Data Wrangling (daw)
FS 2021
Cédric Huwyler

Fachhochschule Nordwestschweiz

Prüfung

Nachname	Vorname

Hinweise

Datum, Zeit, Ort: 08. Juli 2021, 13:15 Uhr, Zimmer 1.313

Prüfungsdauer: 90 Minuten

Rahmenbedingungen

- Sie bestreiten die Prüfung auf Ihrem eigenen Computer und geben am Schluss diese Prüfungsblätter und ein begleitendes Jupyter-Notebook (an `cedric.huwyler@fhnw.ch`) ab.
- Benötigte Datensets werden vom Dozierenden zum Start der Prüfung zur Verfügung gestellt.
- Für die maximale Punktzahl wird ein sauber dokumentierter Lösungsweg im Notebook erwartet. Zusätzliche Erklärungen im Code sollen in Markdown erstellt werden.
- Die Prüfung ist Open-Book und Sie dürfen das Internet zur Informationssuche benutzen.
- Kommunikation mit anderen Studierenden oder aussenstehenden Personen während der Prüfung ist nicht erlaubt. Zuwiderhandlung zieht die Note 1 für alle Beteiligten mit sich.

Benotung

Aufgabe	1	2	Total	Note
Maximale Punktzahl	42	23	65	
Erreichte Punktzahl				

Die Note 6 ist bei 50 Punkten angesetzt.

Aufgabe 1. (42 Punkte)

Kürzlich hat mir jemand gesagt, dass nach dem Vollmond besonders viele Babies zur Welt kämen. Der Mond scheint durchaus Einfluss auf uns zu haben, aber diese Aussage schien mir doch etwas gewagt. Da zum Glück Daten zur täglichen Anzahl der Geburten schon länger erfasst werden, können wir diese Aussage relativ einfach überprüfen. Das Bundesamt für Statistik stellt Daten mit der täglichen Geburtenzahl ab dem Jahr 1969 zur Verfügung. Sie finden die Excel-Datei `'BFS_VitalStatistics.xlsx'` bei den mitgelieferten Dateien.

a) **(23 Punkte)** Lesen Sie die Datei in ein Data Frame und bringen Sie die Daten in das Format

Kanton	**Jahr**	**Monat**	**Tag**	**Anzahl Geburten**
...	...	...	...	...

Stellen Sie sicher, dass Sie alle Verunreinigungen entfernt haben und dass keine fehlenden Werte vorkommen. Setzen Sie den Datentyp `float` für alle Spalten ausser *Kanton* (Pandas hat Probleme mit fehlenden Werten für den Datentyp `int`).

Hinweis: Setzen Sie geeignete Parameter `skiprows`, `nrows` und `usecols` zum Einlesen der Excel-Datei.

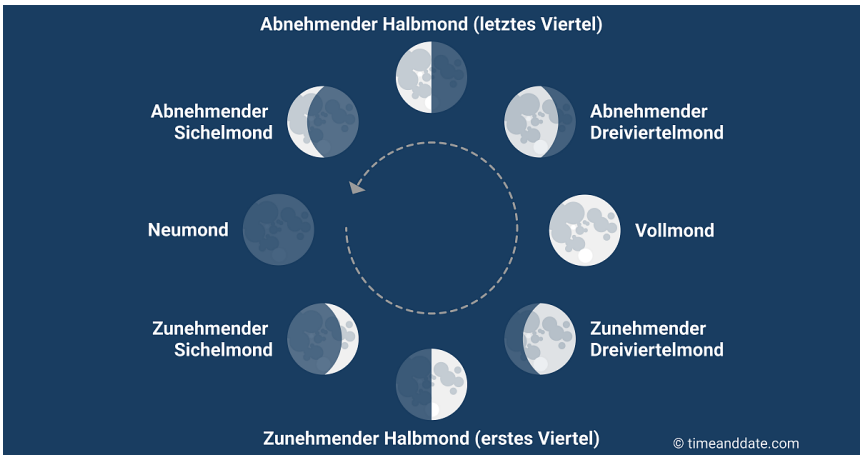
b) **(6 Punkte)** Leiten Sie nun aus den Spalten *Jahr*, *Monat* und *Tag* einen Datumstring ab und bringen Sie diesen in ein geeignetes Datumsformat. Beschränken Sie das Data Frame ausserdem auf Daten nur aus der ganzen Schweiz (Kanton='Switzerland'). Am Schluss sollte ihr Data Frame im folgenden Format vorliegen:

Datum	**Anzahl Geburten**
...	...

