



Google Play Apps



Phân tích & Dự đoán

NGUYỄN DUY LINH - NHÓM 7

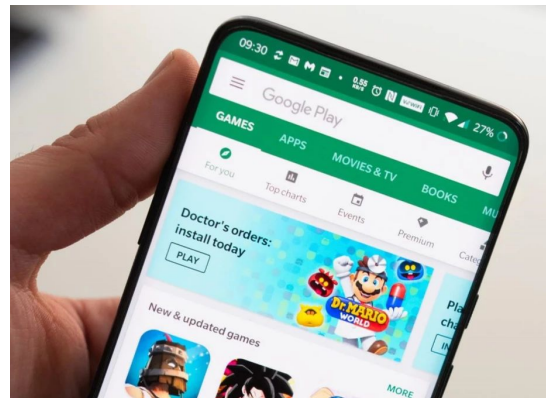




Về Google Play

Google Play là một sản phẩm của Google (từ 2012) - một nền tảng giúp phân phối các ứng dụng cho hệ điều hành Android.

Mô hình hoạt động: nhà phát triển phần mềm đăng tải, quảng bá các ứng dụng của mình; sau đó người dùng duyệt, tìm kiếm và cài đặt các ứng dụng đó vào thiết bị Android của mình.





Google Play

VS



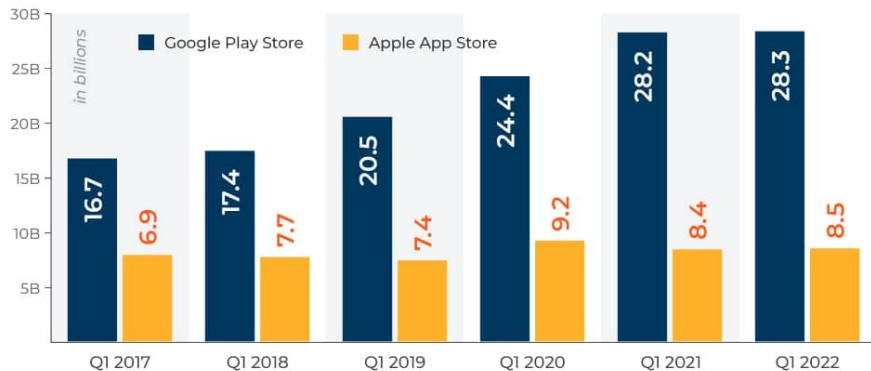
App Store



Google Play Store vs. Apple App Store App Downloads Over a 6-Year Period



TRUELIST



Source: Statista

Thống kê về tổng số lượt
tải trên 2 chợ ứng dụng
phổ biến **Google Play** và
App Store
(giai đoạn 2017 - 2022)



Dataset

Bộ dữ liệu **Google Play Store Apps** được thu thập vào 06/2021, bao gồm khoảng 2,3 triệu ứng dụng được public kể từ lúc nền tảng bắt đầu hoạt động.

[Google Play Store Apps | Kaggle](#)

Bộ dữ liệu bao gồm 24 cột dữ liệu, chứa thông tin cơ bản của từng ứng dụng có trên nền tảng, bao gồm thông tin về:

1. **Tên, ID ứng dụng, tên Nhà phát triển, thông tin liên hệ**
2. **Kích thước ứng dụng, chính sách phát hành (miễn phí, tính phí, chứa quảng cáo, giới hạn độ tuổi,...), danh sách các phiên bản HĐH hỗ trợ**
3. **Ngày phát hành, ngày cập nhật gần nhất**
4. **Số lượt tải, số lượt đánh giá, điểm đánh giá,...**





Dataset

A	B	C	D	E	F	G	H	I	J	K
App Name	App Id	Category	Rating	Rating Count	Installs	Minimum Inst	Maximum	Free	Price	Currency
Gakondo	com.ishakw	Adventure	0	0	10+	10	15	TRUE	0	USD
Ampere Battery Info	com.webse	Tools	4.4	64	5,000+	5000	7662	TRUE	0	USD
Vibook	com.doanti	Productivity	0	0	50+	50	58	TRUE	0	USD
Smart City Trichy Public Servi	cst.stJosepl	Communication	5	5	10+	10	19	TRUE	0	USD

L	M	N	O	P	Q	R	S	T	U	V	W
Size	Minimum Android	Developer	Developer	Developer	Released	Last Updated	Content R	Privacy Po	Ad Support	In App Pur	Editors Ch
10M	7.1 and up	Jean Confi	https://be	jean21101	26-Feb-20	26-Feb-20	Everyone	https://be	FALSE	FALSE	FALSE
2.9M	5.0 and up	Webservei	https://we	webservei	21-May-20	6-May-21	Everyone	https://de	TRUE	FALSE	FALSE
3.7M	4.0.3 and up	Cabin Crew		vnacrewit	9-Aug-19	19-Aug-19	Everyone	https://wv	FALSE	FALSE	FALSE
1.8M	4.0.3 and up	Climate Sn	http://ww	climatesm	10-Sep-18	13-Oct-18	Everyone		TRUE	FALSE	FALSE

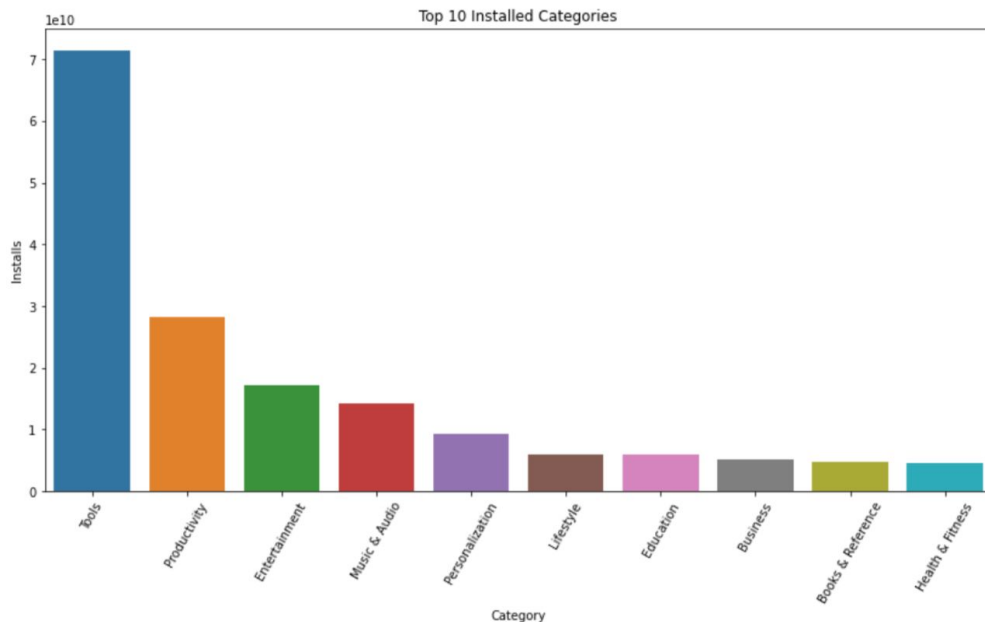




Google Play: Hiện trạng

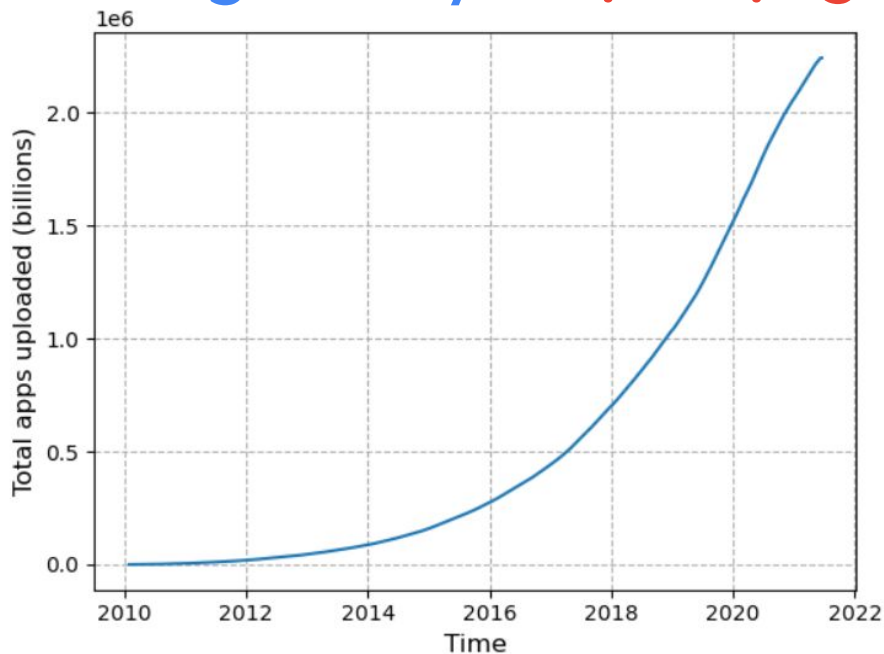
Nhóm ứng dụng có số lượt tải nhiều nhất: Nhóm ứng dụng công cụ.

Tiếp theo đó là các nhóm ứng dụng liên quan tới sáng tạo, giải trí và chia sẻ nội dung.





Google Play: Hiện trạng



Tốc độ tăng trưởng có xu hướng nhanh dần theo các năm (từ những năm 2012, 2013).

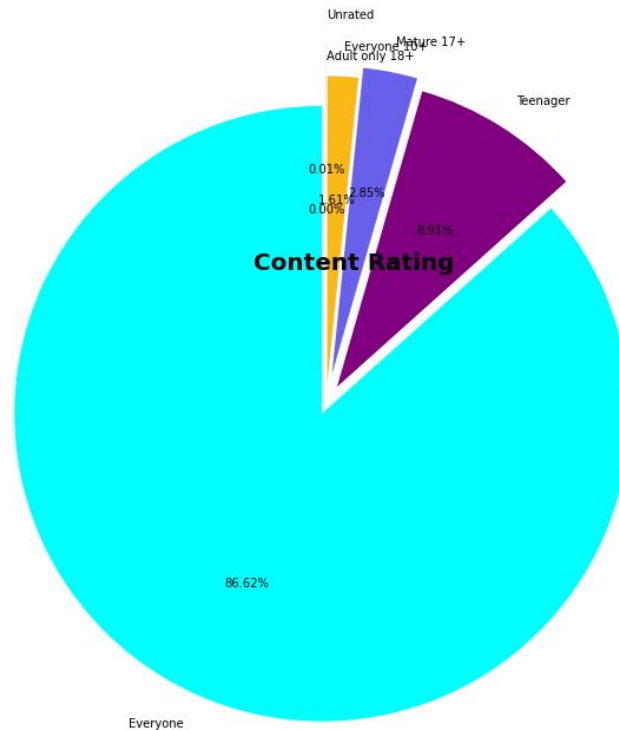
Dự báo có thể chạm mốc 3 triệu ứng dụng vào năm 2024 (xây dựng mô hình dự báo)





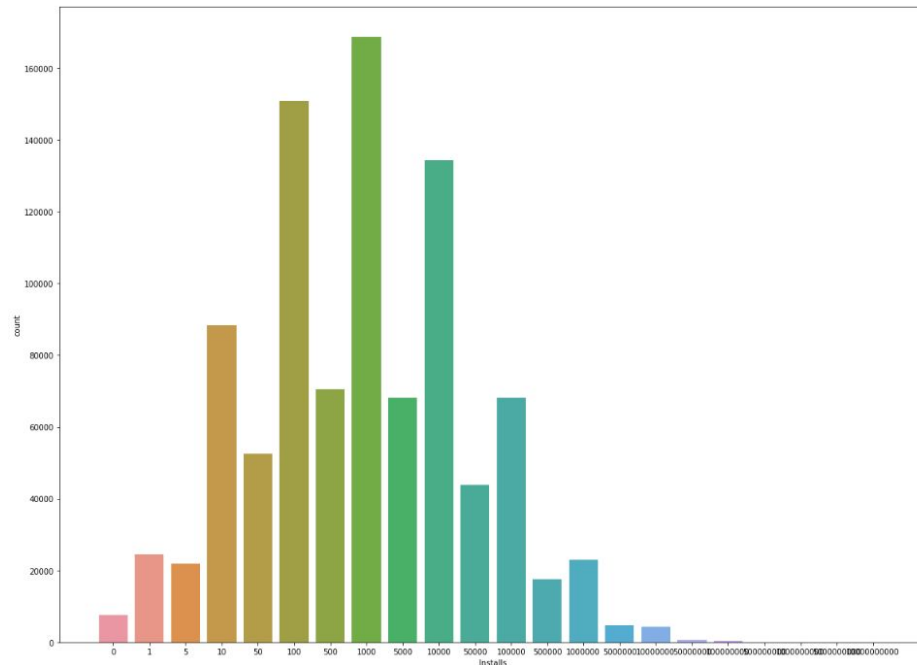
Google Play: Hiện trạng

Giới hạn độ tuổi: khoảng **5/6** số lượng ứng dụng phù hợp cho tất cả mọi người, gần **9%** ứng dụng gán nhãn phù hợp cho trẻ vị thành niên và gần **4%** ứng dụng chỉ phù hợp cho người lớn.



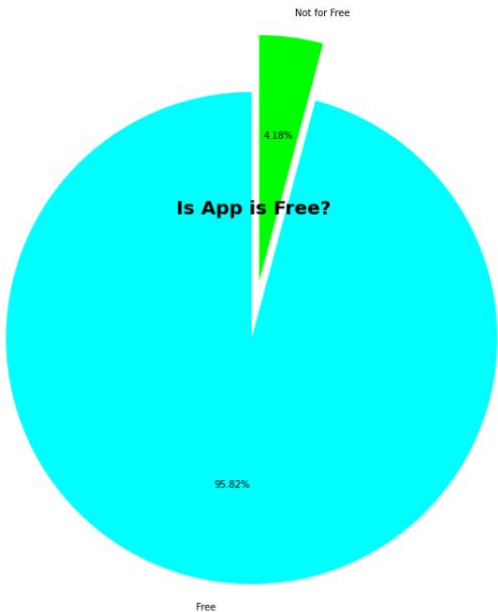


Chỉ một số ứng dụng đạt trên **100.000** lượt tải, và cực kỳ hiếm ứng dụng thành công (từ **1000.000** lượt tải trở lên).





Google Play: Hiện trạng



Chỉ một lượng nhỏ (4%) ứng dụng có tính phí tải đối với người dùng.

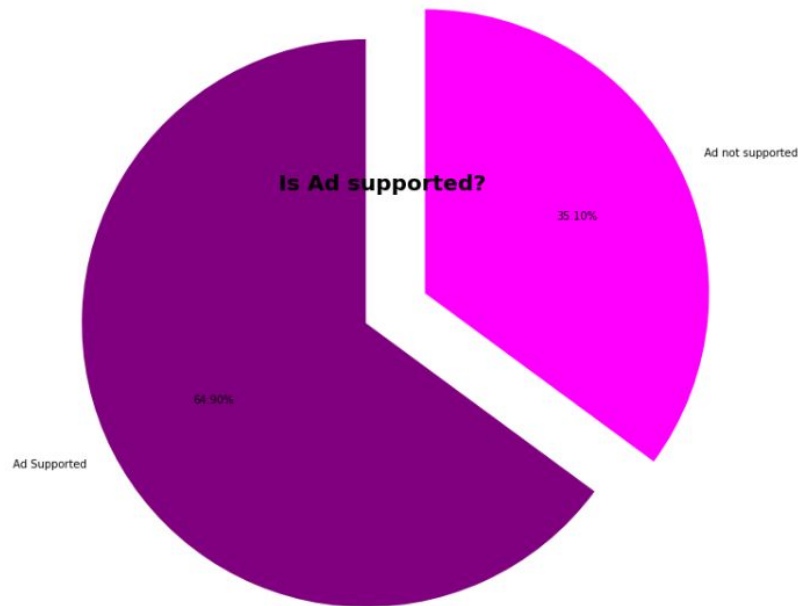
Hầu hết ứng dụng đều có thể cài đặt miễn phí.





Google Play: Hiện trạng

Khoảng **2/3** ứng dụng có chứa quảng cáo để làm nguồn thu nhập.



🏠 Xử lý dữ liệu cho các bài toán dự đoán

Loại bỏ giá trị null:

```
RangeIndex: 2312944 entries, 0 to 2312943
```



```
RangeIndex: 1287191 entries, 0 to 1287190
```

Tạo cột bổ sung:

Days_alive

```
encoded_df['Days_alive'] = (non_null_df['Last Updated']  
                             - non_null_df['Released']).dt.days
```

