DATA ANALYSIS REPORT

Table of Contents

Introduction

The Google Play Store, being one of the largest digital distribution service platforms, hosts millions of apps spanning various categories and genres. Understanding the patterns and trends in this dataset can provide valuable insights to developers, investors, and marketers. The goal of this study is to analyze the Google Play Store dataset, focusing on answering the following questions:

1. Which app categories have the highest average rating?
2. Which categories have the most installations on average?
3. Is there a correlation between rating and number of installs?
4. Do paid apps generally have better ratings than free apps?
5. Which genres are most common in highly rated apps (rating > 4.5)?
6. Which content ratings (e.g. Everyone, Teen, Mature) have the highest-rated apps?
7. What is the average price of apps per category?
8. Is there a relationship between price and installs?
9. Do recently updated apps tend to have better ratings?

Project Files Description

This Project includes:

1. Jupyter Notebook file
2. Presentation PDF

Data Source

The dataset for this project comes from Kaggle, Dataset.

Features

App Category: Category of the app. This could be beauty, business, entertainment, education…etc.
Rating: How users rate the app out of 5, with 1 being the lowest rating and 5 being the highest.
Reviews: The number of users reviews each app has received.
Size: The memory size needed to install the application.
Installs: The number of times each application has been installed by users.
Type: Whether the app is free or a paid app.
Price: The price of the app.

Content Rating: This column specifies the intended audience for the app. Can be for teens, mature audiences, or everyone.
Genres: The sub-category for each app. Example: for the Education category, this could be Education: Pretend Play, for example.
Last Updated: Release date of the most recent update for the app.
Current Ver: The app's current version.
Android Ver: The oldest version of Android OS supported by the app.

Data Cleaning

Data cleaning involves removing duplicates, filling or removing missing values in critical columns(features), changing the data types to reflect the appropriate ones, removing special characters, if needed from numerical columns.

For our dataset, we first removed special characters like + from Installs, $ from Price.
Then, the data types need to be changed for numerical columns from string to float. Last Updated column should be Datetime.
Next step is to remove rows with missing values in the columns, Rating, Reviews, Category, Installs, Type and Price.
Next, we remove rows that have missing 'App' values.
Then we remove columns that are almost empty, meaning a meaningful analysis cannot be performed when a column is not missing values in at least 20% by setting a threshold.

Exploratory Data Analysis

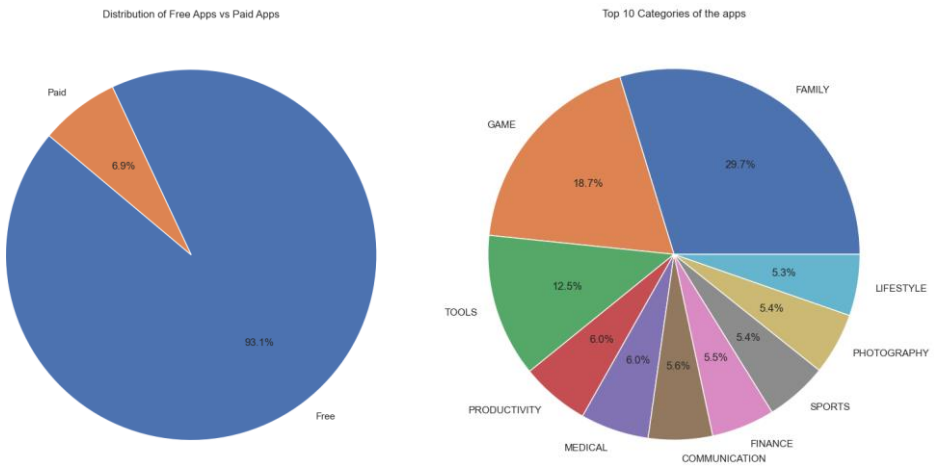
After cleaning the dataset, the number of apps in the dataset is 9366, of which 8719 are free and 647 are paid.

