MODELING SPOTIFY CHURN: A STATISTICAL ANALYSIS OF USAGE DEMOGRAPHICS, AND ENGAGEMENT

GALICIA, MOJICA, PLURAD, RAMIREZ, TOLENTINO

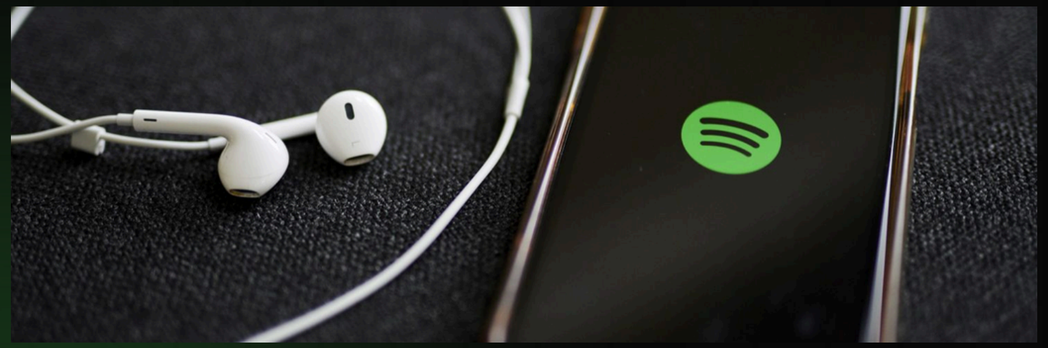
DATASET DESCRIPTION

1.07 -1.49

ALL MUSIC PODCAST ALBUMS

DATASET BASIC DESCRIPTION

SPOTIFY ANALYSIS DATASET 2025



- Synthetic dataset for customer retention analysis
- Anonymized user demographics, behavior, and churn
- 8,000 users | 12 variables
- Obtained from Kaggle
- Useful for predictive modeling

DATA COLLECTION PROCESS

- Algorithmically created to mimic real Spotify user behavior
- Used when real data is unavailable due to privacy or proprietary constraints
- Preserves statistical patterns, but may miss real-world nuances
- High accuracy reflects generator assumptions, not real-world messiness

ALL MUSIC PODCAST ALBUMS

STRUCTURE OF THE DATASET FILE

FILE NAME: SPOTIFY_CHURN_DATASET.CSV

- Stored as a single CSV file
- **Rows:** represent individual users
- Columns: represent user attributes
- **Records (observations):** 8,000
- Variables: 12
- Covers demographics, behavior, and churn outcome

ALL MUSIC PODCAST ALBUMS

VARIABLE DESCRIPTION

1. USER IDENTIFICATION & DEMOGRAPHICS

- user_id
- gender
- age
- country

2. FINANCIAL COMMITMENT & PLATFORM FRICTION

- subscription_type
- ads_listened_per_week

3. ENGAGEMENT & SATISFACTION

- listening_time
- songs_played_perday
- skip_rate,
- device_type
- offline_listening
- is_churned

DATASET PREPROCESSING

ALL MUSIC PODCAST ALBUMS

CLEAN MULTIPLE REPRESENTATIONS

All object values are already standardized and have no spelling or capitalization inconsistencies

```
Gender Objects
['Female' 'Other' 'Male']

Country Objects
['CA' 'DE' 'AU' 'US' 'UK' 'IN' 'FR' 'PK']

Subscription Type Objects
['Free' 'Family' 'Premium' 'Student']

Device Type Objects
['Desktop' 'Web' 'Mobile']
```

DATA TYPE

Data type is correct and valid

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   user_id                8000 non-null   int64
 1   gender                 8000 non-null   object
 2   age                    8000 non-null   int64
 3   country                8000 non-null   object
 4   subscription_type      8000 non-null   object
 5   listening_time         8000 non-null   int64
 6   songs_played_per_day   8000 non-null   int64
 7   skip_rate              8000 non-null   float64
 8   device_type            8000 non-null   object
 9   ads_listened_per_week  8000 non-null   int64
 10  offline_listening      8000 non-null   int64
 11  is_churned             8000 non-null   int64
dtypes: float64(1), int64(7), object(4)
memory usage: 750.1+ KB
```

ALL MUSIC PODCAST ALBUMS

MISSING DATA

No missing data

```
Missing values per variable:

user_id                  0
gender                   0
age                      0
country                  0
subscription_type        0
listening_time           0
songs_played_per_day     0
skip_rate                0
device_type              0
ads_listened_per_week    0
offline_listening        0
is_churned               0
dtype: int64

Any missing data?: False
```

DUPLICATE DATA

```
Any duplicate rows?: False
Number of duplicate rows: 0

No duplicate rows found.
```

INCONSISTENT FORMATTING

```
user_id formatting inconsistent?: Not applicable (int64)
gender formatting inconsistent?: False
age formatting inconsistent?: Not applicable (int64)
country formatting inconsistent?: False
subscription_type formatting inconsistent?: False
listening_time formatting inconsistent?: Not applicable (int64)
songs_played_per_day formatting inconsistent?: Not applicable (int64)
skip_rate formatting inconsistent?: Not applicable (float64)
device_type formatting inconsistent?: False
ads_listened_per_week formatting inconsistent?: Not applicable (int64)
offline_listening formatting inconsistent?: Not applicable (bool)
is_churned formatting inconsistent?: Not applicable (bool)
```

EXPLORATORY DATA ANALYSIS

1.07 -1.49

ALL MUSIC PODCAST ALBUMS

EDA Q1: HOW DOES LISTENING TIME PER DAY VARY ACROSS SUBSCRIPTION TYPES?

PURPOSE

- See if Premium users listen longer than Free or Student users

NUMERICAL SUMMARY

- Examine the mean, median, and standard deviation of listening time per subscriptiontype.

VISUALIZATION

- Boxplot of listening_time by subscription_type

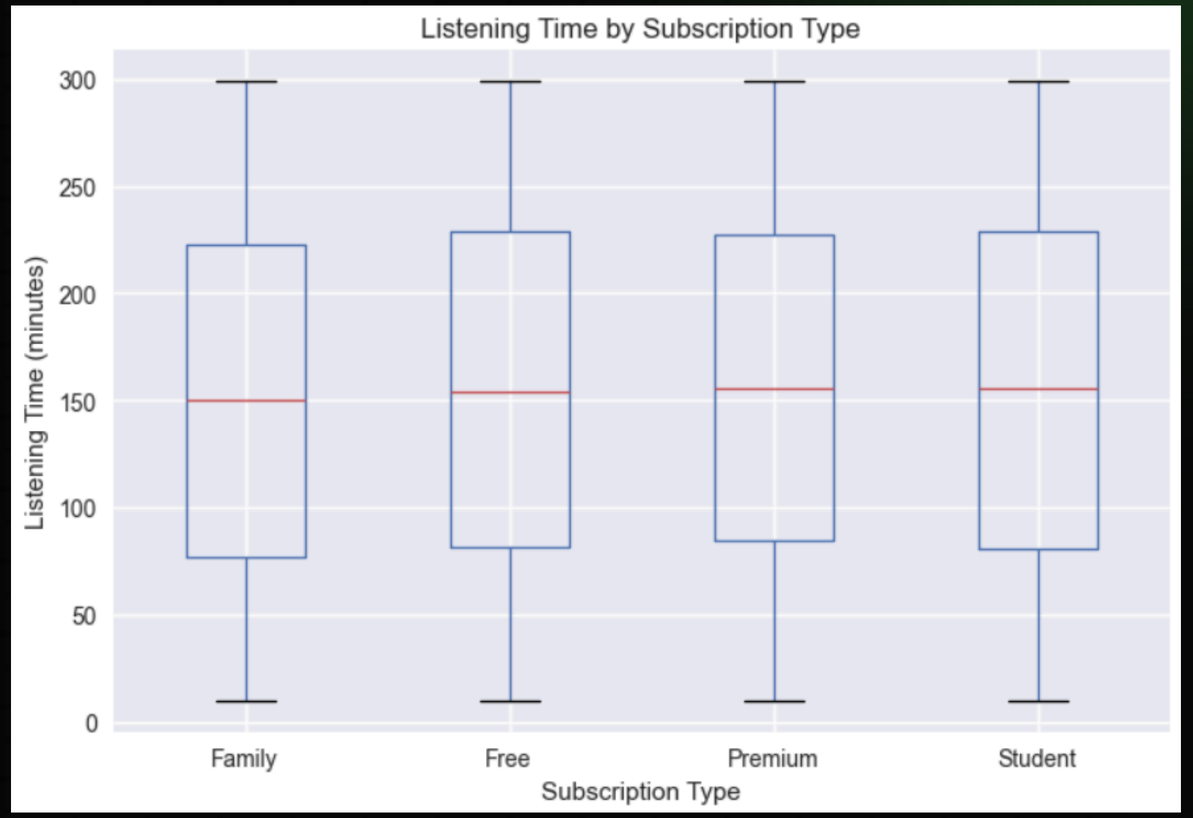
STATISTICAL TEST

- Analysis of Variance (ANOVA)

EDA 1 RESULT

F-statistic: 1.1354954581789256
p-value: 0.3331632951262332 7

Accept the null hypothesis. There is no evidence to suggest that Premium users listen longer than Free or Student users.



ALL MUSIC PODCAST ALBUMS

EDA Q2: DOES THE TYPE OF SPOTIFY SUBSCRIPTION AFFECT THE LIKELIHOOD OF A USER CHURNING?

PURPOSE

- Identify whether churn rates differ between subscription plans.

NUMERICAL SUMMARY

- Contingency table between churned and non-churned users by subscription type

VISUALIZATION

- Bar graph of churn proportions by subscription_type.

STATISTICAL TEST

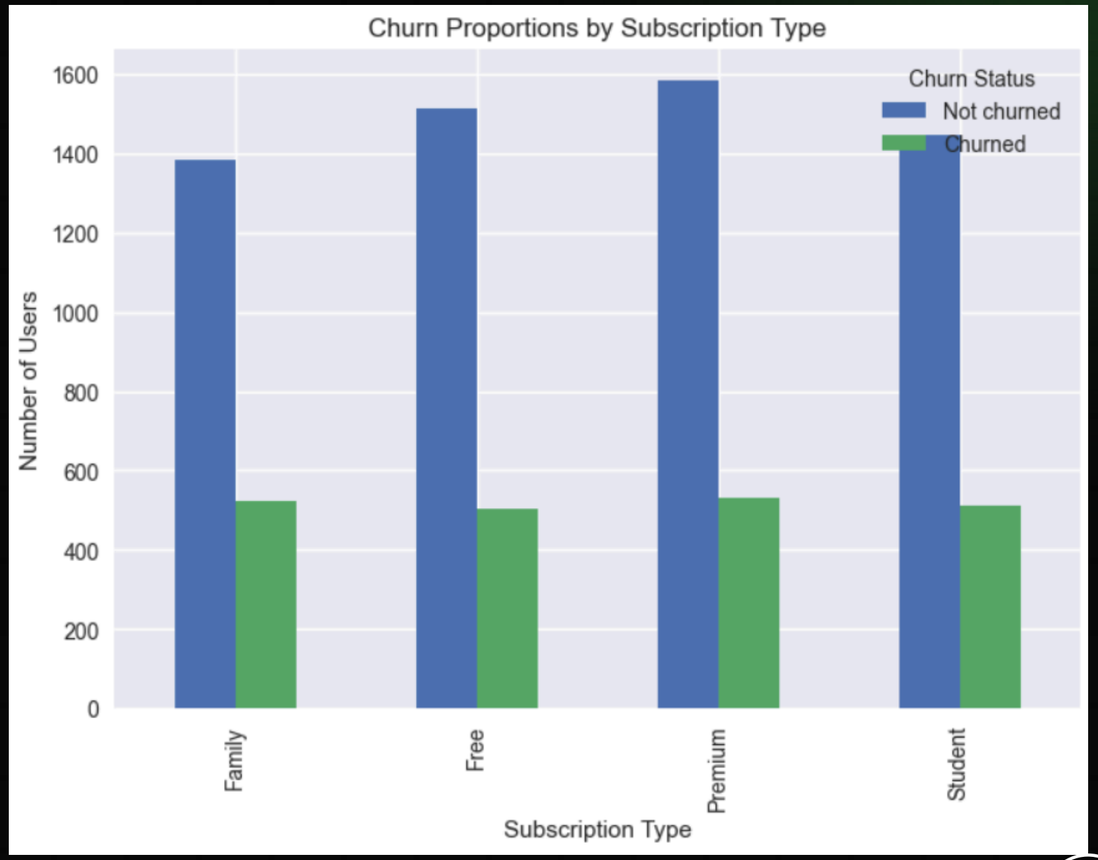
- Chi-square test of independence

ALL MUSIC PODCAST ALBUMS

EDA 2 RESULT

Chi-square Statistic: 4.457518638155985
Degrees of Freedom: 3
P-value: 0.216110972429786

Accept the null hypothesis. There is no significant difference in listening time between different subscription types.



EDA Q3: DO CHURNED USERS DIFFER FROM ACTIVE USERS IN THE NUMBER OF SONGS THEY PLAY PER DAY?

PURPOSE

- Identify if listening activity level (songs/day) is linked to churning.

NUMERICAL SUMMARY

- Count, mean, median, and standard deviation of songs_played_per_day for is_churned = True vs False.

VISUALIZATION

- Side-by-side boxplots of songs_played_per_day by churn status.

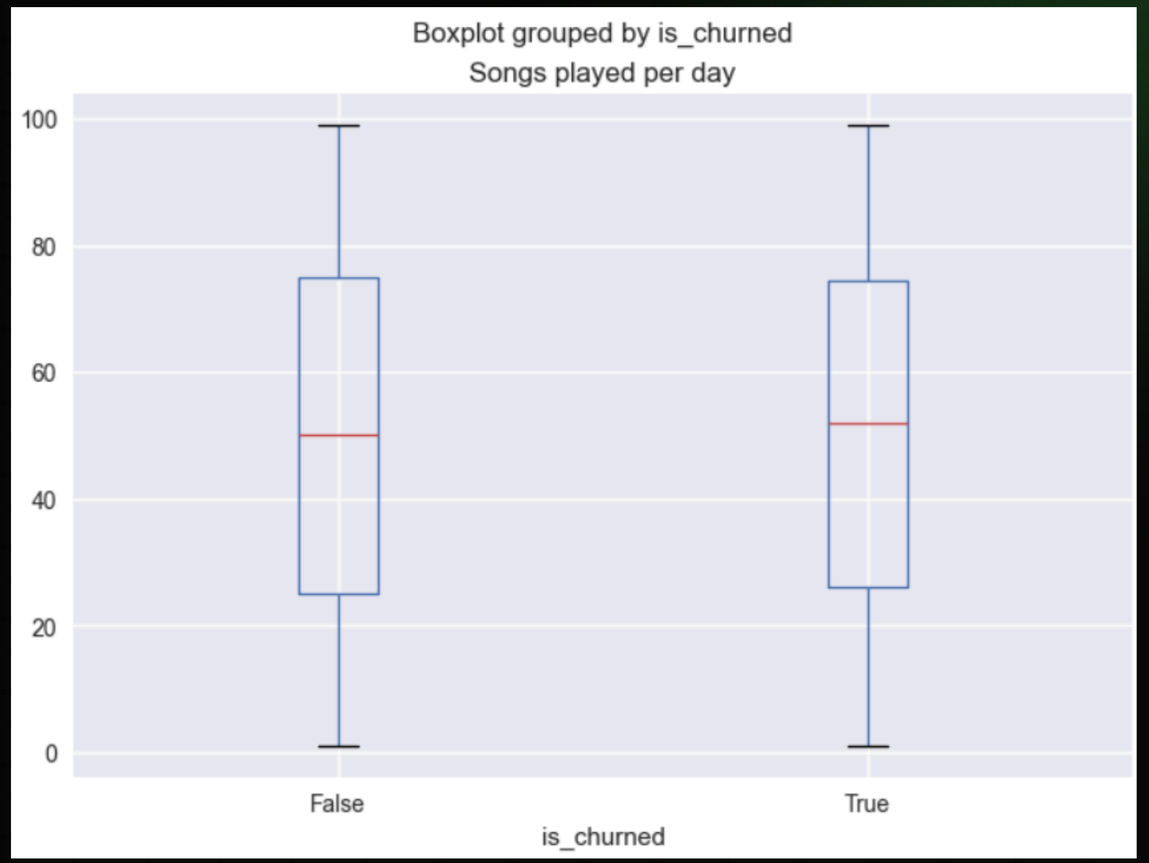
STATISTICAL TEST

- Unpaired t-test

EDA 3 RESULT

t-statistic = 0.8346
Critical t-value: 1.9611
p-value (one-tail): 0.2020 > 0.05

Accept the null hypothesis. There is no statistically significant difference in songs played per day between churned and non-churned users.



ALL MUSIC PODCAST ALBUMS

EDA Q4: IS SKIP RATE DIFFERENT ACROSS AGE GROUPS?

PURPOSE

- Examine whether skip behavior varies by age cohort

NUMERICAL SUMMARY

- Create age groups then report count, Measure of Central Tendency, and, Measure of Dispersion

VISUALIZATION

- Boxplot of skip_rate by age group

STATISTICAL TEST

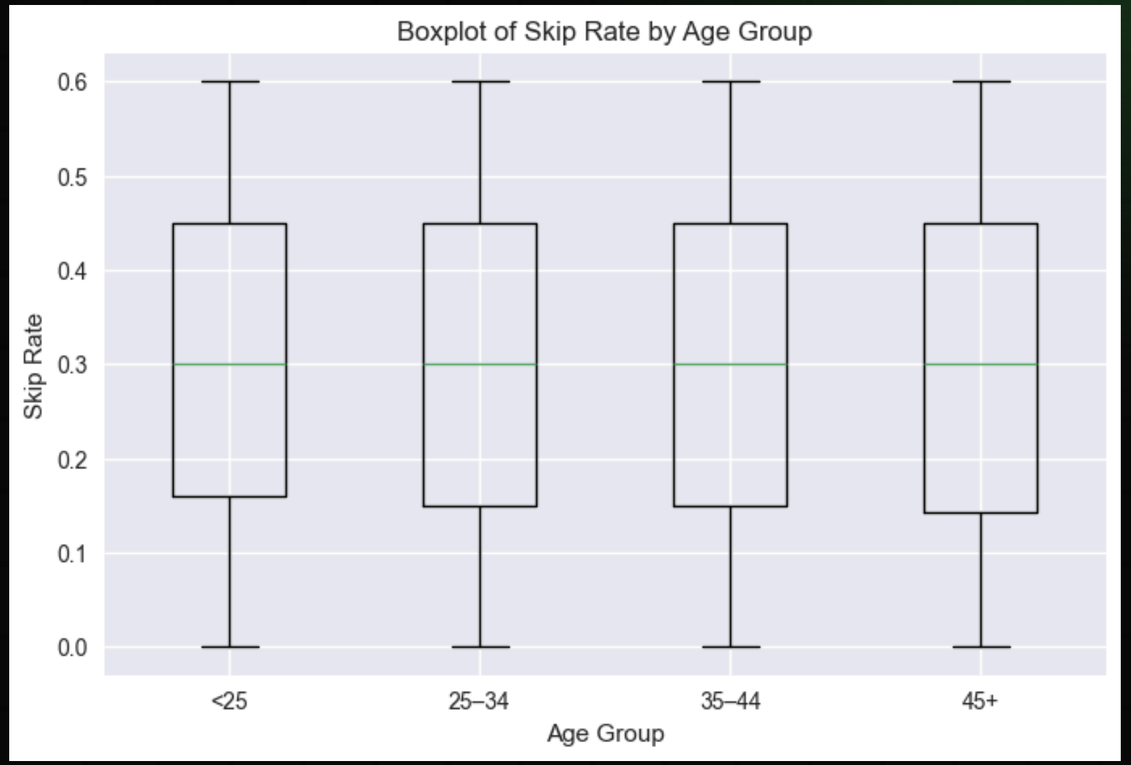
- Analysis of Variance (ANOVA)

ALL MUSIC PODCAST ALBUMS

EDA 4 RESULT

F-statistic: 0.7933744039027895
p-value: 0.49738385770378535
alpha: 0.05

We accept the null hypothesis. There is no significant difference in skip rates across the different age groups.



ALL MUSIC PODCAST ALBUMS

EDA Q5: DOES OFFLINE LISTENING BEHAVIOR DIFFER BETWEEN CHURNED AND ACTIVE USERS?

PURPOSE

- Explore if users who use offline downloads are less likely to churn

NUMERICAL SUMMARY

- Contingency table and proportions of offline_listening by is_churned

VISUALIZATION

- Stacked Bar Chart showing share of offline vs online listeners within churn status

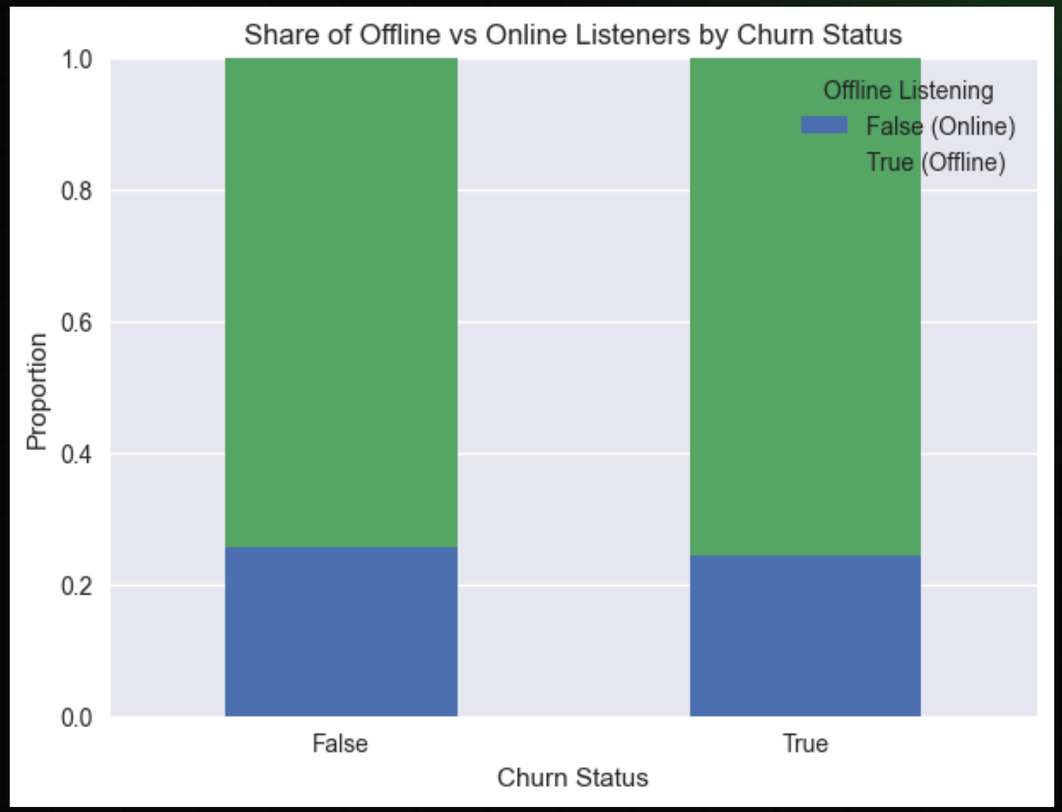
STATISTICAL TEST

- Chi-square test of independence

EDA 5 RESULT

Chi-square Statistic: 1.2351292752394203
Degrees of Freedom: 1
p-value: 0.26641183030820925
alpha: 0.05

Accept null hypothesis. There is no significant relationship between offline listening behavior and churn status.



RESEARCH QUESTION FORMED

ALL MUSIC PODCAST ALBUMS

What is the relationship between a user's skip_rate and the likelihood of customer churn?

- skip_rate measures the percentage of songs skipped and is a direct indicator of user dissatisfaction.
- Unlike other variables (e.g., listening_time, songs_played_per_day), skip_rate stands out as a potential driver of churn.
- Examining skip_rate vs is_churned sets up a binary classification problem for predicting churn.
- This analysis supports future project phases by guiding data modeling and statistical inference toward the most relevant variables.

REFERENCES

Caballar, R. (n.d.). Synthetic Data. IBM. https://www.ibm.com/think/topics/synthetic-data

Dilmegani, C. (2025, September 26). Top 20+ Synthetic Data use cases. AIMultiple. https://research.aimultiple.com/synthetic-data-use-cases/

Artificial intelligence and the growth of synthetic data. (2025, October 15). World Economic Forum. https://www.weforum.org/stories/2025/10/ai-synthetic-data-strong-governance/

Jolly, K. (2025, January 8). Everything you should know about synthetic data in 2025. Daffodil Unthinkable Software Corp. https://insights.daffodilsw.com/blog/everything-you-should-know-about-synthetic-data-in-2025

Wassel, B. (2024, November 11). How synthetic data might shape consumer research. CX Dive. https://www.customerexperiencedive.com/news/synthetic-data-consumer-research-customer-journey-qualtrics/732408/

ALL MUSIC PODCAST ALBUMS

THANK YOU

FOR LISTENING

GALICIA, MOJICA, PLURAD, RAMIREZ, TOLENTINO