Hinweis: Hier können je nach Vorgehen in Teil a) invalide Datumsfelder entstehen. Die Funktion `pd.to_datetime()` hat die Option `errors="coerce"` um die Fehler einfach zu ignorieren und stattdessen `NaN`s zu setzen (die Sie natürlich entfernen müssen).

c) **(2 Punkte)** Visualisieren Sie exemplarisch die Anzahl der Geburten pro Tag in der Schweiz (Kanton='Switzerland') im Jahr 2019. Stellen Sie dabei sicher, dass die x-Achse korrekt als Datum dargestellt wird.

d) **(7 Punkte)** In der Datei `'mondphasen.csv'` finden Sie für jedes Datum seit 1970 die zugehörige Mondphase, benannt nach dem folgenden Schema:



Berechnen Sie nun die durchschnittliche Anzahl der Geburten in der Schweiz pro Mondphase seit 1970. Sehen Sie einen signifikanten Unterschied? (ohne auf statistische Tests zurückzugreifen) Natürlich müssten wir hier eine vertiefte Analyse durchführen, um sauber argumentieren zu können, aber einen ersten Eindruck können wir damit gut gewinnen.

e) **(4 Punkte)** Visualisieren Sie nun die durchschnittliche Anzahl der Geburten pro Wochentag über die Jahre hinweg. Finden Sie ein Muster?

Aufgabe 2. (23 Punkte)

Aufgrund von Engpässen an Covid19-Tests sollen Personen zuerst getestet werden, deren Symptome eher auf Covid19 hinweisen. Sie werden beauftragt, dazu die Symptome mit dem stärksten Charakter zu identifizieren und verwenden dazu den Datensatz einer israelischen Studie (Januar 2021), den Sie in den Begleitdaten als `'corona_tests.csv'` finden.

a) **(1 Punkt)** Lesen Sie die Datei in ein Data Frame ein.

b) **(3 Punkte)** Verschaffen Sie sich einen ersten Überblick über die Daten. Geben Sie dazu die verschiedenen (einzigartigen) Ausprägungen aller Merkmale ausser dem Datum aus.

c) **(4 Punkte)** Setzen Sie für jede Spalte einen sinnvollen Datentyp entsprechend den Ausprägungen. Stellen Sie dabei sicher, dass fehlende Werte auch sauber als solche gekennzeichnet sind.

Hinweis: Der Integer-Datentyp von Pandas kann nicht mit fehlenden Werten umgehen, hier dürfen Sie stattdessen Float verwenden.

d) **(2 Punkte)** Wieviele Werte fehlen pro Spalte? Geben Sie Ihre Resultate absolut und in Prozent aus.

e) **(3 Punkte)** Das Feld `'corona_result'` ist für statistische Analysen einfacher zu verwenden, wenn Sie 'negative' mit einer 0, 'positive' mit einer 1 und 'other' mit einem fehlenden Wert repräsentieren. Passen Sie Ihr Data Frame entsprechend an und verfahren Sie im gleichen Stil mit `'age_60_and_above'`.

f) **(3 Punkte)** Untersuchen Sie die Variable `'test_indication`. Welche der drei möglichen Test-Indikationen korrespondiert prozentual mit am meisten positiven Testergebnissen?

g) **(3 Punkte)** Im folgenden möchten wir nur die Symptome `'cough'`, `'fever'`, `'sore_throat'`, `'shortness_of_breath'` und `'head_ache'` untersuchen. Berechnen Sie das durchschnittliche Auftreten in Prozent dieser Symptome für getestete Personen mit und ohne Covid19. Überlegen Sie sich eine Reihenfolge, wie Sie die einzelnen Symptome priorisieren würden.

g) **(4 Punkte)** Sie möchten gerne ein ML-Modell trainieren, das versucht, aus den gegebenen Input-Features das Testresultat vorherzusagen. Ein sehr grosser Teil der Spalte `'age_60_and_above'` fehlt jedoch, darum möchte man eher auf das Entfernen der Spalten mit den fehlenden Werten verzichten. Nennen Sie zwei konkrete Strategien, wie man mit fehlenden Werten umgehen könnte. Was sind ihre Vor- und Nachteile?