Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin.

# Đồ án thực hành số 02

Trực quan hóa dữ liệu – Data Visualization.

Nhóm 16
Tháng 11, 2024.

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**


# ĐỒ ÁN THỰC HÀNH SỐ 02

**Bộ môn:** Trực quan hóa dữ liệu.

**Tên đề tài:** *"Working with Timeseries Data".*

**STT nhóm:** 16.

**Thành viên:**

1. 22120378 – Nguyễn Ngọc Khánh Trân.
2. 22120384 – Nguyễn Đình Trí.
3. 22120387 – Trần Đức Trí.
4. 22120412 – Nguyễn Anh Tường.

## Thông tin chung:

1. **Bộ môn:** Trực quan hóa dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Bùi Tiến Lên.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22_21.
5. **STT nhóm:** 16.
6. **Danh sách thành viên:**
    a. 22120378 – Nguyễn Ngọc Khánh Trân.
    b. 22120384 – Nguyễn Đình Trí.
    c. 22120387 – Trần Đức Trí.
    d. 22120412 – Nguyễn Anh Tường.
7. **Link repo github:** click here

# MỤC LỤC

**ĐỒ ÁN THỰC HÀNH SỐ 02** ..... 2

**Thông tin chung:** ..... 3

**Section 0:** *Nhận xét và đánh giá đồ án.* ..... 6

**Section 1:** *Data Collection – Thu thập dữ liệu.* ..... 8

**I. Giới thiệu thông tin:** ..... 9

**II. Giới thiệu tập dữ liệu:** ..... 9

**III. Diễn giải tập dữ liệu:** ..... 9

**IV. Quá trình thu thập dữ liệu:** ..... 10

**Section 2:** *Tiền xử lý dữ liệu.* ..... 11

**I. File cần chạy:** ..... 12

**II. Diễn giải quá trình:** ..... 12

**III. Kết luận:** ..... 13

**Section 3:** *Khai thác dữ liệu.* ..... 14

**I. Phân tích lịch sử của giá cổ phiếu:** ..... 15

**1. Sự thay đổi giá cổ phiếu từ trước đến nay:** ..... 15

**2. Phân tích sự biến động của giá cổ phiếu qua các giai đoạn** ..... 17

**3. Phân tích sự biến động của giá cổ phiếu qua các tháng trong năm** ..... 19

**4. Tìm hiểu sự chênh lệch khối lượng giao dịch trung bình giữa 2 thời điểm - giá cổ phiếu biến động lớn và giá cổ phiếu biến động nhỏ qua các tháng trong năm** ..... 21

**II. Phân tích chi tiết những biến động ngắn hạn của cổ phiếu Google:** ..... 22

**1. Trong ba tháng gần nhất, khối lượng giao dịch trung bình nằm ở đầu tuần hay nằm ở những ngày gần cuối tuần là nhiều nhất?** ..... 22

**2. Biến động giá cổ phiếu trong ngày của 3 tháng gần nhất có cho thấy sự đặc trưng nào tiêu biểu không? (Ví dụ giá mở cửa thường thấp hơn giá đóng cửa)?** ..... 24

**3. Có khi nào giá đóng cửa xấp xỉ gần với giá cao nhất hoặc giá thấp nhất trong ba tháng gần nhất hay không? Nếu có, điều này mang ý nghĩa gì đối với doanh nghiệp?** ..... 25

**4. Mối quan hệ giữa khối lượng giao dịch và biên độ dao động (giá cao nhất - giá thấp nhất) hằng ngày trong 3 tháng gần nhất là gì?** ..... 27

**Section 4:** *Xây dựng mô hình và đánh giá.* ..... 29

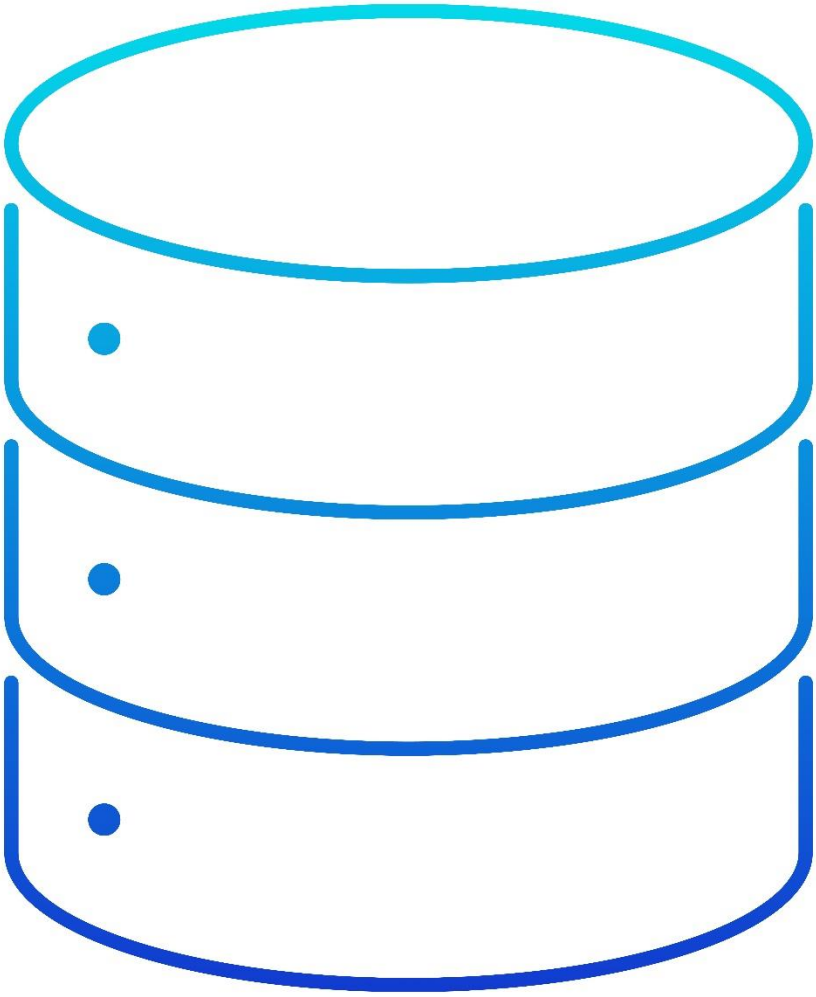
**I. Giới thiệu mô hình và thuật toán sử dụng:** ..... 30

