
```

Concepts

- An r value measures how closely two sequences of numbers are correlated.
- An r value ranges for `-1` to `1` .
- An r value closer to `-1` tells us the two columns are negatively correlated while an r value closer to `1` tells us the columns are positively correlated.
- The r value is also known as Pearson's correlation coefficient.

Resources

- R value (https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)
- pandas.DataFrame.plot() (http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.plot.html)
- Correlation (https://www.mathsisfun.com/data/correlation.html)
- Guess the Correlation (http://guessthecorrelation.com/)