3.1 Testaufgaben:

(A) Finden Sie eine mathematische Formel zur Berechnung der Standardabweichung. Übersetzen Sie diese in Python-Code. NumPy und Pandas haben natürlich Funktionen, mit denen man das direkt berechnen kann. Diese dürfen Sie aber vorerst noch nicht nutzen.

- Datei *hausaufgabe03* im Anhang!!!

(B) Laden Sie Ihren Theaterstück-Datensatz in Pandas. Beispielcode:

```
**import pandas as pd**

**sp_novels = pd.read_csv("../Datensaetze/spanish-novels.tsv", sep="\t")**
```

- Datei *hausaufgabe03* im Anhang!!!

(C) Beschreiben Sie jede Spalte in Ihrem Theaterstück-Datensatz durch

- einen Lageparameter

- einen Streuparamter (wenn sinnvoll)

- das angemessene Skalenniveau

Spalte 1 (ID): Der Lageparameter macht hier keinen Sinn, da es sich bei der ID um eine ansteigende einzigartige Zuweisung handelt. Evtl. könnte man das Maximum heranziehen, allerdings würde hier das Minimum keine Rolle spielen, da eine ID immer bei 1 startet. Über das Maximum ließe sich zumindest feststellen, wie viele maximal verschiedene Auftritte im Theaterstück vorkommen. Das passende Skalenniveau wäre ordinalskaliert, da zwar eine Reihenfolge vorliegt, aber kein mathematischer Zusammenhang vorliegt.

Spalte 2 (label): Das Skalenniveau entspricht der Nominalskala, da hier nur Vor- und Nachnamen behandelt werden. Lage- und Streuparameter spielen keine Rolle, da hier keine Zahlenwerte vorliegen und die vollen Namen - bestehend aus Vor- und Nachname - zu zählen zu nichts führen würde.

Spalte 3 (gender): Ein Lageparameter bzw. Streuparameter ergibt bei einem Theaterstück auch keinen Sinn, dazu müsste man einen Vergleich zu mehreren Stücken herstellen und vorher alle Geschlechtsbezeichnungen (2 mal weiblich, männlich 27 mal) zählen. Die Genderzuordnungen sind nominalskaliert, da hier keine vorgegebene Reihenfolge vorliegt.

Spalte 4 (role): Gleiche Begründung wie bei Spalte 3 (gender), allerdings dann auf Rolle bezogen. (Antagonist: 2, Liebhaberin: 1, Protagonist: 1, andere: 25)

Spalte 5 (importance): Gleiche Begründung wie bei Spalte 3 (gender) und Spalte 4 (role), allerdings dann auf Bedeutsamkeit bezogen. (Primär: 5, Sekundär: 3, Weniger: 21)

Spalte 6 (per_mes_sps): Hierbei handelt es sich um eine Zählung der Sprechakte. Über den Lageparameter lassen sich in z.B. Excel oder über Python Mittelwert(37,75862069), Median(12) bestimmen (Modalwert macht keinen logischen bzw. analytischen Sinn, bei einer Zählung der 1095 Sprechakte, könnte allerdings sinnentfremdet auch angewendet werden). Über den Streuparameter könnten die Sprechakte auf Minimum(0)/Maximum(358), Varianz (4843,189655), Standardabweichung (68,3826229) sowie Interquartilabstände untersucht werden. Zum Skalenniveau: Es liegt eine Intervallskala vor, da diverse Datenpunkte in Zahlenform ausgedrückt werden und ein Zusammenhang der Daten überprüfbar ist.

3.2 Testaufgaben:

(A) Wir messen die Länge von 200 Romanen. Der Median liegt bei 231 Seiten, das Minimum bei 130 Seiten, das Maximum bei 877 Seiten, das 1. Quartil (oder 25% Quartil) bei 193 Seiten, das 3. oder 75% Quartil bei 598 Seiten. Wie viele dieser 200 Bücher sind weniger als 193 Seiten lang?

<= 50

(B) Sie haben drei verschiedene Lageparameter kennen gelernt. In welchen Fällen würden sie welchen davon benutzen und warum?

Arithmetisches Mittel:

- statistischer Durchschnittswert
- bei festen Maßeinheiten (Größe, Gewicht, Distanz etc.)

Median:

- Wert in der Mitte einer nach der Größe geordneten Datenreihe
- falls Werte zu sehr von den anderen Werten abweichen

Modus:

- Wert der in Datensatz am häufigsten vorkommt
- sinnvoll, wenn mehrere Werte mehr als einmal vorkommen