P1: Kreditkartenbetrug

Prof. Dr. Ingo Claßen

Montag, 14. März 2022 14:52

Dataset bereitgestellt von Kaggle:
https://www.kaggle.com/ealtman2019/credit-card-transactions
Transformiert und bereinigt durch mich

Tabelle ccfraud2010

Spalten		
userid	integer	Nutzerkennung
card	integer	Kreditkartennummer
ts	timestamp	Zeitpunkt der Transaktion
amount	float	Betrag der Transaktion
use_chip	varchar	
mname	integer	Verkäufer, als Zahl kodiert
mcity	varchar	
mstate	varchar	
zip	varchar	
mcc	integer	
err1	varchar	
err2	varchar	
err3	varchar	
is_fraud	varchar	'Yes' oder 'No'
ccfid	integer	Anonymer Schlüssel

Aufgabenstellung

- Analyse der Daten mit SQL
- Visualisierung mit Pandas Dataframes
  z.B Barcharts, Histogramme
  https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
- Präsentation der Analyse im LV-Termin

Anzahl Datensätze, Histogramme

Prof. Dr. Ingo Claßen

Montag, 8. August 2022 09:52

- Anzahl Datensätze (AnzahlDS), Anteile
  - AnzahlDS insgesamt
  - AnzahlDS pro Jahr, prozentualer Anteil pro Jahr
  - AnzahlDS Ok/Betrug, prozentualer Anteil Ok/Betrug
  - AnzahlDS pro Land, prozentualer Anteil pro Land
    - Für USA liegen nur die States vor, diese müssen zusammengefasst werden
    - Nur Top 20
- Nutzer (userid)
  - Anzahl
  - min, max und avg Anzahl Transaktionen pro Nutzer
  - Histogramm Transaktionenanzahl pro Nutzer
    - Unterteilung in 40 Bins, d.h bei ca. 50.000 Transaktionen für den Nutzer mit der maximalen Anzahl an Transaktionen verlaufen die Bins in Schritten von 1250
    - Der erste Bin liefert die Anzahl der Nutzer die zwischen 1 und 1250 Transaktionen haben, der zweite Bin die Anzahl derer mit 1251 und 2500 Transktionen usw.
- Unternehmen (mname)
  - Anzahl
  - min, max und avg Anzahl Transaktionen pro Unternehmen
  - Histogramm Transaktionenanzahl pro Unternehmen
  - Wie bei Histogramm für Nutzer, jedoch 3 verschiedene Histogramme
    - Histogramm 1: Bis 10 Transaktionen, Unterteilung in 10 Bins
    - Histogramm 2: Bis 100 Transaktionen, Unterteilung in 100 Bins
    - Histogramm 3: Ab 100.000 Transaktionen, Unterteilung in 7Bins

Analysieren Sie die Daten nach folgenden Kriterien, wobei immer nur Ergebnisse mit einem Bertrugsanteil > 0 in der Ausgabe erscheinen sollen

- Pro Jahr
- Pro Land (beachten Sie die USA-Problematik, es liegen nur Daten zu den States vor)
- Pro err1
- Pro use_chip
- Pro Umsatzbereich
  - Bilden Sie verschiedene Gruppen, z.B amount zwischen 0 und 1000, 1001 und 2000 usw.
  - Experimentieren Sie, welche Gruppenaufteilung gute Informationen in Bezug auf den
- Pro mname (Unternehmen)
  - nur > 50 Betrugsfälle und einer Quote über 10%
- Pro Zeitbreich Abstand Transaktionen gleicher Nutzer und unterschiedliche Stadt des Händlers
  - kleiner 5 Minuten
  - größer gleich 5 Minuten
- Pro mcc

Ausgabe der Abfragen sollen folgende Spalten enthalten

Kriterium	wird im Folgenden erläutert
anz_trx	Anzahl Transaktionen
ant_ok	Anteil (d.h. Anzahl Transaktionen) OK
ant_fraud	Anteil (d.h. Anzahl Transaktionen) Betrug
quote_ok	Prozentualer Anteil OK
quote_betrug	Prozentualer Anteil