**II. Cài đặt mô hình và đánh giá mô hình:** ..... 31

- **Kết luận:** .....33
- **Giải pháp thay đổi:** .....33

**III. Một số hạn chế của mô hình:** .....35

## **Section 0:** *Nhận xét và đánh giá đồ án.*



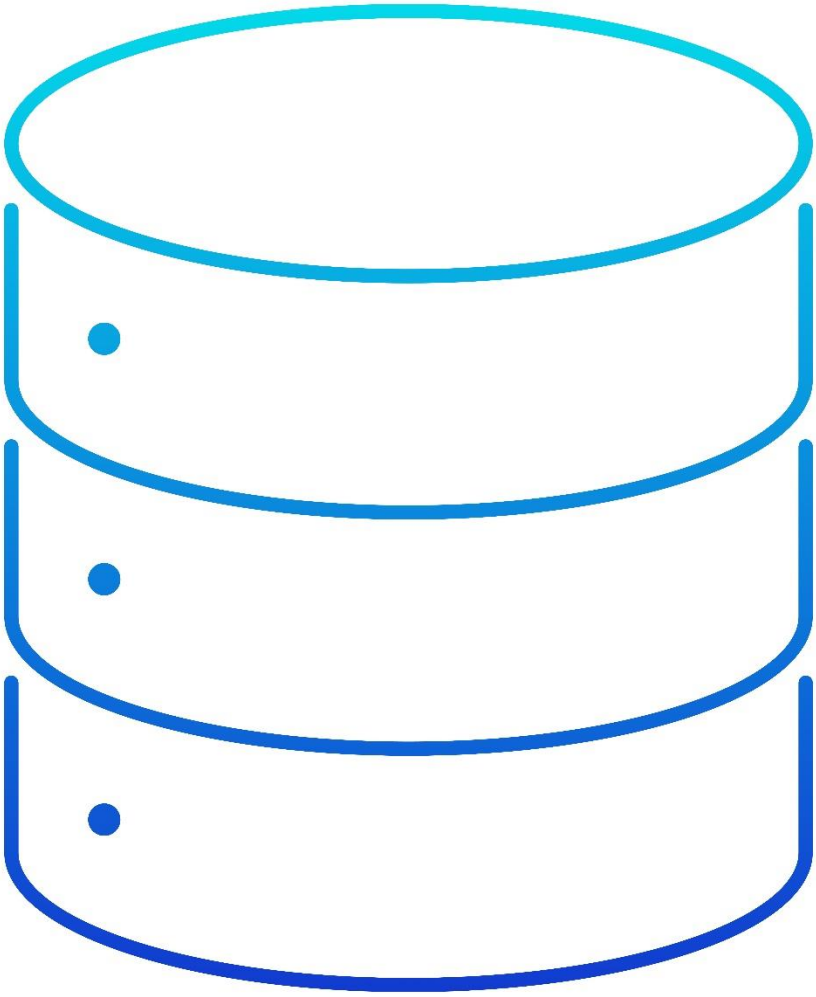
## I. Phân công công việc và tiến độ hoàn thành:

| Công việc | Người thực hiện | Tiến độ | % tổng |
|---|---|---|---|
| Thu thập dữ liệu | Nguyễn Anh Tường | 100% | 5% |
| Tiền xử lý dữ liệu | Nguyễn Anh Tường | 100% | |
| Chọn, diễn giải và trực quan hóa các trường và mối quan hệ ẩn của chúng. | Nguyễn Ngọc Khánh Trân | 33.33% | 50% |
| | Trần Đức Trí | 33.33% | |
| | Nguyễn Đình Trí | 33.33% | |
| Rút ra ý nghĩa logic đằng sau mỗi dữ liệu được trực quan hóa | Nguyễn Ngọc Khánh Trân | 33.33% | 20% |
| | Trần Đức Trí | 33.33% | |
| | Nguyễn Đình Trí | 33.33% | |
| Hãy xem xét nhiều mối quan hệ và nhiều quan điểm khác nhau | Nguyễn Ngọc Khánh Trân | 33.33% | 10% |
| | Trần Đức Trí | 33.33% | |
| | Nguyễn Đình Trí | 33.33% | |
| Báo cáo trình bày theo định dạng và bố cục hợp lý, rõ ràng. | Nguyễn Ngọc Khánh Trân | 20% | 15% |
| | Trần Đức Trí | 20% | |
| | Nguyễn Đình Trí | 20% | |
| | Nguyễn Anh Tường | 40% | |
| Sử dụng các mô hình học máy cơ bản. | Nguyễn Anh Tường | 100% | 5% |
| Hiểu biết chung về mã nguồn đã gửi | Nguyễn Anh Tường | 100% | 5% |

## II. Tổng hợp tiến độ của đồ án:

**110%**

## **Section 1:** *Data Collection – Thu thập dữ liệu.*



## I. Giới thiệu thông tin:

**Nguồn**: kaggle.

**Tên tập dữ liệu**: Google Stock Price ( All time ).

**Bản quyền sử dụng**: không có.

**Mô tả công ty Alphabet, Inc:**

Là một công ty mẹ, tham gia vào hoạt động mua lại và vận hành các công ty khác nhau.

Công ty hoạt động thông qua các phân khúc Google và Other Bets.

a. Phân khúc Google bao gồm các sản phẩm Internet chính như quảng cáo, Android, Chrome, phần cứng, Google Cloud, Google Maps, Google Play, Tìm kiếm và YouTube.
b. Phân khúc Other Bets bao gồm các doanh nghiệp như Access, Calico, CapitalG, GV, Verily, Waymo và X. Công ty được thành lập bởi Lawrence E. Page và Sergey Mikhaylovich Brin vào ngày 2 tháng 10 năm 2015 và có trụ sở chính tại Mountain View, CA.

## II. Giới thiệu tập dữ liệu:

Đây là tập dữ liệu ghi lại thông tin cổ phiếu của google theo thời gian từ ngày 19-08-2004 đến ngày 10-11-2021.

Tập dữ liệu bao gồm 7 trường dữ liệu ( cột ) và 4317 hàng.

## III. Diễn giải tập dữ liệu:

1. **Date –** Ngày **:**
    a. **Ý nghĩa:** Ngày giao dịch.
    b. **Kiểu dữ liệu:** Datetime.
2. **High –** Giá trần**:**
    a. **Ý nghĩa:** Giá cao nhất mà một giao dịch trong ngày đó có thể khớp lệnh.
    b. **Kiểu dữ liệu:** Float.
3. **Low –** Giá sàn**:**
    a. **Ý nghĩa:** Giá thấp nhất mà một giao dịch trong ngày đó có thể khớp lệnh.
    b. **Kiểu dữ liệu:** Float.
4. **Open –** Giá mở cửa**:**
    a. **Ý nghĩa:** Giá đầu tiên khi mở phiên giao dịch.
    b. **Kiểu dữ liệu:** Float.

5. **Close –** Giá đóng cửa**:**
    a. **Ý nghĩa:** Giá giá cuối cùng sau khi chốt cuối phiên giao dịch.
    b. **Kiểu dữ liệu:** Float.
6. **Volume –** Khối lượng giao dịch**:**
    a. **Ý nghĩa:** Tổng số cổ phiếu được giao dịch trong ngày đó.
    b. **Kiểu dữ liệu:** Float.
7. **Adj Close –** Giá đóng cửa điều chỉnh**:**
    a. **Ý nghĩa**: Thể hiện giá trị thực tế của một cổ phiếu vào cuối ngày giao dịch, được điều chỉnh để phản ánh các thay đổi trong cấu trúc vốn hoặc chia cổ tức.
    b. **Kiểu dữ liệu:** Float.

## IV. Quá trình thu thập dữ liệu:

### 1. File cần chạy:

Source\DataCollection&DataPreProcessing\DataCollection.ipynb

### 2. Diễn giải:

**Bước 1:** Tải tập dữ liệu bằng kaggle hub.

**Link tập dữ liệu: Click**

**Bước 2:** Copy tập dữ liệu từ thư mục kaggle trên máy chủ sang thư mục làm việc.

**Kết quả:** Thư mục dataset sẽ tồn tại file **google.csv**.

### 3. Lưu ý:

Cần phải tải kagglehub trước khi sử dụng chương trình – chi tiết trong hướng dẫn ở phần README.md

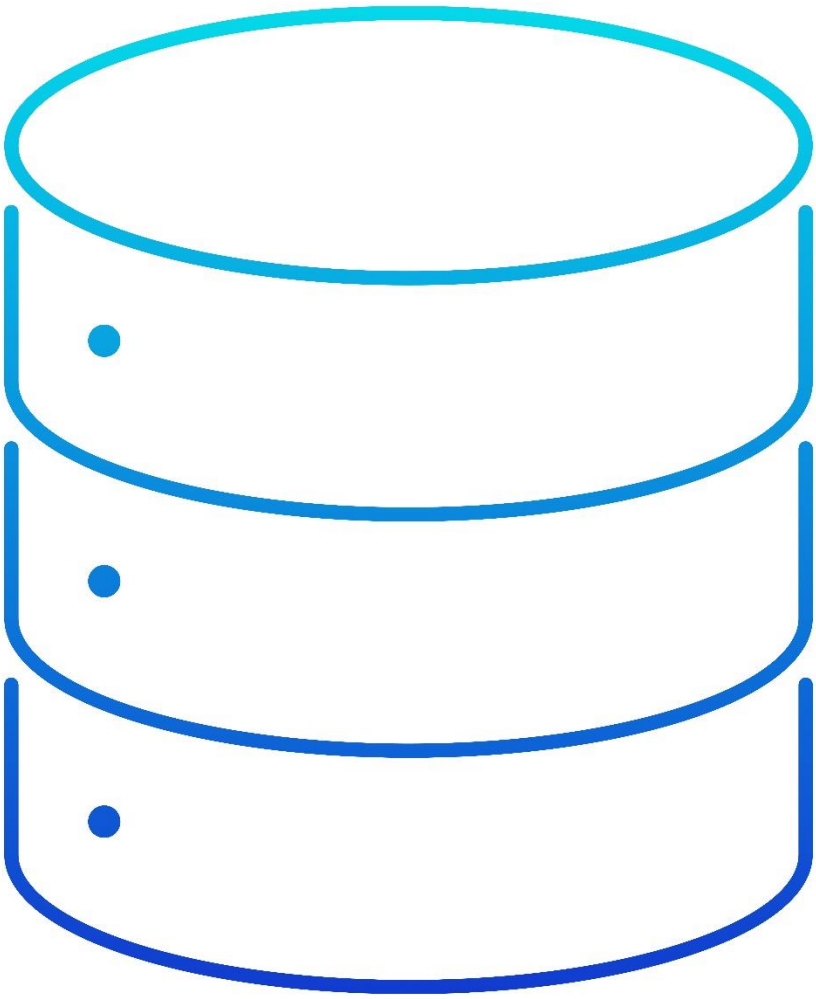
## V. Cách sử dụng và độ phù hợp cho việc giáo dục:

Đây là tập dữ liệu được thu thập bởi chính Yahoo Finance.

Yahoo! Finance là tài sản truyền thông là một phần của Yahoo!, kể từ năm 2017, thuộc sở hữu của Verizon Media. Nó cung cấp tin tức, dữ liệu và bình luận tài chính bao gồm báo giá cổ phiếu, thông cáo báo chí, báo cáo tài chính và nội dung gốc.

Vì đây là một tập dữ liệu không có bản quyền và hợp thức hóa bởi Kaggle cho nên độ phù hợp với mục đích giáo dục là rất lớn, ta có thể tự tin sử dụng.

## **Section 2:** *Tiền xử lý dữ liệu.*



## I. File cần chạy:

Source\DataCollection&DataPreProcessing\DataPreProcessing.ipynb

## II. Diễn giải quá trình:

**Bước 1:** Kiểm tra kiểu dữ liệu của các cột.

```
Date          object
High         float64
Low          float64
Open         float64
Close        float64
Volume       float64
Adj Close    float64
dtype: object
```

**Nhận xét:** Ở đây ta thấy các cột đã đúng kiểu dữ liệu trừ cột Date. Do đó ta phải chuyển cột Date này kiểu dữ liệu datatime.

**Bước 2:** Kiểm tra dữ liệu bị thiếu của các cột.

```
Column: Date - Missing Ratio: 0.0
Column: High - Missing Ratio: 0.0
Column: Low - Missing Ratio: 0.0
Column: Open - Missing Ratio: 0.0
Column: Close - Missing Ratio: 0.0
Column: Volume - Missing Ratio: 0.0
Column: Adj Close - Missing Ratio: 0.0
```

**Nhận xét:** Các cột dữ liệu đều có % thiếu sót dữ liệu là 0.0

**Bước 3:** Kiểm tra và xóa bỏ các dòng bị trùng lập.

```
Duplicated rows: 0
```

**Nhận xét:** Không có dòng nào là giống nhau 100% trong tập dữ liệu này.

**Bước 4:** Kiểm tra điều kiện tồn tại của các dữ liệu số học.

```
Number of elements <0 in columns:
{'High': 0, 'Low': 0, 'Open': 0, 'Close': 0, 'Volume': 0, 'Adj Close': 0}
number of rows with high less than low:
0
Rows need to drop:
None
```

1. Do các cột *[ high, low, open, close, volume, adj close ]* là các cột đại lượng chỉ giá trị tiền tệ hoặc khối lượng nên vì thế điều kiện tiên quyết là phải lớn hơn 0.0

**Nhận xét:** Trong các cột đã xét không có hàng nào vi phạm điều kiện >= 0.

2. Ý nghĩa của cột High và Low lần lượt là giá trần và sàn cho nên ta có thêm điều kiện là High phải lớn hơn hoặc bằng Low.

**Nhận xét:** Trong các cột High, Low không có hàng nào vi phạm điều kiện High > Low.

3. Vì các điều kiện đã thỏa và không có hàng nào bị vi phạm nên số lượng hàng cần xóa bỏ là 0.

**Bước 5:** Kiểm tra điều kiện tồn lại của các kiểu dữ liệu không phải số.

Vì ở đây chỉ có cột Date là có kiểu dữ liệu khác số cho nên ta sẽ xét cột này.

**Mã code:** ta sẽ sử dụng tham số **errors = 'coerce'** trong hàm **to_datetime** của thư viện **pandas**.

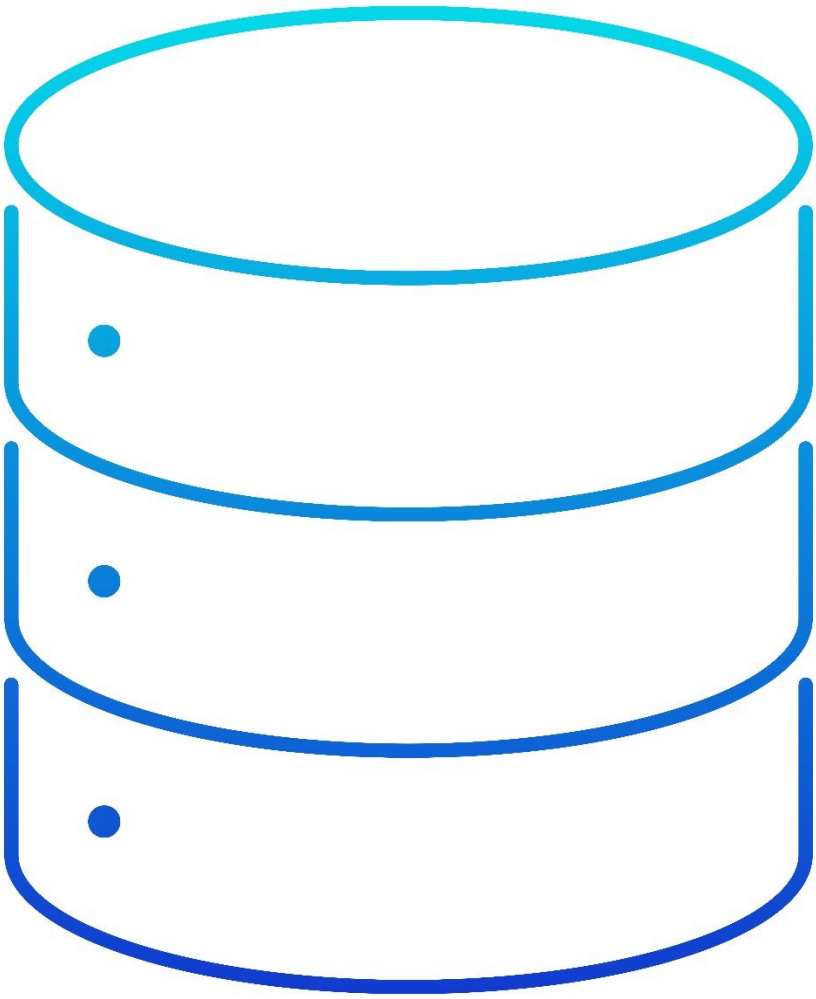
| Date | High | Low | Open | Close | Volume | Adj Close | Valid_Date |
|---|---|---|---|---|---|---|---|

Ta có thể thấy danh sách các hàng cần xóa rỗng, do đó không có dữ liệu ngày bị sai.

## III. Kết luận:

Như vậy tập dữ liệu đã được xử lý xong.

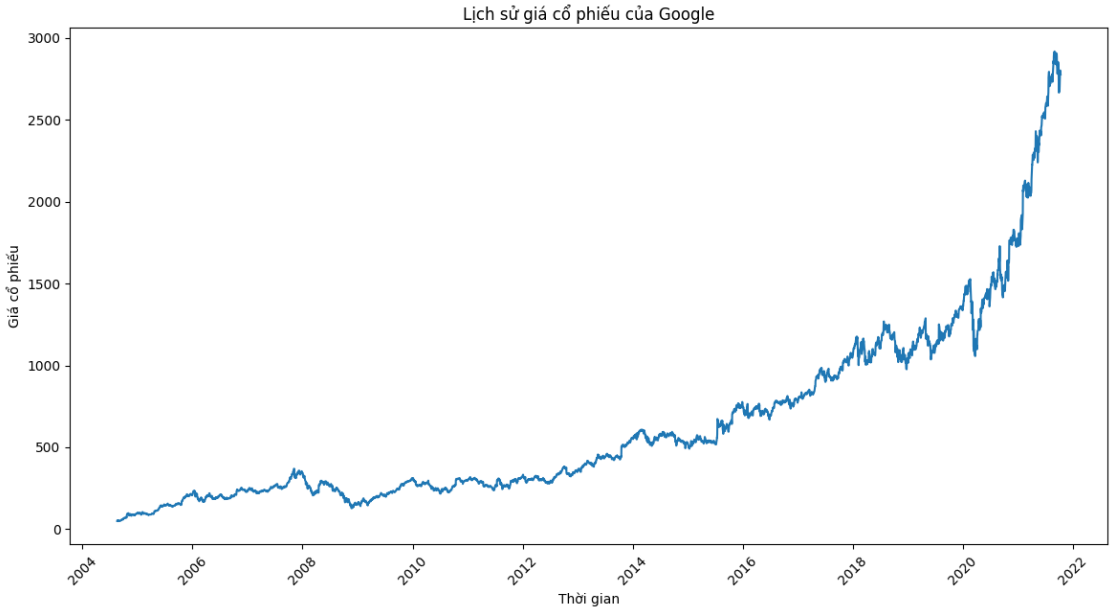
# **Section 3:** *Khai thác dữ liệu.*



# I. Phân tích lịch sử của giá cổ phiếu:

## 1. Sự thay đổi giá cổ phiếu từ trước đến nay:

Chúng ta cùng quan sát qua hình ảnh trực quan dưới đây:




Đây là biểu đồ đường thể hiện giá cổ phiếu qua từng ngày từ năm 2004 đến năm 2021.

- **Tại sao sử dụng biểu đồ đường lại phù hợp cho việc thể hiện xu hướng giá cổ phiếu qua từng ngày từ năm 2004 đến năm 2021?**
  - Biểu đồ đường có thể biểu diễn xu hướng dài hạn rõ ràng, kết nối các giá trị liên tiếp theo thời gian. Chính vì thế mà người dùng có thể dễ dàng trực quan xu hướng tăng hay giảm của giá cổ phiếu qua các năm.
- **Nhận xét và kết luận:**

***1.1. Xu hướng tăng mạnh mẽ theo thời gian:***

- Từ năm 2004 đến khoảng 2015:

+ Giá cổ phiếu tăng ổn định, nhưng mức độ tăng không quá lớn.

+ Thị trường chủ yếu trong giai đoạn tăng trưởng đều, phù hợp với sự phát triển của Google ở giai đoạn đầu với doanh thu chủ yếu từ quảng cáo trực tuyến

- Từ 2016 đến 2022:

+ Giá cổ phiếu bắt đầu tăng nhanh hơn, đặc biệt sau năm 2020, với mức tăng đột biến rõ rệt.

+ Đây là giai đoạn Google (Alphabet) mở rộng sang nhiều lĩnh vực khác như dịch vụ đám mây (Google Cloud), trí tuệ nhân tạo, và doanh thu tăng mạnh từ YouTube.

***1.2. Giai đoạn biến động lớn:***

- Khủng hoảng tài chính 2008-2009:

+ Giá cổ phiếu giảm mạnh, thể hiện qua sự chững lại rõ rệt trong xu hướng tăng.

+ Nguyên nhân chủ yếu do tác động tiêu cực của khủng hoảng kinh tế toàn cầu.

- 2020 (COVID-19):

+ Có một đợt giảm nhẹ trong giá vào đầu năm 2020, nhưng sau đó phục hồi và tăng trưởng rất mạnh.

+ Nguyên nhân:

- Đại dịch thúc đẩy sự chuyển đổi số, tăng nhu cầu sử dụng các dịch vụ của Google như tìm kiếm trực tuyến, đám mây, YouTube.
- Kỳ vọng nhà đầu tư về việc Alphabet sẽ hưởng lợi từ xu hướng này.

