Projektbeschreibung für das Wahlfach "Einführung in die Datenanalyse mit R"

Stand: 07.11.2022

Die Aufgabe

Die Aufgabe besteht darin einen *"durchgehenden"* Report/Bericht aus sich abwechselnden Code-"Analyse"-Blöcken und erklärenden Blöcken zu einer interessanten bzw. interessierenden Fragestellung bezügl. eines Datensatzes zu erstellen.

Es ist wichtig, dass Sie in den unten genannten Abschnitten des Berichts Ihre Gedanken, Beobachtungen und Schlussfolgerungen notieren. Sie sollen Ihre Analyse durchdenken und für Ihr Vorgehen argumentieren - auch wenn Ihre Gedanken zum jeweiligen Zeitpunkt des Abschnittes vorläufig und nicht endgültig sind.

Sie können das Projekt **allein oder in einer Gruppe bis zu zwei Personen** bearbeiten. Wenn Sie in der Gruppe arbeiten, achten Sie darauf die Aufgaben in etwa gleichförmig aufzuteilen und das **beide zugleich die Vorlesung als Wahl- oder Zusatzfach** belegt haben.

Da es im Report um **Ihre Arbeit bzw. Ihre Analyse** geht und der verwendete Datensatz eventuell öffentlich zugänglich ist und möglicherweise schon öffentlich analysiert wurde, ist es sehr wichtig, dass Sie fremde Code- bzw. Analyse-Beiträge auch **als solche entsprechend kenntlich machen und zitieren**. Zitieren Sie auch verwendete Packages, die nicht in der Standardinstallation enthalten sind.

Für das **Bestehen des Zusatzfaches** ist das Erreichen der **Hälfte der maximal möglichen Punktzahl** ausreichend (30 Pkt.).

Datensatz

Wählen Sie selbst einen passenden und Sie interessierenden Datensatz aus:

- Idealerweise ist es ein Datensatz von mindestens mittlerer Größe (ab ca. 1000 Datenpunkten) und enthält eine Mischung aus numerischen und kategorialen Daten.
- Sie können eigene Datensätze verwenden, die Ihnen irgendwo begegnet sind und die notwendigen Bedingungen erfüllen. Ansonsten können Sie auf Quellen von Datensätzen im Internet zurückgreifen, wie z.B.:
    - Datensätze für den Tidy Tuesday: **https://github.com/rfordatascience/tidytuesday**
    - Kaggle-Datensätze: **https://www.kaggle.com/datasets**
    - ...

Aufbau des Reports/Berichtes

1. Definition/Formulierung der Fragestellung (10 Pkt.)

Definieren Sie eine Sie interessierende bzw. interessante Fragestellung im Zusammenhang mit dem Datensatz:

- Was interessiert Sie an dem Datensatz?
- Welche spezifische Fragestellung würden Sie gern mit Hilfe des Datensatzes beantworten?

- Was erwarten Sie, angesichts Ihrer Fragestellung, bezüglich des Datensatzes?

2. Laden der Daten (10 Pkt.)

Laden Sie die Daten in die R-Sitzung und verschaffen Sie sich einen ersten Überblick

- Welche Typen sind enthalten?
- Ist sichergestellt, dass alle Daten den richtigen Typ haben?
- Haben die Daten irgendwelche "Seltsamkeiten" mit denen Sie umgehen müssen, wie z.B. anders codierte `NA`'s, mehrere Tabellen, ... etc.
- Je nach Datensatz können Sie die Daten auch in eine Datenbank laden und dann auf diese in R zugreifen.

Beschreiben Sie, was Sie tun müssen, bevor Sie die Daten im nächsten Abschnitt aufbereiten und bearbeiten können!

3. Bearbeiten/Transformieren der Daten (15 Pkt.)

In diesem Abschnitt sollten Sie alle notwendigen Transformationen/Bereiningungen/... etc. der Daten vornehmen (Data Muning, Data Cleansing), wie z.B.:

- Umcodierung von Daten, z.B. numerisch in kategorial
- Subsetting der Daten
- Joining von Datentabellen - falls nötig. Welcher Join ist notwendig? Warum?
- ...

Verschaffen Sie sich eine Übersicht der transformierten Daten. Sie können hierzu Hilfsmittel wie `glimpse()`, `skim()` und `head()` benutzen, um Ihre Erläuterungen zu veranschaulichen.

Sind die sich ergebenden Daten so, wie Sie es erwartet haben? Warum oder warum nicht?

4. Geeignete Visualisierung und Aggregation der Daten (15 Pkt.)

Fassen Sie die Daten in einer geeigenten Form zur Beantwortung Ihrer formulierten Fragestellung zusammen. Ziehen Sie auch geeignete Visualisierungen der transformierten und/oder aggregierten Daten heran, um Ihre Aussagen entsprechend zu untermauern oder zu veranschaulichen.

Hier könne Sie auch geeignete statistische Verfahren bzw. Modellierungen nutzen, falls diese Ihnen bezüglich Ihrer Fragestellung weiterhelfen.

5. Zusammenfassung und Schlussfolgerung (10 Pkt.)

Fassen Sie hier Ihre Fragestellung und Ihre Erkenntnisse aus Ihrer Analyse zusammen.

Sind Ihre Erkenntnisse das, was Sie erwartet haben? Warum oder warum nicht?

Quellenverzeichnis

Abgabe des Reports

Abzugeben ist ein Zip-Archiv, dass den Datensatz, alle verwendeten Quelldateien, d.h. *R-Skripte* (`.R` / `.r`), die *Quarto-* bzw. *R-Markdown-Files* (`.qmd` / `.Rmd`), eventuell verwendete Datenbank-Files ... etc., sowie die zugehörigen fertig gerenderten `.html`-Files enthält. Wer möchte kann den Report auch noch als statische

PDF-Version hinzufügen.

- **Abgabetermin**: 09.01.2023
- **Abgabeort**: Moodle-Kursraum