There are 33 categories of apps in total and those are listed below:

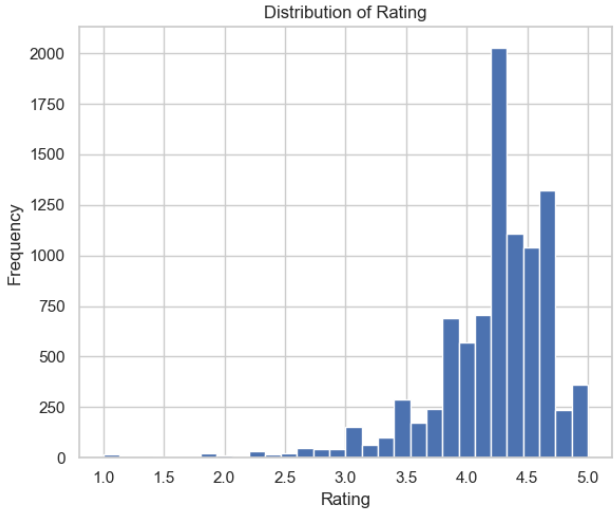
Art_and_design
auto_and_vehicles
beauty
books_and_reference
business
comics
communication
dating
education
entertainment

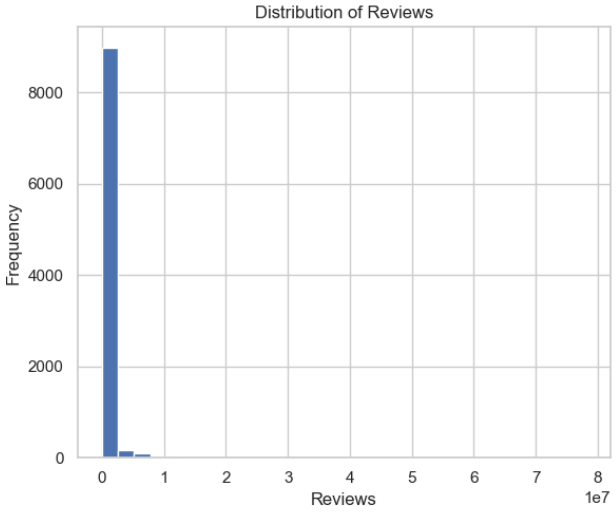
events
finance
food_and_drink
health_and_fitness
house_and_home
libraries_and_demo
lifestyle
game
family
medical
social
shopping
photography
sports
travel_and_local
tools
personalization
productivity
parenting
weather
video_players
news_and_magazines
maps_and_navigation

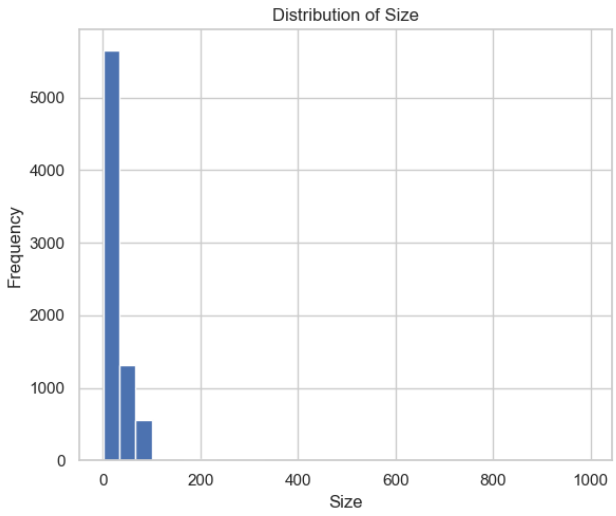
Of all the categories, Family seems to be the most popular category in the store with 1747 apps, nearly 30%.