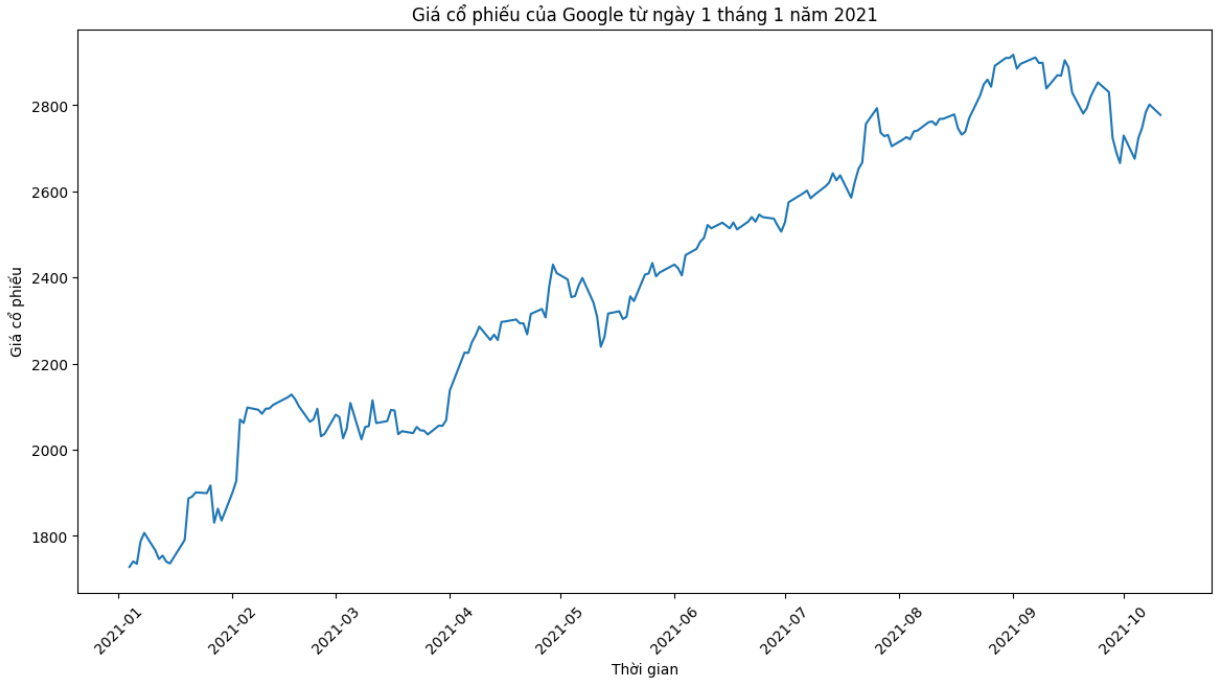
***1.3. Giai đoạn tăng trưởng vượt bậc (2020-2022):***

- Giá cổ phiếu tăng mạnh nhất trong giai đoạn này, đạt đỉnh vào cuối năm 2021, đầu năm 2022.
- Nguyên nhân:

+ Lợi nhuận kỷ lục từ mảng quảng cáo trực tuyến và sự phát triển của các dịch vụ mới.

+ Tâm lý lạc quan của nhà đầu tư với cổ phiếu công nghệ trong bối cảnh thế giới chuyển đổi số.

***Ta cùng quan sát thêm một biểu đồ đường nữa về xu hướng của giá cổ phiếu trong năm 2021 dưới đây:***




- Ở biểu đồ trên, ta có thể thấy giá cổ phiếu thời gian gần đây vẫn có xu hướng tăng từ đầu năm 2021 đến đầu tháng 9, sau đó có xu hướng giảm nhưng không đáng kể, và khả năng sẽ tiếp tục tăng theo dự đoán trước đó.

**<u>Kết luận:</u>**

- Biểu đồ cho thấy xu hướng tăng trưởng dài hạn của giá cổ phiếu Google, với các giai đoạn tăng trưởng mạnh rõ rệt sau các sự kiện quan trọng như:

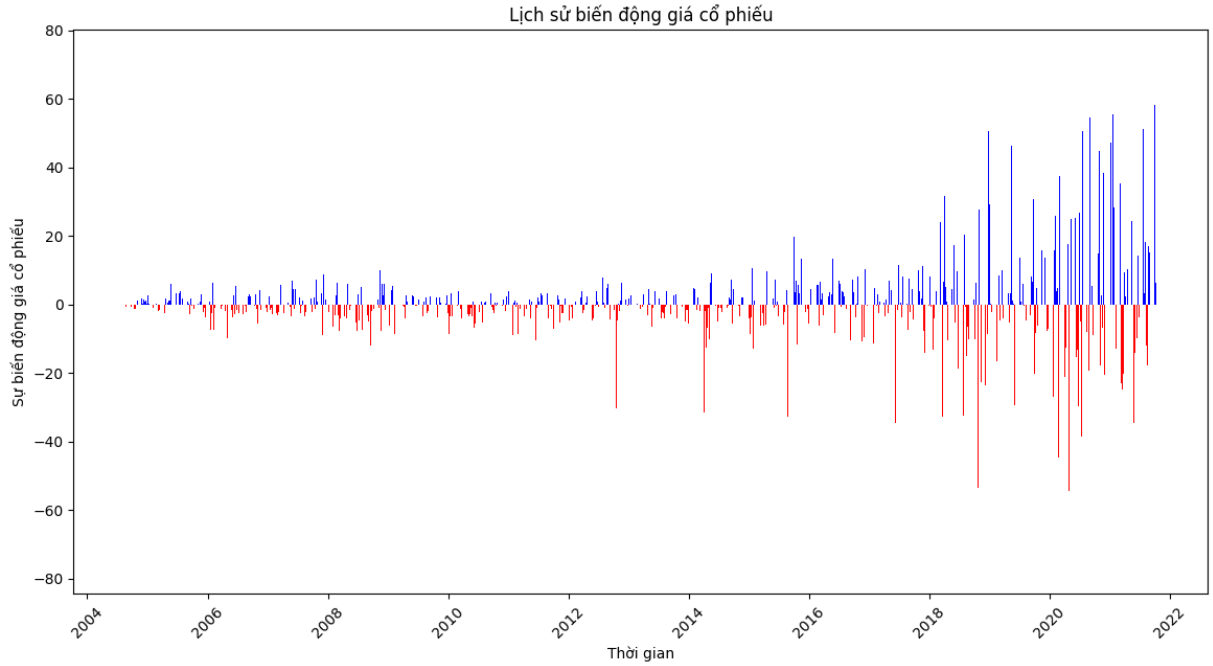
    + Sự mở rộng lĩnh vực kinh doanh của Alphabet.

    + Tác động tích cực của chuyển đổi số do COVID-19.

- Tuy nhiên, cổ phiếu Google cũng chịu ảnh hưởng từ các biến động kinh tế toàn cầu như khủng hoảng tài chính 2008 và giai đoạn đầu của đại dịch COVID-19. Nhìn chung, đây là một cổ phiếu có xu hướng tăng trưởng vượt bậc trong dài hạn.

**2. Phân tích sự biến động của giá cổ phiếu qua các giai đoạn**

Dưới đây là biểu đồ thể hiện sự biến động của giá cổ phiếu, với trên mức 0 là giá cổ phiếu tăng và dưới mức 0 là giá cổ phiếu giảm. Mức biến động của giá cổ phiếu được tính bằng giá đóng – giá mở (Close - Open).




