Data Cleaning Walkthrough: Analyzing and Visualizing the Data: Takeaways

Syntax

- Finding correlations between columns in a dataframe:

```
combined.corr()
```

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

```
combined.plot.scatter(x='total_enrollment', y='sat_score')
```

- Creating a map of New York City:

```
from mpl_toolkits.basemap import Basemap

m = Basemap(

    projection='merc',

    llcrnrlat=40.496044,

    urcrnrlat=40.915256,

    llcrnrlon=-74.255735,

    urcrnrlon=-73.700272,

    resolution='i'

    )

 m.drawmapboundary(fill_color='#85A6D9')

 m.drawcoastlines(color='#6D5F47', linewidth=.4)

 m.drawrivers(color='#6D5F47', linewidth=.4)
```

- Converting a Pandas series to list:

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

- Making a scatterplot using Basemap:

```
m.scatter(longitudes, latitudes, s=20, zorder=2, latlon=True)
```

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 postively correlated.
- The r value is also called Pearson's correlation coefficient.
- Keyword arguenents for `scatter()` method:
    - `s` : Determines the size of the point that represents each school on the map.
    - `zorder` : Determines where the method draws the points on the z axis. In other words, it determines the order of the layers on the map.
    - `latlon` : A Boolean value that specifies whether we're passing in latitude and longitude coordinates instead of x and y plot coordinates.

Resources

- R value
- pandas.DataFrame.plot()
- Correlation
- Guess the Correlation