Dokumentation – Meilenstein 2: Datenbereitstellung

Projekt: Verlustprävention an Selbstbedienungskassen

Ziel

Dieses Projekt hat das Ziel, anhand von Transaktionsdaten aus Selbstbedienungskassen potenziellen Betrug und andere Warenverluste zu erkennen und den dadurch entstandenen Schaden vorherzusagen.

Der zweite Meilenstein nach DASC-PM umfasst die Datenbereitstellung, -prüfung und explorative Analyse zur Vorbereitung der Modellierung

Vorbemerkungen

Methodische Anpassung bei neuen Erkenntnissen

Im Sinne eines iterativen und explorativen Data-Science-Prozesses behalten wir uns vor, Analysen, Annahmen und Modellierungsstrategien im weiteren Projektverlauf anzupassen, sofern sich durch neue Erkenntnisse oder zusätzliche Daten relevante Änderungen ergeben. Diese Flexibilität entspricht den Anforderungen an eine verantwortungsbewusste, datengestützte Entscheidungsfindung und ist mit dem Vorgehensmodell DASC-PM vereinbar.

Mögliche Verzerrung durch fehlende negative Schadensfälle

Im bereitgestellten Datensatz sind ausschließlich Transaktionen mit einem positiven finanziellen Schaden enthalten. In der Realität ist jedoch davon auszugehen, dass auch gegenteilige Fälle auftreten – etwa, wenn Kunden mehr Artikel bezahlen, als sie tatsächlich mitnehmen, sei es durch vergessene Ware oder unvollständige Stornierungen.
Diese „negativen Schäden“ sind in den verfügbaren Daten nicht berücksichtigt, was potenziell zu einer Verzerrung der Modellbewertung führen kann.

Konformität mit gesetzlichen Anforderungen

Wir gehen im Rahmen dieses Projekts davon aus, dass die Kassensysteme der Wertkauf GmbH den Anforderungen der Kassensicherungsverordnung (KassenSichV) entsprechen. Insbesondere wird unterstellt, dass sämtliche Kassenaufzeichnungen durch eine zertifizierte technische Sicherheitseinrichtung (TSE) manipulationssicher protokolliert wurden.

Verwendung gelabelter Daten auf Basis statistischer Repräsentativität

Für sämtliche Analysen und Modellierungen im Rahmen dieses Projekts werden ausschließlich die gelabelten Transaktionen (mit label = FRAUD oder label = NORMAL) berücksichtigt.

'UNKNOWN'-Labels wurden ausgeschlossen.
Die Fokussierung auf diese Teilmenge ist gerechtfertigt, da mithilfe statistischer Tests (Chi-Quadrat-Test für kategoriale Merkmale und t-Test für numerische Merkmale) nachgewiesen wurde, dass die gelabelten Daten in ihrer Merkmalsverteilung als repräsentativ für den Gesamtdatensatz gelten können.

Zielvariablen: Klassifikation und Regressionsziel

Im Zentrum der Klassifikationsaufgabe steht die Zielvariable label, die angibt, ob bei einer kontrollierten Transaktion ein Betrugsfall festgestellt wurde (FRAUD) oder nicht (NORMAL). Ergänzend wird die numerische Variable damage als Regressionsziel herangezogen. Sie quantifiziert den durch die jeweilige Transaktion verursachten finanziellen Schaden in Euro – jedoch nur für solche Fälle, die als FRAUD klassifiziert wurden.
Da damage nur in Verbindung mit erkannter Unregelmäßigkeit (Betrug) sinnvoll interpretierbar ist, wird diese Variable ausschließlich für gelabelte FRAUD-Fälle in der Regressionsanalyse verwendet.

1. DATENBESCHREIBUNG

Die Wertkauf GmbH hat unserer Projektgruppe mehrere strukturierte Datensätze zur Verfügung gestellt, die sich auf Transaktionen an Selbstbedienungskassen beziehen. Die Daten beinhalten Informationen über vollständige Einkäufe, Kontrollklassifikationen sowie ergänzende Informationen zu Produkten und Filialen. Der Zugriff innerhalb unserer Projektgruppe erfolgt über ein passwortgeschütztes GitHub-Verzeichnis.

