Google Play Store Dataset Analysis

Project Overview

This project conducts a comprehensive analysis of the Google Play Store dataset to uncover patterns, trends, and insights that can guide app developers, investors, and marketers in making data-driven decisions. The analysis examines app categories, ratings, installations, pricing strategies, and user engagement metrics across 9,366 applications.

Technologies Used: Python, Pandas, NumPy, Matplotlib, Seaborn, Scipy (ANOVA), Statistical Analysis

Business Problem

The mobile app market is highly competitive with millions of apps vying for user attention. Understanding market dynamics is crucial for:

- **App Developers:** Identifying profitable categories and optimal pricing strategies
- **Investors:** Evaluating market opportunities and app performance metrics
- **Marketers:** Understanding user preferences and engagement patterns
- **Business Analysts:** Tracking industry trends and competitive positioning

Research Questions

This analysis addresses key business 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 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?

Dataset Description

Dataset Size: 9,366 apps after data cleaning **App Distribution:** 8,719 free apps (93.1%) and 647 paid apps (6.9%) **Categories:** 33 distinct app categories **Time Period:** Data last updated in 2018

Methodology

1. Data Cleaning Process

- **Special Character Removal:** Eliminated '+' symbols from Installs column and '$' from Price column
- **Data Type Conversion:** Converted string columns to appropriate numerical formats (float for prices, datetime for Last Updated)

- **Missing Value Treatment:** Removed rows with missing values in critical columns (Rating, Reviews, Category, Installs, Type, Price)
- **Quality Threshold:** Removed columns with >80% missing values to ensure meaningful analysis
- **Duplicate Removal:** Ensured data integrity by removing duplicate entries

2. Exploratory Data Analysis

- Statistical summary of numerical variables
- Distribution analysis of key metrics (Price, Reviews, Ratings)
- Category-wise performance comparison
- Correlation analysis between variables

3. Statistical Testing

- **ANOVA Test:** Analyzed differences between free and paid app installations
- **Correlation Analysis:** Examined relationships between ratings, installs, and pricing
- **Significance Testing:** Validated findings with p-value < 0.05 threshold

Key Findings

Market Landscape

- **Category Dominance:** Family category leads with 1,747 apps (30% of total dataset)
- **Business Model:** Free apps dominate the market (93.1% vs 6.9% paid apps)
- **Price Range:** Paid apps range from $0.99 to $400, with an average of $13.91

Performance Metrics

Top-Performing Categories by Average Rating

1. **EVENTS:** 4.436 average rating
2. **EDUCATION:** 4.389 average rating
3. **ART_AND_DESIGN:** 4.358 average rating
4. **BOOKS_AND_REFERENCE:** 4.346 average rating
5. **PERSONALIZATION:** 4.336 average rating

Installation Leaders by Category

1. **COMMUNICATION:** ~84.4 million average installs (WhatsApp, Facebook Messenger)
2. **SOCIAL:** ~47.7 million average installs (Facebook, Instagram, Snapchat)
3. **VIDEO_PLAYERS:** ~35.6 million average installs
4. **PRODUCTIVITY:** ~33.4 million average installs
5. **GAME:** ~30.7 million average installs

Pricing and Revenue Analysis

- **Highest-Priced Categories:**
    - Finance: $187.68 average price
    - Lifestyle: $108.93 average price
- **Price-Install Relationship:** Strong inverse correlation - higher prices lead to fewer installations
- **Revenue Concentration:** Most apps cluster under $10 with <10,000 installations

Quality and User Engagement

- **High-Quality Apps:** 1,917 apps rated above 4.5 (20.5% of dataset)
- **Content Rating Impact:** 1,581 apps (82.5%) in the 4.5+ rating category have "Everyone" content rating
- **Free vs. Paid Quality:**
    - Free apps: 4.186 average rating
    - Paid apps: 4.267 average rating
    - Overall average: 4.192 rating

Notable Insights

- **Review Champion:** Facebook leads with 78,158,306 reviews
- **Family Category Performance:** Despite popularity, Family apps have average ratings (4.192) similar to overall dataset
- **Premium Gaming:** Minecraft dominates paid Family category apps
- **Educational Excellence:** Education genre has 174 highly-rated apps (most in 4.5+ category)