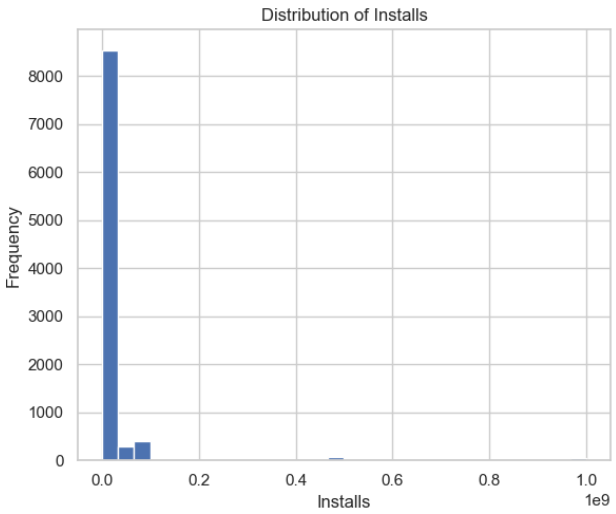


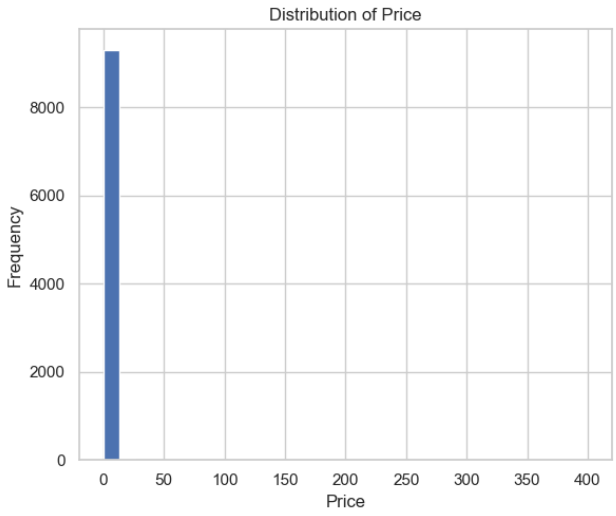
Let’s look at the distribution and the averages of the features like Price, Reviews, Ratings (numerical columns).











Distributions are right-skewed, with means significantly higher than medians in both *Reviews* and *Installs*, indicating a few apps have very large numbers pulling the average up.

Findings

1. The average price of paid apps is $13.91, with the highest price $400 and the lowest price being $0.99.

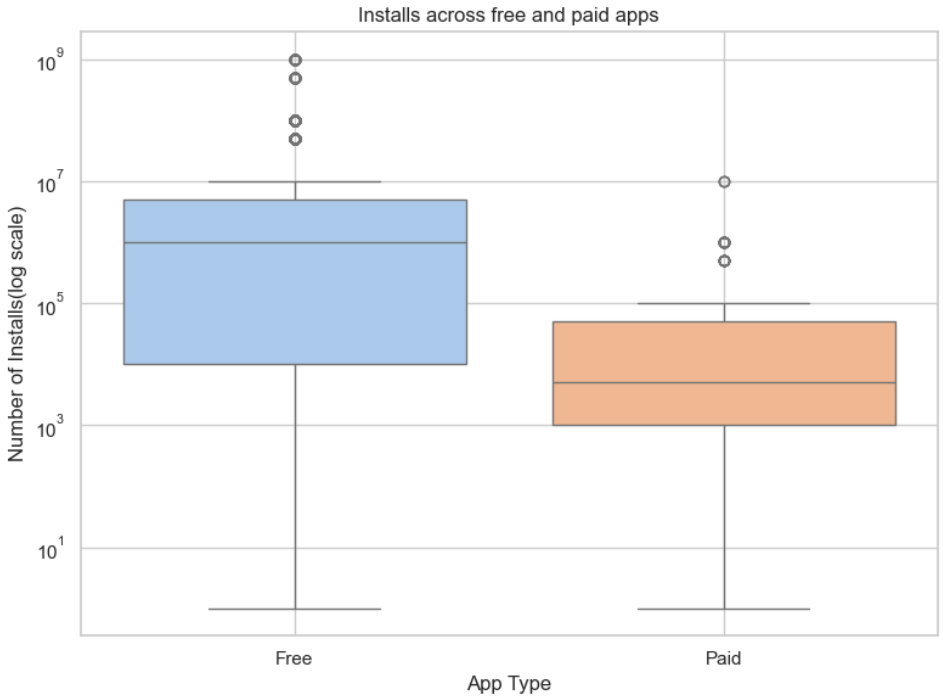
2. Facebook has the highest number of reviews (78158306).

3. Average number of installations: ~17.8 million. COMMUNICATION has the highest number of installations with 99M, which includes apps like Whatsapp, Facebook Messenger, UC Browser, etc. Next follows SOCIAL with 54M installs having apps like Facebook, Instagram, Snapchat, etc. Here's the average number of installations by app category (top 5 shown below):
    1. **COMMUNICATION**: ~84.4 million
    2. **SOCIAL**: ~47.7 million
    3. **VIDEO_PLAYERS**: ~35.6 million
    4. **PRODUCTIVITY**: ~33.4 million
    5. **GAME**: ~30.7 million

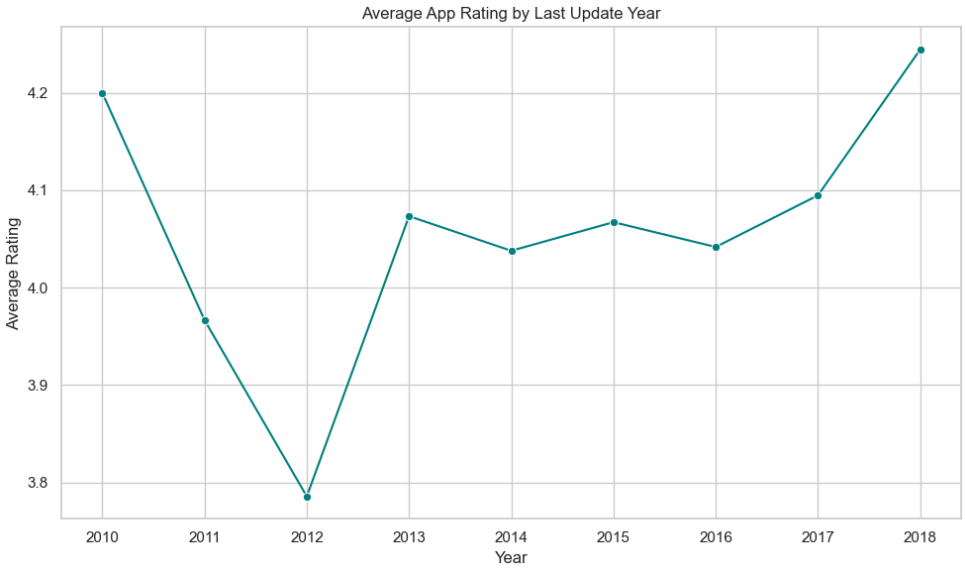
4. The top 5 app categories that have the highest average rating are:
    1. EVENTS: 4.435556
    2. EDUCATION: 4.389032
    3. ART_AND_DESIGN: 4.358065
    4. BOOKS_AND_REFERENCE: 4.346067
    5. PERSONALIZATION: 4.335987

Correlation

There is a very weak correlation between **Rating** and **Installs**. Which shows that apps with most installs don't necessarily have better ratings. Upon checking the correlation between type (free or paid) and installations. As we are performing correlation calculation between a categorical feature and a numerical feature, we cannot use the .corr() method. We can perform the ANOVA statistical test to check the difference in means across groups. As p-value < 0.05, there is a statistically significant difference between installs of Free and Paid apps. Free apps have WAY more installs.



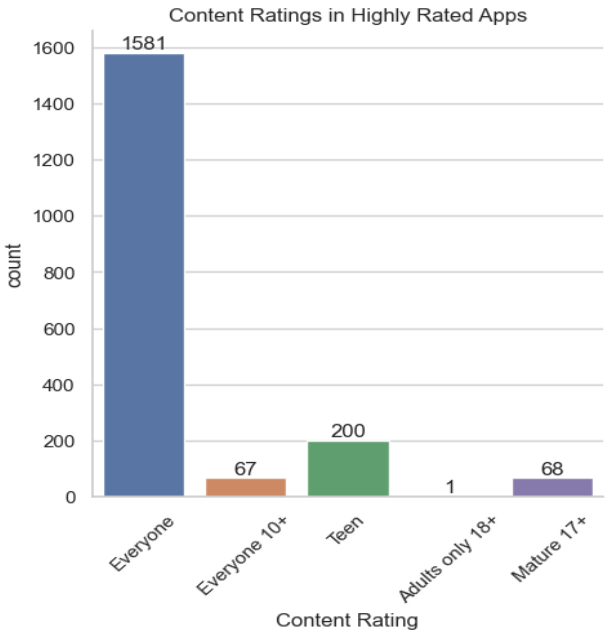
There is a slight positive relationship between **rating** and **recent updates**, but it is not strong enough to be considered significant on its own.