Zusätzlich wurden von der Wertkauf GmbH **Metadaten** als JSON-Datei zur Verfügung gestellt, welche die fachliche Bedeutung der Spalten sowie Datenqualitätsanforderungen dokumentieren. Diese Informationen wurden bei der Datenprüfung und Vorbereitung berücksichtigt.

Die Datenquellen liegen in folgenden Dateien vor:

Datei	Inhalt	Relevanz
transactions_train.parquet	Metadaten zu Transaktionen inkl. Label (NORMAL, FRAUD, UNKNOWN) und Schaden (damage)	zentrale Grundlage für Klassifikation und Regressionsmodelle
transactions_lines_train.parquet	Detaillierte Zeilen zu gekauften Produkten pro Transaktion (z. B. Produkt, Preis, Gewicht, Voiding)	Grundlage für Feature-Engineering auf Artikel-Ebene

Datei	Inhalt	Relevanz
products.csv	Produktstammdaten inkl. Preis, Kategorie, Altersfreigabe etc.	Anreicherung der line-items
stores.csv	Standortinformationen (z. B. Urbanisierung, Bundesland, Öffnungsdatum)	mögliche erklärende Variable
transactions_test.parquet, transactions_lines_test.parquet	Entsprechend der Trainingsdaten, aber ohne Label für spätere Evaluation	wird im Modellvalidierungsschritt verwendet

- **transactions_train.parquet**

Enthält 1.481.783 Transaktionen aus dem Trainingszeitraum, davon 148.025 kontrollierte Transaktionen (gelabelt mit NORMAL oder FRAUD), 4.656 davon mit erkanntem Betrug. Enthält u.a. Zeitstempel, Zahlungstyp, Kassennummer und Kundenfeedback.

- **transactions_lines_train.parquet**

Enthält 15.793.671 einzelne Kassenzeilen (Produkte) zu den Transaktionen, inkl. Produkt-ID, Menge (Stück/Gewicht), Preis, Kamera-Sicherheitsklassifikation und Zeitstempel pro Scanvorgang.

- **products.csv**

Enthält 8.120 Produkte mit Eigenschaften wie Kategorie, Preis, Gewicht, Beliebtheit, Altersfreigabe sowie Gültigkeitszeitraum.

- **stores.csv**

18 Filialen mit Standortinformationen, Bundesland, Urbanisierungsgrad und Datum der Einführung von Selbstbedienungskassen (SBK).

3. DATENQUALITÄT UND -BEREINIGUNG

Die wichtigsten Merkmale wurden auf Ausreißer, fehlende Werte und logische Konsistenz geprüft.

- **Fehlende Werte:**
  - `customer_feedback`: Nur in ca. 105.000 von 1,48 Mio. Fällen vorhanden
  - `damage`: Nur bei kontrollierten Transaktionen vorhanden
  - `weight`: Teilweise fehlend bei Produkten, die nicht nach Gewicht verkauft werden
  - `valid_to`: Fehlend bei Produkten, die derzeit noch aktiv verkauft werden

- **Bereinigungsschritte:**
  - Konvertierung von Zeitspalten zu `datetime`
  - Behandlung von fehlenden Werten (Imputation bei Bedarf)
  - Entfernen oder Transformieren von Ausreißern in numerischen Spalten
  - Zusammenführung von Transaktionen und Produktinformationen für Feature Engineering

4. Datenaufbereitung

Durchgeführte Schritte

- **Weitere Merkmale extrahiert bzw. berechnet:**
  - „transaction_duration“ (in Sekunden) als Differenz aus transaction_end und transaction_start
  - „hour_of_day „aus transaction_start
  - „weekday“ aus transaction_start
  - calc_sales-price als rechnerischer Verkaufspreis als Produkt aus Anzahl und Preis pro Einheit
  - calc_price__difference als Differenz zwischen rechnerischem und tatsächlichem Verkaufspreis

- **Labels bereinigt**: Fokus nur auf kontrollierte Transaktionen (label ≠ UNKNOWN)
- **Join-Vorgänge durchgeführt**:
  - Transaktionen mit Transaktionszeilen (über id ↔ transaction_id)
  - Lines mit Produktdaten (über product_id)
  - Stores mit Transaktionen (über store_id)
- **Fehlende Werte identifiziert und ggf. imputiert**:
  - Vgl. hierzu den Data Audit Report
- **Ausreißer identifiziert (z. B. IQR)**:

Die wichtigsten Merkmale wurden auf Ausreißer, fehlende Werte und logische Konsistenz geprüft**.**

- **Wesentliche Merkmale werden für die spätere Modellbildung in der Feature Map erfasst**

---

5. EXPLORATIVE DATENANALYSE (EDA)

Zur Vorbereitung der Modellierung wurde eine erste umfassende EDA durchgeführt:

Ziel: Strukturen, Auffälligkeiten und potenzielle Features erkennen

-Numerische Merkmale:

Statistische Kenngrößen (Mittelwert, Median, Standardabweichung, Ausreißer), Visualisierungen wie Histogramme und Boxplots. Außerdem: t-Tests zwischen gelabelten und ungelabelten Transaktionen zur Prüfung der Repräsentativität.

- Kategorische Merkmale

Häufigkeitstabellen, Chi-Quadrat-Tests, Visualisierung mit Balkendiagrammen und Heatmaps (z.B. Cramer's V) zur Bewertung von Zusammenhängen mit der Zielvariable label.

- Berechnete Variablen

Die Spalten `transaction_start` und `transaction_end` wurden genutzt, um eine neue Spalte `transaction_duration` (Transaktionsdauer in Sekunden) zu erstellen. Zusätzlich wurden die Tagesstunde und der Monat aus `transaction_start` extrahiert, um potenzielle zeitliche Muster zu erkennen.

6. Datenmanagement

Struktur und Handhabung

- **Dateiformate**: .parquet für große Transaktionen/Lines, .csv für Stammdaten
- **Speicherung und Versionierung**: lokale Ablage, passwortgeschützte Einbindung in Versionskontrolle (GitHub)
- **Datenschutz**: Es sind keine personenbezogenen Daten enthalten – DSGVO-konform
- **Skalierbarkeit**: Alle Schritte in Jupyter Notebooks dokumentiert und modular aufgebaut für spätere Automatisierung

- **Join-Strategien:** Zur Kombination von Transaktionen und Zeileninformationen wurde über die Spalte `transaction_id` ein Join durchgeführt. Weitere Joins mit Produktdaten (`product_id`) und Store-Daten (`store_id`) wurden durchgeführt.

- Hinweis: Die vollständige technische Dokumentation inklusive aller verwendeten Skripte, Transformationsschritte und Explorationsgrafiken wird separat zur Verfügung gestellt.
- Strukturierung: Die Daten wurden in DataFrames organisiert, jeweils für Transaktionen, Produktzeilen, Produkte und Stores.
- Transaktionen mit Label ≠ UNKNOWN wurden gefiltert und separat gespeichert (`labeled`).

---

7. Zusammenfassung und Erkenntnisse

Mehrere Merkmale zeigten deutliche Unterschiede zwischen FRAUD und NORMAL (insbesondere payment_medium und calc_price__difference). Die Datenqualität ist insgesamt gut, es gibt jedoch Einzelfälle mit Abweichungen und Nullwerten.

7. Ausblick auf die Modellierung (Meilenstein 3)

Im nächsten Schritt werden Klassifikationsmodelle (z. B. Random Forest, SVM, Neuronale Netzte) trainiert.

