Matthias Mruzek-Vering, M.Sc.

Hochschule Karlsruhe
University of
Applied Sciences

DSCB 230 - Aufgabenblatt 9

Die Aufgabenblätter werden in dieser Veranstaltung in Jupyter Notebooks veröffentlicht und bearbeitet. Diese finden Sie in der Github Organisation für Data Science 2 unter dem Repository *dscb230-tutorial* (https://github.com/hka-mmv/dscb230-tutorial). Die Musterlösung wird ebenfalls in Form eines Jupyter Notebook in Github hochgeladen.

Aufgabenteil 1: Exploratory Data Analysis - Immoscout

Ziel dieser Aufgabenstellung ist es, den Immoscout24-Datensatz zu untersuchen. Nutzen Sie die Daten aus dem Git-Repository im Verzeichnis 'assets'. Alternativ können Sie sich die Datei immo_data.csv hier herunterladen: https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany (Account benötigt). Die Daten stammen von Immoscout24, der größten Immobilienplattform in Deutschland. Immoscout24 bietet sowohl Miet- als auch Kaufobjekte an, allerdings enthalten die Daten nur Angebote für Mietobjekte. Zu einem bestimmten Zeitpunkt wurden alle verfügbaren Angebote von der Website abgefragt und gespeichert. Dieser Vorgang wurde dreimal wiederholt, so dass der Datensatz Angebote aus den Zeiträumen 2018-09-22, 2019-05-10 und 2019-10-08 enthält.
Der Datensatz enthält die meisten wichtigen Eigenschaften, wie z.B. die Größe der Wohnfläche, die Miete, sowohl Kaltmiete als auch Gesamtmiete (falls zutreffend), die Lage (Straße und Hausnummer, falls vorhanden, Postleitzahl und Bundesland), Energieart usw. Außerdem gibt es zwei Variablen, die längere Freitextbeschreibungen enthalten: Beschreibung mit einem Text, der das Angebot beschreibt, und Ausstattung mit einer Beschreibung aller verfügbaren Einrichtungen, der neuesten Renovierung usw. Die Datumsspalte wurde hinzugefügt, um den Zeitpunkt des Scrapings anzugeben.

- **Aufgabenteil 1:** Identifizieren Sie die Attribute welche zu viele Ausprägungen bewirken könnten und entfernen Sie diese entsprechend. Beschreiben Sie Vor- und Nachteile Ihrer Variante.
- **Aufgabenteil 2:** Ermitteln Sie die Häufigkeitsverteilung der verschiedenen Ausprägungen. Hier wird die Methode .value_counts()von Pandas benötigt, erzeugen Sie jeden Plot seperat in einem Gitter.

Tim Ehmann
Fabian Juny
Raphaele Salvatore Licciardo, B.Sc.

Matthias Mruzek-Vering, M.Sc.

Aufgabenteil 2

Ein Unternehmen hat alle ihre Mitarbeiter mit einer MitarbeiterID, Geschlecht, Namen, Abteilung und Gehalt in einer CSV Datei (employees.csv) gespeichert. Die Geschäftsführung möchte nun sehen, wie viel ein Mitarbeiter*in durchschnittlich pro Abteilung verdient. Stellen Sie die Ergebnisse grafisch als Säulendiagramm dar, die Abteilung mit dem höchsten Durchschnittseinkommen soll dabei ganz links, die mit dem geringsten ganz rechts stehen.

Sie sollen nun auf Wunsch des Gleichstellungsbeauftragten die Gehaltsdifferenz (Differenz zwischen Durchschnittsgehalt Männer und Durchschnittsgehalt Frauen) pro Abteilung herausfinden. Stellen Sie diese grafisch als Säulendiagramm dar, wobei die Abteilung mit der höchsten Differenz ganz links stehen soll, die mit der geringsten ganz recht.

Tim Ehmann
Fabian Juny
Raphaele Salvatore Licciardo, B.Sc.

Aufgabenteil 3

Ihre Aufgabe ist es, mithilfe von zwei Dateien „order_payments_dataset.csv“ und „orders_dataset.csv“ aus dem Ordner „archive“ eines online Shops den Umsatz sowie die Anzahl an Bestellungen in einer bestimmten Stunde in einem Diagramm darzustellen. z.B. wie viele Bestellungen gibt es zwischen 0 und 1 Uhr und welcher Umsatz wird dabei erzielt?

Das Datum und die Uhrzeit der Bestellung finden Sie in der „orders_dataset.csv“ Datei, die Kosten in der „order_payments_dataset.csv“. Folglich müssen Sie diese erstmal zu einem Dataset mithilfe einer gemeinsamen Spalte mergen.

Aufgrund der großen Unterschiede in der Größenordnung soll das Diagramm zwei y-Achsen haben, sodass die Anzahl an Bestellungen und der Umsatz gut dargestellt werden kann.

Quelle der Daten: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce