



- Thema: *Suche nach Datasets – Spotify Churn*
 - Team: jAIm
 - Mitglieder
 - Jan Ritt
 - Imre Obermüller
 - Datum des Arbeitsauftrags: 01.10.2025
-

Inhaltsverzeichnis

- [Workflow](#)
 - [Ergebnisse](#)
 - [Analyse der Aufgabenstellung](#)
 - [Lösungen](#)
 - [Statements](#)
 - [Ergebnis der Statements](#)
 - [Interpretation](#)
 - [Zusammenfassung der Ergebnisse](#)
 - [Detaillierter Bericht](#)
 - [Zusammenfassung](#)
 - [1. Datensatz-Suche und Auswahl](#)
 - [2. Metadaten-Analyse](#)
 - [3. Datenstruktur-Analyse](#)
 - [4. Explorative Datenanalyse \(EDA\)](#)
 - [5. Eignung für Klassifikation](#)
 - [6. Referenzen](#)
 - [Anhang A – Kurzüberblick Felder](#)
-

Workflow

- **Ablauf des Projektes:**
 - Suche passender Datasets →
 - Kriterien prüfen →
 - Metadaten analysieren →
 - Datenstruktur prüfen →
 - EDA durchführen →
 - Bericht/Präsentation erstellen
- **Detailabschnitte:**
 - Datensatz-Suche und Auswahl (Kriterien, Quellen)
 - Metadaten-Analyse (Quelle, Lizenz, Semantik)
 - Datenstruktur-Analyse (Spalten, Typen)
 - EDA (Zielverteilung, erste Beobachtungen)

Ergebnisse

Analyse der Aufgabenstellung

Die Aufgabe verlangt ein CSV-Dataset mit klarer Klassifizierungsspalte. Das ausgewählte Kaggle-Dataset „Spotify Dataset for Churn Analysis“ ist synthetisch, enthält die Zielvariable `is_churned` (0/1) und erfüllt Lizenz-, Format- und Dokumentationsanforderungen.

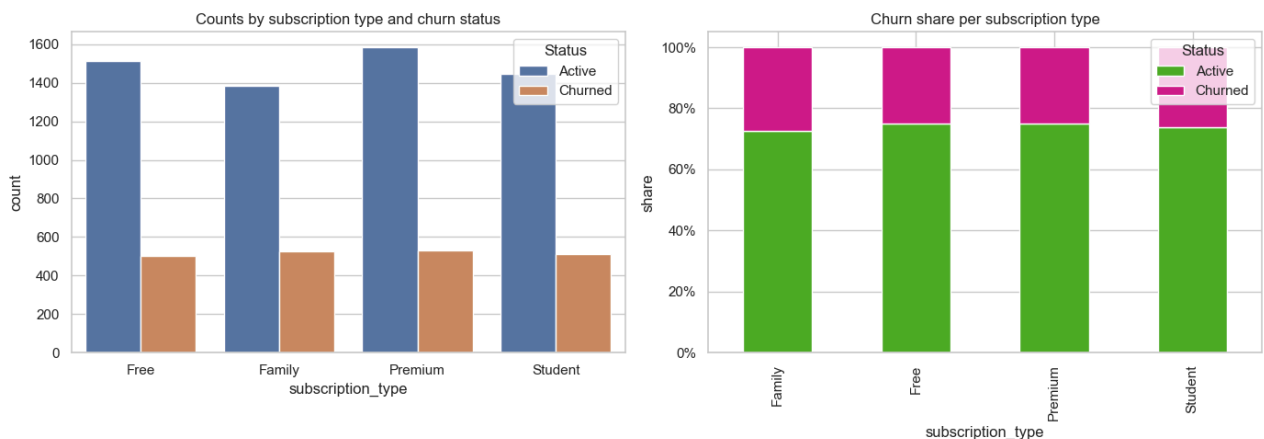
Lösungen

Statements

- Der Datensatz liegt als CSV vor und besitzt die Klassifizierungsspalte `is_churned`.
- Die Feature-Menge ist gemischt (kategorial + numerisch) und EDA-tauglich.
- Metadaten und Lizenz (Apache 2.0) erlauben freie Nutzung für Lehre/Analyse.

