Data Analytics & Visualisation

Projektarbeit

Datensätze

- OpenData (NYC)
- Originalquelle und Dokumentation der Datensätze: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
  - **Beschreibung:** am Ende der Webseite!
- Jedes Team nimmt einen Monat (Yellow Taxi Trip Records)
- Schritte, Bemerkungen und Diskussion der Ergebnisse in Jupyter NB
- Präsentation der Ergebnisse am Ende der LV

Teams (2-3 Mitglieder)

Team	Mitglieder	Monat
1 - Hurtigen Hasen	Anna, Sophia	Feb ‘22
2 - Die Wühlmäuse	Janis, Alex	Mai ‘22
3 - Compiler Cowboyz	Simon, Lukas	Dez ‘22
4 - Die Datendetektive	Felix, Raphael	Sept ‘22
5		
6		
7		

Zeitplan

Fr.	06.10.2023	14:45	20:45	7.0	
Sa.	07.10.2023	13:00	16:15	4.0	
Fr.	20.10.2023	14:00	20:45	8.0	(teilweise freie Projektarbeit)
Sa.	21.10.2023	08:45	15:45	7.0	(teilweise freie Projektarbeit)
Sa.	11.11.2023	14:30	17:45	4.0	virtuelle Präsenz
Fr.	17.11.2023	18:00	20:45	3.5	
Fr.	15.12.2023	18:15	21:00	3.5	(teilweise freie Projektarbeit)
Sa.	16.12.2023	08:45	15:30	8.0	(Präsentation der Projektergebnisse
Fr.	02.02.2024	14:00	15:30	0.0	Klausur (Moodle; Online von Zuhause....)

Aufgaben (1)

- Thema: Einlesen, Bereinigen, Anreichern, Visualisieren
- Einlesen
  - herunterladen des Parquet Files und entpacken (yellow_tripdata_xx.parquet)
  - Dokumentation des Datenmodel in der Parquet beachten
  - laden des Parquet mit Python
- Bereinigen
  - nicht kompletter Datensätze
  - Datensätzen mit ungültigen Werten

Aufgaben (2)

- Anreichern
  - Tip per person
  - Umsatz pro Weglänge (Verhältnis)
  - Gesamtumsatz pro Taximeter (vendor_id)
  - Median von Weglänge aller Fahrten pro Tag und Stunde
- Visualisieren
  - Summe des zurückgelegten Wegs pro Tag über Monat
  - Fahrten pro Tagesstunde (0-23)
  - Gesamtumsatz (fare) pro vendor
  - Box Plot von trip_distance pro Tag

Aufgaben (3)

- Statistische Auswertung und Korrelationen
- Predict price (fare_amount)
  - zusätzlich eine externe Datenquelle nutzen (Kalender/Arbeitstag, Wetter, ...)
- Scatter Plot von predicted vs actual (über Testdatensatz)
  - Entscheidung für Ansatz (Argumentation)
  - Auswahl der unabhängigen Variable(n) (X)
  - Split in Trainings- und Testdatensatz
    - Aufteilung von Training- und Testdatensatz (70/30) oder K-Fold
  - Training des Modells
  - Evaluierung (Genauigkeit)
    - Ergebnis kann auch nicht zielführend sein
    - Diskussion der Evaluierung