Fetch Data Analysis Report

1. Introduction

The purpose of this report is to analyze Fetch's given dataset, identify data quality issues, extract meaningful insights, and provide actionable business recommendations. This analysis integrates exploratory data analysis (EDA) conducted in Python, SQL queries used to extract key metrics, key findings, and visualizations from the data.

2. Data Quality Issues

2.1 Missing Values

- **Users Data:**
    - `BIRTH_DATE`: 1,559 missing values
    - `LANGUAGE`: 28,235 missing values
    - `GENDER`: 2,483 missing values
    - **Strong correlation** (96.79%) between missing `BIRTH_DATE` and missing `GENDER`, requiring further investigation.
- **Transactions Data:**
    - `BARCODE`: 5,762 missing values, possibly due to unscanned items or technical issues.
    - `FINAL_QUANTITY = 0` but still records a `FINAL_SALE` amount, suggesting inconsistencies in quantity and price mapping.
- **Products Data:**
    - Multiple columns with missing values:
        - `CATEGORY_1`: 111
        - `CATEGORY_2`: 1,424
        - `CATEGORY_3`: 60,566
        - `CATEGORY_4`: 778,093
        - `MANUFACTURER`: 226,474
        - `BRAND`: 226,472
        - `BARCODE`: 4,025
    - **Duplicate Products:** 215 records found.

2.2 Inconsistencies in Categorization

- Different versions of the same gender label (e.g., `non_binary` vs. `Non-Binary`).
- Store name variations: (e.g., `TRADER JOE'S` vs. `TRADER JOES`).
- **Category Misclassification:**
    - `Hard Seltzers` grouped under non-alcoholic beverages.
    - `Soda` in restaurants may refer to fountain sodas, different from retail-packaged sodas.

2.3 Data Type Inconsistencies

- `FINAL_QUANTITY` contains non-numeric values such as 'zero' instead of numeric 0, leading to potential calculation errors in sales analysis.

2.4 Final Quantity and Sales Mapping Issues

- `FINAL_QUANTITY = 0` should logically lead to `FINAL_SALE = 0`, but discrepancies were found.

2.5 Brand Standardization Issues

- `L'OREAL PARIS COSMETICS`, `L'OREAL PARIS HAIR COLOR`, and `L'OREAL PARIS HAIR CARE` are all listed as different brands, but they are the same brand.

2.6 Business Implications of Data Quality Issues

- **Incomplete user data (Birth Date, Gender, Language)** can impact customer segmentation, personalization, and targeted marketing efforts, limiting Fetch's ability to offer tailored promotions.
- **Missing barcode information in transactions** may result in inaccurate sales tracking, affecting inventory decisions and brand partnerships.
- **Inconsistencies in Final Quantity and Sales Mapping** could lead to revenue miscalculations, making financial forecasting unreliable.
- **Duplicate records in products** can skew analytics, leading to incorrect business insights and poor decision-making.

3. Outstanding questions about the data

- Are the missing values in `BIRTH_DATE, GENDER,` and `LANGUAGE` due to user opt-out policies, or is it a data collection issue?
- What is the product hierarchy? What do the different categories represent in the product dataset?
- Is a receipt assigned a new receipt_ID if re-uploaded?
- Does the Final sale reflect the item's final price paid after discounts/coupons or the original price before adjustments?

4. Exploratory Data Analysis (EDA) Findings

4.1 Age Distribution Analysis

- The distribution is right-skewed, meaning there are more younger users than older ones.
- There are multiple peaks, suggesting distinct age groups in the data.
- **Peak Age Groups:**
    - A significant concentration of users falls within the 20-30 and 40-50 age ranges.
    - Another smaller peak appears around 60+, indicating a notable presence of older users.
- Some users appear to have extreme ages (e.g., above 100), which could indicate data entry errors or unusual cases.
- **Diversity of Users:** The spread of ages suggests a broad range of users from young adults to older individuals.
- Fetch should tailor marketing strategies and promotions to target different age groups.

4.2 User Signups Over Time

- The number of signups increased steadily from 2017 to 2022, with the sharpest growth between 2019 and 2022.
- The highest peak in signups was in 2022, followed by a decline in 2023 and early 2024.
- The rapid acceleration between 2019 and 2021 may have been influenced by increased digital adoption during COVID-19.
- **Key Business Insights:**

    - Fetch should analyze 2022's success to determine what drove high signups (ads, referral programs, partnerships).
    - Investigate the 2022 decline—was it due to product changes, marketing reductions, or external factors?
    - The recent spike in 2024 signups should be assessed—if it results from promotions, Fetch should sustain momentum with continuous engagement strategies.
    - Consider predictive modeling to forecast future signup trends based on historical patterns.

5. SQL Insights

5.1 Top 5 Brands by Receipts Scanned Among Users 21+

- I identified the most frequently scanned brands among users aged 21+, providing insights into high-engagement products.
- **Key Finding:** DOVE, NERDS CANDY, COCA-COLA, HERSHEY'S, and SOUR PATCH KIDS had the highest number of receipts scanned.
- **Business Implication for Fetch:** These brands demonstrate strong customer engagement. Fetch can explore potential brand partnerships and promotions.

5.2 Identifying Fetch's Power Users

- **Assumption:** Power users are defined as those contributing to 80% of total receipts, total revenue, or total transactions (Pareto principle).

- **Methodology:** I used three different approaches to identify power users:

    - **Receipt-Based Approach:** Users who contribute to 80% of total uploaded receipts.
    - **Revenue-Based Approach:** Users who contribute to 80% of total revenue.
    - **Transaction-Based Approach:** Users who contribute to 80% of total transactions.
- **Key Findings:**

    - A few users contribute to 80% of total receipts, indicating a small but highly engaged user base.
    - A similarly small number of users drive 80% of total revenue and transactions.
    - The receipt-based, revenue-based, and transaction-based approaches all showed that **a minority of users generate the majority of Fetch's activity**.

- **Business Implication for Fetch:**
    - Fetch should prioritize **loyalty programs, exclusive rewards, and personalized offers** for these users to retain high-value customers.
    - Understanding what keeps these users engaged can help convert more casual users into power users.
    - Fetch can use targeted retention efforts to encourage **more users to transition into the power user category**.

5.3 Leading Brand in the Dips & Salsa Category

- **Key Finding:** The analysis identified **Tostitos** as the leading brand based on revenue and sales volume.
- **Business Implication for Fetch:**
    - This presents an opportunity for **strategic brand partnerships and targeted promotions** in the Dips & Salsa category.
    - Fetch can use this insight to **recommend relevant promotions to users who frequently purchase from this category**.

6. Trends and Interesting Insights

6.1 User Inactivity Trends

- **50% of inactive users disengage within the first 30 days**, suggesting that it's a critical period for re-engagement.
- **Another 50% churn between 31-90 days**, indicating a gradual drop-off in user activity.
- **No users are inactive beyond 90 days** due to data limitations- the current data set is limited to recent transactions.

Business Implications for Fetch:

- Fetch should implement targeted campaigns (discounts, notifications, reminders) within **the first 30 days** to retain users.
- Introduce loyalty programs or personalized incentives for users at **the 31-90 day mark** for user retention.

6.2 Seasonal Trends in Receipt Uploads

- A significant spike in receipt uploads occurs during summer compared to fall, suggesting higher shopping activity in warmer months.
- Fetch should analyze whether this trend is linked to seasonal promotions, holiday events, or increased outdoor activities driving purchases.

Business Implications for Fetch:

- Develop seasonal marketing campaigns to maximize engagement during peak shopping periods.
- Introduce targeted promotions during lower-activity seasons (e.g., fall) to balance engagement throughout the year.

7. Request for Action

To further improve the analysis and address outstanding issues, Fetch should consider the following actions:

- Include **comprehensive seasonal data** (winter, spring, summer, and fall) to identify full-year shopping behaviors and seasonal shopping patterns.
- Understanding the strategies used during peak signup years (2019-2022) can help explain fluctuations in user acquisition and engagement.
- App usage metrics, email click-through rates, and notification responses could help determine how engagement strategies influence user retention.
- Provide long-term user engagement data to build better churn prediction models.

8. Conclusion

This analysis identified key data quality issues, transaction trends, and user engagement patterns, providing actionable insights for Fetch.

Key Findings:

- **Early Churn Risk:** 50% of inactive users disengage within 30 days, stressing the need for early re-engagement efforts.
- **Seasonal Trends:** Receipt uploads peak in summer, suggesting a need for targeted seasonal marketing.
- **Power Users' Impact:** A small group of users contributes significantly to transactions and revenue, making loyalty programs essential.
- **Data Limitations:** The absence of long-term data restricts deeper churn analysis and predictive modeling.

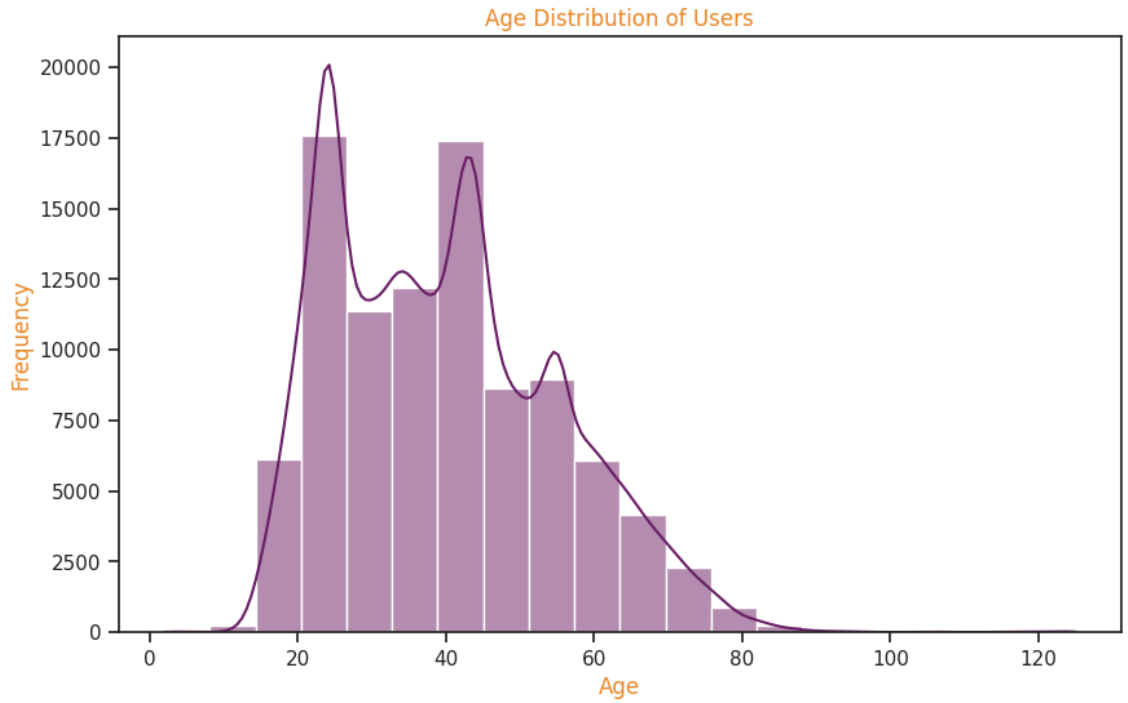
Recommendations for Fetch:

- **Improve Data Quality** – Address missing values, inconsistencies, and standardization issues.
- **Provide More Data** – Access to historical transactions will enable better analysis of user retention and seasonal trends.
- **Enhance Retention Strategies** – Implement personalized marketing, loyalty programs, and re-engagement efforts.
- **Leverage Predictive Modeling** – Use advanced analytics to forecast churn and optimize customer engagement.

