ĐỒ ÁN KHOA HỌC DỮ LIỆU

THÀNH VIÊN:

1. 18120368 – CAO LÊ MINH HIẾU
2. 18120311 – THÁI TẤN ĐẠT

Đề tài: dự đoán sự thành công của dự án trên KickStarter.com

Explore **502,789 projects**

Project We Love

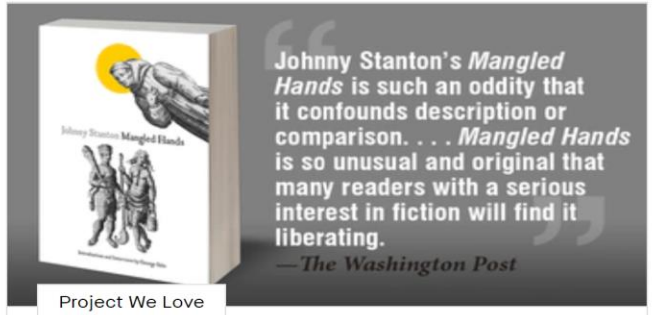
Wild Inks & Paints: A Seasonal Palette — Make 100

Another art handbook—this time exploring foraged plants & other gifts from nature. For pro artists as wel...

by Joumana Medlej

£14,493 pledged
1,449% funded
26 days to go

Publishing · London, UK



Project We Love

Reviving a unique neglected American novel from the 1980s

Publishing a new expanded edition of Johnny Stanton's 1985 novel, MANGLED HANDS

by Rick Schober

$3,828 pledged
222% funded
7 days to go

Fiction · Arlington, MA

Project We Love

LaserPecker 2-Super Fast Handheld Laser Engraver & Cutter

Ultra-fast 5W laser engraver & cutter for art, business & leisure use

by LaserPecker

HK$ 32,813,543 pledged
42,323% funded
14 days to go

DIY Electronics · Hong Kong, Hong Kong

Đề tài: dự đoán sự thành công của dự án trên KickStarter.com

Mô tả đồ án

Đồ án này được tạo ra như một công cụ trợ giúp trong hoạt động gọi vốn cộng đồng, đặc biệt là trong lĩnh vực ***công nghệ***. Thông qua công cụ này, các nhà sáng lập biết được đâu là những yếu tố kĩ thuật nào đem đến thành công cho hoạt động gọi vốn, giúp họ có sự chuẩn bị tốt trước và trong quá trình gọi vốn. Các nhà đầu tư có công cụ đánh giá các dự án, phân tích và lựa chọn dự án phù hợp.

CÁC GIAI ĐOẠN

1. Thu thập dữ liệu.
2. Khám phá dữ liệu
3. Tiền xử lý dữ liệu
4. Mô hình hóa dữ liệu
5. Nhìn lại quá trình làm đồ án
6. Tài liệu tham khảo

Thu thập dữ liệu

Các thư viện cần sử dụng ở giai đoạn Thu thập dữ liệu.

Đường dẫn để crawl dữ liệu.

```
from bs4 import BeautifulSoup
import requests
import json
import pandas as pd
import time
import datetime as dt
import re
```

```
start_url = f'http://www.kickt
```

Thu thập dữ liệu

1. Request lên đường link next_url của kicktraq.com
2. Kiểm tra có get link được không.
3. Parse html để lấy danh sách dự án và đường dẫn tiếp theo.

```
next_url = start_url
pid = 3200
print(df.shape)

while next_url is not None:
    url = next_url
    req = requests.get(url)

    tries = 3
    while not req.ok and tries > 0:
        print('Request to ', url, ' failed')
        time.sleep(1)
        tries -= 1
        req = requests.get(url)

    if tries <= 0:
        continue
    else:
        tries = 3

    print('Requested URL: ', url)

    soup = BeautifulSoup(req.content, 'html.parser')
    project_list = soup.find('div', {'id': 'project-list'})
    projects = project_list.find('div', {'class': 'projects'}).find_all('div', {'class': 'project'})
    next_url = project_list.find('div', {'class': 'paging'}).find('a', {'class': 'prn'}, text='Next >')

    if next_url is not None:
        next_url = ''.join([f'http://www.kicktraq.com/search/', next_url['href'], f'&sort=new'])
```

Thu thập dữ liệu

Parse từ kicktraq.com

```
for proj in projects:
    info = proj.find('div', {'class': 'project-infobox'})
    a = info.find('h2').find()
    link = f'http://www.kickstarter.com' + a['href']
    name = a.text
    info = proj.find('div', {'class': 'project-infobits'})
    pledgilizer = info.find('div', {'class': 'project-pledgilizer'})
    if pledgilizer is None:
        continue

    status = pledgilizer.find('div', {'class': 'project-pledgilizer-top'}).find().text.lower()
    if status != 'closed':
        continue
    else:
        funded = pledgilizer.find('div', {'class': 'project-pledgilizer-mid'}).find().text[:-1]
        if int(funded) >= 100:
            status = 'success'
        else:
            status = 'failed'

    funding_info = info.find('div', {'class':'project-details'}).text.split('\n\t\t\t\t\t\t')
    moneys = funding_info[2].split(' ')
    goal = moneys[3].replace(',', '')
    pledged = moneys[1].replace(',', '')
    backers = funding_info[1].split(' ')[1]

    dates = funding_info[3].split(' -> ')
    start_date = dates[0].split(': ')[1]
    end_date, year = dates[1].split(' (')
    year = year[:-1]
```

Thu thập dữ liệu

Parse từ trang kickstarter.

```
kick_req = requests.get(link)
tries = 3
while not kick_req.ok and tries > 0:
    print('Request to Kickstarter ', link, ' failed')
    time.sleep(1)
    tries -= 1
    kick_req = requests.get(link)

if tries <= 0:
    continue
else:
    tries = 3

kick_soup = BeautifulSoup(kick_req.content, 'html.parser')
navigators = kick_soup.find('div', {'class': 'campaign-side-nav project-nav__links'})

# Hidden projects
if navigators is None:
    continue

comments = navigators.find('a', {'id': 'comments-emoji'})
comments = comments.find().text if comments is not None else 0

updates = navigators.find('a', {'id': 'updates-emoji'})
updates = updates.find().text if updates is not None else 0

faqs = navigators.find('a', {'id': 'faq-emoji'}).find()
faqs = faqs.text if faqs is not None else '0'

tiers = kick_soup.find('div', {'class': 'NS_projects__rewards_list'}).find('ol').find_all('li', recursive=False)
tier_count = len(tiers)
tier_min = 0
tier_max = 0
if tier_count > 0:
    tier_cost_min = tiers[0].find('div', {'class': 'pledge__info'}).find('span', {'class': 'money'}).text.split(' ')[-1].repl
    tier_cost_max = tiers[-1].find('div', {'class': 'pledge__info'}).find('span', {'class': 'money'}).text.split(' ')[-1].rep

while not tier_cost_min.isnumeric():
    tier_cost_min = '0' if len(tier_cost_min) == 0 else tier_cost_min[1:]
while not tier_cost_max.isnumeric():
    tier_cost_max = '0' if len(tier_cost_max) == 0 else tier_cost_max[1:]
tier_min = int(tier_cost_min)
tier_max = int(tier_cost_max)
```

Thu thập dữ liệu

Lưu dữ liệu vô csv.

```
new_row = {'Id': pid, 'Name': name, 'Url': link, 'Goal': goal, 'Pledged': pledged,
           'Launch': start_date, 'End': end_date, 'Year': year,
           'Comments': comments, 'Updates': updates, 'Faqs': faqs,
           'Backers': backers, 'Tiers': tier_count, 'TierMin': tier_min, 'TierMax': tier_max,
           'Status': status}
df = df.append(new_row, ignore_index=True)
pid += 1
if pid % 200 == 0:
    df.to_csv('data.csv', index=None)
    print('Backed up ', pid, ' indices')
if pid == 1200:
    break
```

Khai phá dữ liệu

Thư viện được sử dụng trong phần này.

```
# Import essential depencies
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
import requests
from datetime import datetime

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.compose import ColumnTransformer, make_column_transformer
from sklearn.model_selection import cross_val_score
from sklearn.neural_network import MLPClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import zero_one_loss
from sklearn.linear_model import LinearRegression, LogisticRegression


%matplotlib inline
```

Khai phá dữ liệu

Tổng quan về dữ liệu

Cấu trúc dữ liệu: 7173 điểm dữ liệu, không có hiện tượng mất, thiếu dữ liệu.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7173 entries, 0 to 7172
Data columns (total 16 columns):
Id              7173 non-null int64
Name            7173 non-null object
Url             7173 non-null object
Goal            7173 non-null object
Pledged         7173 non-null object
Launch          7173 non-null object
End             7173 non-null object
Year            7173 non-null int64
Comments        7173 non-null object
Updates         7173 non-null int64
Faqs            7173 non-null int64
Backers         7173 non-null int64
Tiers           7173 non-null int64
TierMin         7173 non-null int64
TierMax         7173 non-null int64
Status          7173 non-null object
dtypes: int64(8), object(8)
memory usage: 896.8+ KB
```

Khai phá dữ liệu

Tổng quan về dữ liệu

Giải thích ý nghĩa thuộc tính

Ý nghĩa các thuộc tính dữ liệu:

- Id: Id của dự án. Ta thấy Id có giá trị trùng với index nên đây có thể là Id do người lấy dữ liệu tự đánh
- Name: Tên dự án
- Url: Đường dẫn đến dự án
- Goal: Số vốn mục tiêu kêu gọi (bao gồm đơn vị tiền tệ)
- Pledged: Số vốn nhận được (bao gồm đơn vị tiền tệ)
- Launch: Ngày bắt đầu gọi vốn
- End: Ngày kết thúc gọi vốn
- Year: Năm gọi vốn
- Comments: Số lượng bình luận
- Updates: Số lần cập nhật trạng thái dự án
- Faqs: Số câu hỏi trong mục FAQ (Frequent ask questions)
- Backers: Số nhà đầu tư
- Tiers: Số lượng gói đầu tư mà nhà sáng lập đưa ra
- TierMin: Gói đầu tư nhỏ nhất
- TierMax: Gói đầu tư lớn nhất
- Status: Trạng thái dự án

Khai phá dữ liệu

Tổng quan về dữ liệu

Thống kê dữ liệu trùng lặp: không có dữ liệu mất hay trùng lặp. Từ đó có thể thấy quá trình Thu thập dữ liệu được thực hiện tốt

```
df.duplicated().sum()
```

```
0
```

Khai phá dữ liệu

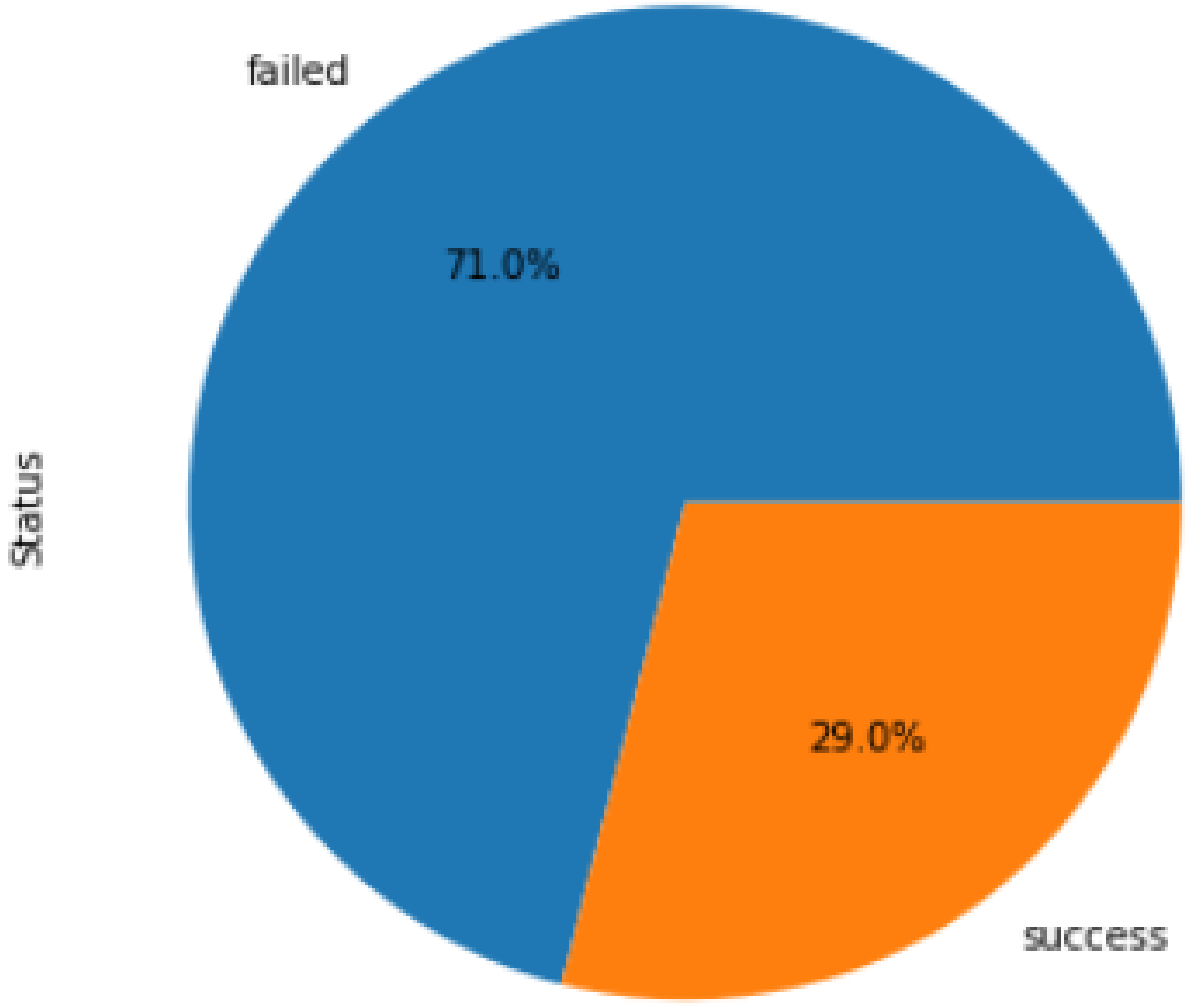
Phân tích dữ liệu

- **Đặt giả thuyết**
- Trước khi bắt đầu, chúng ta sẽ tiến hành đặt ra các giả thuyết về dữ liệu. Đây là một cách rất hiệu quả để đưa ra được một góc nhìn cụ thể cho vấn đề mà chúng ta muốn giải quyết.
- Ở đây, vì chúng ta muốn tìm ra công thức cho sự thành công nên các giả thuyết của chúng ta sẽ xoay quanh vấn đề này. Các giả thuyết có thể có là:
  - Kêu gọi số vốn quá lớn sẽ làm giảm khả năng thành công của dự án
  - Bắt đầu gọi vốn vào ngày nghỉ sẽ giúp tăng khả năng thành công.
  - Khoảng thời gian gọi vốn càng ngắn thì dự án càng nhận được ít tiền đầu tư
  - Cập nhật trạng thái liên tục sẽ giúp các nhà đầu tư tin tưởng đầu tư vào dự án
  - Đưa ra các câu trả lời cho FAQ giúp tăng độ tin cậy cũng như nguồn tiền đổ vào dự án
  - Đặt ra càng nhiều gói đầu tư thì khả năng gọi vốn thành công càng cao
  - ...
- Bên cạnh đó, trong quá trình phân tích và khám phá dữ liệu. Chúng ta sẽ tiếp tục đặt ra các giả thuyết mới.

Khai phá dữ liệu

Phân tích dữ liệu

- Bắt đầu với cột **Status** – cột trạng thái, nhãn của bộ dữ liệu.
- Hiểu được cột này sẽ cho chúng ta cái nhìn chính xác về vấn đề kinh tế đặt ra.
- Từ biểu đồ ta nhận thấy phần lớn dự án gọi vốn thất bại. Cho thấy mức độ khó khan của hoạt dộng gọi vốn cộng đồng nói riêng, đầu tư nói chung khó khan như thế nào.



Khai phá dữ liệu

Phân tích dữ liệu

- Ta sẽ tiếp tục phân tích cột `Name`. Giống như trong lĩnh vực báo chí, tiêu đề hay, thu hút có thể thu hút một lượng lớn người đọc. Ta hoàn toàn có thể đặt ra giả thuyết rằng cách đặt tên dự án cũng sẽ ảnh hưởng thu hút được vốn đầu tư.

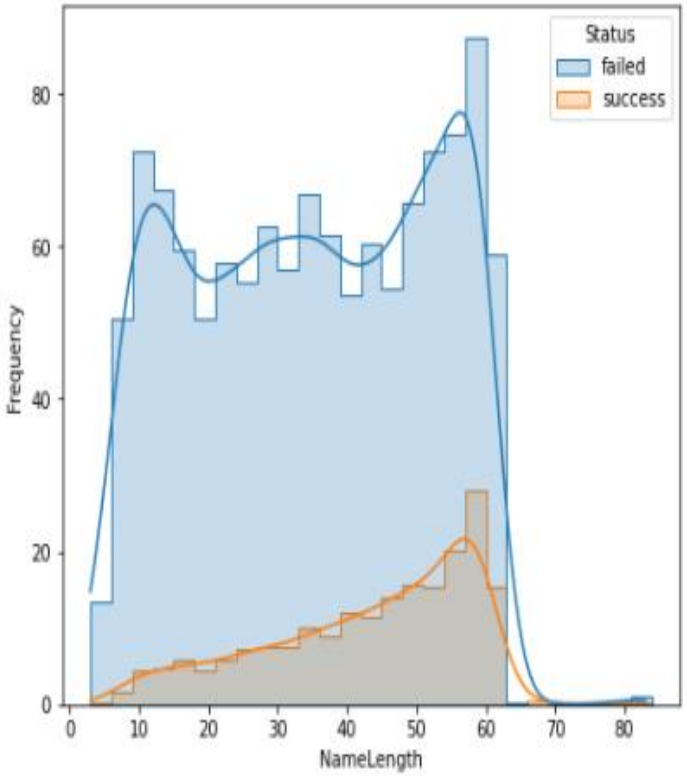
- Trong khuôn khổ đồ án, chúng ta sẽ bỏ qua các kĩ thuật xử lý ngôn ngữ phức tạp. Nhưng chúng ta vẫn sẽ chiết tách hai thuộc tính mới cho cột `Name` là **độ dài** và **số từ sử dụng.**

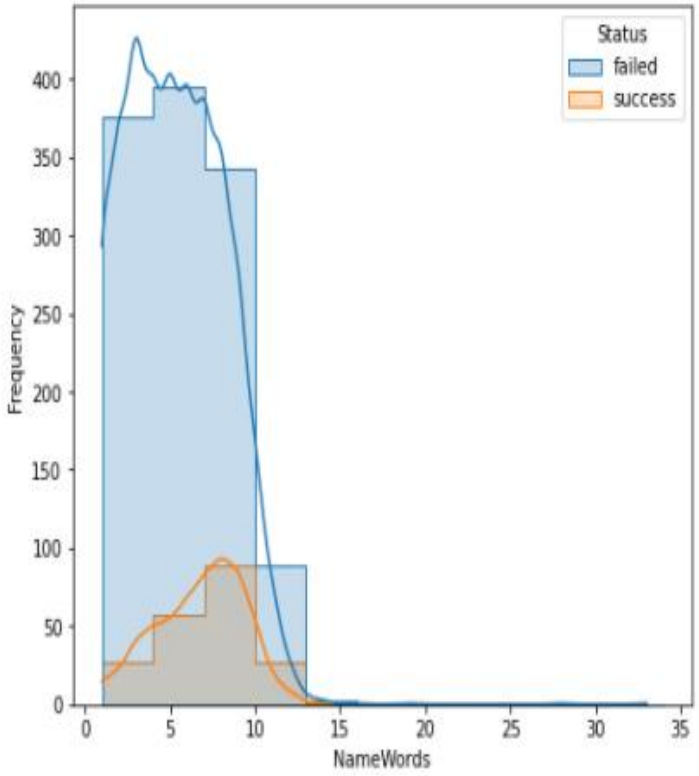
	Id	Name	NameLength	NameWords	Status
358	357	VR Forming Suit	15	3	failed
3105	3104	SussMyBike | Smart suspension set up tool	41	7	success
3310	3309	Persuasive Technologies	23	2	failed
5068	5067	Stars of Apollo	15	3	failed
2082	2081	Peacock: Make Twitter Fun Again	31	5	failed

Khai phá dữ liệu

Phân tích dữ liệu

- Dựa vào đồ thị, ta thấy độ dài tên các dự án thất bại phân bố trải dài tương đối đều (có dấu hiệu trũng ở giữa, cao hơn ở hai phía). Trong khi các dự án thành công thì tăng dần. Từ đó tỉ lệ thành công / thất bại theo độ dài tên cũng tăng dần độ dài tên tăng.
- Để tiện tính toán, xử lý. Từ giờ chúng ta sẽ quy định `**success**` là 1 và `**failed**` là 0





Khai phá dữ liệu

Phân tích dữ liệu

- Đối với cột `**Goal**` và cột `**Pledged**`, chúng ta thấy dữ liệu bao gồm cả đơn vị tiền tệ. Điều này gây khó khăn cho phân tích và xử lý các dữ liệu số.
- Bên cạnh đó đơn vị tiền tệ cũng là một thuộc tính có thể mang thông tin hữu ích. Nó cho thấy phần nào tình hình thị trường vốn của các quốc gia. Một giả thuyết có thể được đặt ra là liệu có đồng tiền nào được ưa chuộng hơn trong đầu tư.
- Do đó, chúng ta sẽ tách các thông tin về số và tiền tệ ra thành các thuộc tính khác nhau.

```
df['Currency'].value_counts()
```

```
$       5621
€        724
£        594
HK$       89
kr        76
MX$       54
¥          6
â,¬        5
Â£         4
Name: Currency, dtype: int64
```

Khai phá dữ liệu

Phân tích dữ liệu

- Trong danh sách này, có hai kí hiệu tiền tệ lạ là **`â,¬`** và **`Â£`**. Chúng ta sẽ kiểm tra các dự án sử dụng loại tiền tệ này.
- Tất cả các trang đều lỗi. Như vậy, các loại tiền tệ này là lỗi trong quá trình lấy dữ liệu. Và số lượng của chúng cũng rất nhỏ trên toàn bộ tập dữ liệu.
- Chúng ta sẽ xóa bỏ các dòng có dữ liệu lỗi đi. Sau đó, đổi các loại tiền tệ còn lại sang dạng tên viết tắt.

```
df[(df['Currency'] == 'Â£')|(df['Currency'] == 'â,¬')]['Url']
```

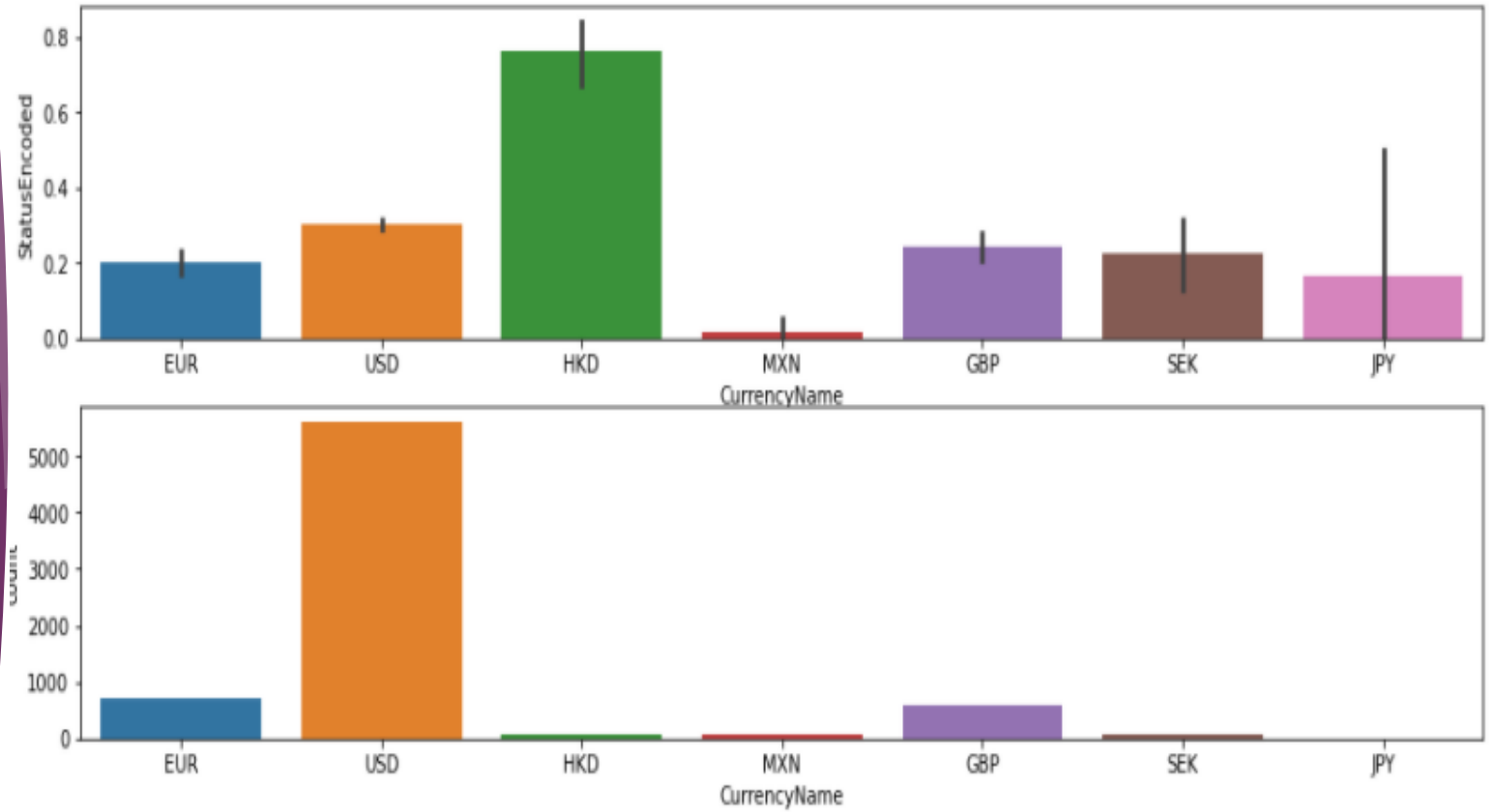
```
942     http://www.kickstarter.com/projects/roseprj/is...
943     http://www.kickstarter.com/projects/2131326048...
946     http://www.kickstarter.com/projects/smart-ivy/...
949     http://www.kickstarter.com/projects/ramoncaraz...
950     http://www.kickstarter.com/projects/hyperloopu...
1243    http://www.kickstarter.com/projects/1140759607...
1246    http://www.kickstarter.com/projects/1206849453...
1249    http://www.kickstarter.com/projects/583173617/...
1250    http://www.kickstarter.com/projects/12283819/q...
Name: Url, dtype: object
```

```
USD    5621
EUR     724
GBP     594
HKD      89
SEK      76
MXN      54
JPY       6
Name: CurrencyName, dtype: int64
```

Khai phá dữ liệu

Phân tích dữ liệu

- Ở đây ta thấy, các đồng tiền khác nhau có khả năng khác nhau trong việc thu hút vốn. Ví dụ như đồng `HKD` có tỉ lệ gọi vốn thành công bằng đồng tiền này rất cao, chứng tỏ thị trường vốn ở Hồng Kông rất sôi động, nhưng các lựa chọn đầu tư cộng đồng còn chưa đa dạng.



Khai phá dữ liệu

Phân tích dữ liệu

- Để chuẩn hóa dữ liệu, tiện cho việc so sánh. Chúng ta sẽ tất cả các con số gọi vốn và đầu tư về cùng một đơn vị tiền tệ. Vì `USD` là đồng tiền thanh toán quốc tế và số lượng mẫu sử dụng `USD` chiếm đa số, nên chúng ta sẽ sử dụng đồng tiền này làm thước đo.

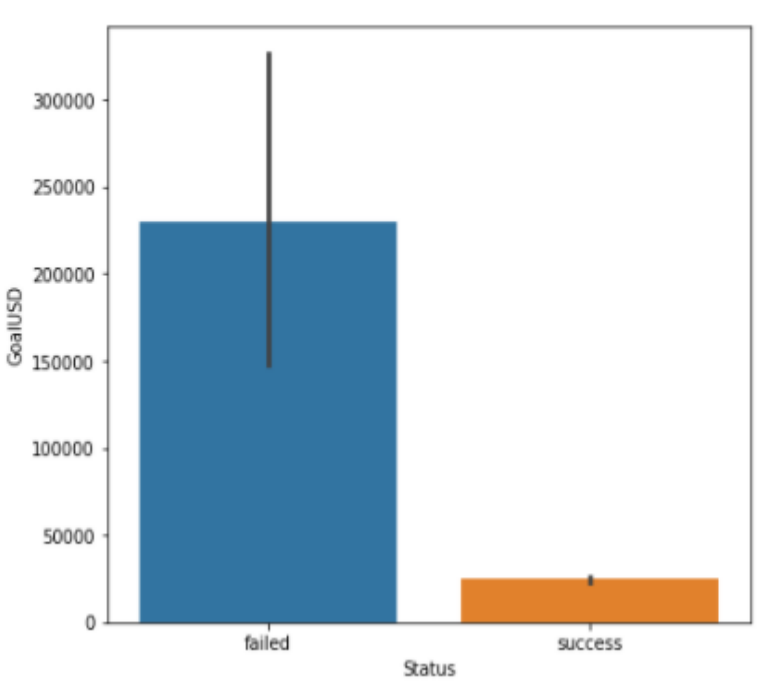
{'EUR': 1.2124, 'USD': 1.0, 'HKD': 0.1289718632, 'MXN': 0.0504288364, 'GBP': 1.3631202006, 'SEK': 0.1199695228, 'JPY': 0.0096062119}

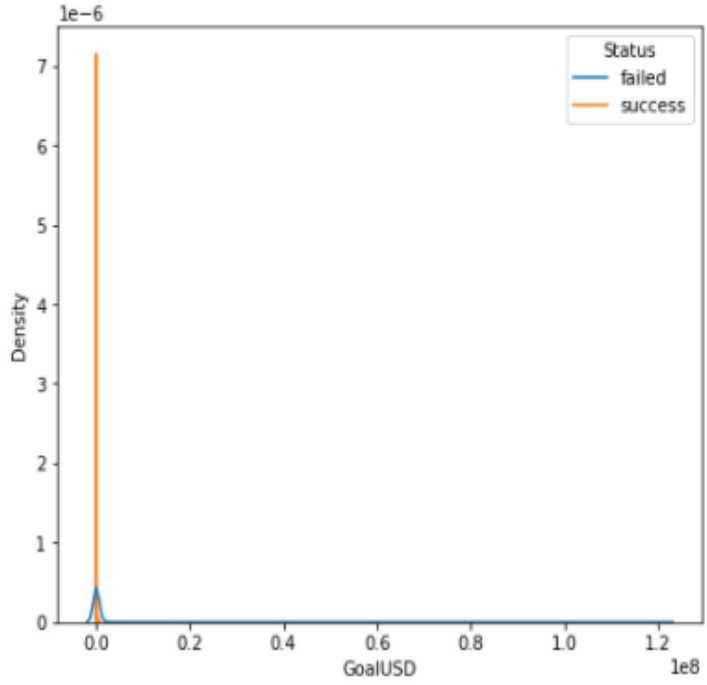
Khai phá dữ liệu

Phân tích dữ liệu

- Bây giờ chúng ta sẽ xem xét sự liên hệ giữa `Goal` và `Status`.
- Như chúng ta thấy, con số kêu gọi rất quan trọng trong mô hình. Các dự án thành công đều có mức kêu gọi phân bố trong khoảng thấp. Trung bình khoảng vốn kêu gọi của các dự án thành công thấp hơn rất nhiều so với các dự án thất bại.
- Chúng ta sẽ bỏ qua cột `Pledged` ở đây. Vì đơn giản là cột `Pledged` có thể suy ra trực tiếp kết quả.

```
Median of success goal:  10000.0
Median of failed goal:   25000.0
```





Khai phá dữ liệu

Phân tích dữ liệu

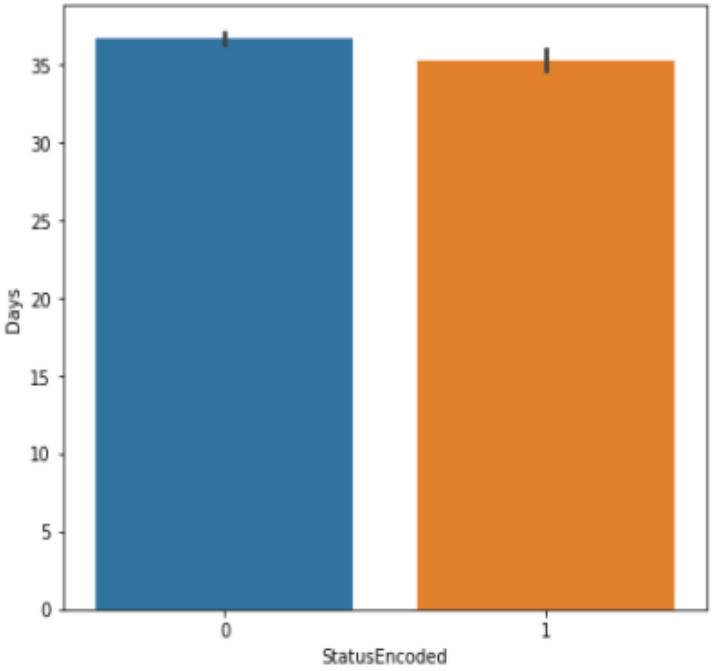
- Ngày bắt đầu và ngày kết thúc cũng có thể cho chúng ta những thông tin hữu ích về bộ dữ liệu. Đầu tiên chúng ta sẽ kiểm tra mối tương quan giữa khoảng thời gian gọi vốn và kết quả gọi vốn. Bên cạnh đó, chúng ta cũng kiểm tra giả thuyết liệu bắt đầu gọi vốn vào ngày cuối tuần có ảnh hưởng đến kết quả không.

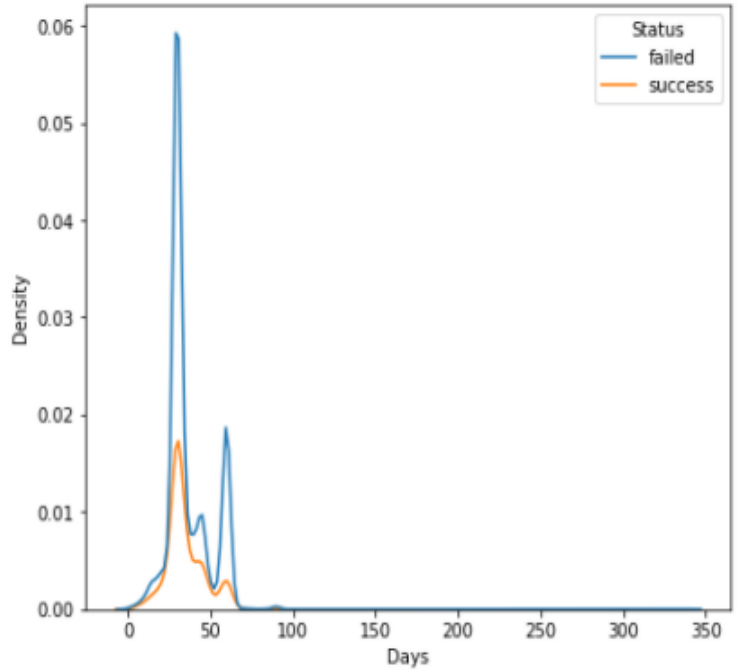
	Launch	End	Year	Days	Weekday	Status
6440	December 2nd	January 16th	2013	45.0	0	failed
5235	July 17th	August 16th	2014	30.0	3	failed
6033	October 20th	November 19th	2013	30.0	6	failed
5380	June 16th	August 15th	2014	60.0	0	success
3478	September 12th	November 11th	2015	60.0	5	failed

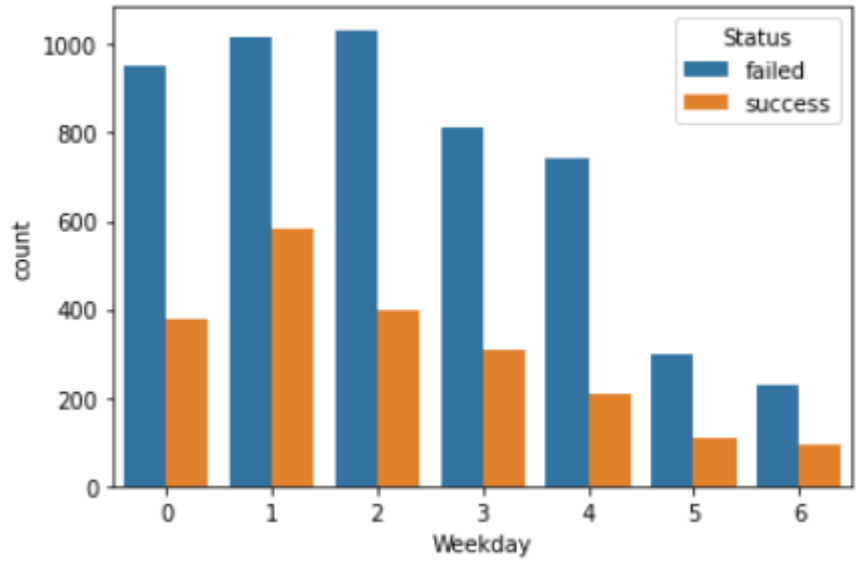
Khai phá dữ liệu

Phân tích dữ liệu

- Dựa vào biểu đồ, chúng ta thấy các thông tin về ngày tháng không cho dấu hiệu "quá" rõ ràng về mối quan hệ với khả năng thành công của dự án gọi vốn.



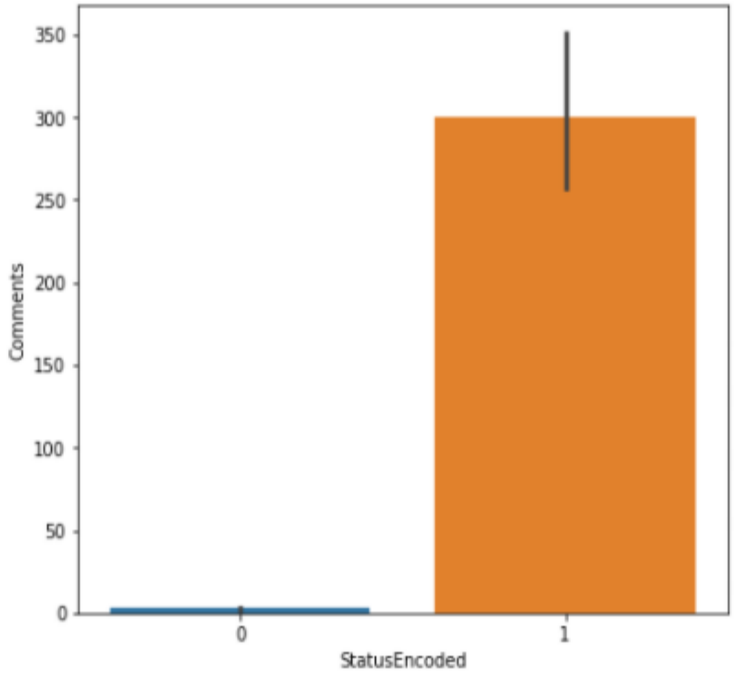


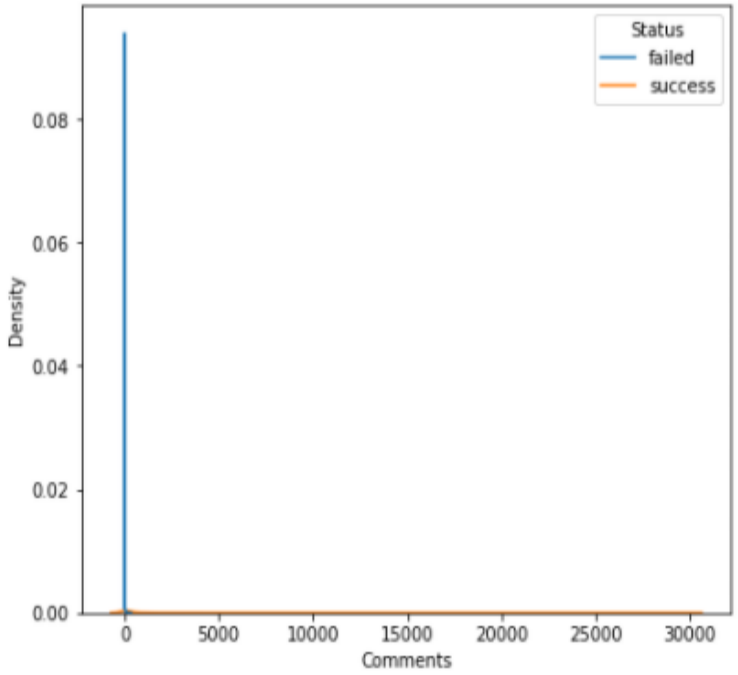


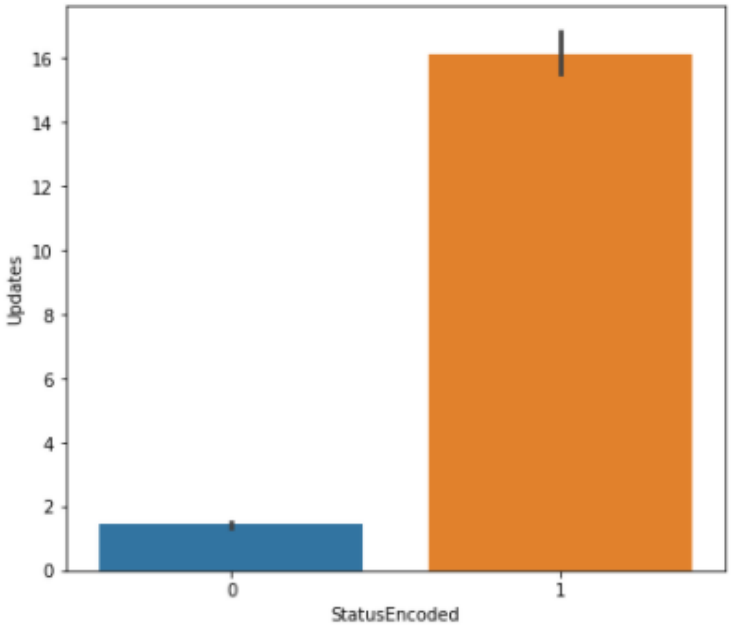
Khai phá dữ liệu

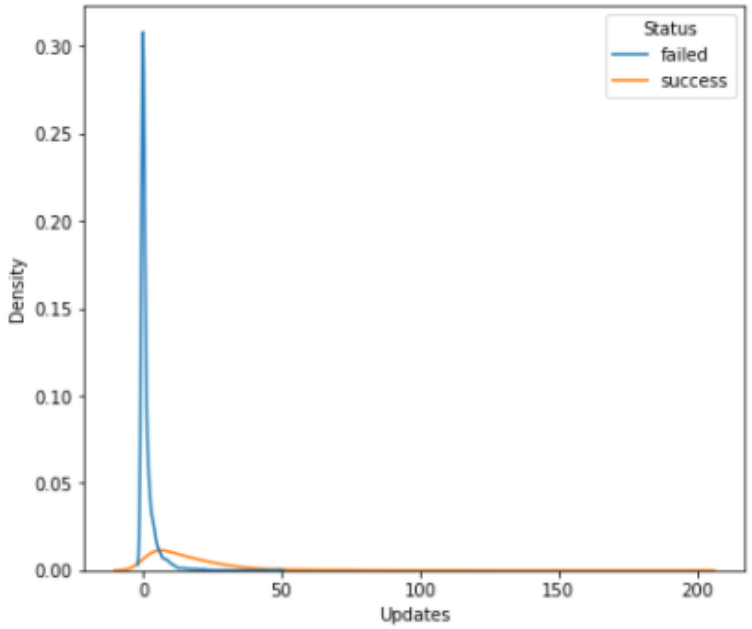
Phân tích dữ liệu

- Tiếp theo, chúng ta sẽ lần lượt kiểm tra mối tương quan giữa số lượng bình luận, số lượng FAQ, số lần cập nhật, số nhà đầu tư. đối với kết quả gọi vốn.
- Trước hết, chúng ta nhận thấy có một cột không chứa dữ liệu số là cột `Comments`. Lí do là khi các con số lớn hơn 1000, dữ liệu sẽ có dấu `,`. Chúng ta cần bỏ dấu `,` đi và chuyển sang dạng số.





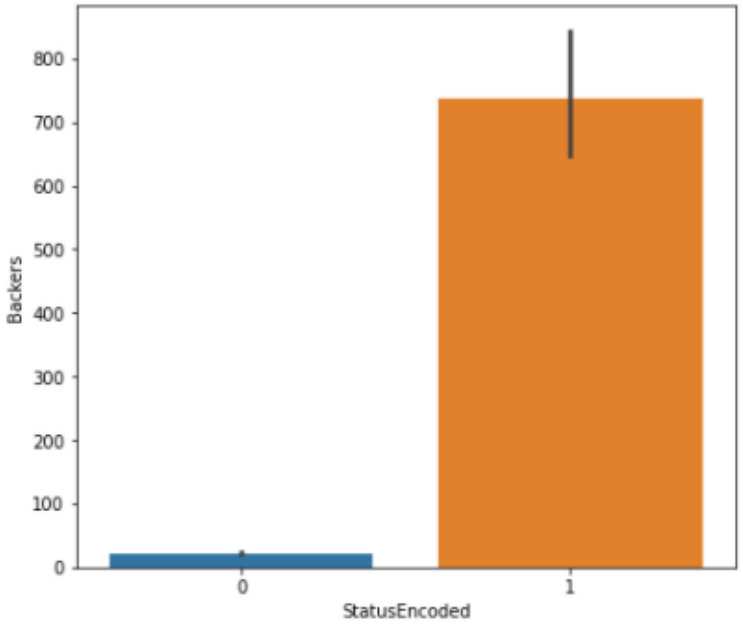


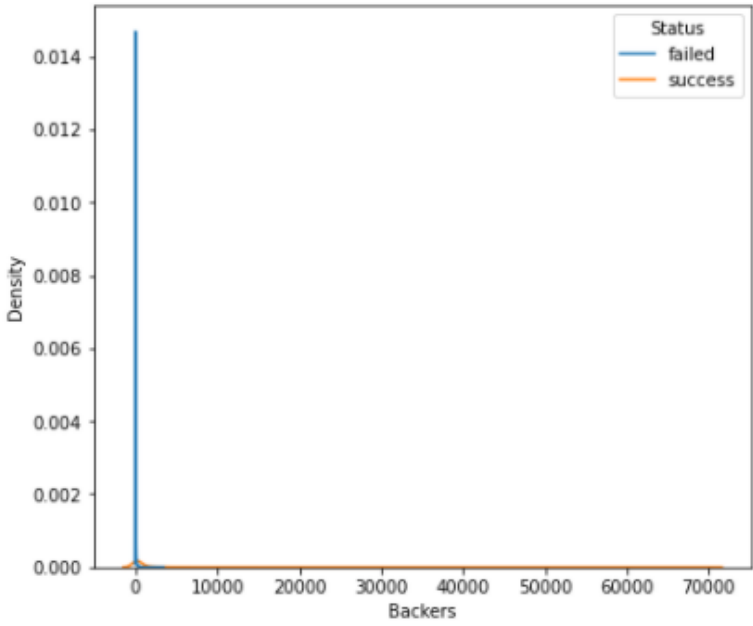


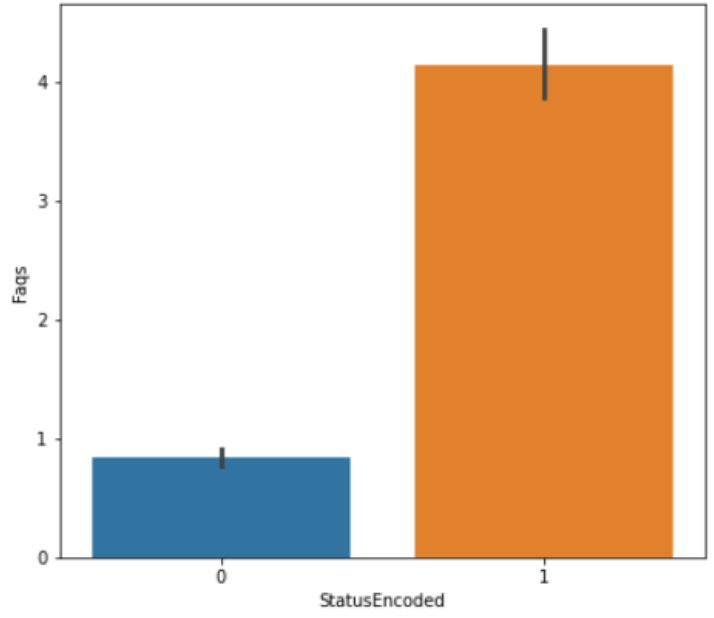
Khai phá dữ liệu

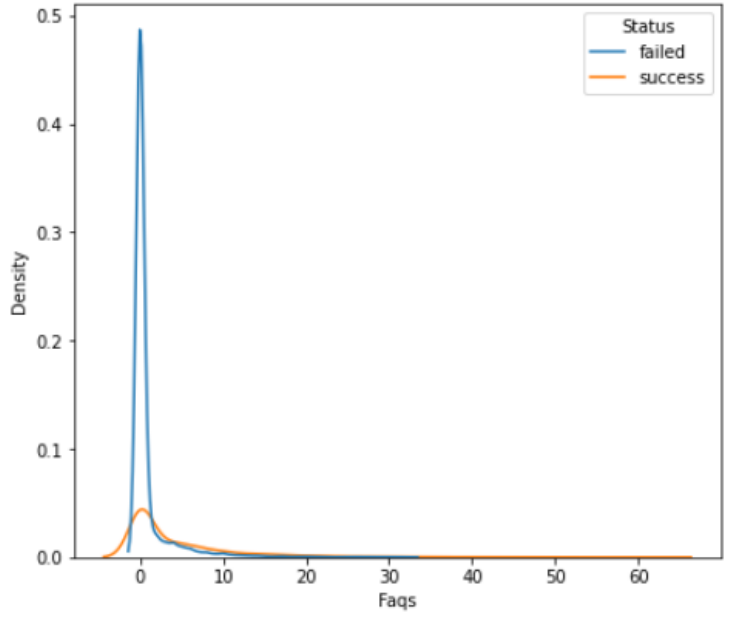
Phân tích dữ liệu

- Như chúng ta đã thấy, đối với các con số này, số càng lớn thì tỉ lệ thành công càng cao. Trong khi phân phối của các số này của các dữ liệu thất bại có đỉnh gần như bằng không thì ở bên phía thành công, đỉnh nằm ở vị trí xa 0 và sườn dốc phân phối khá thoải.





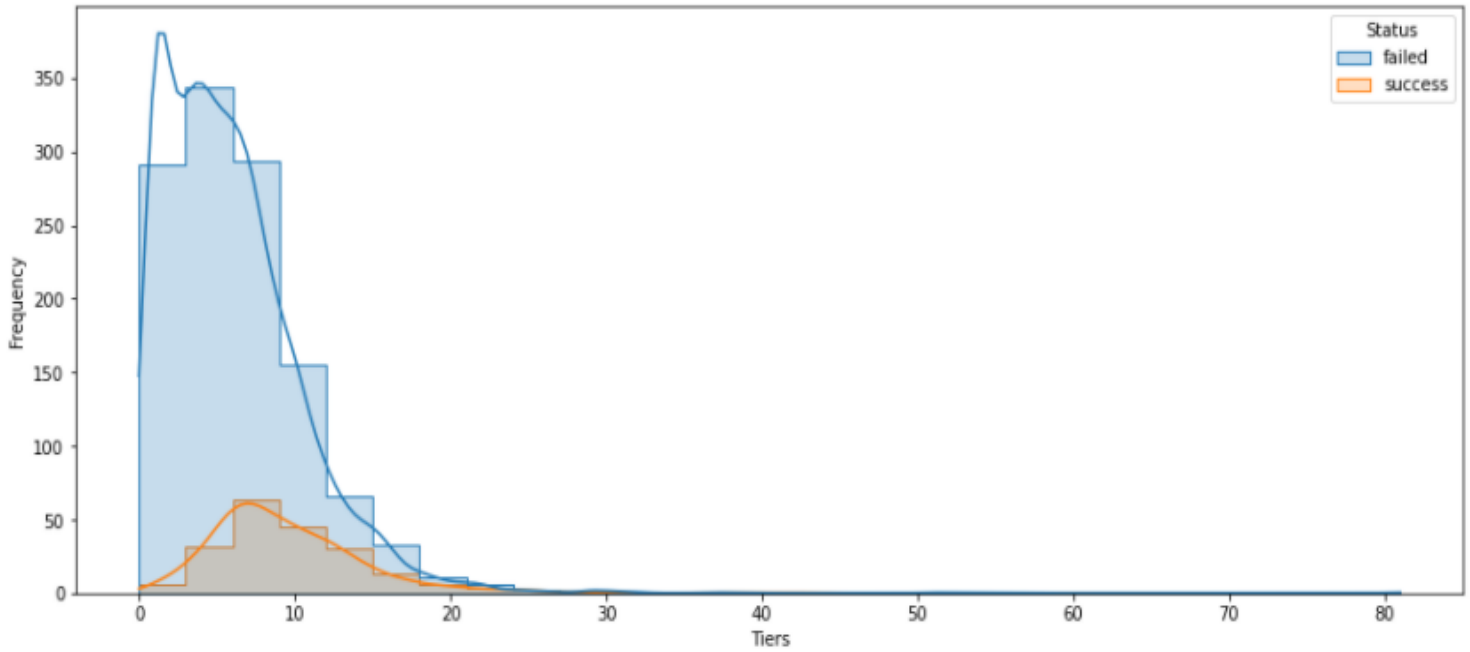


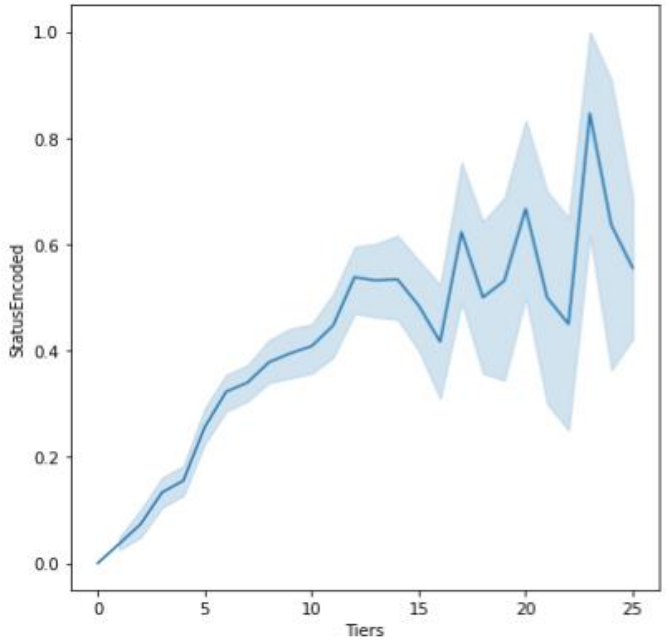


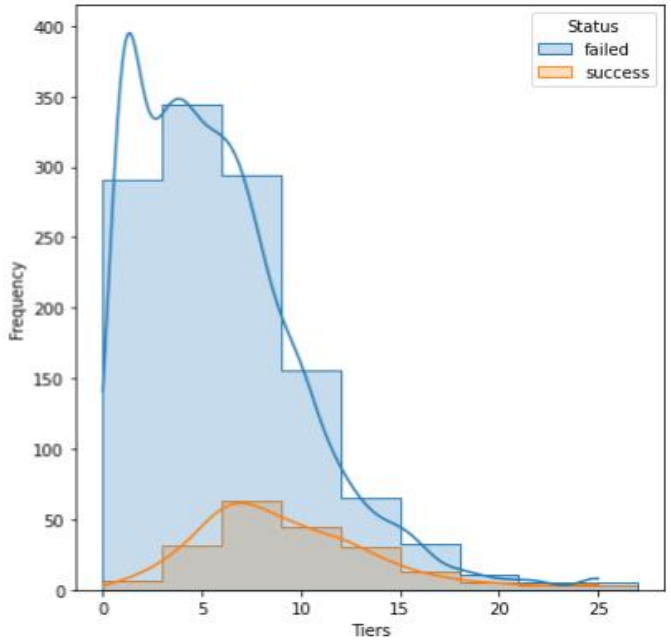
Khai phá dữ liệu

Phân tích dữ liệu

- Bây giờ chúng ta sẽ xem xét giả thuyết liệu cách thiết lập các gói đầu tư có ảnh hưởng đến kết quả gọi vốn hay không.
- Dựa vào phân phối ở trên, chúng ta thấy có một vài điểm dữ liệu ở ngoại biên làm ảnh hưởng đến khả năng đánh giá. Thay vì cắt bọ chúng, chúng ta sẽ giới hạn chúng ở một giá trị nhất định (ở đây là 25)
- Với biểu đồ, chúng ta nhận thấy khi xây dựng đa dạng gói đầu tư, kết quả gọi vốn đã có thể tăng đáng kể.



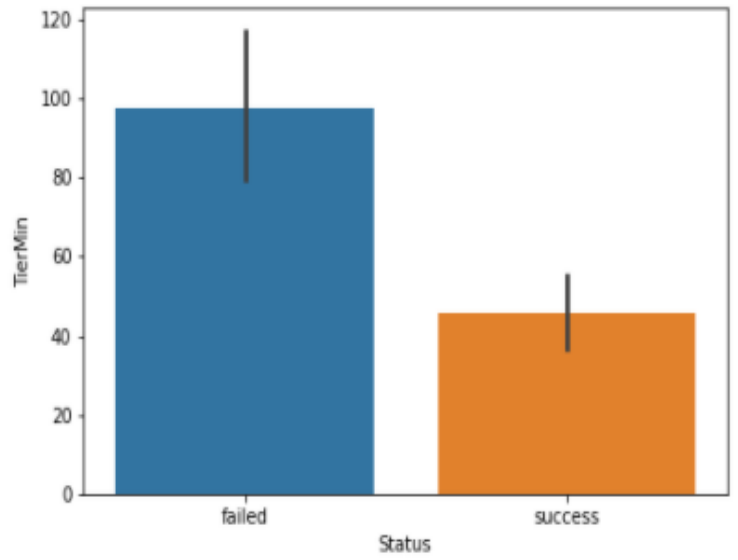


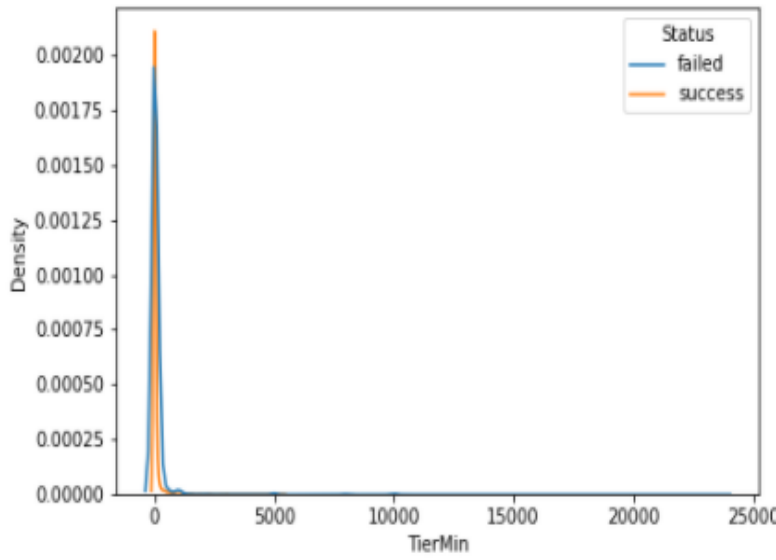


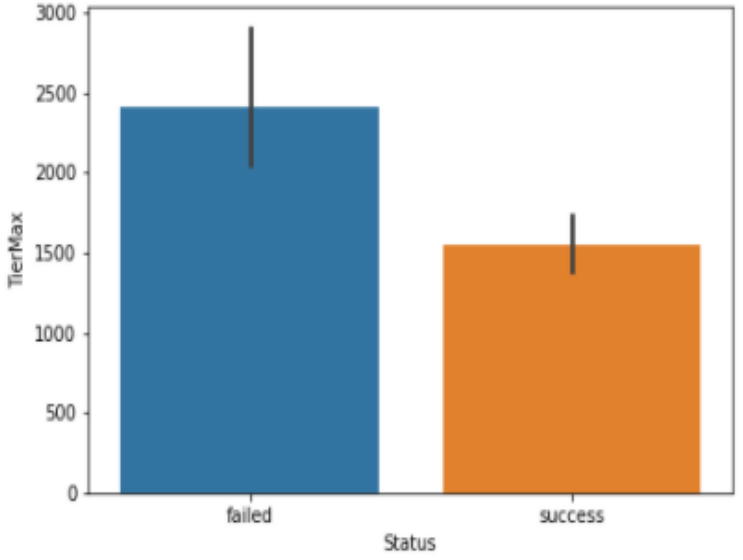
Khai phá dữ liệu

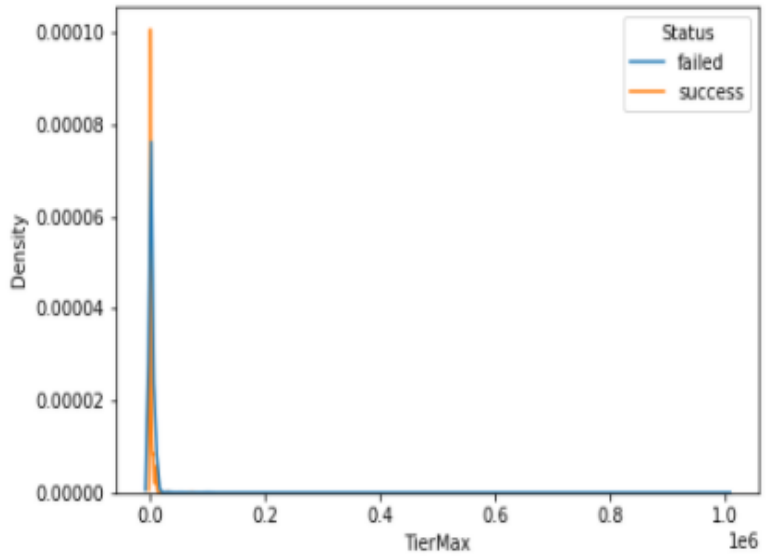
Phân tích dữ liệu

- Với dữ liệu về gói đầu tư tối đa và gói đầu tư tối thiểu, có thể thấy có mối quan hệ với kết quả gọi vốn. Tuy nhiên sự liên hệ này có thể giải thích thông qua mục tiêu gọi vốn (`Goal`). Bởi lẽ, mục tiêu càng cao thì giá trị các gói đầu tư cũng lớn theo tương ứng.









Tiền xử lý dữ liệu

- **Dựa vào kết quả khai phá dữ liệu ở trên, chúng ta sẽ tiến hành tiền xử lý dữ liệu.**
- **Đầu tiên, ta đọc dữ liệu thô đã được làm sạch (chỉ xóa các dòng lỗi do lấy dữ liệu, không thêm hay thay đổi gì)**
- **Bây giờ chúng ta tách dữ liệu thành hai bộ huấn luyện train và thẩm định validation. Mỗi bộ gồm có dữ liệu đầu vào và đầu ra. Chúng ta cũng sẽ mã hóa cột kết quả sang dạng 0 và 1 để thuận tiện về sau.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7164 entries, 0 to 7163
Data columns (total 16 columns):
Id           7164 non-null int64
Name         7164 non-null object
Url          7164 non-null object
Goal         7164 non-null object
Pledged      7164 non-null object
Launch       7164 non-null object
End          7164 non-null object
Year         7164 non-null int64
Comments     7164 non-null int64
Updates      7164 non-null int64
Faqs         7164 non-null int64
Backers      7164 non-null int64
Tiers        7164 non-null int64
TierMin      7164 non-null int64
TierMax      7164 non-null int64
Status       7164 non-null object
dtypes: int64(9), object(7)
memory usage: 895.6+ KB
```

```
Train data: (5014, 15) (5014,)
Test data: (2150, 15) (2150,)
```

Tiền xử lý dữ liệu

- **Đầu tiên, chúng ta sẽ thêm vào các cột mà chúng ta đã phân tích ở trên.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7164 entries, 0 to 7163
Data columns (total 23 columns):
Id              7164 non-null int64
Name            7164 non-null object
Url             7164 non-null object
Goal            7164 non-null object
Pledged         7164 non-null object
Launch          7164 non-null object
End             7164 non-null object
Year            7164 non-null int64
Comments        7164 non-null int64
Updates         7164 non-null int64
Faqs            7164 non-null int64
Backers         7164 non-null int64
Tiers           7164 non-null int64
TierMin         7164 non-null int64
TierMax         7164 non-null int64
NameLength      7164 non-null int64
NameWords       7164 non-null int64
Currency        7164 non-null object
GoalValue       7164 non-null int32
PledgedValue    7164 non-null int32
CurrencyName    7164 non-null object
GoalUSD         7164 non-null float64
PledgedUSD      7164 non-null float64
dtypes: float64(2), int32(2), int64(11), object(8)
memory usage: 1.2+ MB
```

Tiền xử lý dữ liệu

- **Bây giờ chúng ta sẽ loại bỏ các cột mà chúng ta đã phân tích là không cần dùng đến.**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7164 entries, 0 to 7163
Data columns (total 8 columns):
Comments      7164 non-null int64
Updates       7164 non-null int64
Faqs          7164 non-null int64
Backers       7164 non-null int64
Tiers         7164 non-null int64
NameLength    7164 non-null int64
NameWords     7164 non-null int64
GoalUSD       7164 non-null float64
dtypes: float64(1), int64(7)
```

Tiền xử lý dữ liệu

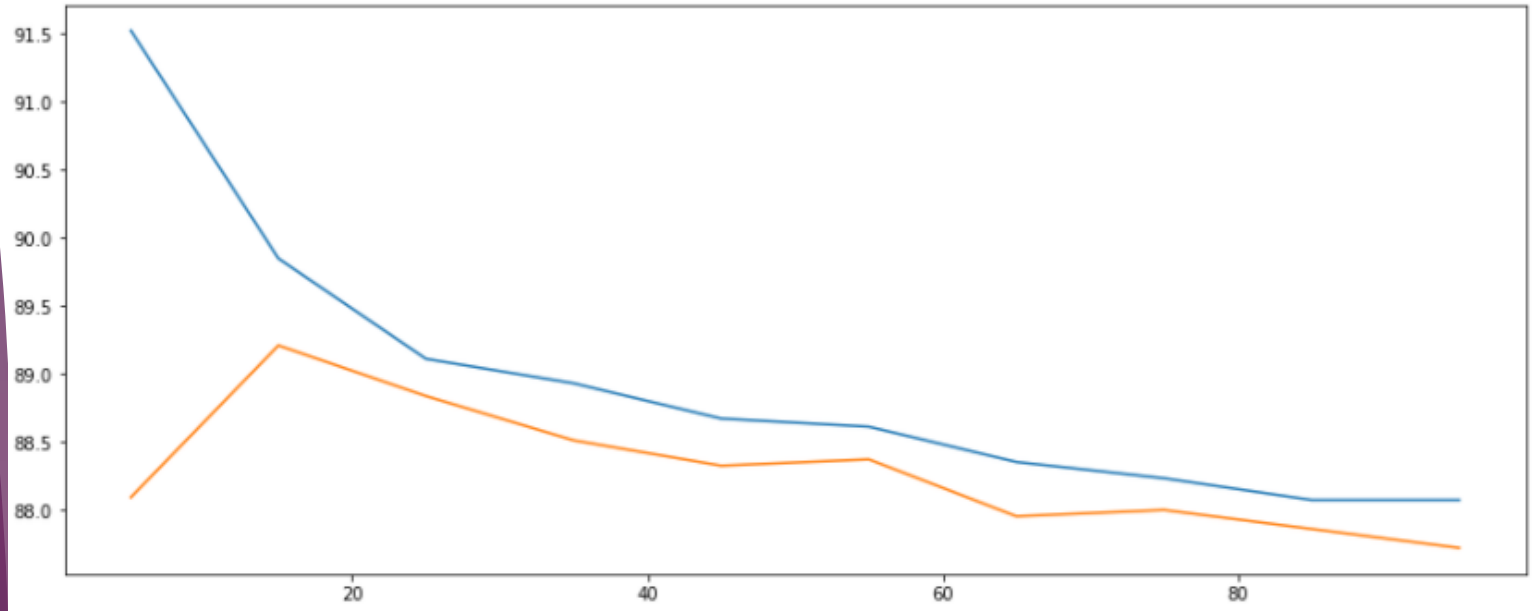
- **Bây giờ chúng ta sẽ tiến hành đặt ngưỡng cho cho các dữ liệu số có đuôi phân bố dài giống như cách chúng ta làm với `Tiers` ở trên. Chúng ta sẽ lấy ngưỡng là quantile 95% (Ngưỡng 95% dữ liệu)**
- **Điều thú vị là các cột của chúng ta đều ở dạng số. Chúng ta sẽ căn chỉnh (scale) dữ liệu lại để tăng tốc mô hình.**
- **Sau đó chúng ta sẽ kết hợp các bước tiền xử lý vào một pipeline để tái sử dụng về sau**

	Comments	Updates	Faqs	Backers	Tiers	NameLength	NameWords	GoalUSD
0	12.0	4.0	4.0	19.0	10.0	55	12	250000.0
1	6.0	1.0	1.0	15.0	9.0	9	1	250000.0
2	0.0	0.0	0.0	6.0	2.0	59	8	250000.0
3	0.0	0.0	0.0	1.0	3.0	33	5	250000.0
4	2.0	0.0	0.0	22.0	8.0	29	4	250000.0

```
Pipeline(steps=[('featureengineer', FeatureEngineer()),
                ('columndropper', ColumnDropper()),
                ('valuelimiter', ValueLimiter()),
                ('standardscaler', StandardScaler())])
```

Xây dựng mô hình

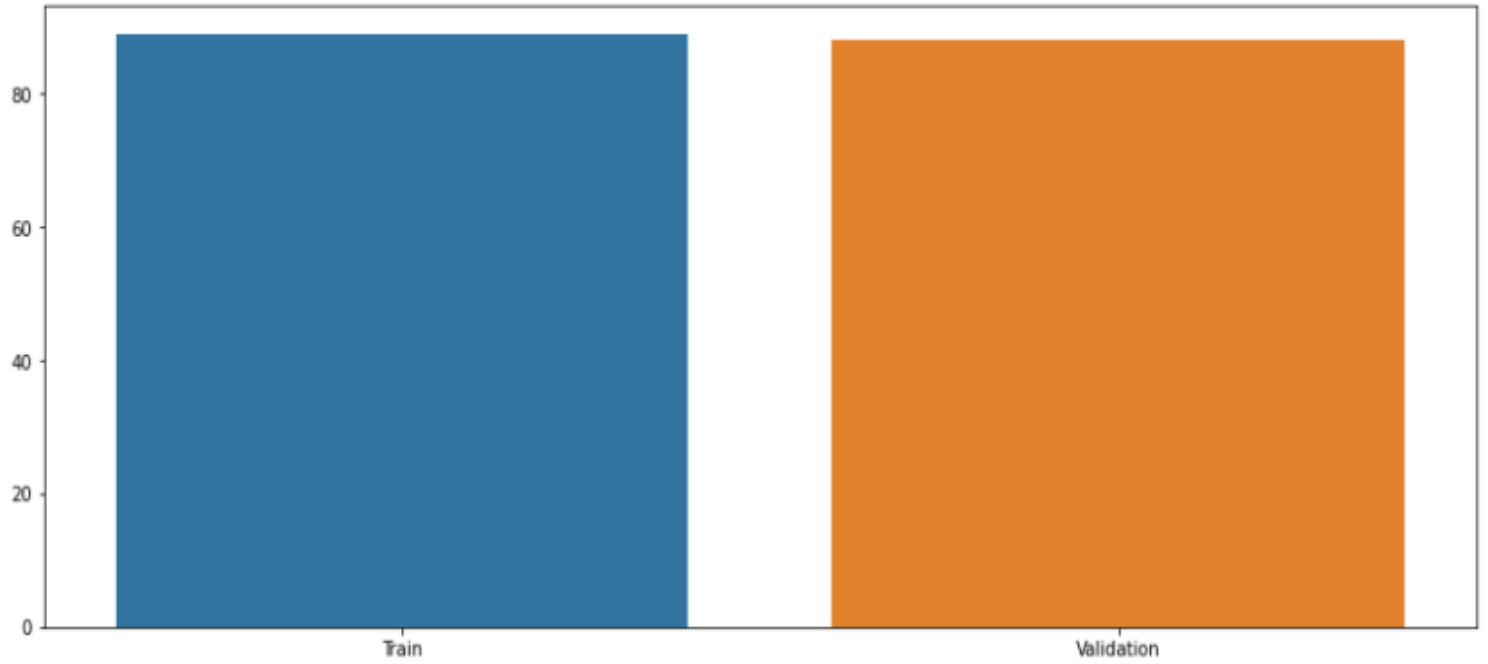
- **Ta sẽ tiến hành xây dựng nhiều mô hình khác nhau.**
- **Thoạt nhìn, các biến đều là số nên chúng ta cảm thấy dữ liệu rất phù hợp cho mô hình KNN.**



Xây dựng mô hình

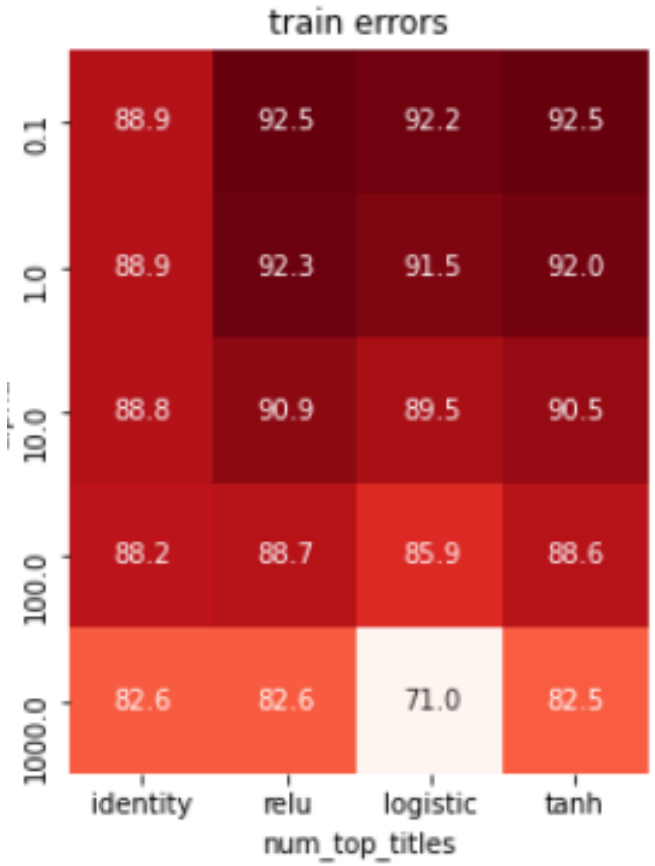
- **Chúng ta nhận thấy có rất nhiều giá trị có quan hệ tuyến tính với kết quả. Chúng ta sẽ thử mô hình hồi quy tuyến tính.**
- **Mô hình hồi quy tuyến ính cho hiệu quả tốt.**

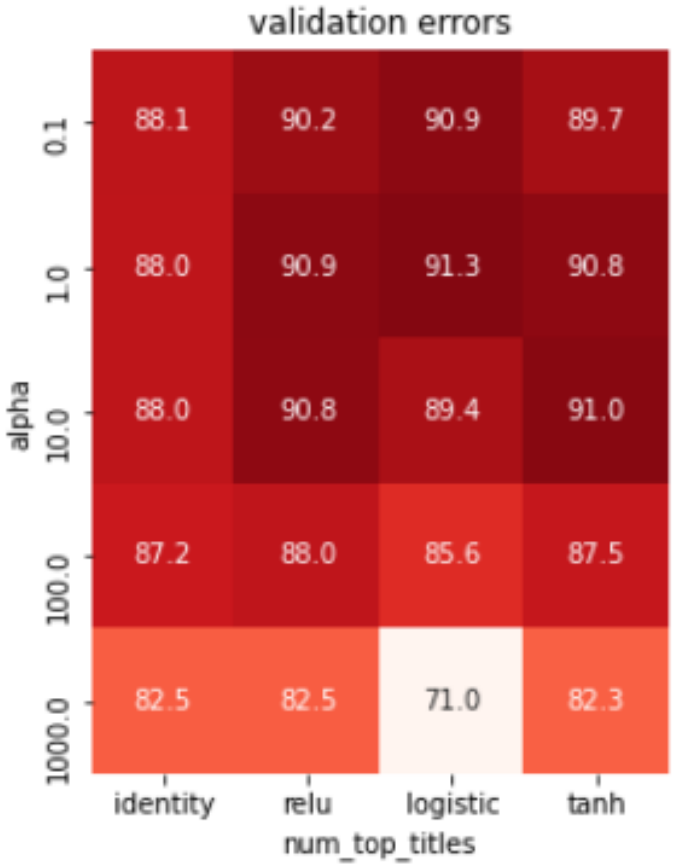
```
Train scores:       88.87116074990028
Validation scores:  88.04651162790698
```



Xây dựng mô hình

- **Bây giờ chúng ta sẽ thử nghiệm trên một mô hình phi tuyến là MLP**





Tài liệu tham khảo

Nhìn lại quá trình làm đồ án

Leader tài tình văn hoa: Thái Tấn Đạt – một người thầy của Cao Lê Minh Hiếu.
Thu thập dữ liệu: lý do Cao Lê Minh Hiếu thành cú đêm và được Thái Tấn Đạt chỉ bảo.