Die Modelle werden anhand eines gewichteten Kostenmodells (Bewertungsfunktion) in enger Absprache mit der Wertkauf GmbH bewertet, das Fehlklassifikationen monetär gewichte.t

---

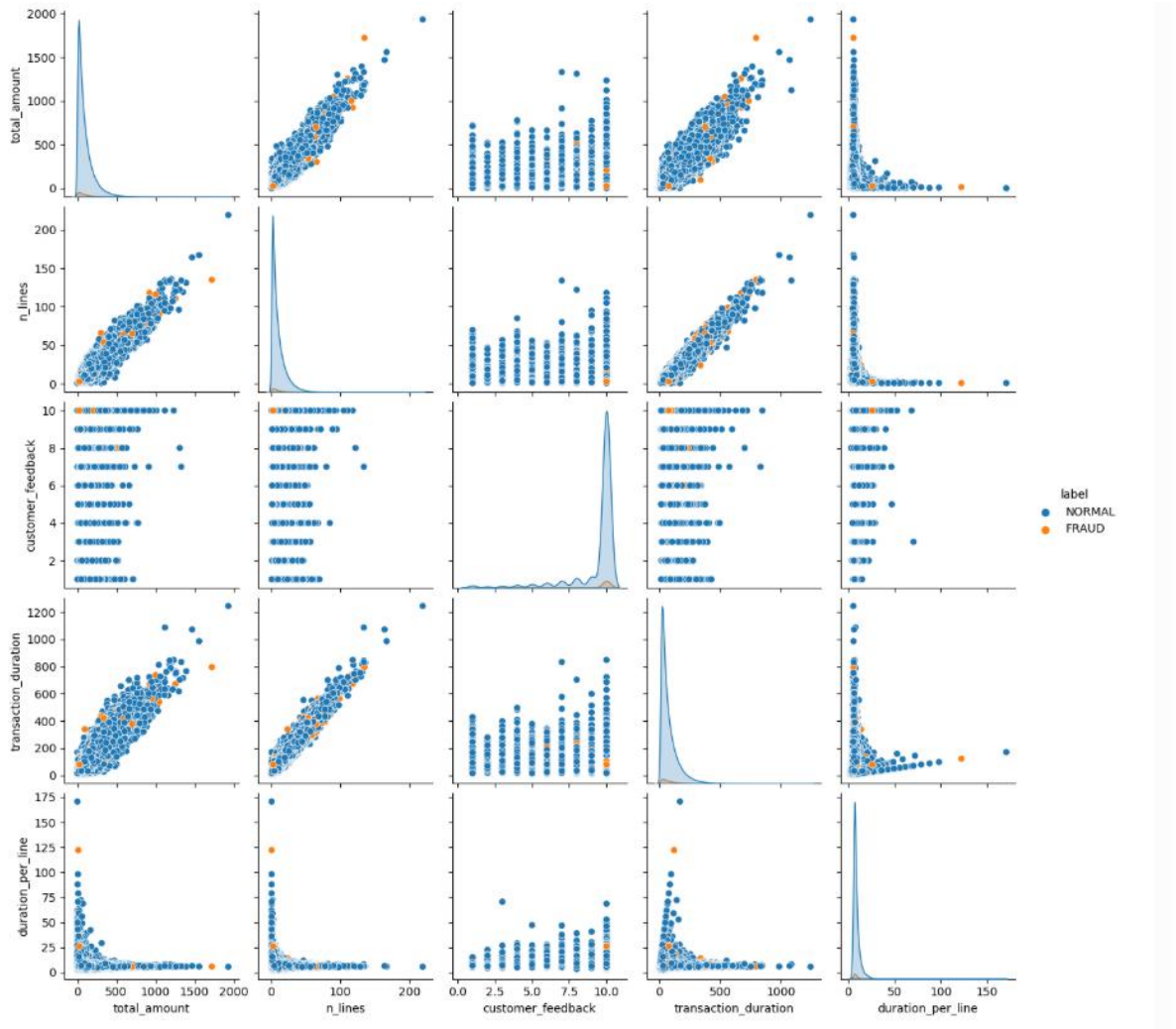
Fazit:

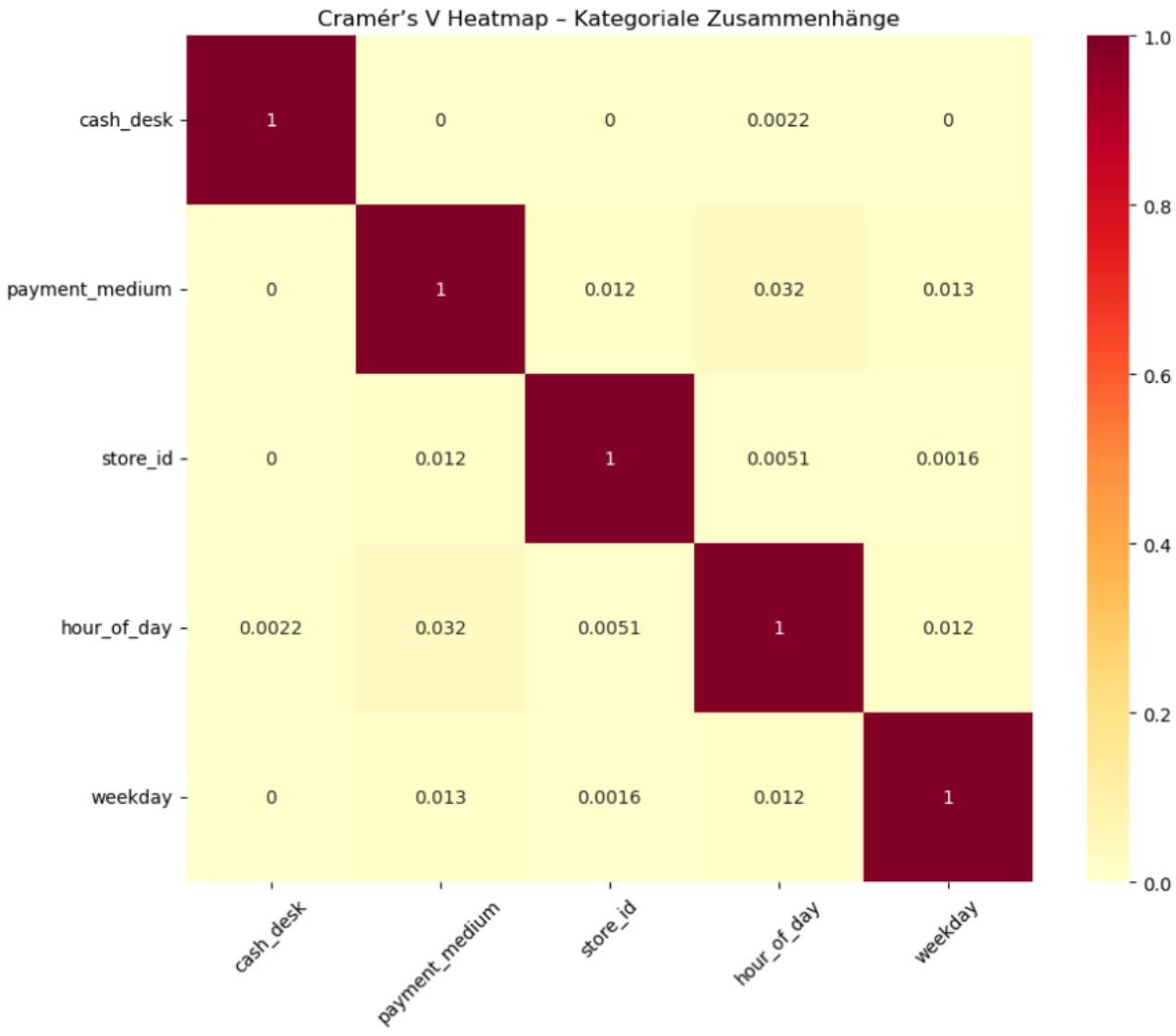
Die bereitgestellten Daten wurden erfolgreich integriert, bereinigt und analysiert. Es liegen valide und repräsentative Merkmalsräume für die Modellbildung vor. Die nächste Phase – **Modellentwicklung** – kann auf einer stabilen, sauberen Datenbasis aufsetzen.