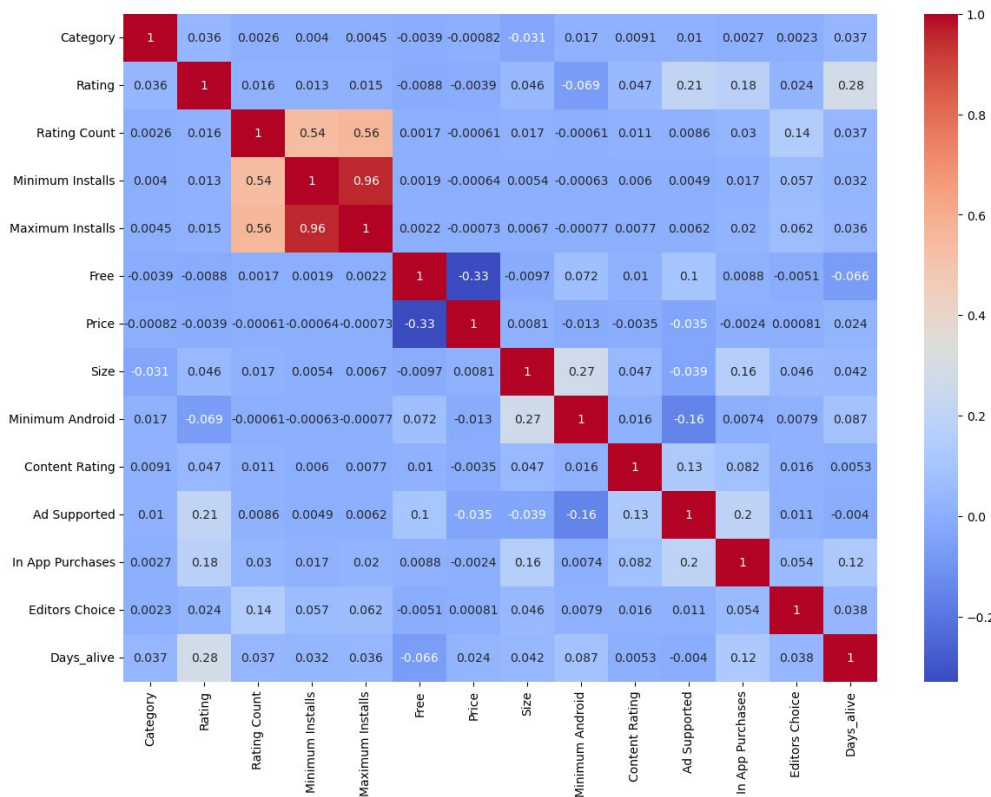
Encoding:

App Name	str
Category	str -> int
Rating	float
Rating Count	float
Minimum Installs	float
Maximum Installs	int
Free	bool
Price	float
Size	str -> float
Minimum Android	str -> int
Released	str -> date
Last Updated	str -> date
Content Rating	str -> int
Ad Supported	bool
In App Purchases	bool
Editors Choice	bool



Tương quan giữa các tham số

Nhận xét: các tham số có mức độ tương quan với nhau không nhiều.



🏠 Dự đoán số lượt tải ứng dụng

1. Lựa chọn một số tham số có mức tương quan tốt nhất:

```
4 X = encoded_df[['Category', 'Free', 'Price', 'Size', 'Minimum Android',
3 |               |               |               |               |               |
2 |               |               |               |               |               |
1 |               |               |               |               |               |
5 y = encoded_df['Maximum Installs']
```

2. Chuẩn hóa dữ liệu:

```
scaler = StandardScaler()
1 scaled_X = scaler.fit_transform(X)
```

3. Chia data thành 2 tập dữ liệu train : test với tỉ lệ 70 : 30

4. Sử dụng các mô hình hồi quy:

- LinearRegression
- DecisionTreeRegressor
- XGBRegressor



Dự đoán số lượt tải ứng dụng

Kết quả: đánh giá thông qua R^2 score

Linear Regression:

```
r2_score of training data -228.0942829619768  
r2_score of testing data -185.78226620877604
```

Decision Tree:

```
r2_score of training data 0.5416680629711631  
r2_score of testing data -0.8140093955350642
```

XG Boost:

```
r2_score of training data -2.2859412985073293  
r2_score of testing data -3.602786036838979
```





Dự đoán điểm đánh giá cho ứng dụng

Kết quả: đánh giá thông qua R^2 score

Linear Regression:

```
r2_score of training data -5.120196245590867  
r2_score of testing data -5.104771173944716
```

Decision Tree:

```
r2_score of training data 0.9424843646316237  
r2_score of testing data 0.9421417842651739
```

XG Boost:

```
r2_score of training data -2.2859412985073293  
r2_score of testing data -3.602786036838979
```



🏠 Dự đoán mức độ tăng trưởng của nền tảng trong tương lai

Xây dựng mô hình dự báo số lượt tải trong tương lai:

Mô hình ARIMA:

MSE: 59230.70
RMSE: 243.37

```
Giá trị dự báo=731962.226134, Giá trị thực tế=732038.000000
Giá trị dự báo=733021.827980, Giá trị thực tế=733022.000000
Giá trị dự báo=733962.145394, Giá trị thực tế=733993.000000
...
Giá trị dự báo=2241910.937037, Giá trị thực tế=2241890.000000
Giá trị dự báo=2241910.381532, Giá trị thực tế=2241891.000000
```





Thank You