Ergebnis der Statements

- Zielverteilung (siehe Abbildung):



Interpretation

Die Verteilung von `is_churned` ist für binäre Klassifikation geeignet. Die Kombination aus Nutzungsintensität (`listening_time`), Interaktionen (`skip_rate`) und Kontextmerkmalen (`subscription_type`, `device_type`) liefert eine sinnvolle Grundlage für Baseline-Modelle (z. B. logistische Regression, Bäume).

Zusammenfassung der Ergebnisse

Das Dataset erfüllt alle Muss-Kriterien, die EDA zeigt eine konsistente Zielvariable und aussagekräftige Prädiktoren. Es eignet sich für Lehrzwecke (Datenaufbereitung, Visualisierung, Klassifikation) und kann ohne rechtliche Hürden verwendet werden.

Detaillierter Bericht

Zusammenfassung

Wir haben ein geeignetes, frei verfügbares Tabellendataset zur Klassifikationsaufgabe „Churn-Vorhersage“ ausgewählt: das „Spotify Dataset for Churn Analysis“ (synthetisch generiert). Der Datensatz erfüllt die in der Aufgabenstellung geforderten Kriterien (CSV-Format, Klassifizierungsspalte), ist gut dokumentiert (Metadaten vorhanden) und für EDA sowie den Aufbau einfacher Baseline-Modelle geeignet. Die Zielvariable `is_churned` klassifiziert Nutzer in „aktiv“ (0) und „gekündigt“ (1). Erste EDA-Ergebnisse (Notebook) zeigen eine sinnvolle Merkmalsstruktur aus numerischen und kategorialen Feldern. Der Datensatz ist damit für Lehrexperimente zu Datenaufbereitung, Visualisierung und Klassifikation angemessen.

1. Datensatz-Suche und Auswahl

- Quelle(n): data.gv.at, kaggle.com
 - Auswahl: Kaggle – Spotify Dataset for Churn Analysis (Apache-2.0-Lizenz)
 - Begründung:
 - CSV-Format und klarer Klassifikations-Target (`is_churned`)
 - Ausreichende Feature-Vielfalt (Nutzer-, Nutzungs- und Geräteattribute) für EDA und Baselines
 - Frei zugänglich, mit Metadaten und klarer Lizenz
-

2. Metadaten-Analyse

- **Titel:** Spotify Analysis Dataset 2025
 - **Katalog/Quelle:** Kaggle – spotify-dataset-for-churn-analysis
 - **Ersteller (Creator):** nabiha zahid
 - **Lizenz:** Apache 2.0 (<https://www.apache.org/licenses/LICENSE-2.0>)
 - **Publikation/Update:** veröffentlicht 2025-08-29; zuletzt geändert 2025-08-28
 - **Zugriff:** kostenfrei zugänglich; Live-Dataset
 - **Datencharakter:** synthetisch generiert für EDA/ML (kein realer Zeitraum)
 - **Distribution:** ZIP-Archiv mit `spotify_churn_dataset.csv`
 - **Zeilen-Semantik:** eine Zeile = ein Spotify-Nutzer
 - **Zielvariable:** `is_churned` (0 = aktiv, 1 = gekündigt)
-

3. Datenstruktur-Analyse

Der Datensatz ist tabellarisch (CSV) mit numerischen und kategorialen Merkmalen. Vorgesehene Felder (intended types):

- `user_id` (Text)
- `gender` (Text)
- `age` (Integer)
- `country` (Text)
- `subscription_type` (Text)
- `listening_time` (Integer, Minuten/Tag)
- `songs_played_per_day` (Integer)
- `skip_rate` (Float)
- `device_type` (Text)
- `ads_listened_per_week` (Integer)
- `offline_listening` (Integer-Flag)
- `is_churned` (Integer-Flag, Target)

