Data Cleaning Walkthrough: Takeaways

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