Abgabe 1 - Gruppenprojekt

Kontext und Zielsetzung

Im Kurs haben Sie bereits wichtige theoretische Konzepte und praktische Fähigkeiten erworben. Nun steht ein Gruppenprojekt bevor, bei dem Sie diese Kenntnisse anwenden und vertiefen können. Hauptziel ist es, Sie mit realen Datensätzen und Szenarien vertraut zu machen und Ihnen die Möglichkeit zu geben, Ihre Fähigkeiten in Datenhandling, statistischer Analyse und Datenvisualisierung unter Beweis zu stellen. Darüber hinaus erwarte ich, dass Sie nicht nur Daten analysieren, sondern auch lernen, Ihre Ergebnisse und Erkenntnisse auf eine klare und überzeugende Weise zu kommunizieren. Ich hoffe, dass dieses Projekt Ihnen hilft, das Potenzial und die Bedeutung von Data Science zu erkennen und Sie dazu ermutigt, Ihre Fähigkeiten und Ihr Wissen in diesem Bereich weiter auszubauen.

Die Abgabe besteht aus **zwei Teilaufgaben**:

1. Fragestellung entwicklen und Daten aufbereiten (10 Punkte)

- Entwickeln einer eigenen Fragestellung
- Entwickeln sie Hypothesen (im Kontext der Fragestellung), die Sie überprüfen wollen
- Aufbereitung des Datensatzes:
    - Skriptbasiert
    - Laden, ergänzen und manipulieren der Daten
    - Überprüfung und Bereinigung des Datensatzes; Fokus auf fehlende und fehlerhafte Werte
    - Dokumentation von hinzugefügten Features (Variablen) sowie der Datensatzbereinigung
    - Daten ggf. skalieren (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing)
    - Dokumentation auch von eventl. offenen Fragen zum Datensatz

2. Deskriptive Statistiken und Visualisierung (10 Punkte)

Erstellen Sie eine Reihe von grundlegenden Statistiken sowie Visualisierungen!

- Deskriptive Statistiken zu den von Ihnen verwendeten Hauptvariablen sind Pflicht -- Was "relevant" ist, entscheiden Sie!
- Erzeugen Sie mindestens drei *verschiedene* Plots -- Was "relevant" ist, entscheiden Sie!
- Insgesamt sollten Sie mehr als drei Plots erzeugen -- Visualisierung ist ein wichtiger Bestandteil der Datenanalyse!
- Begründen Sie stets, weswegen Sie sich für eine Statistik bzw. einen Plot entscheiden. Was wollen Sie damit zeigen?
- Interpretieren Sie stets das Ergebnis vor dem Hintergrund Ihrer Hypothesen und Fragestellung.
- Berechnen Sie Lage-, Streuungsmaße, z.B. mittlere Konsumausgaben für Sortiment “A” -- ggf. auch für Gruppen von Personen

- Führen Sie ine Korrelationsanalyse aus, z.B. um zu untersuchen, ob die Korrelation zw. Variable x und y tatsächlich positiv ist
- Wählen Sie passenden Plots aus, um Ihre Fragestellung zu untersuchen. Wählen z.B. aus den folgenden Plottypen:
    - Histogramm
    - (faktorisierte) kumulierte Dichtefunktionen
    - (faktorisierte) Boxplots
    - barplots
    - Korrelationsmatrix
    - Heatmap
    - (faktorisierte) Scatter- bzw. PairPlots
    - Zeitreihenplots
- Achten Sie auf eine verständliche, vollständige und lesbare Beschriftung der Diagramme
    - Titel und Achsenbeschriftungen hinzufügen
    - ggf. Farben etc.
    - Skalenbeschriftung
    - Schriftgröße
    - Legenden hinzufügen
- Nutzen Sie Textblöcke, um Ihre Überlegungen und die Ergebnisse schrittweise (= pro Diagramm) zu beschreiben! Meist genügt ein einzelner Satz.
- Tipp für die Formatierung der Textblöcke: **Markdown**

Datensatz:

Finden Sie in der zip-Datei “daten_abgabe_1.zip” auf moodle

Formalien

- Format der Einreichung als **zip-Datei** auf moodle.

Die zip-Datei umfasst dabei:

- Python-Notebook (.ipynb)
- Datendatei in einem geeigneten Format (z.B. csv, xlsx)
- ggf. weitere Datendateien, falls Sie weitere externe Daten verwendet haben
- Abgabe in Gruppen. Eine Einreichung pro Gruppe genügt!
- **Abgabe am 28.4.2024**

Bewertungkriterien

- Motivation, Strukturiertes Herangehen, Story-Telling, Nutzung angemessener Statistiken und Visualisierungen
- Komplexe Daten und Analysen sollen durch erzählerische Elemente verständlich und greifbar gemacht werden. Es geht darum, Daten so aufzubereiten und zu präsentieren, dass sie eine klare,

einprägsame “Geschichte” erzählen, die es ermöglicht, Zusammenhänge zu erkennen und Entscheidungen zu treffen. Achten Sie darauf, dass die Sprache klar und verständlich ist. Vermeiden Sie zu viel Fachjargon und erklären Sie alle verwendeten Begriffe und Konzepte.

Vorlage zur Orientierung

Gute Vorlage zur Orientierung für Sie: http://rstudio-pubs-static.s3.amazonaws.com/278252_34411a1b6f514be5a4a8afec04601e1a.html

Was gefällt daran?

- gute Motivation
- strukturiertes herangehen
- gutes Story-Telling

Individuelle Bewertung innerhalb der Gruppen

Für eine genauere Bewertung können Sie die individuellen Leistungsanteile innerhalb der Gruppe selbst verteilen (**ganz am Ende im Notebook**). Dafür werden von jeder Gruppe folgende Informationen benötigt:

Eine Prozentzahl von jedem Gruppenmitglied für jedes andere Gruppenmitglied.

Beispiel A

Anke hat einen Großteil der Arbeit erledigt und gleichzeitig alle anderen Gruppenmitglieder einbezogen und gleichzeitig individuelle Arbeitspakete verteilt und in das Gesamtergebnis integriert. Sie hat auch einen Großteil der Präsentation übernommen.

Jedem Mitglied in dieser 4er-Gruppe stehen 3 Punkte (n - 1) zur Verfügung. Anke verteilt ihre Punkte gleichmäßig auf alle anderen Mitglieder. Die Anderen haben Anke jeweils zwei Punkte gegeben und den dritten Punkt auf die anderen drei Mitglieder für ihre individuellen Beiträge.

=> Von der Gruppennote (2,7)

Beispiel B

Werden diese Informationen nicht eingereicht, erhalten alle Gruppenmitglieder dieselbe Note.

Beispiel für die Bewertung einer Gruppenleistung (mit der Gruppennote vom obigen Beispiel) Fall A Person 1 nimmt an allen Veranstaltungen teil und Personen 2 Personen 3 und 4 lassen sich in der Vorlesung berieseln und nehmen nicht an den Übungen teil. => Alle Gruppenmitglieder erhalten eine 2,7

==> Angedacht: ~20% Variation durch gegenseitige Bewertung der Gruppenmitglieder (z.B., 2,0 – 2,7 innerhalb einer Gruppe möglich)