Statistical Analysis Results

Correlation Findings

- **Rating vs. Installs:** Very weak correlation - popular apps don't necessarily have better ratings
- **App Type vs. Installations:** ANOVA test ($p < 0.05$) confirms statistically significant difference - free apps have significantly more installs
- **Update Frequency vs. Rating:** Slight positive relationship, but not statistically significant

Distribution Characteristics

- **Right-Skewed Distributions:** Reviews and Installs show extreme outliers pulling averages above medians
- **Price Concentration:** Most apps priced under $10 with few high-priced outliers

- **Update Patterns:** Sharp increase in app updates from 2017 to 2018, particularly in Family and Game categories

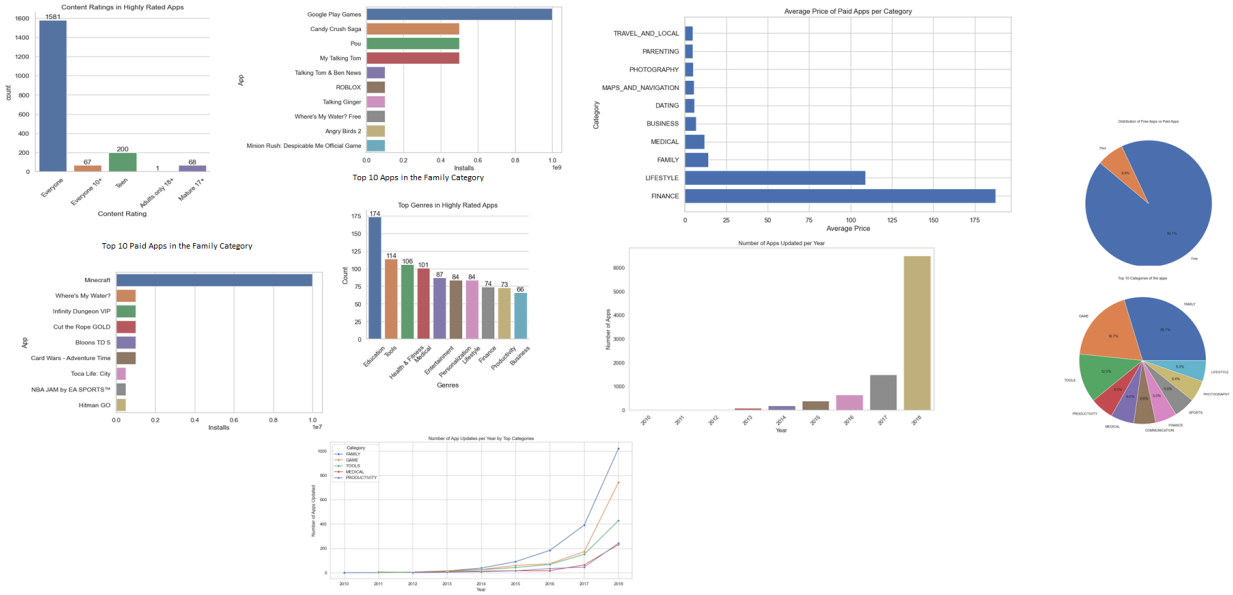
Technical Implementation

```
python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway

# Data cleaning process
def clean_playstore_data(df):
    # Remove special characters
    df['Installs'] = df['Installs'].str.replace('+', '').str.replace(',',
'')
    df['Price'] = df['Price'].str.replace('$', '')

    # Convert data types
    df['Installs'] = pd.to_numeric(df['Installs'], errors='coerce')
    df['Price'] = pd.to_numeric(df['Price'], errors='coerce')
    df['Last Updated'] = pd.to_datetime(df['Last Updated'])

    # Remove missing values
    critical_columns = ['Rating', 'Reviews', 'Category', 'Installs',
'Type', 'Price']
    df = df.dropna(subset=critical_columns)

    return df

# Statistical analysis
def analyze_app_performance(df):
    # Category-wise ratings
    category_ratings =
df.groupby('Category')['Rating'].mean().sort_values(ascending=False)

    # Installation analysis
    category_installs =
df.groupby('Category')['Installs'].mean().sort_values(ascending=False)

    # ANOVA test for free vs paid
    free_installs = df[df['Type'] == 'Free']['Installs']
```

```
    paid_installs = df[df['Type'] == 'Paid']['Installs']
    f_stat, p_value = f_oneway(free_installs, paid_installs)

    return category_ratings, category_installs, p_value
```