- **Tại sao sử dụng biểu đồ cột đối xứng lại phù hợp cho việc miêu tả sự biến động của cổ phiếu (với trên mức 0 là giá cổ phiếu tăng và dưới mức tăng là giá cổ phiếu giảm)?**
    - Biểu đồ cột đối xứng giúp người xem dễ dàng nhận biết được sự chênh lệch giữa giá mở cửa và giá đóng cửa. Nếu giá đóng cửa cao hơn giá mở cửa, cột sẽ nằm ở trên trục 0 (màu xanh), và ngược lại, cột sẽ nằm ở dưới trục 0 (màu đỏ).
    - Mặt khác, người xem còn có thể so sánh sự biến động tăng giảm rõ ràng theo từng ngày trong cùng một biểu đồ.

- **Nhận xét và kết luận:**
    - Biến động tăng dần theo thời gian:

+ Từ giai đoạn 2004 đến khoảng 2017, mức độ biến động (cả tăng và giảm) của giá cổ phiếu Google tương đối thấp và ổn định.

+ Từ 2018 trở đi, đặc biệt vào giai đoạn 2020-2021, biến động bắt đầu tăng mạnh với biên độ lớn hơn (cả giá tăng mạnh và giảm mạnh).

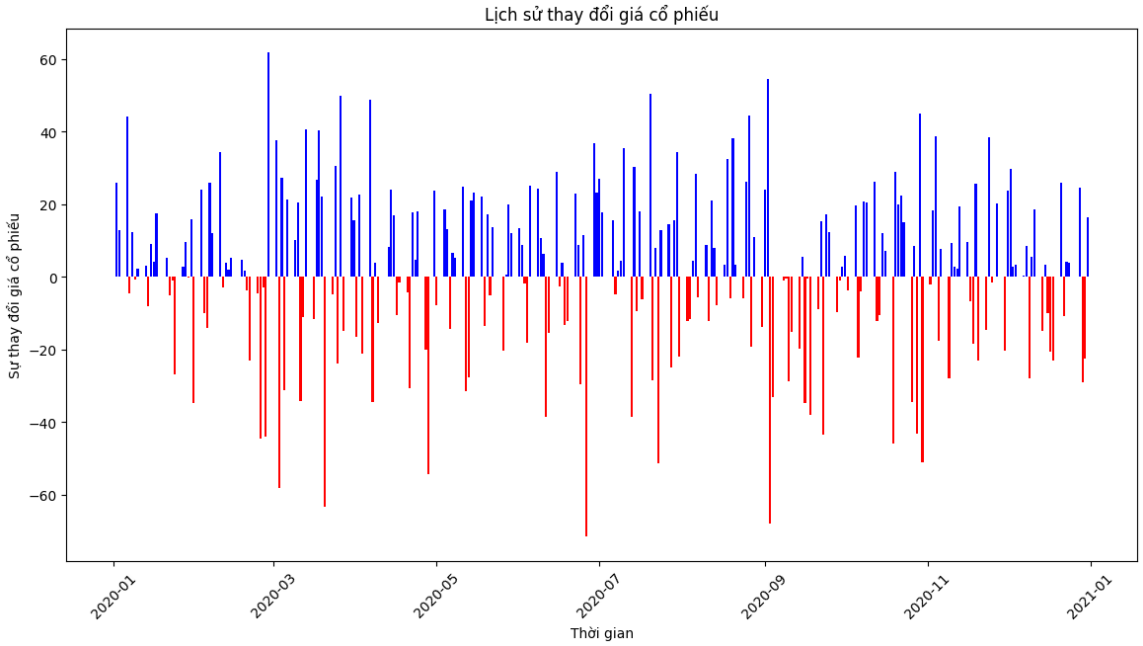
- Xu hướng dương chiếm ưu thế:

+ Các thanh màu xanh (biến động tăng) xuất hiện nhiều hơn, đặc biệt từ 2020 trở đi. Điều này cho thấy xu hướng tích cực của giá cổ phiếu.

- Những giai đoạn giảm mạnh:

+ Có một số thanh đỏ dài, đặc biệt vào giai đoạn khủng hoảng (có thể liên quan đến các sự kiện lớn như khủng hoảng tài chính 2008 hoặc COVID-19 2020). Điều này phản ánh mức giảm đột ngột của giá cổ phiếu trong các sự kiện lớn.

***Giai đoạn biến động mạnh mẽ của giá cổ phiếu (2020) do COVID-19***

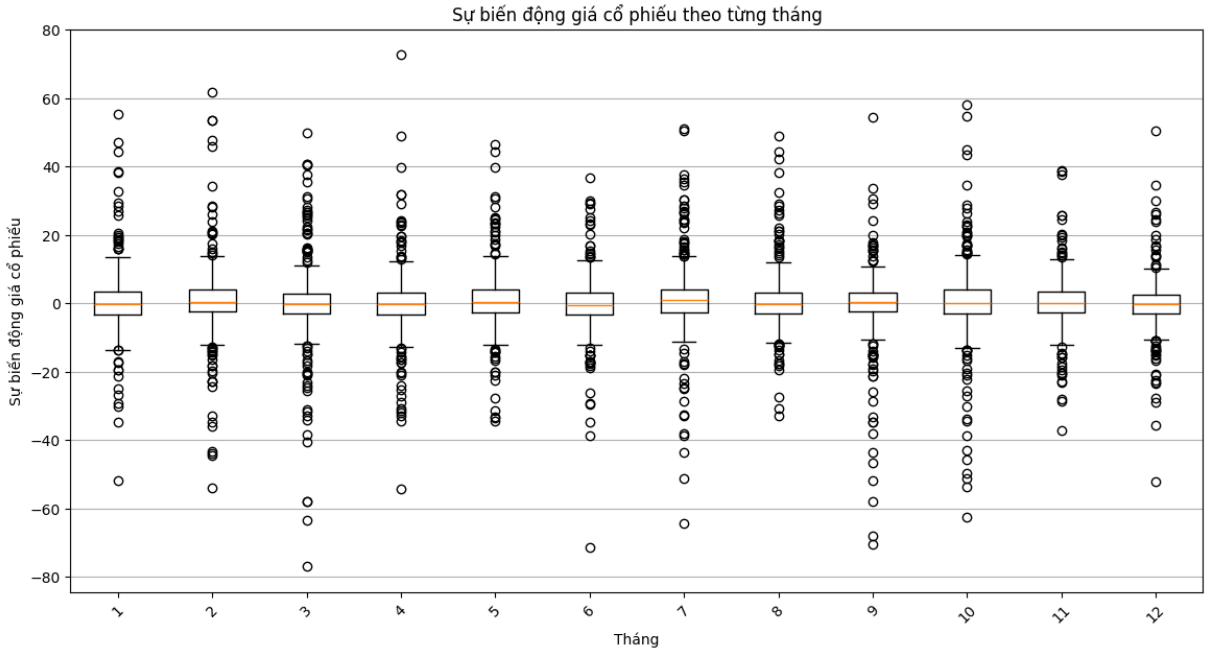



**<u>Kết luận:</u>**

- Các biểu đồ cho thấy sự biến động ngày càng lớn của giá cổ phiếu Google theo thời gian. Sự tăng trưởng của công ty cùng với các sự kiện toàn cầu như khủng hoảng tài chính và đại dịch COVID-19 là những yếu tố chính tác động đến mức độ biến động. Biểu đồ cũng cho thấy xu hướng tổng thể tích cực với nhiều thanh xanh hơn thanh đỏ.

