Infosheet Data Science Kurs

***axhline*** - Fügt eine horizontale Linie in den Plot ein. Spezifiziere x und y Koordinaten, an denen die Linie gezeichnet werden soll. Automatisch gesetzte Parameter sind [aus pyplot package]

***dataframe*** - Eine zweidimensionale Tabelle, in der Daten gespeichert werden können. Sie besteht aus Zeilen und Spalten.

***describe*** - Generiert eine beschreibende Statistik für eine DataFrame. Anwendung: DataFrame.describe(). Zeigt count, mean, std, min, 25%, 50%, 75%, max.

***drop / dropna*** - Schließe Zeile(n) oder Spalten aus DataFrame aus. Dropna ist ein Sonderfall, bei dem alle Zeilen oder Spalten ausgeschlossen werden, die fehlende Werte haben. Anwendung: (bei axis = 0 Zeilen ausgeschlossen, bei axis = 1 Spalten)

[aus pandas]

***fit*** - Angewendet auf ein lineares Modell Objekt. Mit den übergebenen Werten führt das Model die Rechnungen aus, um das Modell zu erstellen. Übergeben werden die Trainingsdaten und die Zielwerte. Also zum Beispiel die Höhe als Trainingsdaten und die Temperatur als Zielvariable, die vorhergesagt werden soll. [aus sklearn.linear\_model package]

***groupby*** - Eine Funktion, mit der man Daten in einem DataFrame gruppieren und aggregieren kann. Die Funktion nimmt eine oder mehrere Spalten im DataFrame als Eingabe und gruppiert die Daten nach den Werten in dieser Spalte (zum Beispiel nach der Gradzahl der Landtemperatur). Danach kann zum Beispiel der Mittelwert für jede Gruppe berechnet werden. [aus pandas]

***head*** - Hiermit kannst du dir den Kopf (engl. head) deines DataFrames anzeigen lassen. Aufgerufen folgendermaßen: DataFrame.head(). Automatisch werden dir die ersten fünf Zeilen angezeigt. Du kannst aber innerhalb der Klammern auch eine Zahl reinschreiben wie zum Beispiel 10 und dann siehst du die ersten 10 Zeilen

***legend*** - Wenn du Sachen innerhalb des Plots kennzeichnen willst, musst du bei der Erstellung des Plots den Parameter Label mit übergeben. Aufgerufen wird das folgendermaßen: plt.plot(X, y, label = “dein Label”). Um die Beschreibung der Label dann auch im Plot sehen zu können, wird plt.legend() aufgerufen.

***library*** - Im Deutschen wird library auch Programmbibliothek genannt. Es ist eine Sammlung von Unterprogrammen / Funktionen / Routinen, die Lösungswege für thematisch zusammengehörende Problemstellungen anbieten. So gibt es zum Beispiel eigenen Libraries um Daten zu Visualisieren oder für machine learning Modelle. Bibliotheken sind im Unterschied zu Programmen keine eigenständig lauffähigen Einheiten, sondern sie enthalten Hilfsmodule, die von Programmen angefordert werden.

***linear regression*** - Eine lineare Regression ist ein statistisches Modell, welches genutzt wird, um die lineare Beziehung zwischen einer abhängigen Variable (Zielvariable) und einer oder mehreren unabhängigen Variablen (Prädiktoren) zu schätzen. Man versucht eine gerade Linie zu finden, die die Datenpunkte auf der Grundlage der Prädiktoren so gut beschreibt. Man kann das Modell dann nutzen, um basierend auf den Prädiktoren Vorhersagen für die Zielvariable zu machen. Man kann zum Beispiel versuchen, die Temperatur (Zielvariable) aufgrund der Höhe und des Breitengrades (Prädiktoren) vorherzusagen.

***loss*** - Eine loss function oder auch Fehlerfunktion/Kostenfunktion ist immer ein Maß für die Abweichung eines Vorhersagemodells von den tatsächlichen Beobachtungen. Er wird verwendet, um das Modell während des Trainingsprozesses zu bewerten und zu optimieren. Das Ziel ist es, dass die Fehlerfunktion so gering wie möglich ist.