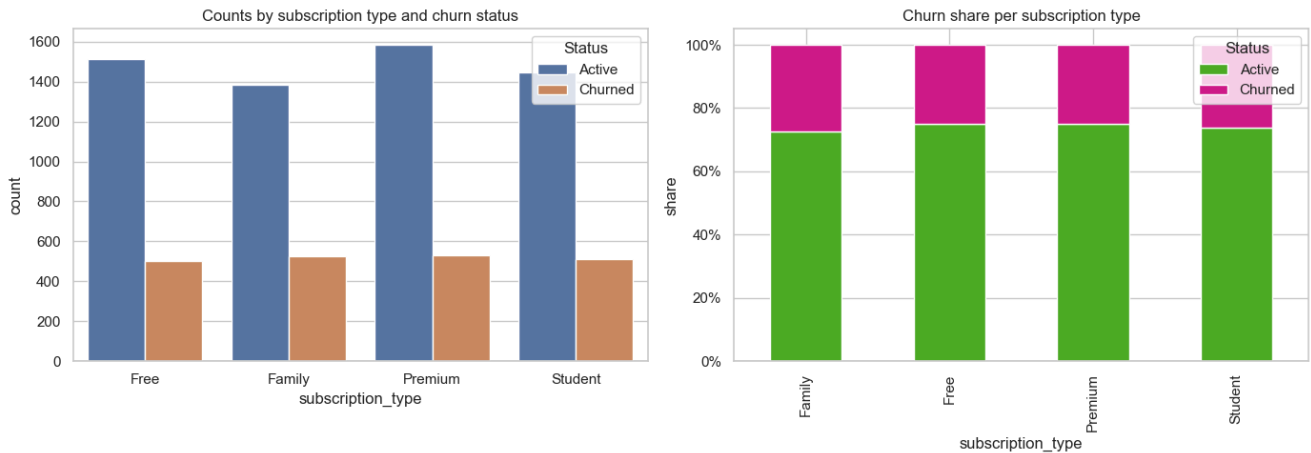
Hinweise zur Nutzung:

- Kategoriale Merkmale können via One-Hot-Encoding oder Target-Encoding verarbeitet werden.
- Numerische Merkmale sollten auf Ausreißer geprüft und ggf. skaliert werden.
- Die Zielvariable ist binär; geeignete Metriken sind z. B. Accuracy, F1, ROC-AUC.

4. Explorative Datenanalyse (EDA)

Die EDA wurde in `notebooks/spotify_churn_eda_executed.ipynb` durchgeführt und dokumentiert.

Zentrale Visualisierung der Zielverteilung:



Beobachtungen (aus der EDA, qualitativ zusammengefasst):

- Die Zielvariable `is_churned` ist klar definiert (0/1) und für Klassifikation geeignet.
- Die Feature-Menge deckt Nutzungsintensität (z. B. `listening_time`), Interaktion (`skip_rate`), Produktvariante (`subscription_type`) und Kontext (`device_type`, `ads_listened_per_week`) ab.
- Für Baselines sind einfache Vorverarbeitungsschritte ausreichend; weiterführend könnte Feature-Engineering (z. B. Interaktionen, Binning) nützlich sein.

5. Eignung für Klassifikation

Begründung der Eignung:

- Binäres Ziel (`is_churned`) und mehrere erklärende Variablen mit plausibler Beziehung zur Churn-Neigung.
 - Synthetische Daten vermeiden Datenschutzprobleme und sind für didaktische Zwecke geeignet.
 - CSV-Format erleichtert den Einstieg in Datenaufbereitung und Modellierung.
-

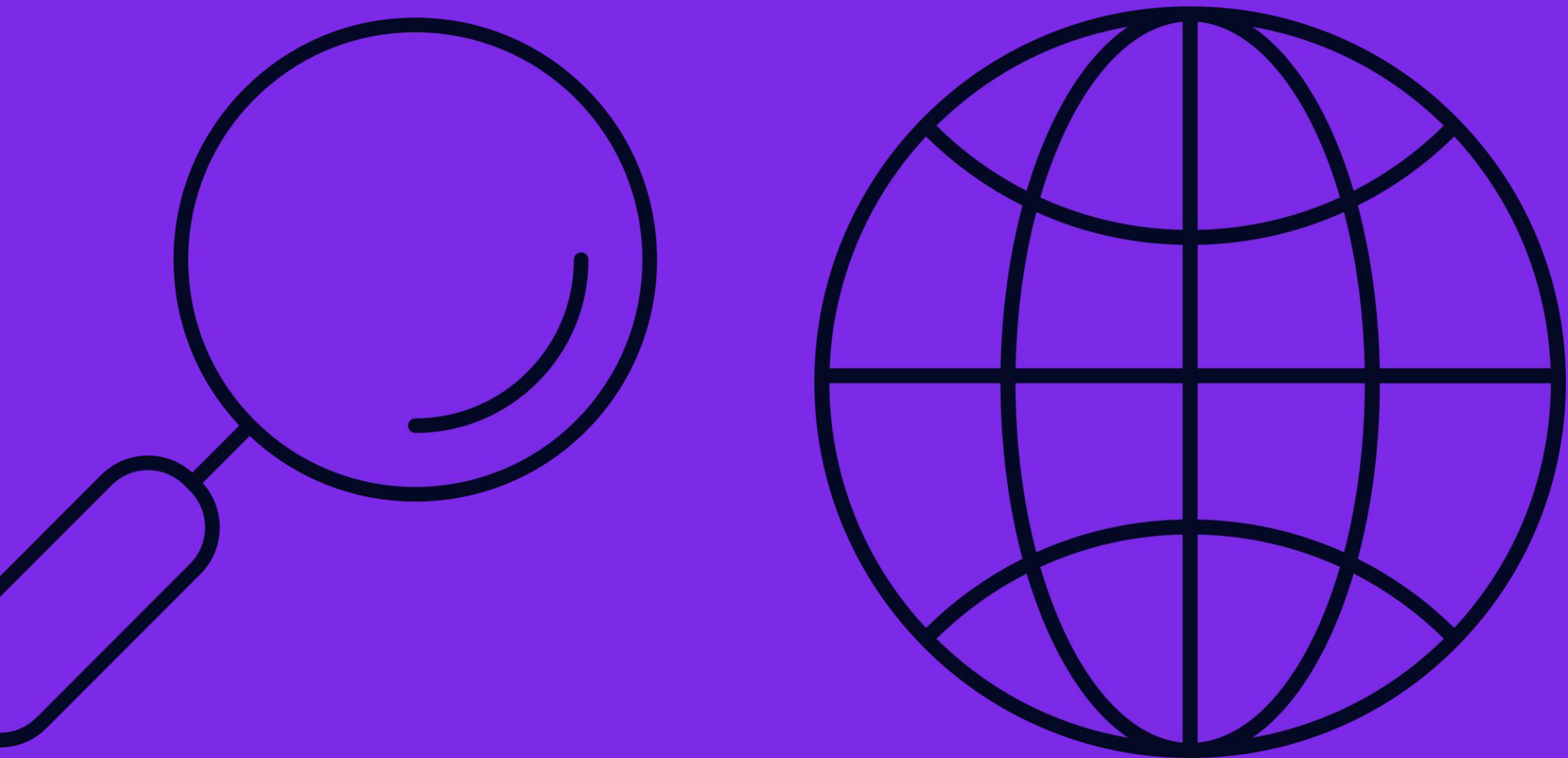
6. Referenzen

- Kaggle Katalogeintrag: <https://www.kaggle.com/datasets/nabihazahid/spotify-dataset-for-churn-analysis>
 - Lokale Metadaten: `spotify-dataset-for-churn-analysis-metadata.json`
 - Datendatei: `spotify_churn_dataset.csv`
 - EDA-Notebook: `notebooks/spotify_churn_eda_executed.ipynb`
 - Präsentation: `spotify_presentation.pdf`
-

Anhang A – Kurzüberblick Felder

Kurzzusammenfassung der Felder (siehe auch Abschnitt 3):

