Scatter Plots and Correlations: Takeaways

Syntax

- Transforming a Series to datetime:

```
dataframe['column_name'] = pd.to_datetime(dataframe['column_name'])
```

- Rotating x- and y-tick labels:

```
plt.xticks(rotation=45)
plt.yticks(rotation=30)
```

- Plotting a scatter plot:

```
plt.scatter(x_coordinates, y_coordinates)
plt.show()
```

- Measuring the Pearson's r between two Series:

```
dataframe['col_1'].corr(dataframe['col_2'])
```

- Measuring the Pearson's r between all columns of a DataFrame:

```
dataframe.corr()
```

- Measuring the Pearson's r for a subset of columns of a DataFrame:

```
dataframe.corr()[['col_1', 'col_2', 'col_3']]
```

Concepts

- The little lines on each axis to show unit lengths are called **ticks**, and the corresponding labels are **tick labels**. The x-axis has x-ticks, and the y-axis has y-ticks.
- In time series data, we sometimes see specific patterns occurring regularly at specific intervals of time — this is called **seasonality**.
- Weather, holidays, school vacations and other factors can often cause seasonality. Identifying seasonality can be useful for businesses.
- In a broad sense, when two columns are related in a specific way and to a certain degree, we call this relationship **correlation**. We can best visualize correlation using a **scatter plot**.
- Two positively correlated columns tend to change in the same direction. On a scatter plot, a positive correlations shows an upward trend.
- Two negatively correlated columns tend to change in opposite directions. On a scatter plot, a negative correlation shows a downward trend.
- Not all pairs of columns are correlated.
- We can measure correlation strength using **Pearson's r**. Pearson's r measures how well the points fit on a straight line.

- Pearson's r values lie between -1.00 and +1.00. When the positive correlation is perfect, the Pearson's r is equal to +1.00. When the negative correlation is perfect, the Pearson's r is equal to -1.00. A value of 0.0 shows no correlation.
- The sign of the Pearson's r only gives us the direction of the correlation, not its strength.
- When we're working with categorical variables that have been encoded with numbers, we need to interpret the sign of the correlation with caution.
- Correlation does not imply causation: proving causality requires more than just correlation, and we can't say that X is the cause of Y simply because columns X and Y are strongly correlated.

Resources

- Anatomy of a graph
- A short article on scatter plots by The Data Visualization Catalogue
- A nice article on scatter plots by MathIsFun
- A nice article on correlation by MathIsFun