***merge*** - Wenn du zwei Datensätze (df1 und df2) miteinander verbinden möchtest, nutzt man die merge Funktion aus der Pandas Library. Du musst noch mit angeben, welche Spalte in den beiden Datensätzen gleich ist, also auf Basis welcher die beiden Datensätze verbunden werden sollen. Außerdem gibst du an, ob du nur Datenpunkte behalten möchtest, die in beiden Datensätzen vorkommen (how = ‘inner’), oder auf jeden Fall alle Daten behalten möchtest, die im linken/oder zu erst genannten Dataframe sind (how = ‘left’),... Im Code sieht das dann so aus

df1

|  | name | age | hobby |
| --- | --- | --- | --- |
| 1 | Sophie Meier | 17 | art |
| 2 | Markus Schmidt | 15 | piano |
| 3 | Lea Luise | 16 | soccer |

df2

|  | name | city |
| --- | --- | --- |
| 1 | Lea Luise | Hamburg |
| 2 | Markus Schmidt | Augsburg |

->

df\_all

|  | name | age | hobby | city |
| --- | --- | --- | --- | --- |
| 1 | Sophie Meier | 17 | art | None |
| 2 | Markus Schmidt | 15 | piano | Augsburg |
| 3 | Lea Luise | 16 | soccer | Hamburg |

Wenn wir geschrieben hätten, wäre Sophie rausgefallen aus dem df\_all, da sie nicht in df2 vorkommt.

***numpy*** - Eine Library für die Programmiersprache Python, um Vektoren, Matrizen und Arrays einfach zu handhaben. SIe wird üblicherweise am Anfang eines Programms folgendermaßen importiert:

***pandas*** - Eine Library für die Python Programmiersprache, die vor allem viel in Data Science genutzt wird, da man besonders gut Daten in Datensätzen speichern kann und auf diese die verschiedensten Funktionen anwenden kann. Sie wird am Anfang eines Skirpt folgendermaßen importiert:

***plot*** - Um Daten zu visualisieren, werden diese oft in einem Schaubild angezeigt. Dies wird plotten von Daten genannt. Die einfachste Weise einen Plot zu erstellen ist

***plotly*** - Ein Package der matplotlib library. Diese bietet viele Möglichkeiten Daten einfach zu visualisieren und wird meist am Anfang durch importiert.

***predict*** - Angewendet auf ein lineares Modell Objekt. Als Parameter wird ein Wert übergeben und anhand dessen Vorhersagewerte getroffen. Zum Beispiel wird ein Array mit der Höhe von verschiedenen Orten übergeben und die Ausgabe ist dann ein Array mit vorhergesagten Temperaturen für diese Orte.

***reset index*** - Jeder DataFrame hat eine Index. Dieser steht ganz links und zählt normalerweise die Zeilen. [aus pandas package]

***reshape*** - Für manche Funktionen wird manchmal eine andere Struktur benötigt. So kann zum Beispiel das Array nicht in der aktuellen Struktur (4,) sondern in (4, 1) benötigt werden. Hierfür nutzt man reshape und erhält als Ergebnis des Ausdrucks folgenden Output:

array( [[0]],

[1],

[2],

[3]])

Dies geht auch in die andere Richtung mit (1, -1) als Parametern.

***scatter*** - Eine Sonderform eines Plots. Punkte mit x und y Koordinaten werden durch Punkte in einem Plot abgebildet.

***show*** - Am Ende eines Plots muss stehen, um unseren Plot im Notebook anzuzeigen. Am Anfang wird der Plot mit erstellt.

***sklearn*** - Machine Learning Package, welches genutzt werden kann, um vorprogrammierte Modelle (wie lineare Regression, Cluster-Algorithmen,..) zu nutzen.

***suptitle*** - Um einen Titel oberhalb eines Plots einzufügen muss vor plt.show() ausgeführt werden

BEISPIEL PLOTTEN:

