BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HỒ CHÍ MINH

ĐỒ ÁN CƠ SỞ

TÌM HIỂU VỀ PHÂN TÍCH DỮ LIỆU TRONG KINH DOANH

Ngành: **HỆ THỐNG THÔNG TIN QUẢN LÝ**

Chuyên ngành: **HỆ THỐNG THÔNG TIN QUẢN LÝ**

Giảng viên hướng dẫn : Trần Lê Vĩnh Bảo

Sinh viên thực hiện : Đoàn Quang Tuấn

MSSV: 2087400064 Lớp: 20DHTA1

TP. Hồ Chí Minh, 2023

Mục lục

1 Giới thiệu đề tài

1.1 Giới thiệu đề tài

- Chúng ta đang bước vào kỷ nguyên số với nhu cầu lưu trữ và khai thác các nguồn dữ liệu (Big Data) ngày một lớn. Trở thành một nhà phân tích dữ liệu hoặc đảm nhiệm các vị trí liên quan đến lĩnh vực phân tích dữ liệu là công việc có ý nghĩa quan trọng với bất kì tổ chức, doanh nghiệp nào.
- Dữ liệu lớn (Big Data) là một trong bốn nền tảng quan trọng nhất của cuộc cách mạng công nghệ 4.0 cùng với Internet vạn vật - IoT (Internet of Things), Trí tuệ nhân tạo - AI (Artificial Intelligence), Chuỗi khối-Block chain. Big Data được hiểu là những dữ liệu khổng lồ, là nguồn tài sản thông tin có dung lượng lớn và đa dạng, có vận tốc cao, đòi hỏi các hình thức xử lý thông tin có hiệu quả về chi phí, để nâng cao việc đưa ra quyết định và tối ưu hóa quy trình. Nói cách khác, Big Data là một tệp dữ liệu khổng lồ không thể phân tích được bằng các công cụ và phần mềm thông thường. Tầm quan trọng của dữ liệu lớn không nằm ở lượng dữ liệu mà chúng ta có, nó nằm ở việc chúng ta làm gì với những dữ liệu đó.
- Hầu hết các doanh nghiệp, tổ chức sẽ sử dụng nguồn dữ liệu lớn phân tích để tìm ra câu trả lời cho các câu hỏi liên quan đến việc giảm chi phí, giảm thời gian, phát triển sản phẩm mới, dịch vụ tối ưu và ra quyết định thông minh. Khi việc phân tích nguồn dữ liệu lớn được hỗ trợ tối đa, con người có thể hoàn thành tốt một số công việc như xác định nguyên nhân gốc rễ của những thất bại, tạo các chương trình khuyến mại hợp lý dựa trên thói quen của khách hàng đối với công việc kinh doanh, tính toán được những rủi ro gặp phải, phát hiện hành vi gian lận trước khi nó có ảnh hưởng, ...
- Là một nhánh của Phân tích dữ liệu, Phân tích dữ liệu kinh doanh (Business Data Analytics) là một ngành có tính liên ngành giữa công nghệ thông tin và kinh tế. Công việc tập trung vào việc thu thập, khai thác, quản lý và xử lý bộ dữ liệu - Big Data để từ đó đưa ra các nhận định, dự đoán xu hướng hoạt động của tương lai. Phân tích dữ liệu kinh doanh có thể bao gồm phân tích dữ liệu thăm dò, phân tích dữ liệu xác nhận, phân tích dữ liệu định lượng và phân tích dữ liệu định tính (tập trung vào các dữ liệu như video, hình ảnh và văn bản), ... Đây là công việc có ý nghĩa và có tầm quan trọng lớn đối với bất cứ tổ chức hoặc doanh nghiệp nào, đặc biệt trong các lĩnh vực như ngân hàng, tài chính, đầu tư, bảo hiểm, du lịch, quốc phòng, hàng không vũ trụ và y học, …
- Và đề tài của chúng ta sẽ tìm hiểu về thông tin và phân tích của 3 bộ dữ liệu: Amazon cell phone, Amazon top 50, Netflix. Và thông qua đó đề xuất một số phương pháp thống kê hoặc mô hình machine learning phù hợp với các bộ dữ liệu này từ kiến thức đã học.

1.2 Bộ dữ liệu đã chọn

Amazon cell phone: Dữ liệu về các dòng điện thoại trên Amazon. Gồm 2 file là items và reviews.

- Items: Chứa thông tin về các dòng sản phẩm.
  - asin: mã sản phẩm
  - brand: hãng sản xuất điện thoại
  - title: tên phiên bản điện thoại
  - url: liên kết đến trang web bày bán điện thoại
  - image: liên kết đến hình ảnh điện thoại
  - rating: đánh giá trung bình của điện thoại
  - reviewUrl: liên kết đến với trang đánh giá điện thoại
  - totalReviews: tổng số lượt đánh giá về điện thoại
  - price: giá điện thoại
  - originalPrice: giá gốc của điện thoại
- Reviews: Chứa bình luận về các dòng sản phẩm đó. 2 file liên kết bằng trường asin.
  - asin: mã sản phẩm
  - name: tên người đánh giá
  - rating: số điểm đánh giá
  - date: ngày đánh giá
  - verified: xác minh
  - title: tiêu đề bài đánh giá
  - body: mô tả chi tiết bài đánh giá
  - helpfulVote: bài đánh giá có hữu ích với người khác không

Amazon top 50: Dữ liệu về 50 cuốn sách bán chạy nhất trên Amazon trong giai đoạn 2009-2019.

- name: tên cuốn sách
- author: tác giả cuốn sách
- user rating: đánh giá của người mua
- reviews: bài phê bình
- price: giá sách
- year: năm xuất bản
- genre: thể loại sách

Netflix: Gồm 2 file:

- Credits.csv: Chứa thông tin về nhân sự của bộ phim, gồm các cột: id, tên phim, đạo diễn và các diễn viên của bộ phim đó.
  - id: ID tiêu đề trên JustWatch.
  - title: tên tiêu đề
  - show_type: chương trình truyền hình hoặc phim
  - description: mô tả ngắn gọn
  - release_year: năm phát hành
  - age_certification: chứng nhận độ tuổi
  - runtime: thời lượng của tập hoặc phim
  - genres: danh sách các thể loại
  - production_countries: quốc gia sản xuất phim
  - seasons: số mùa nếu nó là Show
  - imdb_id: id tiêu đề trên IMDB.
  - imdb_score: điểm trên IMDB.
  - imdb_votes: bình chọn trên IMDB.
  - tmdb_popularity: mức độ phổ biến trên TMDB.
  - tmdb_score: điểm trên TMDB.
- Titles.csv: Chứa thông tin về các bộ phim như thể loại, năm phát hành, …
  - person_ID: ID người trên JustWatch.
  - id: ID tiêu đề trên JustWatch.
  - name: Tên diễn viên hoặc đạo diễn.
  - character_name: Tên nhân vật.
  - role: DIỄN VIÊN hoặc ĐẠO DIỄN.

2 Cơ sở lý thuyết

2.1 Tổng quan về thống kê mô tả

- Thống kê mô tả là gì: Thống kê mô tả được hiểu là các hệ số mô tả ngắn gọn hay tóm tắt một tập dữ liệu nhất định, cũng có thể là đại diện cho toàn bộ hoặc một mẫu của một tổng thể. Thống kê mô tả được chia thành đo lường xu hướng tập trung và đo lường biến động. Đo lường xu hướng tập trung có giá trị trung bình, trung vị và yếu vị, trong khi các đo lường biến động gồm độ lệch chuẩn, phương sai, giá trị nhỏ nhất và giá trị lớn nhất, độ nhọn và độ lệch.
- Đặc điểm của số liệu thống kê mô tả: Thống kê mô tả ra đời đã giúp mô tả và hiểu được các tính chất của một bộ dữ liệu cụ thể bằng cách đưa ra các tóm tắt ngắn về mẫu và các thông số của dữ liệu. Loại thống kê mô tả phổ biến nhất là các thông số xu hướng tập trung gồm: giá trị trung bình, trung vị và yếu vị, các thông số này được sử dụng ở hầu hết các cấp độ toán học và thống kê. Giá trị trung bình được tính bằng cách cộng tất cả các số liệu trong tập dữ liệu sau đó chia cho số lượng dữ liệu trong tập. Thống kê mô tả được sử dụng để nhằm mục đích có thể cung cấp những thông tin định lượng phức tạp của một bộ dữ liệu lớn thành các mô tả đơn giản.
- Các thuật ngữ liên quan:
  - Giá trị trung bình:
    - Giá trị trung bình trong tiếng Anh là Mean. Giá trị trung bình được hiểu là bình quân toán học đơn giản của một tập hợp gồm hai hoặc nhiều số.
    - Giá trị trung bình của một tập hợp số đã cho có thể được tính theo nhiều cách, gồm có phương pháp trung bình số học sử dụng tổng các số trong chuỗi và phương pháp trung bình hình học.
    - Tuy nhiên, tất cả các phương pháp chính để tính trung bình đơn giản của một chuỗi số bình thường đều tạo ra kết quả xấp xỉ bằng nhau.

- Trung vị:
  - Trung vị trong tiếng Anh là Median. Trung vị được hiểu là số nằm giữa trong một tập dữ liệu có các số được sắp xếp. Để nhằm mục đích có thể xác định giá trị trung vị trong một chuỗi số, trước tiên các số phải được sắp xếp theo thứ tự giá trị từ thấp nhất đến cao nhất hoặc cao nhất đến thấp nhất.
  - Trung vị có thể được sử dụng để nhằm xác định giá trị trung bình gần đúng hoặc giá trị trung bình, tuy nhiên không được nhầm lẫn trung vị với giá trị trung bình thực tế.
  - Nếu tập dữ liệu có số lượng điểm dữ liệu là lẻ, trung vị là số nằm ở giữa có cùng một số lượng điểm dữ liệu ở bên dưới và bên trên. Nếu tập dữ liệu có số lượng điểm dữ liệu là chẵn, để tìm giá trị trung vị cần xác định cặp điểm dữ liệu ở giữa sau đó cộng 2 số này lại và chia cho hai.
  - Trung vị được sử dụng thay cho giá trị trung bình khi có các điểm ngoại lai trong chuỗi dữ liệu, các điểm ngoại lai có thể làm lệch giá trị trung bình của các giá trị. Trung vị của một chuỗi ít bị ảnh hưởng bởi các điểm ngoại lai hơn giá trị trung bình.
- Yếu vị:
  - Yếu vị trong tiếng Anh là Mode.
  - Một tập hợp các số có thể có một hoặc nhiều hơn một yếu vị hoặc không có yếu vị nào cả. Các khái niệm thống kê phổ biến khác theo xu hướng đo lường trung tâm gồm có giá trị trung bình hay bình quân của một tập dữ liệu và trung vị, giá trị nằm ở giữa trong một tập dữ liệu.
  - Yếu vị có thể có cùng giá trị với giá trị trung bình và trung vị, nhưng không phải lúc nào cũng đúng như vậy.
- Độ lệch chuẩn:
  - Độ lệch chuẩn, hay độ lệch tiêu chuẩn (Standard Deviation) là một đại lượng thống kê dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số.
  - Có thể tính ra độ lệch chuẩn bằng cách lấy căn bậc hai của phương sai. Khi hai tập dữ liệu có cùng giá trị trung bình cộng, tập nào có độ lệch chuẩn lớn hơn là tập có dữ liệu biến thiên nhiều hơn.
  - Trong trường hợp hai tập dữ liệu có giá trị trung bình cộng không bằng nhau, thì việc so sánh độ lệch chuẩn của chúng không có ý nghĩa. Độ lệch chuẩn còn được sử dụng khi tính sai số chuẩn. Khi lấy độ lệch chuẩn chia cho căn bậc hai của số lượng quan sát trong tập dữ liệu, sẽ có giá trị của sai số chuẩn.
- Phương sai:
  - Trong lý thuyết xác suất và thống kê, phương sai của một biến ngẫu nhiên là một độ đo sự phân tán thống kê của biến đó, nó hàm ý các giá trị của biến đó thường ở cách giá trị kỳ vọng bao xa.
  - Phương sai của biến ngẫu nhiên giá trị thực là moment trung tâm, nó còn là nửa bất biến (cumulant) thứ hai của nó. Phương sai của một biến ngẫu nhiên là bình phương của độ lệch chuẩn.

2.2 Giới thiệu về Python

Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình và là ngôn ngữ lập trình dễ học; được dùng rộng rãi trong phát triển trí tuệ nhân tạo. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Vào tháng 7 năm 2018, van Rossum đã từ chức lãnh đạo trong cộng đồng ngôn ngữ Python sau 30 năm làm việc.
Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (Machine Learning). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.

2.3 Giới thiệu về DataFrame

Pandas.DataFrame (Gọi tắt là DataFrame). là một mảng hai chiều có gắn nhãn.
DataFrame có một số đặc điểm sau:

- DataFrame là một mảng hai chiều.
- DataFrame có thể xem như là nhiều Series có chung label (index) được ghép kế tiếp nhau.
- Dữ liệu trong một cột là đồng nhất.
- Tuy nhiên hai cột khác nhau có thể có kiểu dữ liệu khác nhau.
- Có thể thay đổi kích thước DataFrame bằng cách thêm bớt dòng cột

2.4 Giới thiệu về thư viện matplotlib

Như chúng ta đã biết Python được sử dụng nhiều nhất trong lĩnh vực phân tích dữ liệu, mà trong khoa học dữ liệu, việc trực quan hóa thông qua các đồ thị, biểu đồ giúp cho chúng ta hiểu được các mối quan hệ trong dữ liệu dễ dàng hơn rất nhiều. Matplotlib là một thư viện sử dụng để vẽ các đồ thị trong Python, chính vì vậy nó là thư viện cực phổ biến của Python

2.5 Google Colab là gì ?

Colab (hay còn gọi là "Colaboratory") cho phép bạn viết và thực thi Python trong trình duyệt với các lợi ích sau:

- Không yêu cầu cấu hình
- Quyền truy cập miễn phí vào GPU
- Chia sẻ dễ dàng

Cho dù bạn là sinh viên, nhà khoa học dữ liệu hay nhà nghiên cứu AI (trí tuệ nhân tạo), Colab đều giúp bạn hoàn thành công việc dễ dàng hơn.
Khoa học dữ liệu: Với Colab, bạn có thể khai thác toàn bộ sức mạnh của các thư viện Python phổ biến để phân tích và trực quan hóa dữ liệu. Ô chứa mã ở bên dưới sử dụng numpy để tạo một số dữ liệu ngẫu nhiên và sử dụng matplotlib để trực quan hóa dữ liệu đó. Để chỉnh sửa mã này, bạn chỉ cần nhấp vào ô đó và bắt đầu chỉnh sửa.

2.6 Một số phương thức tổng hợp cơ bản trong pandas

.count(). Số lượng phần tử khác NaN. Trong Python trả về số lần xuất hiện của chuỗi con trong khoảng [start, end]. Đếm xem chuỗi str này xuất hiện bao nhiêu lần trong chuỗi string hoặc chuỗi con của string nếu bạn cung cấp chỉ mục ban đầu start và chỉ mục kết thúc end.
.min(). Giá trị nhỏ nhất. Xuất ra giá trị nhỏ nhất của một cột dữ liệu mà mình yêu cầu
.max(). Giá trị lớn nhất. Xuất ra giá trị lớn nhất của một cột dữ liệu mà mình yêu cầu
.mean(). Giá trị trung bình. Tính các giá trị trung bình cho mỗi nhóm dữ liệu và là một trong những thước đo quan trọng nhất và được sử dụng rộng rãi.
.std(). Độ lệch chuẩn. Là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số
.describe(). Thống kê mô tả của Series và DataFrame. Trả về một DataFrame mới với số hàng được hiển thị ra các thông số như giá trị trung bình, độ lệch chuẩn, min, max và tỉ lệ phần trăm của các cột.

2.7 Nhập dữ liệu

Định dạng CSV (Comma Separated Values). Là một loại định dạng văn bản đơn giản mà trong đó, các giá trị được ngăn cách với nhau bằng dấu phẩy. Định dạng CSV thường xuyên được sử dụng để lưu các bảng tính quy mô nhỏ như danh bạ, danh sách lớp, báo cáo … Thông thường, một file csv có đuôi là .csv
Cách đọc file .csv trong pandas. Để đọc file .csv trong python, ta có thể dùng đến hàm pandas.read_csv() như sau:

```
Pandas.read_csv(<đường_dẫn_đến_file>, [các_tham_số_khác])
```

Trong đó, đường_dẫn_đến_file có thể là đường dẫn một tập tin trong máy (local) hoặc là một url (remote)

Một số tham số của pandas.read_csv(). Vì dữ liệu thô có muôn hình vạn trạng, nên pandas.read_csv() cung cấp đến hơn 40 tham số để giúp đỡ quá trình đọc dữ liệu, mà trong dữ liệu này của chúng ta là những tham số được ngăn cách bởi dấu phẩy ‘,’ nên chúng ta có thể không cần các_tham_số_khác ở đằng sau.

2.8 Khái niệm về điểm IMDB

Điểm IMDB: IMDB là từ viết tắt của Internet Movie Database (Cơ ở dữ liệu điện ảnh trên Internet). Đây là thư viện điện ảnh, nơi cung cấp thông tin về các bộ phim, diễn viên, đạo diễn, chủ đề điện ảnh, truyền hình và video game. IMDB còn là nơi tổng hợp những nhận xét, đánh giá, phê bình các yếu tố như bối cảnh, kịch bản, hiệu quả hình ảnh, kỹ thuật quay... hay xếp hạng phim. Bên cạnh đó, IMDB cũng đưa ra thang điểm từ 0 - 10 để khán giả đánh giá, bình luận đối với mỗi bộ phim (gọi là thang điểm IMDB). Bộ phim nào nhận được điểm

IMDB cao tức là nó nhận được nhiều sự ủng hộ từ khán giả. Đó là lí do vì sao điểm IMDB quyết định một bộ phim có hay không.
Cách chọn phim theo điểm IMDB:

- IMDB có thang điểm 10, do đó, những bộ phim có thang điểm thấp dưới 5 thì chúng ta nên cân nhắc kỹ càng trước khi xem.
- Những bộ phim từ 5 - 6 điểm ở mức bình thường,chúng ta có thể xem qua trailer, giới thiệu phim hoặc review rồi quyết định xem
- Những bộ phim từ 7 - 8 điểm ở mức khá hay, chúng ta nên dành thời gian để xem nó.
- Những bộ phim từ 8 điểm trở lên chắc chắn là chúng ta không nên bỏ qua. Những bộ phim này sẽ mang đến cho chúng ta những trải nghiệm tuyệt vời với tính nghệ thuật cao.

Một lưu ý nhỏ là thang điểm IMDB dù co tính khách quan cao do được đánh giá bởi giới chuyên môn và khán giả nhưng vẫn bị chi phối bởi yếu tố cá nhân, do đó, chỉ mang tính chất tham khảo. Niềm vui, tính giải trí và những bài học nhân văn sau sắc khi xem phim mới là cái đích cuối cùng để bạn cảm thấy bộ phim có đáng để xem hay không.

3 Kết quả

3.1 Các thư viện cần thiết

```
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import missingno as msno
from wordcloud import WordCloud
colors = ['#494BD3', '#E28AE2', '#F1F481', '#79DB80', '#DF5F5F',
              '#69DADE', '#C2E37D', '#E26580', '#D39F49', '#B96FE3']
from plotly.offline import init_notebook_mode, iplot
from ast import literal_eval
```

3.2 Amazon Cell Phone

Đọc dữ liệu vào.

```
items = pd.read_csv('https://raw.githubusercontent.com/qtuan79/tuanfile/main/20191226-items.csv')
reviews = pd.read_csv('https://raw.githubusercontent.com/qtuan79/tuanfile/main/20191226-reviews.csv')
```

Tổng quan về dữ liệu.

```
items.info()
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 720 entries, 0 to 719
    Data columns (total 10 columns):
     #   Column         Non-Null Count  Dtype
    ---  ------         --------------  -----
     0   asin           720 non-null    object
     1   brand          716 non-null    object
     2   title          720 non-null    object
     3   url            720 non-null    object
     4   image          720 non-null    object
```

```
 5   rating          720 non-null    float64
 6   reviewUrl       720 non-null    object
 7   totalReviews    720 non-null    int64
 8   price           720 non-null    float64
 9   originalPrice   720 non-null    float64
dtypes: float64(3), int64(1), object(6)
memory usage: 56.4+ KB
```

Chúng ta có thể thấy rằng dữ liệu này có một biến brand chỉ có 716 dữ liệu và có thể suy đoán rằng có một vài dữ liệu NaN đã bị xen lẫn vào đây.
Mình sẽ kiểm tra dữ liệu lỗi:

```
items[pd.isna(items.brand)]
```

Index	Asin	Brand	Title	Rating
0	B0000SX2UC	NaN	Dual-Band / Tri-Mode Sprint PCS Phone w/ Voice Activated Dialing & Bright White Backlit Screen	3.0
144	B01EWKHIAI	NaN	Microsoft Lumia 950 32GB Dual Sim NAM RM-1118 GSM Factory Unlocked - US Warranty (Black)	3.9
471	B07JHXX5YR	NaN	ROG Phone Gaming Smartphone ZS600KL-S845-8G512G - 6” FHD+ 2160x1080 90Hz Display - Qualcomm Snapdragon 845 - 8GB RAM - 512GB Storage - LTE Unlocked Dual SIM Gaming Phone - US Warranty	3.9
631	B07T3KMJW8	NaN	Redmi 7A 2+16Gb Black EU	3.8

Bảng 3.2.1. Bảng trích lọc các dữ liệu NaN
Sau khi kiểm tra dữ liệu lỗi mình có thể thấy rằng những dữ liệu này chưa được phân ra thuộc thương hiệu nào và chúng ta có thể xóa nó như sau.

```
items = items.dropna(subset=['brand'])
reviews.info()
```

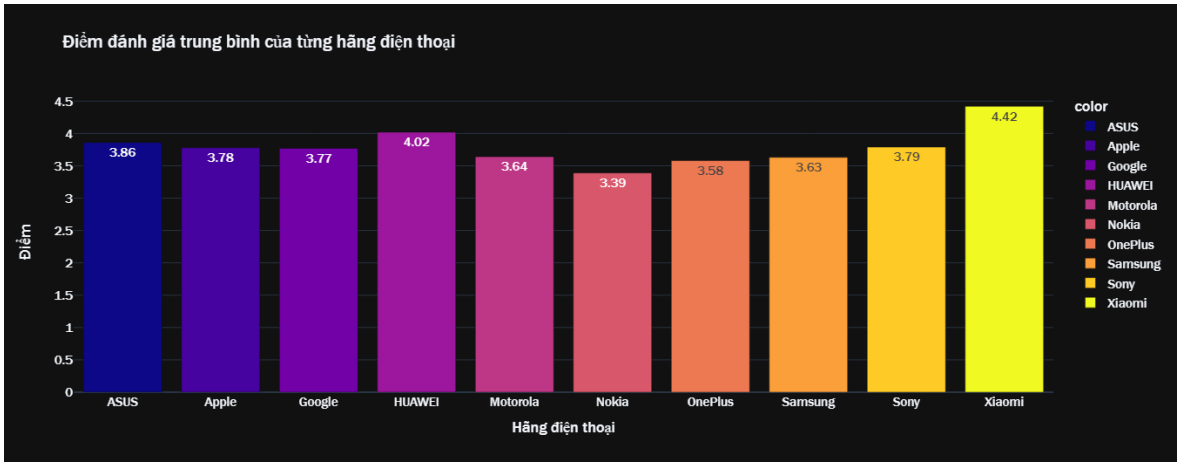
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67986 entries, 0 to 67985
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   asin          67986 non-null  object
 1   name          67984 non-null  object
 2   rating        67986 non-null  int64
 3   date          67986 non-null  object
 4   verified      67986 non-null  bool
 5   title         67972 non-null  object
 6   body          67965 non-null  object
```

```
 7   helpfulVotes  27215 non-null  float64
dtypes: bool(1), float64(1), int64(1), object(5)
memory usage: 3.7+ MB
```

Theo như trực quan dữ liệu chúng ta thấy rằng sẽ có 67986 cột dữ liệu và có một vài dữ liệu NaN ở các cột name, title, body và helpfulVotes. Nhưng theo quan sát thực tế thì name, title và body có thể người dùng quên không điền vào hoặc họ không điền dẫn đến lúc lấy dữ liệu ra phần dữ liệu để trống dẫn đến việc máy tự hiểu là NaN; còn helpfulVotes có thể do cái bài đánh giá này của người dùng không có lượt đánh giá nào của những người dùng khác rằng nó sẽ hữu ích vậy nên ô này đã được để trống rất nhiều nên có thể sẽ ảnh hưởng đến các kết quả lúc chúng ta thống kê mô tả dữ liệu này.

Phân tích.

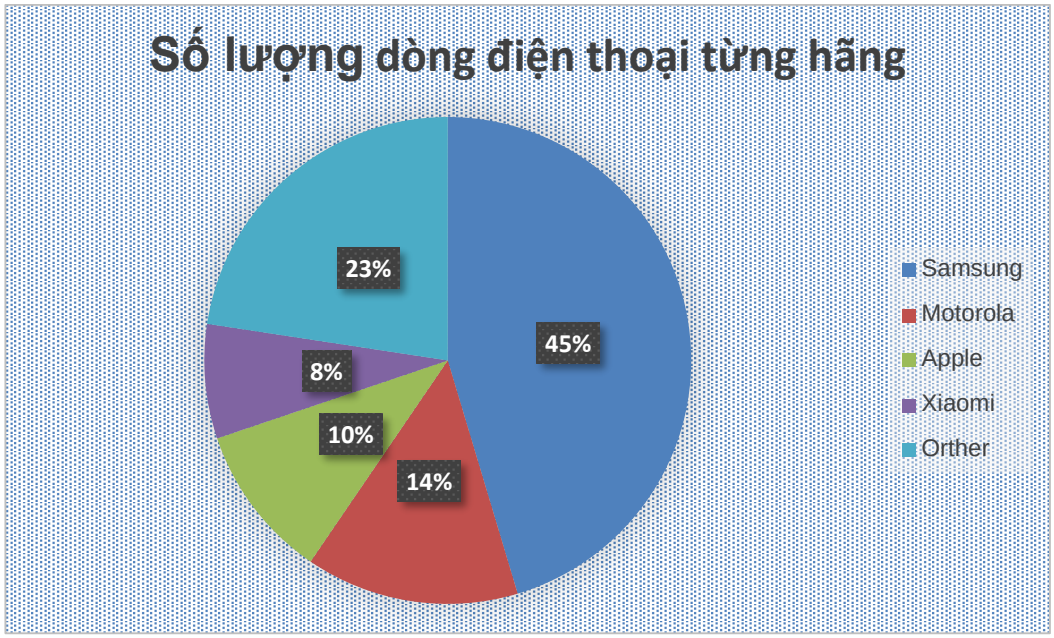
1. *Điểm đánh giá trung bình của từng hãng điện thoại*



Hình 3.2.1. Điểm đánh giá trung bình của từng hãng điện thoại.

Như chúng ta nhìn vào bảng điểm trung bình của từng hãng điện thoại ở trên Hình 3.2.1 thì chúng ta có thể thấy rằng Xiaomi là hãng điện thoại có điểm số cao nhất trong 10 hãng điện thoại có ở trên và điều này có thể cho thấy rằng mức độ hài lòng về các sản phẩm của Xiaomi đối với khách hàng và hãng điện thoại có số điểm đánh giá thấp nhất chính là Nokia bởi vì Nokia họ đã không làm tốt và không cạnh tranh được với các hãng điện thoại khác ở trong phân khúc cùng với việc chậm chạp trong việc đổi mới công nghệ, phát triển các sản phẩm không mang tính đột phá và làm cho người dùng thất vọng và quay đầu với các sản phẩm ra mắt của Nokia. Và các hãng điện thoại khác còn lại hầu hết đều ở trên mức 3.5 và có nhiều hãng có số điểm ở gần mức 4.0 vậy nên số điểm này đều khá là ổn và đều ở mức độ hài lòng khá tốt đối với khách hàng sử dụng sản phẩm.

2. *Tỉ trọng điện thoại các hãng bán trên Amazon.*



Hình 3.2.2. Tỉ trọng các dòng điện thoại của từng hãng bán trên Amazon

Nhìn vào biểu đồ phần trăm số lượng dòng điện thoại từng hãng bán trên Amazon chúng ta có thể thấy rằng Samsung là hãng điện thoại chiếm tỉ trọng cao nhất với 45% và thấp nhất là các dòng điện thoại đến từ các hãng khác khi các hãng khác còn lại chỉ chiếm có 23%. Nó có thể cho chúng ta thấy rằng Samsung là hãng điện thoại có nhiều dòng điện thoại và nhiều phân khúc giá khác nhau để đánh vào nhiều đối tượng khách hàng sử dụng.

3. *Giá trung bình của từng hãng điện thoại*

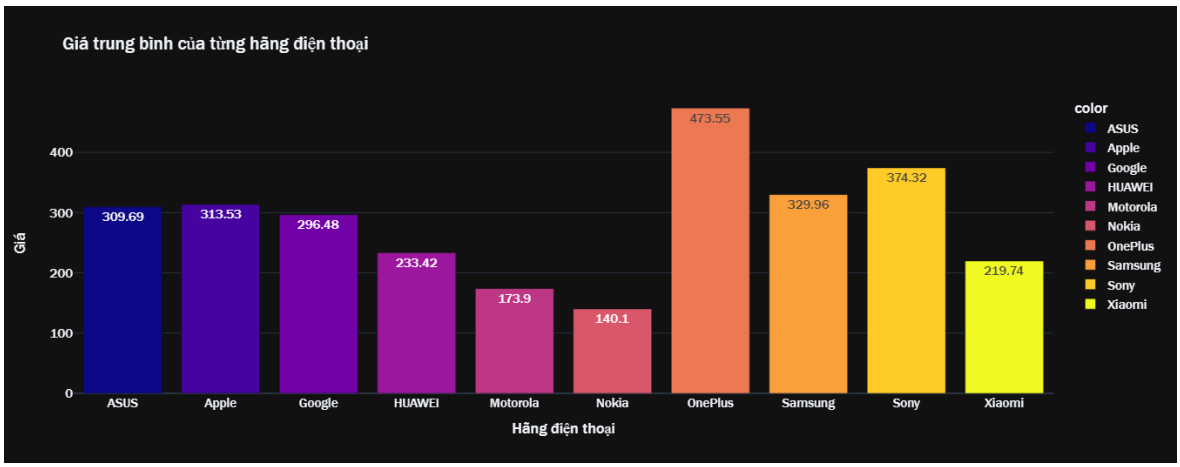
Khi kiểm tra dữ liệu mình thấy rằng có những điện thoại có mức giá là 0 mà mức giá đó là không thể có được. Vậy nên chúng ta sẽ phải xóa đi những hàng dữ liệu mà có mức giá là 0

```
items = items.drop(items[items.price == 0].index)#xóa những điện thoại có
 price = 0
```

Và chúng ta sẽ nhóm nó lại theo từng hãng và tính mức giá trung bình của từng hãng điện thoại.

```
BrandMeanPrice = items.groupby('brand')[['price']].mean().reset_index().r
ound(2)
```

Sau đó chúng ta sẽ vẽ biểu đồ để trực quan hóa dữ liệu.

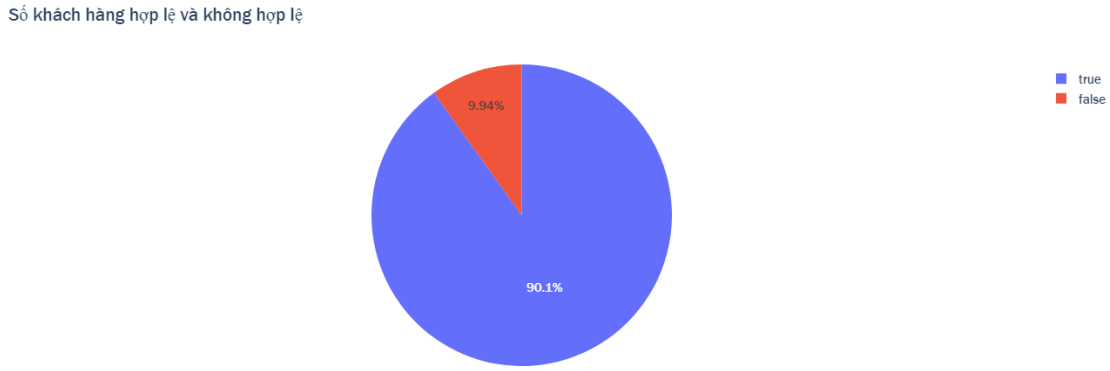


Hình 3.2.3. Biểu đồ thể hiện mức giá trung bình của từng hãng điện thoại

Nhìn vào biểu đồ chúng ta có thể thấy rằng điện thoại OnePlus có mức giá trung bình là cao nhất trong 10 hãng điện thoại là 473.55$ và điện thoại có mức giá trung bình thấp nhất đó là Nokia 140.1$ và hầu hết các hãng điện thoại còn lại điện thoại của họ nằm trong mức giá ở trong khoảng 290$ - 330$.

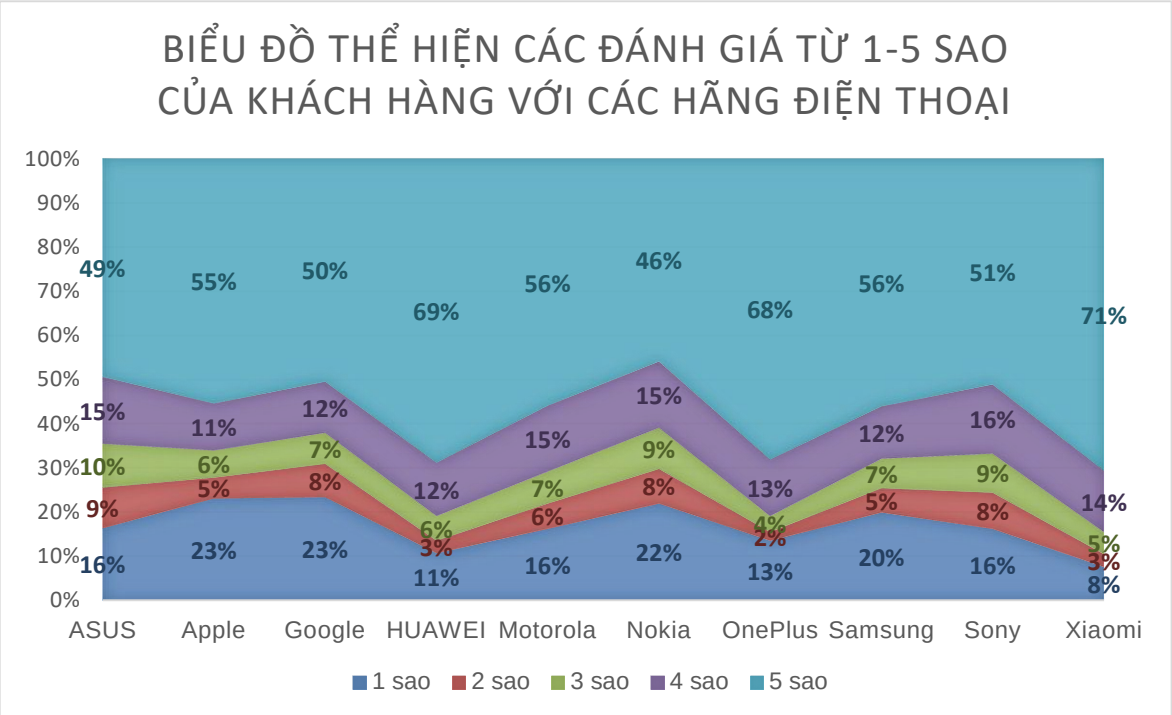
Và chúng ta dựa vào Hình 3.2.1 và Hình 3.2.3 thì có thể thấy rằng Xiaomi họ đang làm rất tốt trong việc đánh vào phân khúc giá rẻ của điện thoại với mức giá chỉ giao động trung bình là 219.74$ và đánh giá mức độ hài lòng về điện thoại của khách hàng đối với Xiaomi là rất cao (4.42/5.0) điều này cho thấy rằng Xiaomi họ đang làm rất tốt trong việc thu hút khách hàng và tạo dựng thương hiệu của họ mặc dù họ là hãng điện thoại tham gia sau khi các ông lớn như Apple, Samsung, HUAWEI đã thâu tóm hầu hết các thị trường trên thế giới.

4. *Tỉ lệ khách hàng hợp lệ và không hợp lệ*



Hình 3.2.4. Tỉ lệ khách hàng hợp lệ và không hợp lệ. True: Hợp lệ False: Không hợp lệ
Biểu đồ cho chúng ta biết được những người đánh giá hợp lệ và không hợp lệ, hợp lệ là

5. *Tổng hợp các đánh giá của người dùng từ 1-5 với các hãng điện thoại*



Hình 3.2.5. Biểu đồ thể hiện các đánh giá từ 1-5 của khách hàng với các hãng điện thoại. 1 sao: Các đánh giá 1 sao, 2 sao: Các đánh giá 2 sao, 3 sao: Các đánh giá 3 sao, 4 sao: Các đánh giá 4 sao và 5 sao: Các đánh giá 5 sao.
Nhìn vào Hình 3.2.5 chúng ta có thể thấy số lượt khách hàng đánh giá 5 sao cho các dòng sản phẩm 5 sao của các hãng điện thoại chiếm tỉ lệ trung bình là hơn 50% và có ba hãng điện thoại HUAWEI, Xiaomi và OnePlus có tỉ lệ đánh giá 5 sao là lớn nhất với tỉ lệ từ 68% đổ lên và cao nhất là Xiaomi với 71% người dùng đánh giá là 5 sao. Điều này cho thấy rằng các sản phẩm đến từ HUAWEI, Xiaomi và OnePlus đang làm cho khách hàng cảm thấy rất hài lòng đối với sản phẩm đến từ họ. Còn Samsung, Nokia, Apple và Google là các hãng điện thoại chiếm lượt đánh giá 1 sao nhiều nhất khi các hãng điện

thoại này các tỉ lệ đánh giá 1 sao là ở trong khoảng từ 20% đến 23% điều này cũng có thể cho thấy rằng cứ 5 người mua sản phẩm của họ thì có 1 người không hài lòng với sản phẩm của họ sau khi sử dụng so với các hãng điện thoại khác.

3.3 AMAZON TOP 50 BESTSELLING BOOKS 2009 – 2019

Đọc dữ liệu vào.

```
booksell = pd.read_csv('https://raw.githubusercontent.com/qtuan79/tuanfile/main/bestsellers_with_categories_2022_03_27.csv')
```

Tổng quan về dữ liệu.

```
booksell.info()#kiểm tra có dữ liệu
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Name         700 non-null    object 
 1   Author       700 non-null    object 
 2   User Rating  700 non-null    float64
 3   Reviews      700 non-null    int64  
 4   Price        700 non-null    int64  
 5   Year         700 non-null    int64  
 6   Genre        700 non-null    object 
dtypes: float64(1), int64(3), object(3)
memory usage: 38.4+ KB
```

Nhìn vào đây chúng ta có thể thấy rằng dữ liệu gồm có 7 cột và mỗi cột có 700 dòng và không có dữ liệu NaN, kiểu dữ liệu từng cột đều phù hợp với dữ liệu để thao tác.

Phân tích.

1. *Top 10 cuốn sách có lượt đánh giá cao nhất*

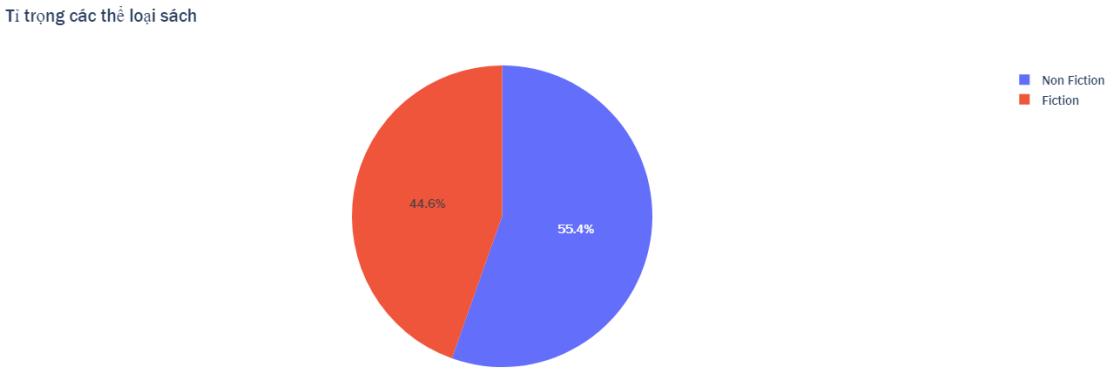
```
booksellbest = booksell.groupby('Name')[['User Rating']].mean().reset_index()
booksellbest.sort_values(['User Rating'], ascending= False).head(10)
```

	Name	User Rating
220	Obama: An Intimate Portrait	4.9
391	The Very Hungry Caterpillar	4.9
159	I Love You to the Moon and Back	4.9
40	Big Shot Diary of a Wimpy Kid Book 16	4.9
252	Rush Revere and the Brave Pilgrims: Time-Trave...	4.9
156	Humans of New York : Stories	4.9
44	Brown Bear, Brown Bear, What Do You See?	4.9
253	Rush Revere and the First Patriots: Time-Trave...	4.9
54	Chicka Chicka Boom Boom (Board Book)	4.9
142	Harry Potter and the Sorcerer's Stone: The Ill...	4.9

Bảng 3.3.1. Top 10 cuốn sách có lượt đánh giá cao nhất.

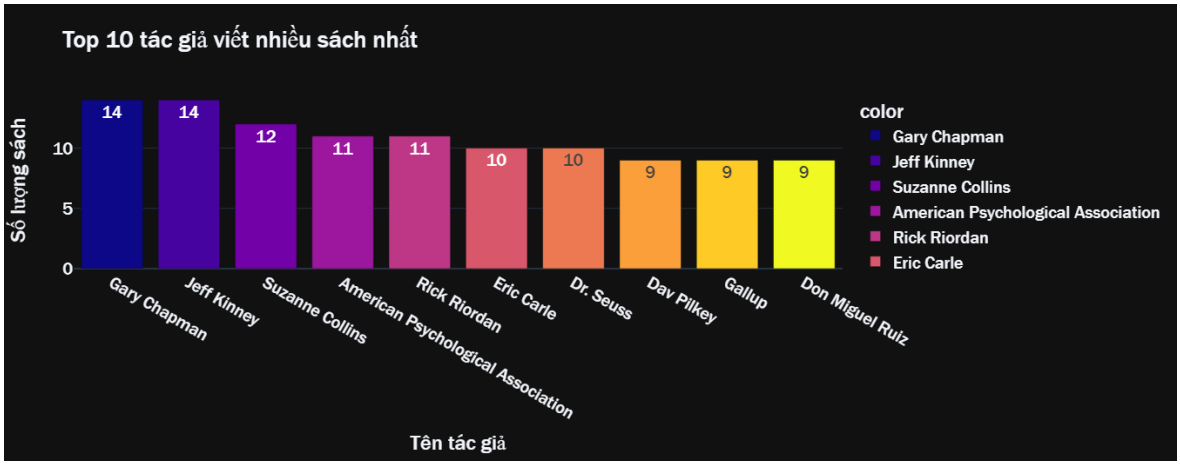
Nhìn vào bảng chúng ta có thể thấy 10 cuốn sách có lượt đánh giá cao nhất và 10 cuốn sách này đều có lượt đánh giá ở mức 4.9/5.0 và nó ở số điểm rất ấn tượng vì được mọi người đánh giá rất cao.

2. *Tỉ trọng các thể loại sách*



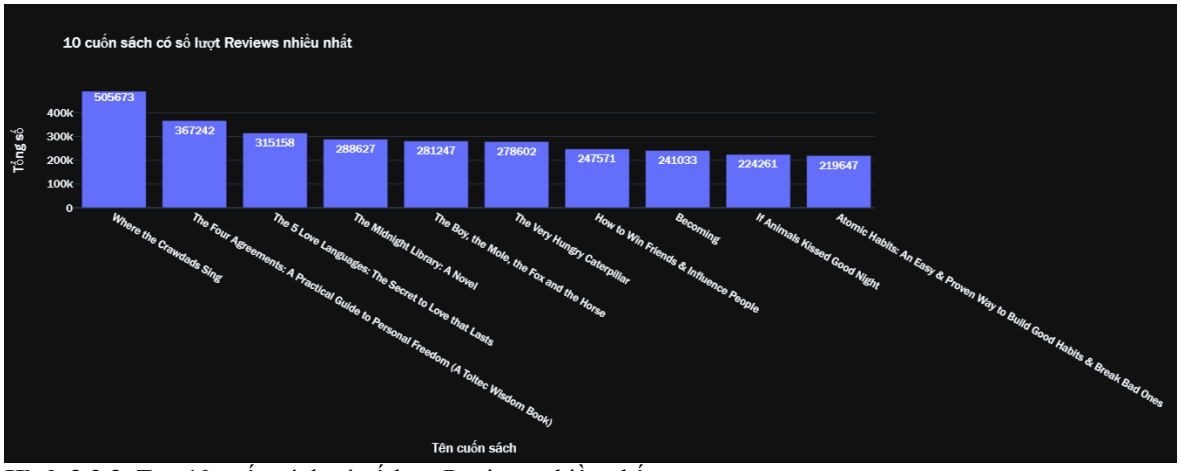
Hình 3.3.1. Tỉ trọng của các thể loại sách. Non Fiction là sách phi hư cấu, Fiction là sách viễn tưởng
Như biểu đồ hình 3.3.1 chúng ta thấy rằng thể loại sách phi hư cấu chiếm 55,4% và sách viễn tưởng chiếm 44,6% vì vậy chúng ta có thể thấy rằng các thể loại sách này chiếm không quá chênh lệch với nhau quá nhiều trong các thể loại.

3. *Top 10 tác giả viết nhiều sách nhất.*



Hình 3.3.2. Top 10 tác giả viết nhiều sách nhất
Nhìn vào Hình 3.3.2 chúng ta có thể thấy được 10 tác giả viết nhiều sách nhất và dẫn đầu đó là hai tác giả Gary Chapman và Jeff Kinney với số sách viết ra là 14 cuốn sách

4. *Top 10 cuốn sách được viết phê bình nhiều nhất*



Hình 3.3.3. Top 10 cuốn sách có số lượt Reviews nhiều nhất

Đứng đầu sách có số lượt phê bình nhiều nhất là cuốn sách "Where the Crawdads Sing" với hơn 500k lượt phê bình và 9 cuốn sách còn lại có số lượt phê bình trong khoảng từ 219k đến 367k lượt phê bình đến từ độc giả.

3.4 Netflix

Đọc dữ liệu vào.

```
titles = pd.read_csv('https://raw.githubusercontent.com/qtuan79/tuanfile/main/titles.csv')
credits = pd.read_csv('https://raw.githubusercontent.com/qtuan79/tuanfile/main/credits.csv')
```

Tổng quan về dữ liệu.

```
titles.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5850 entries, 0 to 5849
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   id                    5850 non-null   object 
 1   title                 5849 non-null   object 
 2   type                  5850 non-null   object 
 3   description           5832 non-null   object 
 4   release_year          5850 non-null   int64  
 5   age_certification     3231 non-null   object 
 6   runtime               5850 non-null   int64  
 7   genres                5850 non-null   object 
 8   production_countries  5850 non-null   object 
 9   seasons               2106 non-null   float64
 10  imdb_id               5447 non-null   object 
 11  imdb_score            5368 non-null   float64
 12  imdb_votes            5352 non-null   float64
 13  tmdb_popularity       5759 non-null   float64
 14  tmdb_score            5539 non-null   float64
dtypes: float64(5), int64(2), object(8)
memory usage: 685.7+ KB
```

Chúng ta có thể thấy trong titles có 5850 dòng dữ liệu và có 15 cột dữ liệu và các cột đều có kiểu dữ liệu phù hợp nên không cần phải chuyển đổi thêm.

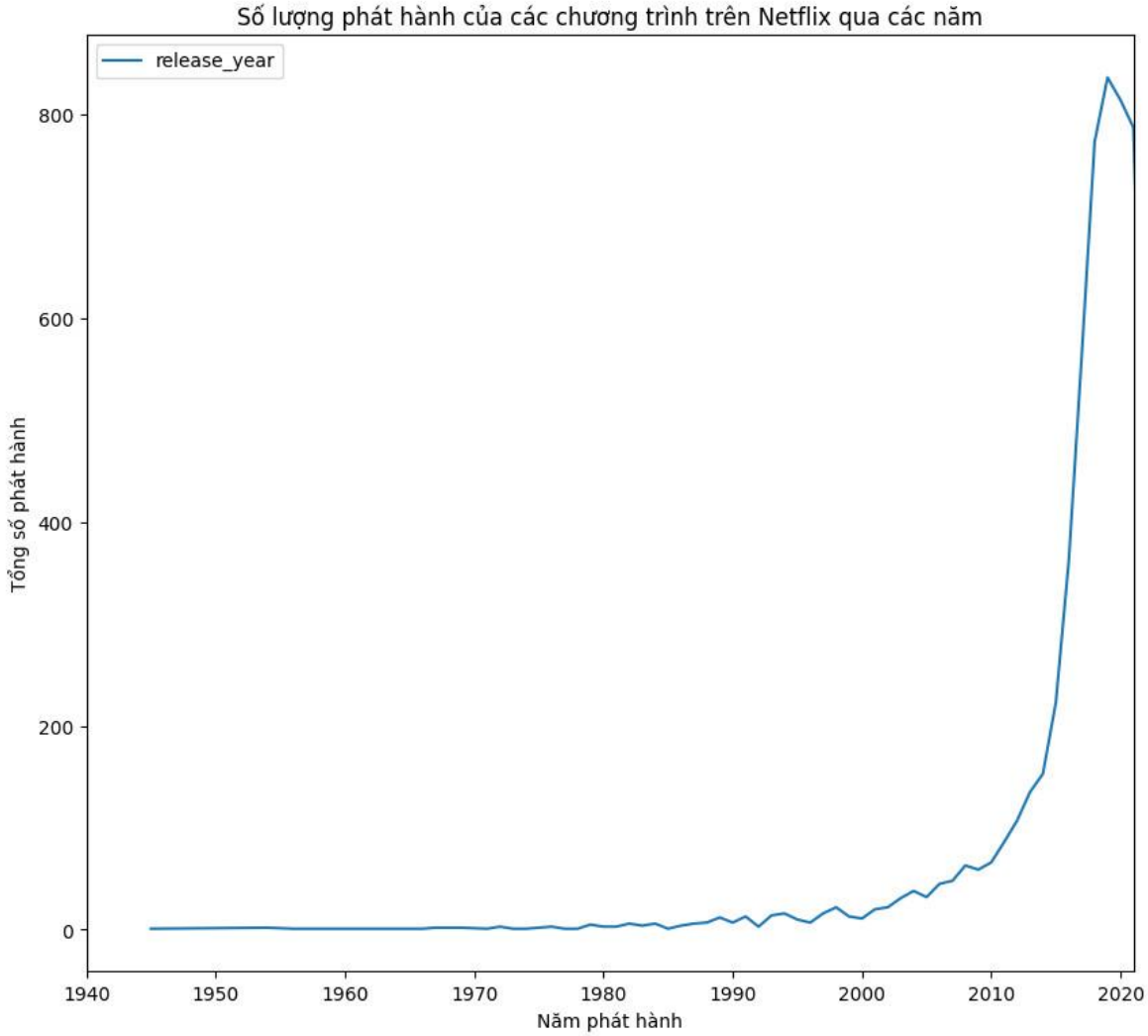
```
credits.info()
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 77801 entries, 0 to 77800
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   person_id  77801 non-null  int64
 1   id         77801 non-null  object
 2   name       77801 non-null  object
 3   character  68029 non-null  object
 4   role       77801 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.0+ MB
```

Chúng ta có thể thấy trong credits có 77801 dòng dữ liệu và có 5 cột, mỗi cột đều có kiểu dữ liệu phù hợp và không cần chuyển đổi thêm.

Phân tích.

1. *Số lượng phát hành của các chương trình trên Netflix qua các năm*

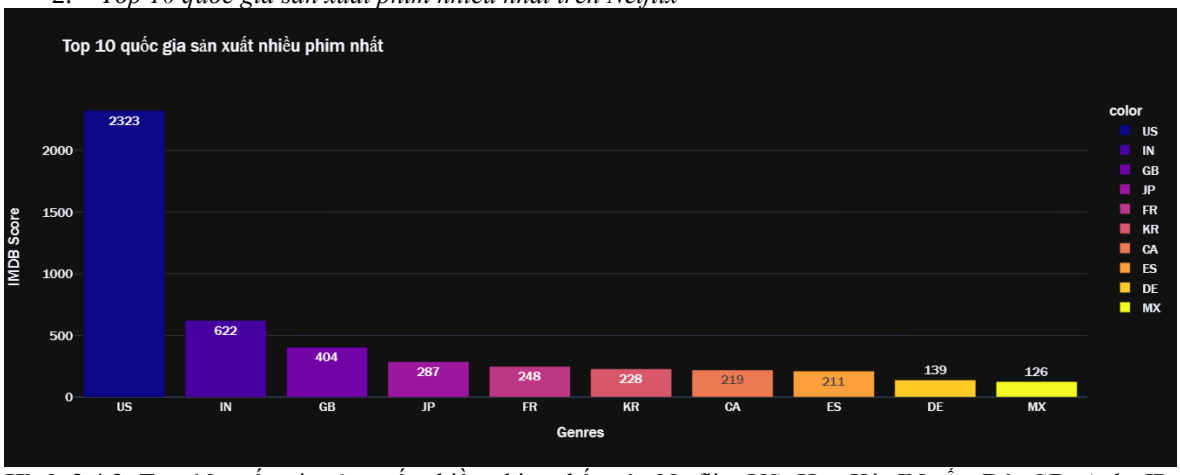


Hình 3.4.1. Số lượng phát hành của các chương trình trên Netflix qua các năm

Nhìn vào hình chúng ta có thể thấy được rằng các chương trình của Netflix từ những năm 1940 cho đến 1990 có số lượng phim phát hành ra theo từng năm rất ít hầu như chỉ là một đường thẳng, từ những năm 1990 đến 2000 số lượng phim của Netflix có sự tăng trưởng nhanh hơn nhưng không đều theo từng năm, từ năm 2000 đến năm 2010 số lượng phim trên Netflix có sự tăng trưởng khá đều vì biểu đồ có hướng đi lên, còn từ năm 2010 đến năm 2020 số lượng phim trên Netflix tăng trưởng một cách bùng nổ bằng biểu đồ đi lên một cách

dựng đứng tuy nhiên vào những năm khoảng độ từ năm 2018 đến 2020 có sự sụt giảm nhẹ về số lượng phim phát hành trên Netflix.

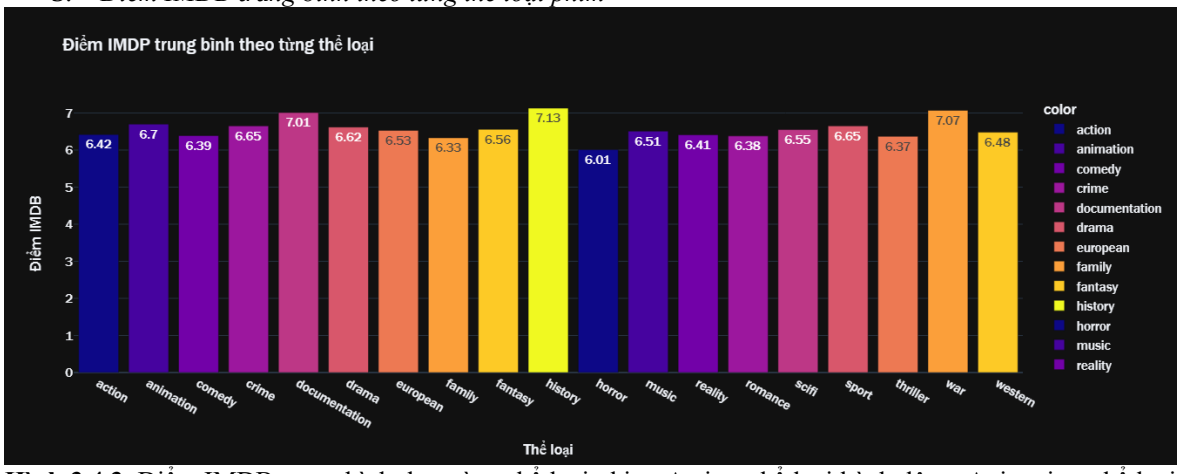
2. *Top 10 quốc gia sản xuất phim nhiều nhất trên Netflix*



Hình 3.4.2. Top 10 quốc gia sản xuất nhiều phim nhất trên Netflix. US: Hoa Kỳ, IN: Ấn Độ, GB: Anh, JP: Nhật Bản, FR: Pháp, KR: Hàn Quốc, CA: Canada, ES: Tây Ban Nha, DE: Đức, MX: Mexico

Nhìn vào biểu đồ chúng ta có thể thấy rằng nền công nghiệp phim đến từ Hoa Kỹ là rất lớn với 2323 chương trình có trên Netflix và nhiều gấp 3 lần so với số lượng chương trình đến từ Ấn Độ và gấp gần 20 lần so với số lượng chương trình đến từ Mexico (vị trí thứ 10). Số lượng chương trình đến từ các nước ngoại trừ Ấn Độ, Anh đều có số lượng chỉ từ 126 đến 287 chương trình và nếu tính tổng số chương trình của các quốc gia này cộng lại thì vẫn chưa bằng số lượng chương trình đến từ Hoa Kỳ. Như vậy chúng ta có thể kết luận rằng nền công nghiệp giải trí cụ thể là các chương trình truyền hình đến từ Hoa Kỳ rất là phát triển và có sự đóng góp rất nhiều về số lượng chương trình cho Netflix.

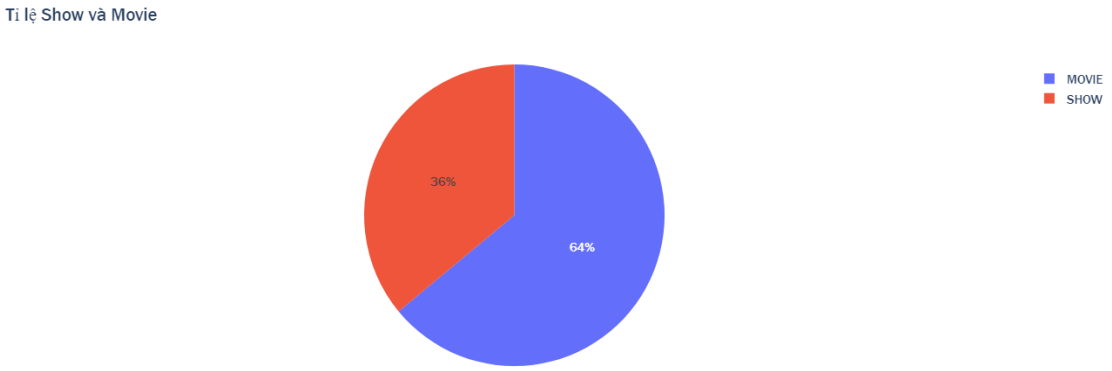
3. *Điểm IMDB trung bình theo từng thể loại phim*



Hình 3.4.3. Điểm IMDB trung bình theo từng thể loại phim. Action: thể loại hành động, Animation: thể loại phim hoạt hình, Comedy: thể loại hài kịch, Crime: thể loại tội phạm, Documentation: thể loại tài liệu, Drama: thể loại kịch tính, European: thể loại phim đến từ Châu Âu, Family: thể loại gia đình, Fantasy: thể loại kì ảo phép thuật, History: thể loại phim lịch sử, Horror: thể loại phim kinh dị, Music: thể loại nhạc phim, Reality: thể loại phim thực tế, Romance: thể loại phim lãng mạn, Scifi: thể loại khoa học viễn tưởng, Sport: thể loại phim thể thao, Thriller: thể loại phim giật gân, War: thể loại phim về chiến tranh, Western: thể loại phim viễn tây.

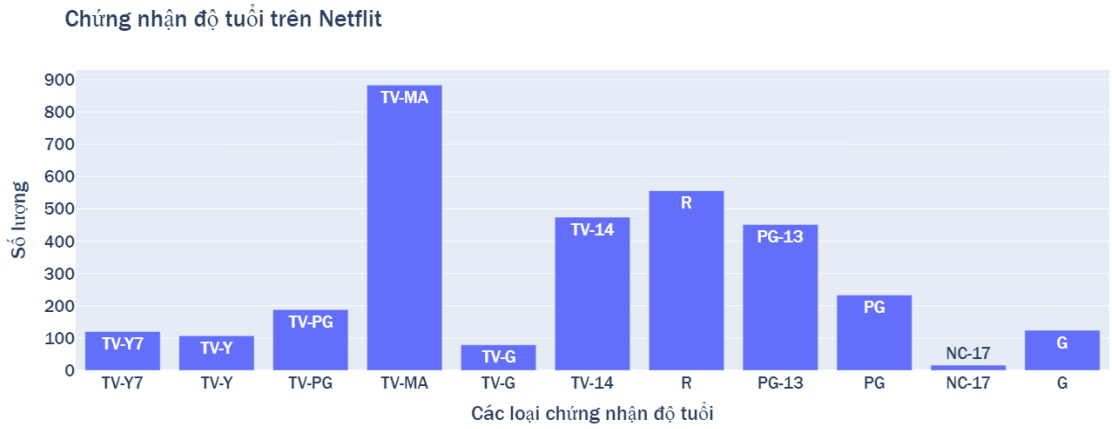
Nhìn vào biểu đồ chúng ta thấy các thể loại phim Documentation, History và War đều có điểm IMDB lớn hơn 7.0 và phim đến từ thể loại Horror có số điểm IMDB thấp nhất chỉ 6.01 vì vậy chúng ta sẽ thấy được đánh giá của các khán giả hoặc các chuyên gia về bộ phim qua đó người xem có thể dễ dàng lựa chọn những bộ phim hay để xem một cách nhanh chóng và dễ dàng nhờ vào số điểm IMDB này.

4. *Tỉ lệ giữa Show và Movie*



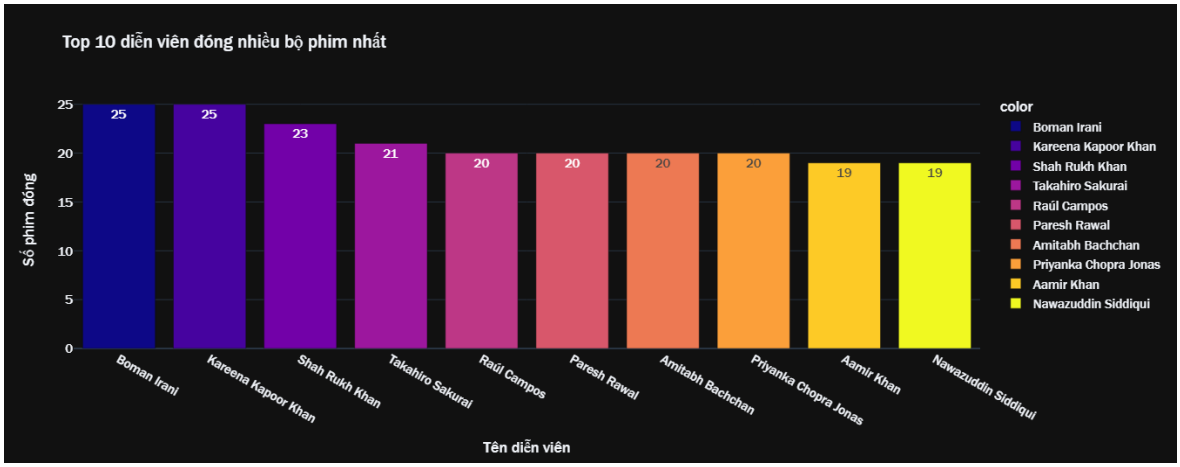
Hình 3.4.4. Tỉ lệ Show và Movie. Show: các chương trình truyền hình hoặc trực tiếp, Movie: các phim chiếu rạp.
Nhìn vào hình 3.4.4 chúng ta có thể thấy rằng số lượng chương trình Movie chiến tỉ lệ là 64% so với số lượng chương trình truyền hình Show là 36% và Movie chiếm gần 2/3 trong tổng số các chương trình có trên Netflix.

5. *Số lượng phim theo chứng nhận độ tuổi*



Hình 3.4.5. Số lượng phim theo chứng nhận độ tuổi. G (General Audiences): Phim dành cho mọi lứa tuổi, PG (Parental Guidance Suggested): Phim có thể có một số chi tiết (hình ảnh, từ ngữ) không phù hợp với trẻ nhỏ. Bố mẹ cần cân nhắc khi cho con cái xem phim, PG-13 (Parents Strongly Cautioned): Phim có một số chi tiết không phù hợp với trẻ dưới 13 tuổi, R (Restricted): Thanh thiếu niên dưới 17 tuổi không được xem phim nếu không có sự đồng ý của người lớn, NC-17 (No One 17 and Under Admitted): Phim hoàn toàn không dành cho khán giả dưới 17 tuổi, do có nhiều yếu tố gây ảnh hưởng xấu đến nhân cách, đạo đức, khuyến khích hành vi phạm tội, TV-14: Một số nội dung không phù hợp với trẻ dưới 14 tuổi, TV-G: Chương trình thích hợp cho mọi độ tuổi, TV-MA: Chương trình dành cho người lớn, TV-PG: Phim dành cho lứa tuổi từ 13 tuổi trở xuống, TV-Y: Phim dành cho trẻ từ 2 - 6 tuổi, TV-Y7: Phim dành cho lứa tuổi từ 7 trở lên.
Nhìn vào biểu đồ chúng ta có thể thấy tổng số lượng phim cho người lớn (TV-MA) là chiếm số lượng nhiều nhất với khoảng gần 900 bộ và phim hoàn toàn không dành cho lứa tuổi dưới 17 (NC-17) là chiếm số lượng ít nhất.

6. *Top 10 diễn viên đóng nhiều bộ phim nhất*



Hình 3.4.6. Top 10 diễn viên đóng nhiều bộ phim nhất.
Nhìn vào hình 3.4.6 chúng ta có thể thấy hai diễn viên góp mặt trong nhiều chương trình trên Netflix nhất đó là Boman Irani và Kareena Kapoor Khan với số lần góp mặt là 25 lần và các vị trí còn lại các diễn viên đều có mặt trong khoảng từ 19 cho đến 23 chương trình truyền hình.

4 Kết luận

Sau khi dùng các loại phân tích bằng cách gộp các dữ liệu có trong cột và biểu diễn bằng biểu đồ cột, biểu đồ miền hoặc biểu đồ tròn, chúng ta có thể khám phá được nhiều điều mới về dữ liệu mà trước đó khi đọc dữ liệu vào chỉ có mỗi dữ liệu thô. Chúng ta có thể từ việc phân tích các dữ liệu đó và đưa ra một vài đề xuất cho mình hoặc cho người khác.
Đối với các dữ liệu của Amazon cell phone chúng ta có thể thấy được rõ về điểm đánh giá về các sản phẩm của người dùng đối với các hãng điện thoại, số lượng của các dòng sản phẩm của các hãng điện thoại bán ở trên amazon cùng với các mức giá trung bình của các sản phẩm của các hãng điện thoại và tỉ lệ của các đánh giá của khách hàng với các hãng điện thoại để từ đó chúng ta có thể nhìn thấy rằng hãng nào đang làm tốt trong việc chiều theo người dùng với các tính năng và các công nghệ được áp dụng trên sản phẩm làm cho khách hàng hài lòng như thế nào. Ngoài ra chúng ta còn có thể áp dụng phương pháp word embedding kết hợp với mô hình Convolutional Neural Network của deep learning để đưa ra cảm nghĩ của khách hàng sau khi sử dụng sản phẩm.
Đối với dữ liệu amazon top 50 bestselling books 2009-2019 chúng ta có thể thấy các cuốn sách được đánh giá cao nhất cùng với các thể loại sách có mặt trên amazon và các tác giả đã nào đã viết và bán nhiều sách nhất trên amazon, chúng ta còn có thể xem được đâu là cuốn sách được các đọc giả quan tâm và để lại nhiều lượt phê bình nhất.
Cuối cùng với dữ liệu Netflix chúng ta có thể xem được quá trình phát triển của Netflix và sự bùng nổ của Netflix ở thời đại công nghệ số đã lan rộng ra hầu như gần hết toàn địa cầu và chỉ trong vòng 10 năm trở lại từ 2010 đến 2020 số lượng chương trình trên Netflix đã tăng một cách chóng mặt. Không những thế bộ dữ liệu còn cho chúng ta thấy được các nước có cho mình số lượng chương trình có mặt trên Netflix nhiều nhất và biết được các điểm trung bình IMDB theo các thể loại để tìm ra đâu là thể loại mà được đánh giá cao nhất và biết được chương trình thuộc thể loại nào là nên xem nhất. Ngoài ra chúng ta còn có thể xem được rằng các chương trình trên Netflix có các chương trình nhắm vào đối tượng nào là nhiều nhất. cùng với việc chúng ta biết được diễn viên nào góp mặt nhiều nhất trong các bộ phim.

Tài liệu tham khảo

Python Homepage. Retrieved 04 19, 2023, from http://vi.wikipedia.org/wiki/Python_(ngôn ngữ lập trình)
Python là gì? Homepage. Retrieved 04 19, 2023, from https://aws.amazon.com/vi/what-is/python/
Thống kê mô tả là gì Homepage. Retrieved 04 19, 2023, from https://luatduonggia.vn/thong-ke-mo-ta-la-gi-dac-diem-va-cac-thong-so-trong-thong-ke-mo-ta/
Điểm IMDB là gì? Homepage. Retrieved 04 19, 2023, from https://chanhtuoi.com/diem-imdb-la-gi-p4658.html
Colab là gì? Homepage. Retrieved 04 19, 2023, from https://colab.research.google.com/

pandas.DataFrame.explode Homepage. Retrieved 04 19, 2023, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.explode.html
ShaneLynn Homepage. Retrieved 04 19, 2023, from https://www.shanelynn.ie/bar-plots-in-python-using-pandas-dataframes/