**3. Phân tích sự biến động của giá cổ phiếu qua các tháng trong năm**

Ta cùng xem xét qua biểu đồ box plot thể hiện phân bố sự biến động giá cổ phiếu qua các tháng trong năm:




- **Tại sao sử dụng biểu đồ box plot lại phù hợp trong việc thể hiện phân bố sự biến động giá cổ phiếu qua các tháng trong năm?**
    - Biểu đồ box plot trực quan sự phân bố giá cổ phiếu theo các tháng trong năm theo hình thức: chia dữ liệu thành các phần và hiển thị các điểm thống kê quan trọng (như sự biến động trung bình của các tháng trong năm). Việc này rất hữu ích trong việc tổng quan biến động giá cổ phiếu và nhận biết được các tháng có sự biến động lớn hay không.
    - Mặt khác, biểu đồ box plot còn nhận diện được các giá trị ngoại lai, từ đó đưa ra những góc nhìn để quản lý rủi ro cho những biến động bất thường này.
- **Nhận xét và kết luận:**

Từ biểu đồ trên, ta có thể quan sát được các tiêu chí sau:

***3.1. Phân phối biến động theo tháng:***

- Hộp (box) trong biểu đồ cho thấy phân phối của sự biến động giá trong từng tháng. Vị trí của median (đường ngang trong hộp) gần mức 0 cho thấy sự biến động trung bình trong tháng gần như bằng 0, tức là giá cổ phiếu có sự thay đổi lớn nhưng không có xu hướng tăng hay giảm mạnh vào hầu hết các tháng.
- Whiskers cho thấy phạm vi biến động trong tháng (từ giá trị cực tiểu đến cực đại trong phạm vi bình thường). Nếu một phần tử nằm ngoài whiskers (được gọi là outliers), đó có thể là các thay đổi giá bất thường trong một tháng nào đó.

***3.2. Sự biến động âm và dương:***

- Sự biến động theo tháng thể hiện các giá trị âm (giảm giá) và dương (tăng giá), với các tháng có giá trị âm (biến động giảm giá) rơi vào khoảng dưới 0, và các tháng có giá trị dương (biến động tăng giá) nằm trên 0.
- Nhìn vào biểu đồ, ta có thể thấy một số tháng có mức giảm giá mạnh, ví dụ như các tháng Mar, Oct, với các outliers ở mức thấp, cho thấy sự sụt giảm mạnh hơn so với các tháng khác.

***3.3. Biến động mạnh trong các tháng:***

- Một số tháng như Jan, Feb, and Dec có sự biến động khá mạnh (có nhiều outliers ngoài phạm vi whiskers), cho thấy trong các tháng này cổ phiếu có thể gặp phải các thay đổi đột ngột trong giá trị, có thể do các yếu tố ngoại cảnh, thông tin thị trường, hoặc các sự kiện quan trọng.

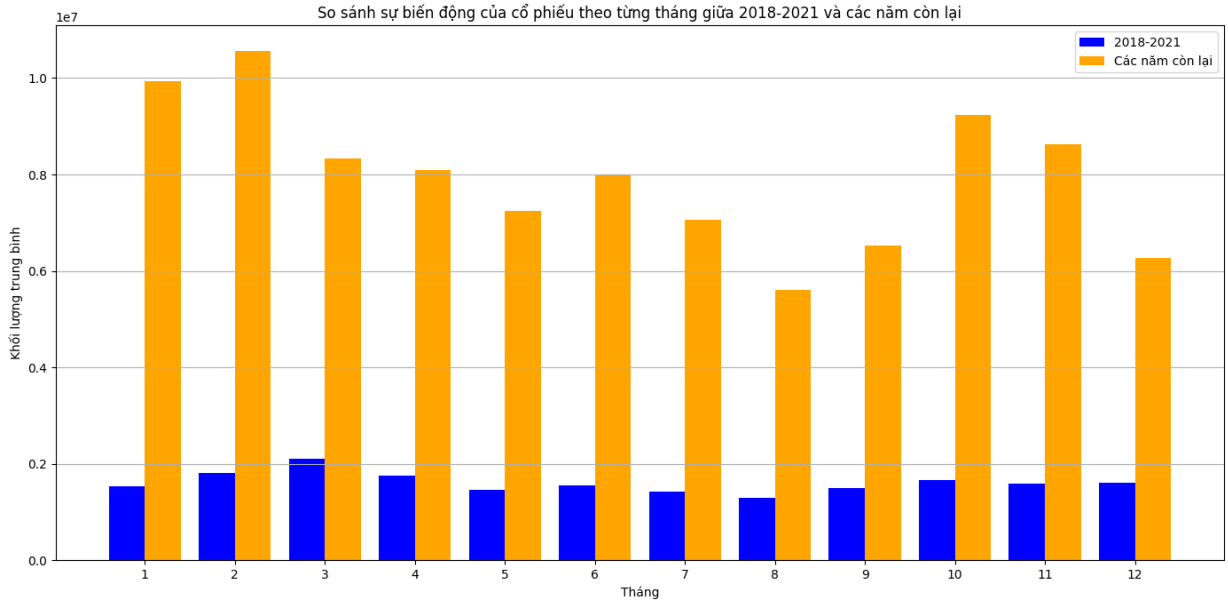
***3.4. Các quý của cổ phiếu:***

- Quý I (Jan, Feb, Mar): Sự biến động khá mạnh, với các outliers lớn, đặc biệt là tháng Mar, cho thấy sự sụt giảm lớn trong quý đầu năm.
- Quý II (Apr, May, Jun): Biến động trong khoảng giữa và không có sự thay đổi quá mạnh mẽ, tuy nhiên vẫn có vài tháng có outliers.
- Quý III (Jul, Aug, Sep): Có sự ổn định hơn với sự biến động không lớn, tuy nhiên tháng Sep có vẻ có sự thay đổi mạnh mẽ hơn.
- Quý IV (Oct, Nov, Dec): Biến động mạnh hơn ở các tháng Oct và Dec, đặc biệt Oct có outliers với sự giảm giá đột ngột.

**Kết luận:**

- Biểu đồ cho thấy sự biến động của cổ phiếu trong năm khá lớn, đặc biệt là trong các tháng có outliers. Các tháng như Mar, Oct, Dec cho thấy mức giảm giá đáng kể, trong khi các tháng khác thể hiện sự ổn định hơn.
- Phân tích này giúp ta hiểu rõ hơn về sự thay đổi của giá cổ phiếu theo mùa và có thể cung cấp thông tin hữu ích để quyết định thời điểm đầu tư hoặc giao dịch.

