Working with Missing Data: Takeaways

Syntax

- Replacing matching values with a single value:

```
s.mask(s == var, value_to_replace)
```

- Replacing matching values with corresponding values from a series:

```
s1.mask(s == var, series_to_replace)
```

- A function to create a null matrix

```
def plot_null_matrix(df, figsize=(18,15)):
    # initiate the figure
    plt.figure(figsize=figsize)
    # create a boolean dataframe based on whether values are null
    df_null = df.isnull()
    # create a heatmap of the boolean dataframe
    sns.heatmap(~df_null, cbar=False, yticklabels=False)
    plt.show()
```

- A function to create a null correlation heatmap

```
def plot_null_correlations(df):
    # create a correlation matrix only for columns with at least
    # one missing value
    cols_with_missing_vals = df.columns[df.isnull().sum() > 0]
    missing_corr = df[cols_with_missing_vals].isnull().corr()
    # create a triangular mask to avoid repeated values and make
    # the plot easier to read
    missing_corr = missing_corr.iloc[1:, :-1]
    mask = np.triu(np.ones_like(missing_corr), k=1)
    # plot a heatmap of the values
    plt.figure(figsize=(20,12))
    ax = sns.heatmap(missing_corr, vmin=-1, vmax=1,
                     cmap='RdBu', mask=mask, annot=True)
    # round the labels and hide labels for values near zero
    for text in ax.texts:
        t = float(text.get_text())
        if -0.05 < t < 0.01:
            text.set_text('')
        else:
            text.set_text(round(t, 2))
    plt.show()
```

Concepts

- Imputation is the process of replacing missing values with other values.

- Imputing can be a better option than simply dropping values because you retain more of your original data.
- You might find values for imputation by:
    - Deriving the value from related columns.
    - Using the most common non-null value from a column.
    - Using an placeholder for missing values.
    - Augmenting factual data (e.g. location data) using an external resource.
- Using plots can help identify patterns in missing values which can help with imputation.

Resources

- pandas documentation