# 🗒️ Favorita – Week 1 Summary: Data Preparation & Exploratory Analysis (Guayas Region)

## Projektkontext und Zielsetzung

In der ersten Woche stand die Datenaufbereitung und ein grundlegendes Verständnis der Favorita-Verkaufsdaten im Mittelpunkt.
Ziel war es, die Datenbasis so vorzubereiten, dass sie in den kommenden Wochen für Modellierung und Forecasting genutzt werden kann.
Dafür habe ich mich auf die **Guayas-Region** konzentriert – eine Region mit elf Stores, die im Datensatz repräsentativ für typische Nachfrage- und Promotionsmuster steht.

Ich wollte die folgenden Fragen beantworten:

- Wie ist der allgemeine Verkaufsverlauf über die Jahre 2013–2017?
- Gibt es klare saisonale oder wöchentliche Muster?
- Welche Stores und Items sind die umsatzstärksten?
- Wie wirken sich Promotionen auf die Verkäufe aus?
- Welche externen Faktoren (z. B. Ölpreis, Transaktionen) haben eine messbare Beziehung zu den Verkäufen?

---

## Datenquellen und Aufbereitung

Ich habe sechs Haupttabellen verarbeitet:
`train.csv`, `items.csv`, `stores.csv`, `transactions.csv`, `oil.csv` und `holiday_events.csv`.
Da die Trainingsdaten sehr groß sind (über 5 GB), habe ich sie **chunk-weise** eingelesen, um Arbeitsspeicher zu sparen, und direkt beim Laden auf **Guayas-Stores** gefiltert.

Wichtige Entscheidungen:

- Guayas wurde gezielt gewählt, weil es eine mittlere Anzahl an Filialen (11 Stores) und eine hohe Datenqualität aufweist.
- Um die Analyse fokussiert zu halten, habe ich mich auf die **Top-3-Produktfamilien** konzentriert:
  **Grocery I**, **Beverages** und **Cleaning**.
  Diese Familien decken wesentliche Konsumgütergruppen ab und zeigen oft

saisonale Schwankungen.

- Negative Werte in `unit_sales` wurden als fehlerhafte Einträge interpretiert und auf 0 gesetzt (84 Fälle).
- Die Ölpreisserie (`dcoilwtico`) enthielt Lücken; ich habe sie linear interpoliert und anschließend vor- und rückwärts aufgefüllt, um kontinuierliche Werte zu erhalten.
- Für jede Store×Item-Serie habe ich den Kalender auf tägliche Frequenz aufgefüllt (`asfreq('D')`), um vollständige Zeitreihen ohne Lücken zu gewährleisten.

Nach diesen Schritten enthielt der Guayas-Subset rund **1,28 Mio Zeilen**, und nach dem Dichtziehen der Zeitreihen wurden daraus über **22 Mio Tageswerte**.

---

## Feature-Engineering und Struktur

Um die Zeitreihen modellierbar zu machen, habe ich verschiedene Kalender- und Rolling-Features ergänzt:

- Jahr, Monat, Tag und Wochentag (`day_of_week`)
- 7-Tage-Rolling-Mean (`target_7d_avg`)
- Z-Scores zur Identifikation von Ausreißern (Werte > 5 σ)
- Im späteren ML-Dataset (für Week 2/3) zusätzlich:
    - Lags (1, 7, 14, 30 Tage)
    - Rolling Mean/Std (7-Tage-Fenster)
    - Zyklische Kodierung von Monat und Wochentag (sin/cos)
    - Transaktionen pro Store×Tag
    - Ölpreis (WTI) pro Tag

Diese Struktur bildet die Grundlage für die Modellierung in den Folgewellen und ermöglicht eine saubere Feature-Separation ohne Daten-Leakage.

---

## Explorative Datenanalyse (EDA)

### 1. Verteilung und Ausreißer

Die Verteilung der Verkaufszahlen (`unit_sales`) ist stark rechtsschief:
Ein Großteil der Werte liegt unter 50, einige wenige Ausreißer gehen jedoch über 4000 hinaus.
Die Z-Score-Analyse bestätigte rund **3.400 Ausreißer** über 5 σ, die zumeist durch Sonderaktionen oder externe Faktoren erklärbar sind.
Für die Modellierung habe ich diese Werte zunächst beibehalten, um reale Peaks nicht zu verlieren.

### 2. Zeitlicher Verlauf

Die aggregierten Tagesverkäufe zeigen zwischen 2013 und 2017 einen klaren Aufwärtstrend.
Auffällig ist eine wiederkehrende saisonale Struktur – mit leichten Dellen im Sommer und besonders hohen Peaks rund um die Jahresendmonate (Dezember).
Dieser Trend deutet auf wachsende Kundenzahlen oder eine Expansion der Stores hin.