Key Visualizations Created

1. **Category Distribution Bar Charts**
2. **Distribution Histograms**
3. **Price vs. Installs Scatter Plots**
4. **Time Series Analysis**
5. **Correlation Heatmaps**



Business Implications

For App Developers

- **Category Selection:** Focus on Education, Events, and Art & Design for higher ratings
- **Pricing Strategy:** Keep prices under $10 for broader market reach
- **Content Rating:** Target "Everyone" rating for maximum reach
- **Update Strategy:** Regular updates show slight positive correlation with ratings

For Investors

- **High-Volume Opportunities:** Communication and Social categories show massive install potential

- **Premium Segments:** Finance and Lifestyle categories command higher prices
- **Quality Indicators:** Paid apps generally maintain higher ratings (4.267 vs 4.186)

For Marketers

- **Market Saturation:** Family and Game categories are highly competitive
- **User Engagement:** Focus on review generation strategies (Facebook's 78M reviews)
- **Timing:** 2018 showed significant update activity - plan marketing around update cycles

Challenges and Solutions

Data Quality Issues

- **Challenge:** Mixed data types and special characters in numerical columns
- **Solution:** Systematic regex-based cleaning and type conversion pipeline

Statistical Complexity

- **Challenge:** Analyzing categorical vs. numerical relationships
- **Solution:** Applied ANOVA testing instead of standard correlation for Type vs. Installs

Outlier Management

- **Challenge:** Extreme outliers in price and install distributions
- **Solution:** Used both mean and median analysis to account for skewed distributions

Future Enhancements

1. **Sentiment Analysis** - Analyze user reviews for qualitative insights
2. **Predictive Modeling** - Build models to predict app success metrics
3. **Competitive Analysis** - Deep-dive into category-specific competition
4. **Seasonal Trends** - Analyze temporal patterns in app performance
5. **User Behavior Analysis** - Study install-to-rating conversion patterns

Conclusion

This comprehensive analysis of 9,366 Google Play Store apps reveals a market dominated by free applications with significant opportunities in premium categories. The findings show that while Communication and Social categories drive massive installations, Educational and Events categories achieve higher user satisfaction ratings. The inverse relationship between price and installations emphasizes the importance of strategic pricing, while the slight quality advantage of paid apps suggests sustainable premium positioning opportunities.

The analysis demonstrates proficiency in data cleaning, statistical analysis, and business intelligence - essential skills for data-driven decision making in the mobile app industry. The project showcases ability to transform raw data into actionable insights that can guide strategic business decisions.

Technical Skills Demonstrated

Data Analysis

- **Advanced Data Cleaning:** Handled mixed data types, special characters, and missing values
- **Statistical Testing:** Applied ANOVA, correlation analysis, and significance testing
- **Business Intelligence:** Converted technical findings into strategic recommendations

Programming & Tools

- **Python:** Pandas for data manipulation, NumPy for numerical computations
- **Statistical Analysis:** Scipy for statistical tests, comprehensive EDA
- **Data Visualization:** Matplotlib and Seaborn for professional visualizations

Domain Knowledge

- **Mobile App Market:** Understanding of app store dynamics and user behavior
- **Revenue Models:** Analysis of freemium vs. premium app strategies
- **Market Segmentation:** Category-wise performance analysis and competitive positioning

Repository Structure

```
Google-Play-Store-Analysis/
├── google-play-store-analysis.ipynb   # Main analysis notebook
├── data/                              # Dataset files
├── visualizations/                    # Generated charts and graphs
├── README.md                          # Project documentation
└── requirements.txt                   # Dependencies
```