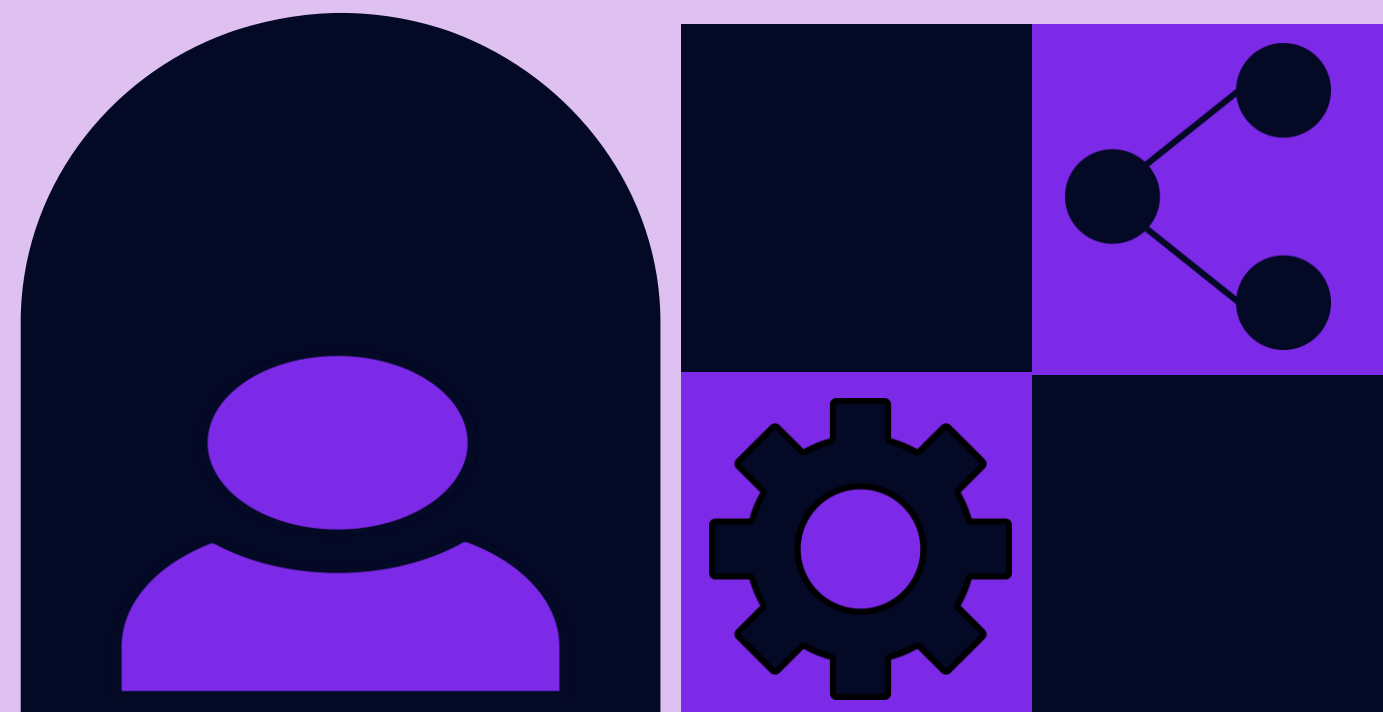
Feld	Typ (intended)	Bedeutung
user_id	Text	Nutzer-ID
gender	Text	Geschlecht
age	Integer	Alter
country	Text	Land
subscription_type	Text	Abo-Typ (Free/Premium/Family/Student)
listening_time	Integer	Minuten pro Tag
songs_played_per_day	Integer	Songs pro Tag
skip_rate	Float	Anteil übersprungener Songs
device_type	Text	Gerätetyp (Mobile/Desktop/Web)
ads_listened_per_week	Integer	Werbeeinblendungen pro Woche
offline_listening	Integer (Flag)	Offline-Modus genutzt
is_churned	Integer (Flag)	Zielvariable: 0 aktiv, 1 gekündigt



Team: jAlm (Jan Ritt, Obermüller Imre) Date: 10/01/2025

Spotify Analysis

Dataset 2025



Data Science & AI

Search for data & basic analysis



Workflow



Find data

Find data from [www.Kaggle.com](https://www.kaggle.com)
or www.data.gv.at



Analyze metadata

Which metadata exists and what
is the purpose of each?



Examine data structure

Which rows and columns exist?
Classification-column?



Findings

Churn rate and class balance:

- About 26% churn (2,071 churned vs 5,929 active out of 8,000). The target is moderately imbalanced, so report metrics beyond accuracy..

Usage intensity vs churn:.

- Boxplots indicate churned users have lower typical usage (lower listening_time and songs_played_per_day) than active users, suggesting engagement is a strong signal for churn.