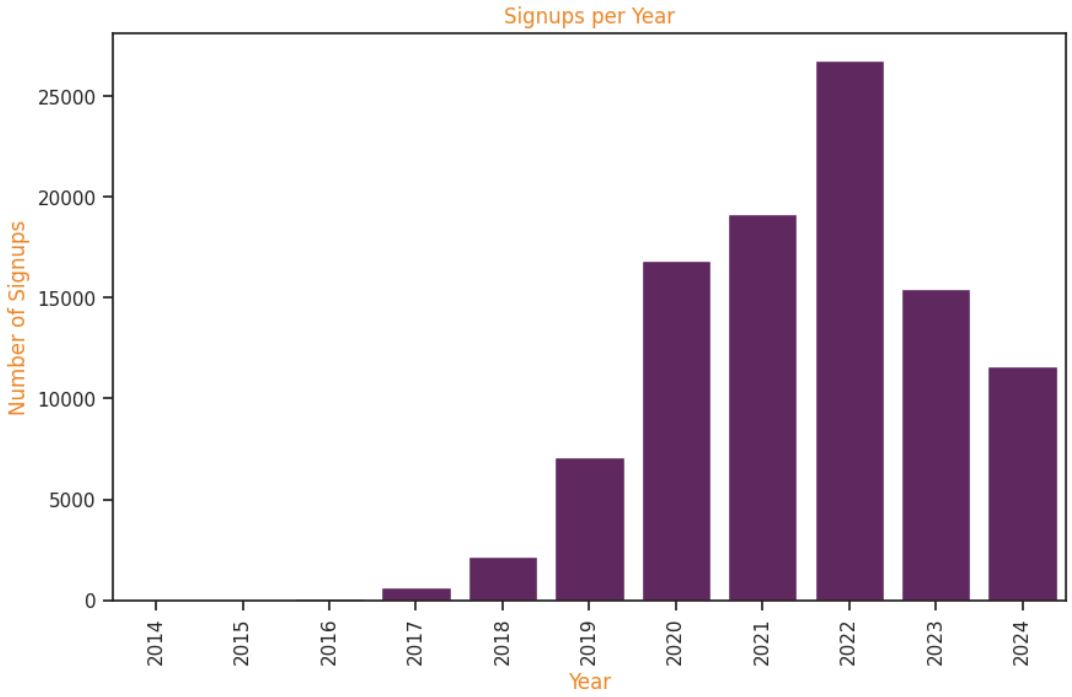
By acting on these recommendations, Fetch can improve data reliability, enhance customer retention, and drive long-term business growth.