### 3. Saisonale Muster

Eine Heatmap der monatlichen Verkäufe zeigte eine deutliche Zunahme von 2014 bis 2016.
Insbesondere das Jahr 2016 war außergewöhnlich stark, während 2017 im Spätsommer abbrach, vermutlich bedingt durch das Ende des Trainingszeitraums.

### 4. Wochentagsmuster

Die Verkäufe verteilen sich sehr ungleichmäßig über die Woche.
Samstag und Sonntag heben sich deutlich ab – beide liegen über dem Wochenmittel, während Dienstag und Donnerstag am schwächsten abschneiden.
Das bestätigt ein starkes **Wochenend-Kaufverhalten**, das bei Forecasts berücksichtigt werden sollte.

### 5. Promotion-Effekt

Nur **0,27 %** aller Tage sind Promotions-Tage – also sehr seltene Ereignisse.
Trotzdem liegt der durchschnittliche Absatz an diesen Tagen bei rund **12 unit_sales**, während Nicht-Promo-Tage im Mittel nur **0,36** erreichen.
Das entspricht einem **Uplift von etwa +11,7 unit_sales** – ein sehr starker Effekt.
Diese Variable wird daher in der Modellierung eine wichtige Rolle spielen.

### 6. Exogene Faktoren

Ich habe zwei externe Einflüsse überprüft:

- **WTI-Ölpreis:** Die Korrelation mit den täglichen Totals liegt bei etwa **−0,64**.
  Das deutet darauf hin, dass höhere Ölpreise tendenziell mit geringeren Verkaufszahlen einhergehen – vermutlich indirekt über Transport- oder

Konsumkosten.

- **Transaktionen:** Die tägliche Gesamtzahl der Transaktionen pro Store zeigt eine **positive Korrelation von ≈ 0,66** mit den Verkäufen.
  Das ist intuitiv: mehr Kund:innen bedeuten mehr Verkäufe.
  Dieses Feature wird daher explizit in das ML-Modell integriert.

**7. Top Stores und Items**

- **Store 51** ist der umsatzstärkste Store der Region, gefolgt von Store 24 und 34.
- Das meistverkaufte Produkt ist **Item 257847**, mit deutlichem Abstand vor den anderen Serien.
- Der stärkste Tagespeak lag am **22. Juni 2013** – ohne Zusammenhang zu Feiertagen.

**8. Wöchentliche Saisonalität**

Eine Heatmap über Wochen (ISO-Woche × Wochentag) zeigt regelmäßige Wellenmuster, die zum Wochenende hin zunehmen.
Die Wiederholungen deuten auf eine klare wöchentliche Saisonalität hin, die sich gut mit Lag-Features modellieren lässt.

---

## Ergebnisse und Erkenntnisse

Zusammenfassend lässt sich sagen:

- Die Verkaufszahlen in Guayas steigen über die Jahre hinweg signifikant.
- Es gibt deutliche saisonale Muster auf Wochen- und Monatsebene.
- Promotionen haben einen außergewöhnlich starken Einfluss auf den Absatz.
- Externe Faktoren wie Transaktionen sind aussagekräftig; Ölpreise nur bedingt.
- Einzelne Stores (v. a. Store 51) und bestimmte Produkte dominieren die Region.

Die Daten zeigen, dass **zeitliche und verhaltensbasierte Features** – insbesondere Wochentag, Promotion, Rolling Means und Transaktionen – entscheidend für die Modellqualität sein werden.

---

## Übergabe an Week 2/3

Zum Abschluss habe ich zwei Datensätze erzeugt:

1. **`df_prepared_guayas_favorita.csv`**
   → vollständige, bereinigte Tagesreihen aller Guayas-Stores (2013–2017)

2. **`guayas_Q1_2014_ml_ready_favorita.parquet`**
   → fokussierter Datensatz für den Modell-Trainingszeitraum (Jan–März 2014) mit Lags, Rollings, zyklischen Features, Transaktionen und Ölpreis

Diese strukturierte Übergabe ermöglicht, in Woche 2 direkt mit Feature-Selektion, Modellierung (z. B. XGBoost / LSTM) und MLflow-Tracking weiterzuarbeiten.

---

## Persönliche Reflexion

Was ich aus dieser Woche mitnehme, ist vor allem der Wert einer sauberen, gut dokumentierten Datenaufbereitung.
Die Arbeit mit 5 GB Rohdaten und mehreren Quellen hat gezeigt, dass **gute Struktur und Lesbarkeit** entscheidend sind, bevor überhaupt modelliert wird.
Ich habe gelernt, bewusst mit Speicher und Zeit umzugehen (Chunk-Loading, Sampling) und mir angewöhnt, jeden Schritt zu validieren.

Inhaltlich fand ich besonders spannend zu sehen, wie deutlich menschliches Verhalten in Daten sichtbar wird – etwa das Wochenend-Kaufmuster oder der extreme Promotion-Effekt.
Diese Muster zu entdecken und zu erklären, ist für mich einer der Gründe, warum ich Data Analytics so faszinierend finde.