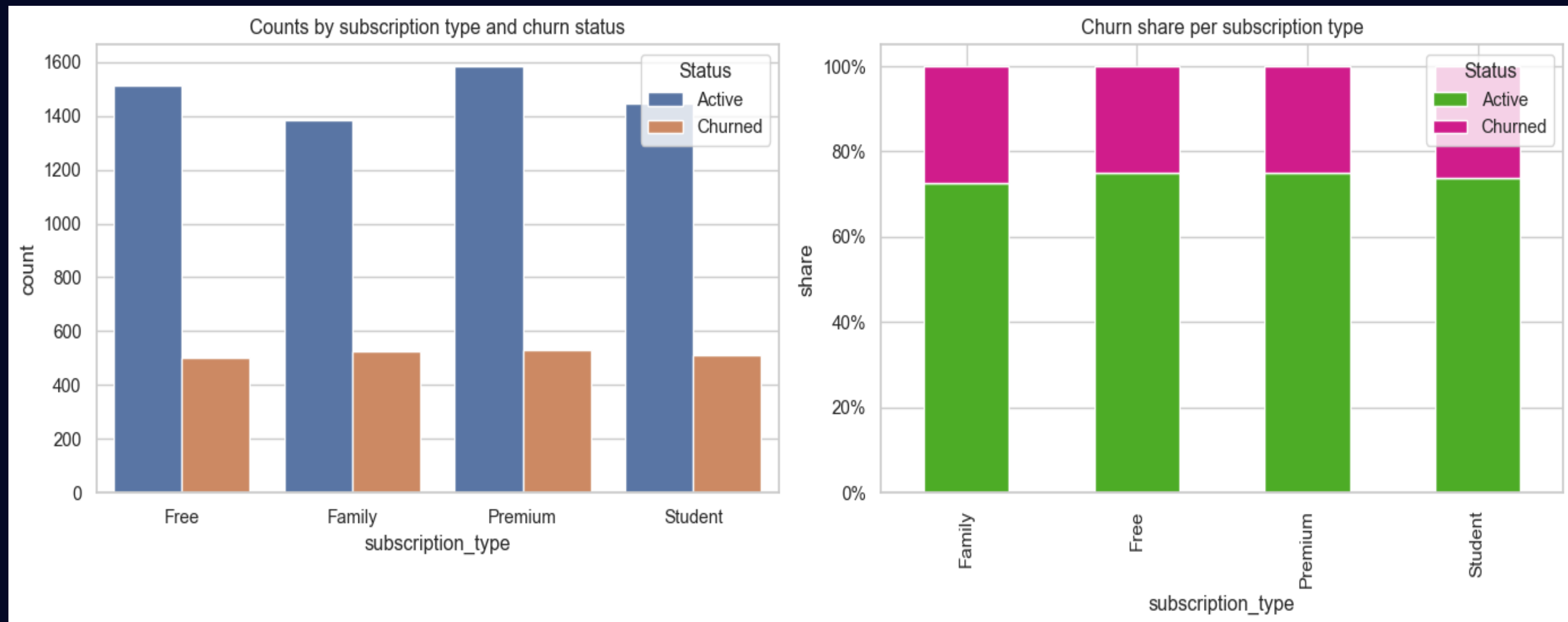
Subscription type matters:

- The churn-by-subscription visuals show notable differences across plans; churn share is highest for lower-tier plans and lowest for higher-tier plans, indicating plan segmentation is predictive.