Relationship between Installs, Ratings and Reviews:

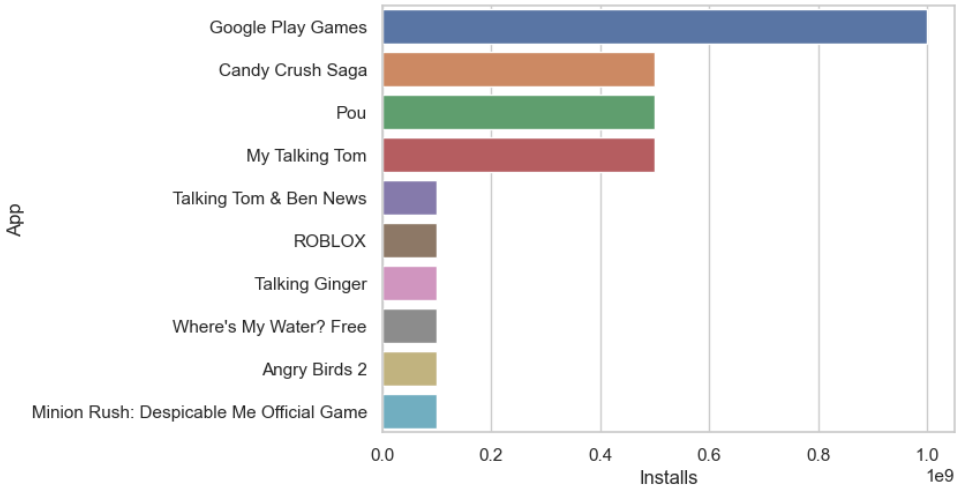
Data Analysis

There are 1917 apps that are rated above 4.5, of which 1581 apps have ‘Everyone’ as Content Rating. The average rating of the Family Apps is 4.192 which is almost the same as the average rating of all apps (4.191).



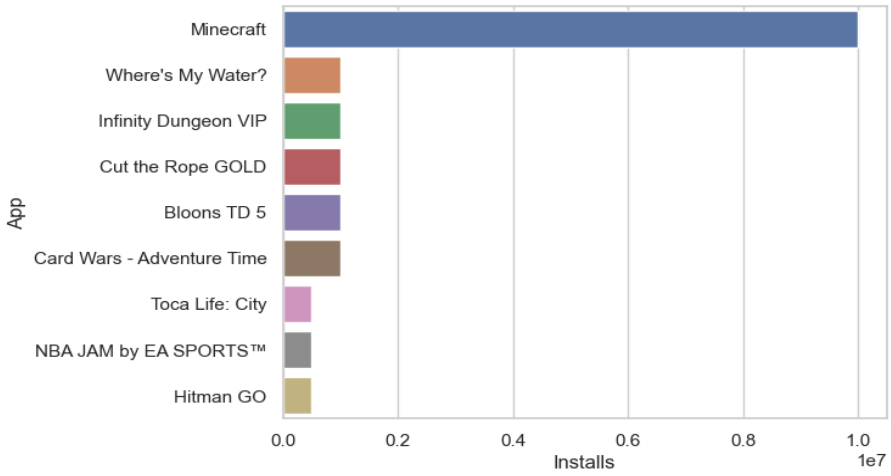
Top 10 Apps in the Family Category

Game Apps have the maximum installations in the Family Category.



Top 10 Paid Apps in the Family Category

Minecraft is the most popular app with a huge number of installations in the Family Category for paid apps.