9. APPENDIX

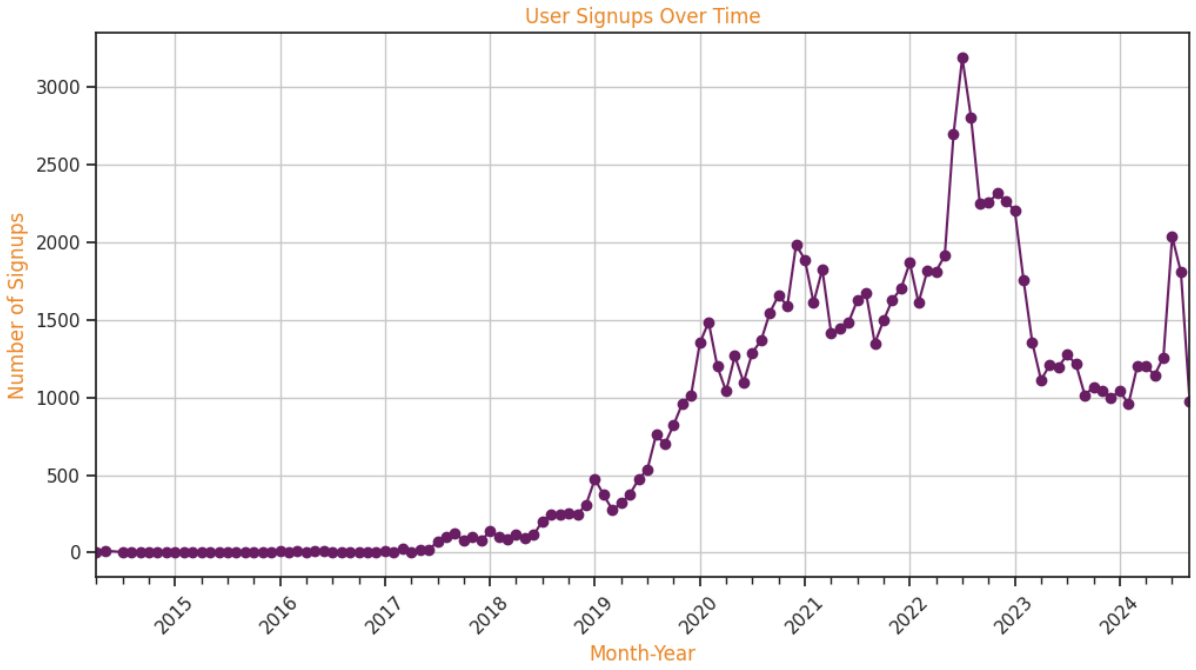
9.1 Age Distribution of Users



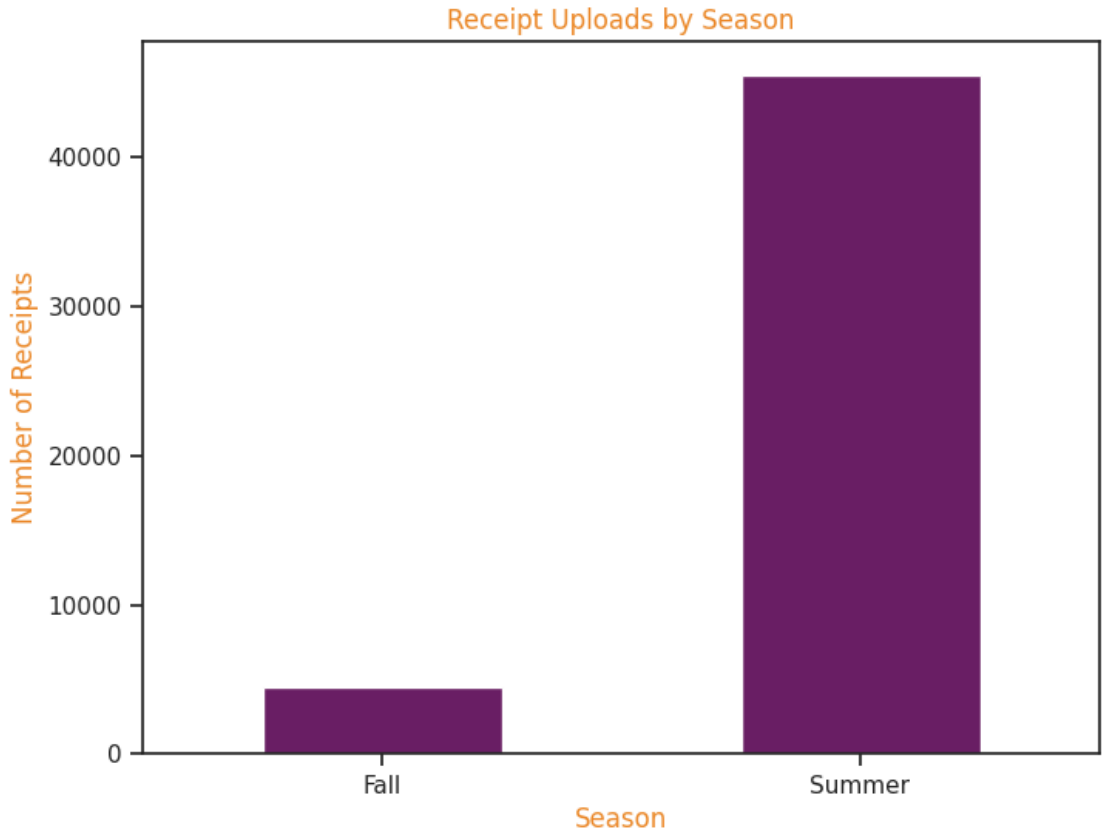
9.2 Bar chat signups per year (compare different years)



9.3 User-signs over time (shows change over time)



9.4 Seasonal analysis of receipt uploads.



9.5 Percentage of Inactive Users by Time Range