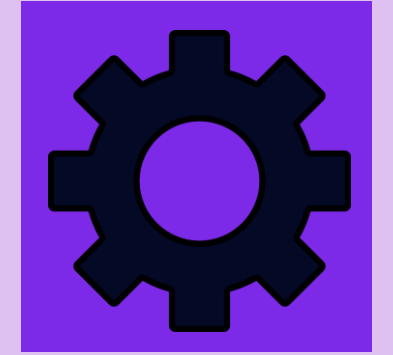
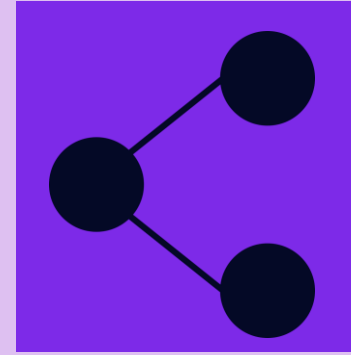
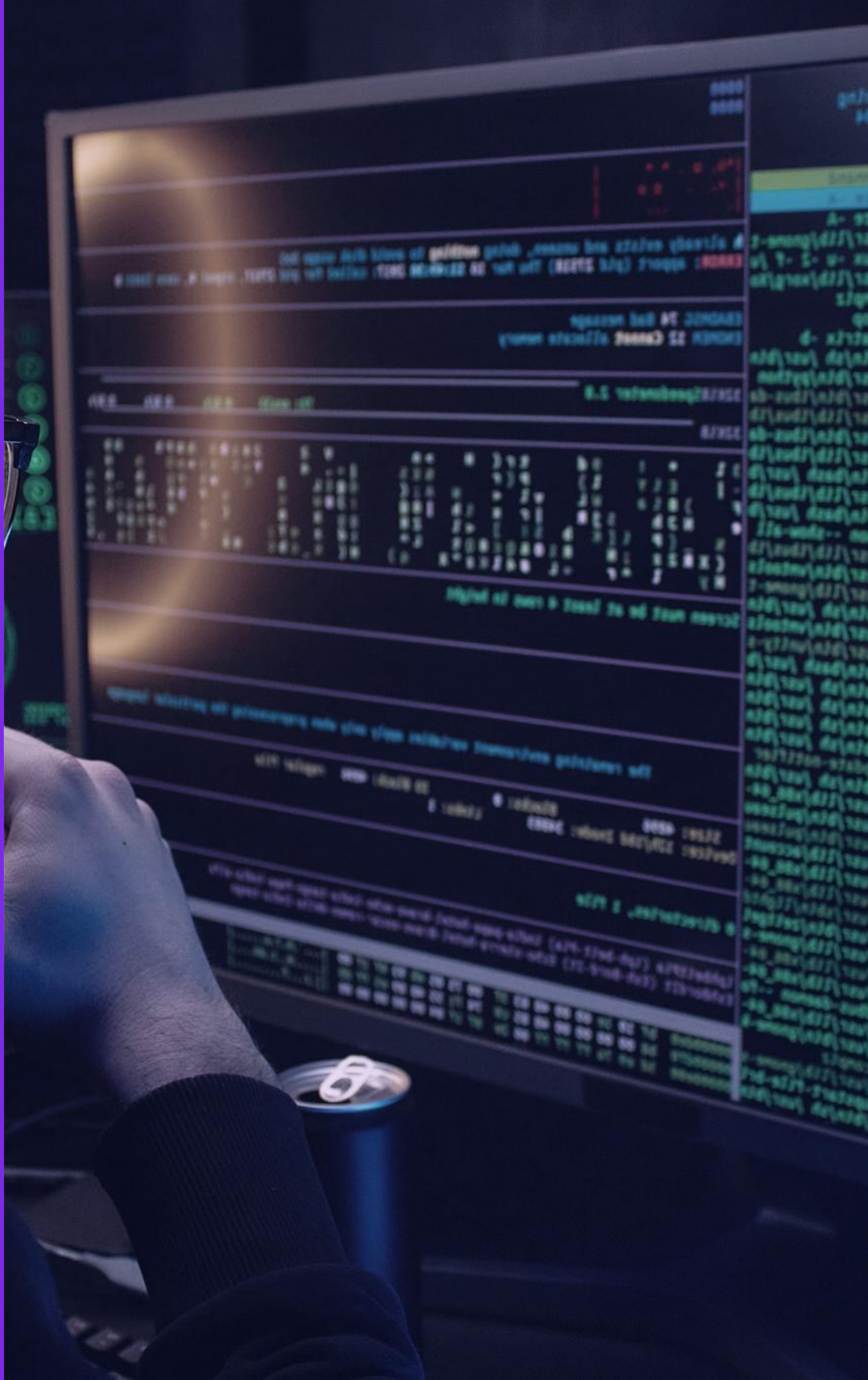


Active vs. Churned

Counts of active vs. churned per subscription_type



Summary

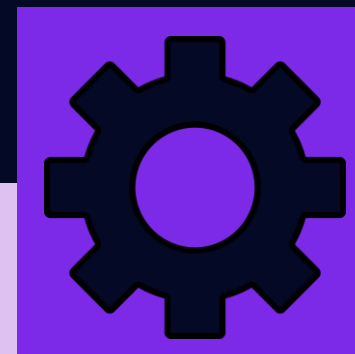
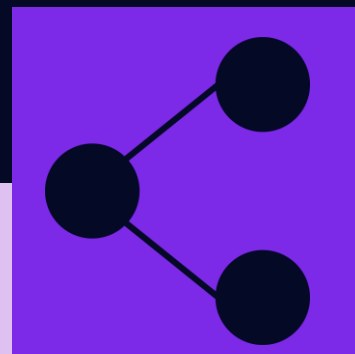


We validated the Kaggle Spotify churn dataset, created and executed an EDA notebook. Generated visuals: target distribution, engagement by churn, churn by subscription.

Key takeaways:

Churn is ~26%, churned users show lower engagement, and lower-tier plans have higher churn.

Thank you!



Any questions? Feel free to ask!



HTL Leonding
next level