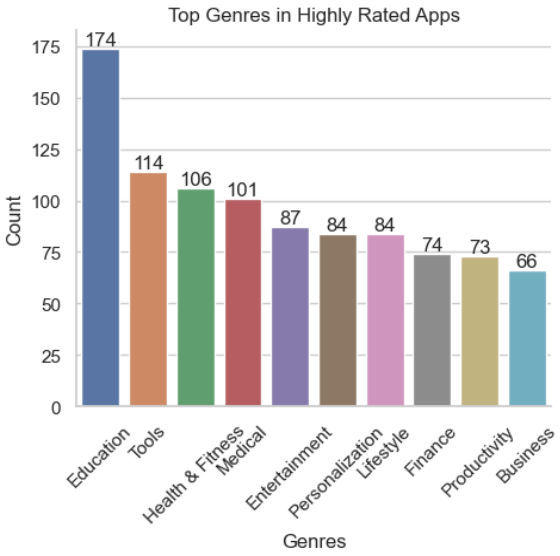


The average ratings for the apps are as follows:

Average rating of free apps: 4.186
Average rating of paid apps: 4.267
Average rating of apps: 4.192

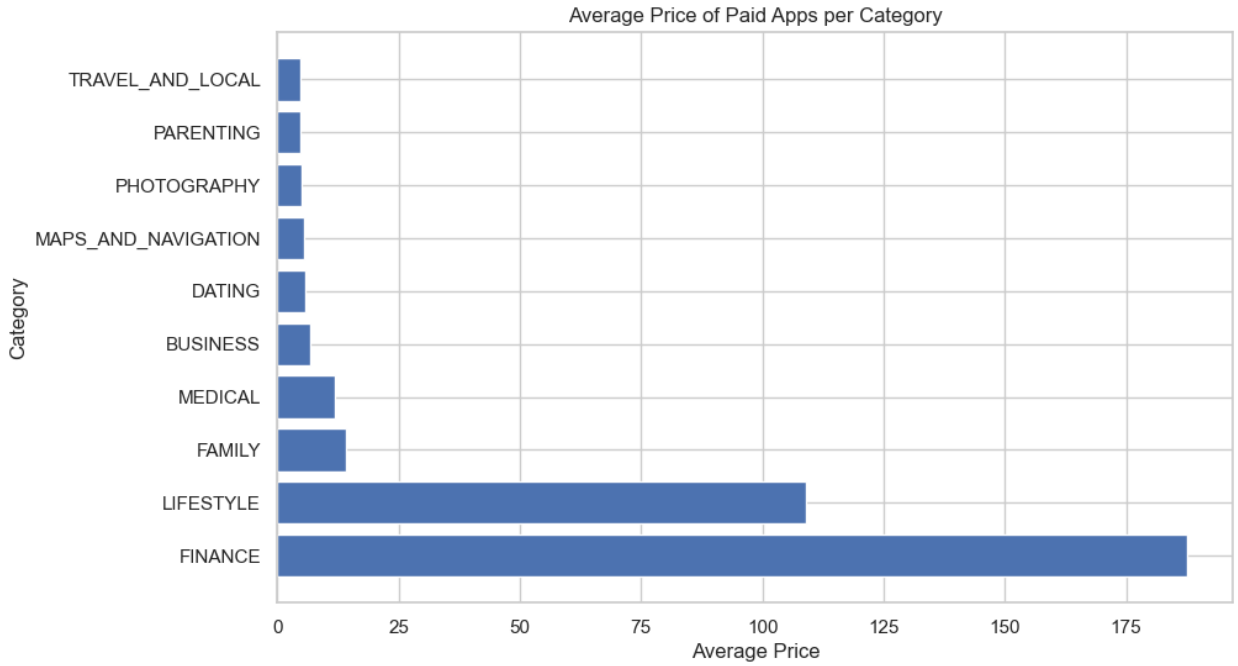
This shows that the free apps, with an average rating of 4.186, are rated lower than paid apps with 4.267 average rating. The paid apps are rated higher than the average score of all apps. The top 5 Genres that have highly rated apps with counts are:

1. Education - 174
2. Tools - 114
3. Health & Fitness - 106
4. Medical - 101
5. Entertainment – 87



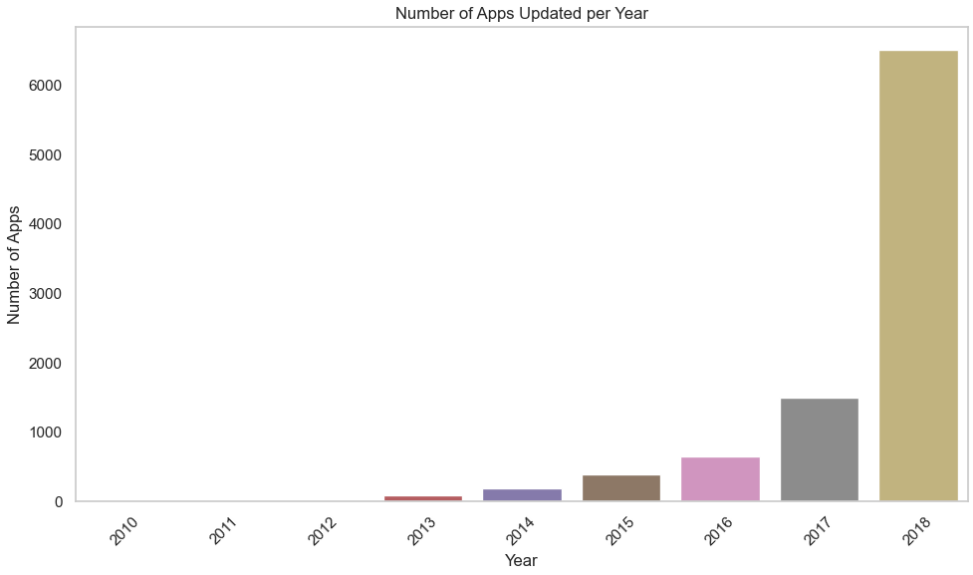
When we delve into Education Genre, the free App, ***Learn English with Wlingua*** is rated 4.7 with 10M installations. The App, ***Learn languages, grammar & vocabulary with Mem*** is free and 10M installations with 1,107,948 reviews.

The Finance Category has the maximum average price in the paid apps($187.68), followed by Lifestyle with $108.93.



Most apps are clustered under $10 and under 10,000 installations. A few high-priced apps are extreme outliers. Overall trend: higher price → fewer installs

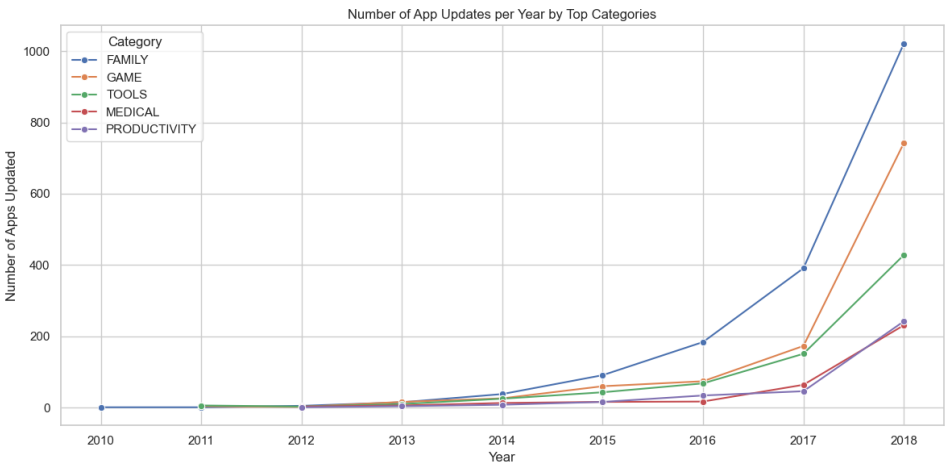
Let's visualize how app updates have changed over the years. Looks like, most of the apps are updated in 2018(current, as the dataset is last updated in 2018).



Let’s visualize app updates over time, broken down by categories like:

✅ App category (e.g. GAME, TOOLS, EDUCATION)

✅ App type (Free vs Paid)



Family and Game apps have sharp increase in updates from year 2017 to 2018. In contrast, Productivity apps have consistent growth.