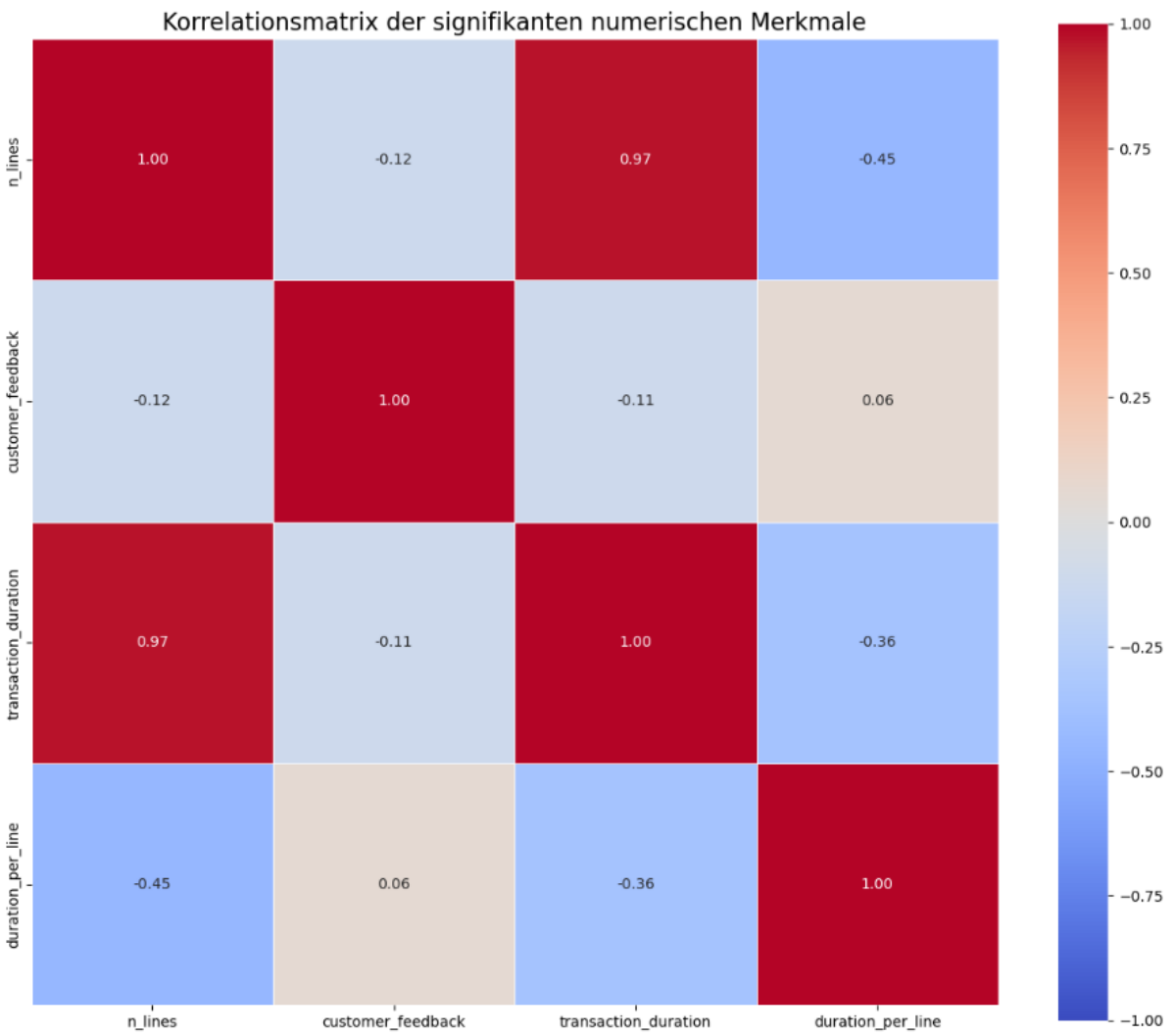
Anlagen

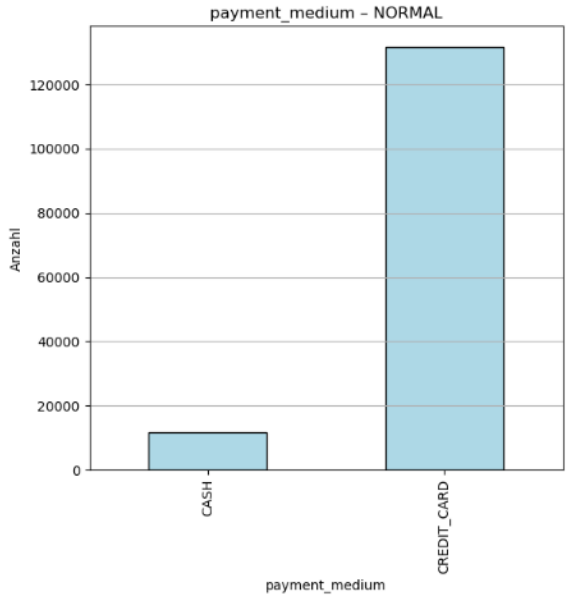
Ergebnisse der Analyse (Visualisierungen & Grafiken):

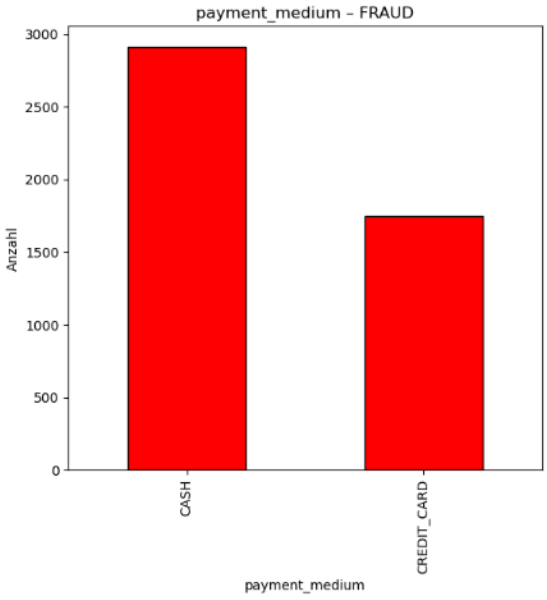
Pairplot

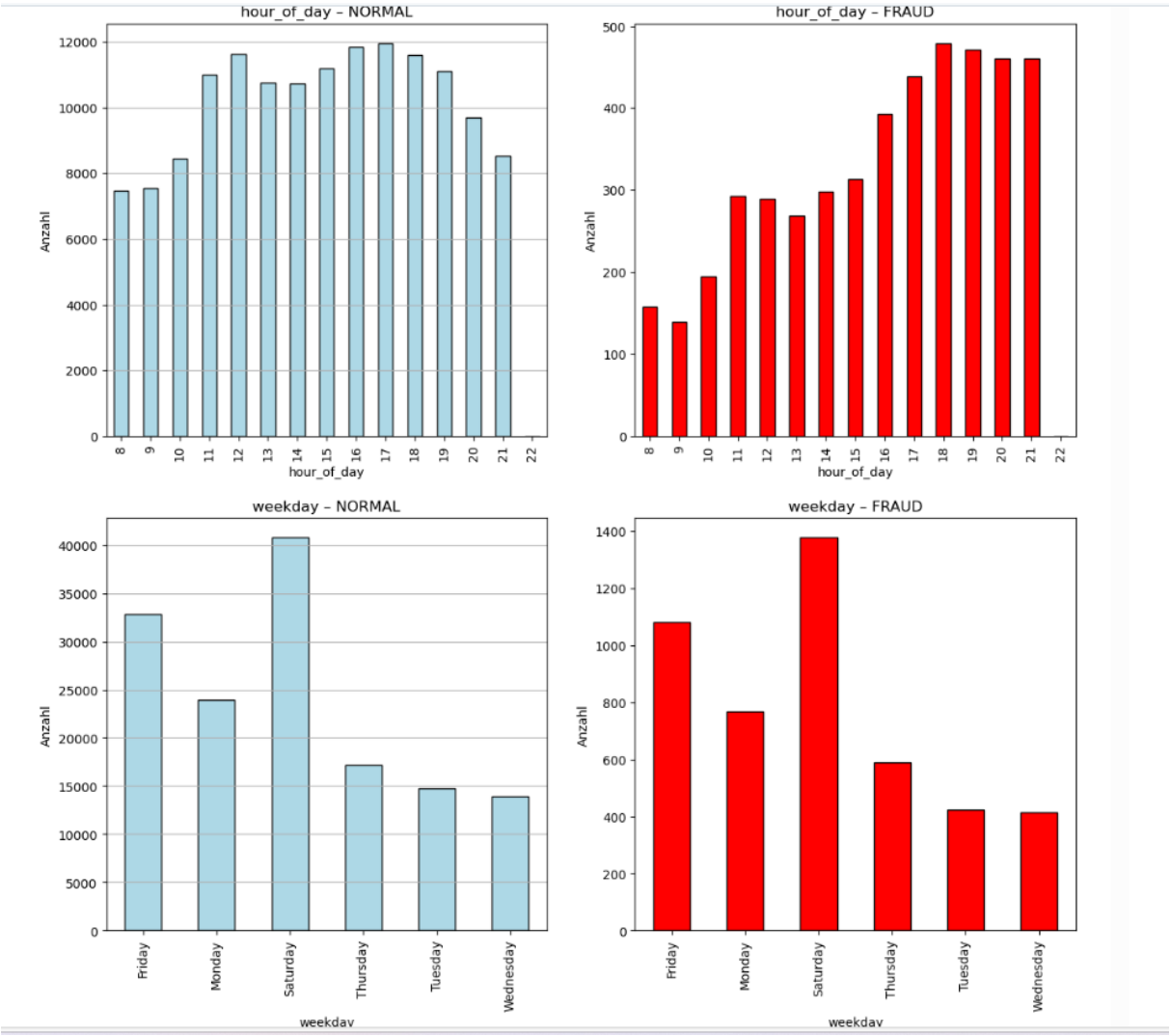



Cramér’s V Heatmap – Kategoriale Zusammenhänge
cash_desk
payment_medium
store_id
hour_of_day
weekday
1
0
0
0.0022
0
0
1
0.012
0.032
0.013
0
0.012
1
0.0051
0.0016
0.0022
0.032
0.0051
1
0.012
0
0.013
0.0016
0.012
1
1.0
0.8
0.6
0.4
0.2
0.0


Korrelationsmatrix der signifikanten numerischen Merkmale
n_lines
customer_feedback
transaction_duration
duration_per_line
1.00
-0.12
0.97
-0.45
-0.12
1.00
-0.11
0.06
0.97
-0.11
1.00
-0.36
-0.45
0.06
-0.36
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


payment_medium – NORMAL
Anzahl
120000
100000
80000
60000
40000
20000
0
CASH
CREDIT_CARD
payment_medium


payment_medium – FRAUD
Anzahl
3000
2500
2000
1500
1000
500
0
CASH
CREDIT_CARD
payment_medium


hour_of_day – NORMAL
Anzahl
hour_of_day
hour_of_day – FRAUD
Anzahl
hour_of_day
weekday – NORMAL
Anzahl
Friday
Monday
Saturday
Thursday
Tuesday
Wednesday
weekday
weekday – FRAUD
Anzahl
Friday
Monday
Saturday
Thursday
Tuesday
Wednesday
weekday