**4. Tìm hiểu sự chênh lệch khối lượng giao dịch trung bình giữa 2 thời điểm - giá cổ phiếu biến động lớn và giá cổ phiếu biến động nhỏ qua các tháng trong năm**




- **Tại sao sử dụng biểu đồ cột ghép để thể hiện sự chênh lệch khối lượng giao dịch trung bình giữa 2 thời diểm - giá cổ phiếu biến động lớn và giá cổ phiếu biến động nhỏ qua các tháng trong năm là phù hợp?**
    - Biểu đồ cột ghép cho phép người xem so sánh hai nhóm dữ liệu song song một cách rõ ràng. Khi đó, người xem có thể so sánh trực tiếp 2 thời điểm trên với nhau trong các tháng.
    - Mặt khác, người xem còn có thể nắm được xu hướng tăng giảm của khối lượng giao dịch trung bình của cả 2 thời điểm theo thứ tự các tháng. Từ đó, người xem dễ dàng đưa ra góc nhìn của bản thân về xu hướng và sự chênh lệch giữa 2 thời điểm trên.
- **Nhận xét và kết luận:**

Nhìn vào biểu đồ trên, ta có một số nhận xét sau:
- Khối lượng trung bình thấp hơn ở giai đoạn 2018-2021: Trong tất cả các tháng, khối lượng trung bình của giai đoạn 2018-2021 (màu xanh dương) đều thấp hơn đáng kể so với khối lượng trung bình của các năm khác (màu cam).
- Sự ổn định trong giai đoạn 2018-2021: Khối lượng trung bình cổ phiếu qua các tháng của giai đoạn 2018-2021 khá đồng đều, không có sự biến động lớn giữa các tháng.
- Khối lượng cao hơn ở các năm khác: Trong nhóm "các năm khác," khối lượng trung bình cổ phiếu cao vượt trội, với mức cao nhất vào các tháng như tháng 2, tháng 3, tháng 6, và tương đối ổn định trong các tháng còn lại.
- Chênh lệch rõ rệt: Chênh lệch lớn nhất giữa hai nhóm thời gian xảy ra vào các tháng đầu năm (tháng 1, tháng 2) và tháng cuối năm (tháng 11, tháng 12), cho thấy hoạt động giao dịch cổ phiếu trong "các năm khác" thường sôi động hơn vào những thời điểm này.

**Kết luận:**

- Sự khác biệt về khối lượng trung bình cổ phiếu giữa hai giai đoạn có thể phản ánh thay đổi trong các yếu tố thị trường như chính sách kinh tế, tâm lý nhà đầu tư, hoặc điều kiện kinh tế chung. Giai đoạn 2018-2021 có thể bị ảnh hưởng bởi các yếu tố như đại dịch COVID-19, dẫn đến khối lượng giao dịch thấp hơn.

## KẾT LUẬN TỔNG QUAN VỀ KHOẢNG THỜI GIAN DÀI HẠN CỦA CỔ PHIẾU GOOGLE:

- Giá cổ phiếu Google cho thấy một xu hướng tăng trưởng dài hạn và mạnh mẽ, với một mức tăng trưởng ổn định từ năm 2004 đến nay. Sự tăng trưởng dài hạn và mạnh mẽ này đã phản ánh vô cùng rõ nét sự thành công của Google khi doanh nghiệp mở rộng lĩnh vực kinh doanh sang nhiều mảng khác nhau, trên nhiều lĩnh vực khác nhau.
- Tuy nhiên, doanh nghiệp cũng trải qua những giải đoạn biến động lớn như khủng hoảng tài chính vào năm 2008 –2009 và đại dịch COVID-19 (2020, 2021). Đây là những thời điểm mà cổ phiếu Google chịu áp lực giảm giá mạnh do khủng hoảng kinh tế. Tuy nhiên sau đó, doanh nghiệp đã phục hồi nhanh chóng nhờ vào sự vững chắc trong mô hình kinh doanh của chính mình.
- Giai đoạn 2020-2022 chứng kiến mức tăng trưởng giá cổ phiếu cao nhất của Google nhờ vào sự tin tưởng tích cực từ các nhà đầu tư. Điều này càng củng cố thêm rằng Google là một tập đoàn doanh nghiệp lớn có xu hướng tăng trưởng mạnh mẽ.
- Sự biến động giá cổ phiếu của Google cũng thay đổi theo các tháng trong năm (với tháng 3, tháng 10, tháng 12 là có xu hướng biến động mạnh mẽ hơn so với các tháng khác).
- Đồng thời, khối lượng giao dịch cũng có thay đổi theo thời điểm. Trong giai đoạn biến động lớn (2018 –2021), khối lượng giao dịch giảm, nhưng trong các năm khác, khối lượng giao dịch thường sôi động hơn.

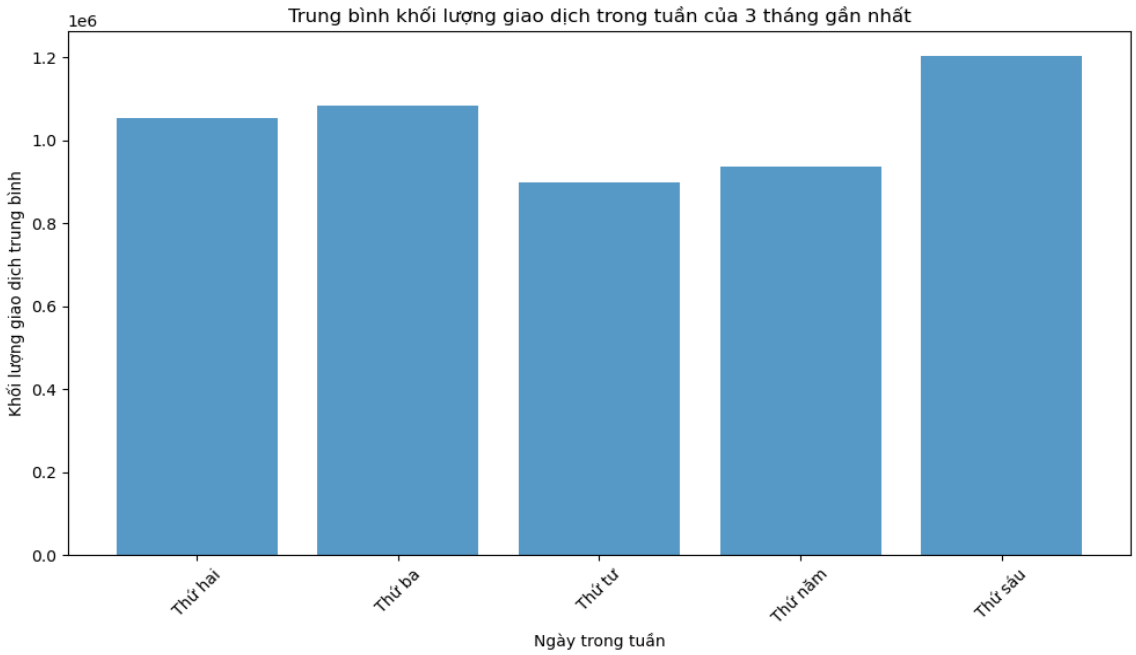
## II. Phân tích chi tiết những biến động ngắn hạn của cổ phiếu Google:

- Sau khi thực hiện tìm ba tháng gần nhất trong tập dữ liệu, ta xét 3 tháng gần nhất, đó là tháng 8/2021, 9/2021, 10/2021 (với tổng số ngày là 50 ngày).

### 1. Trong ba tháng gần nhất, khối lượng giao dịch trung bình nằm ở đầu tuần hay nằm ở những ngày gần cuối tuần là nhiều nhất?

- Chia ba tháng trên thành 7 ngày trong tuần, và thực hiện tìm **khối lượng giao dịch trung bình của các thứ trong các tuần** để xác định khoảng thời gian mà khối lượng giao dịch trung bình là nhiều nhất.

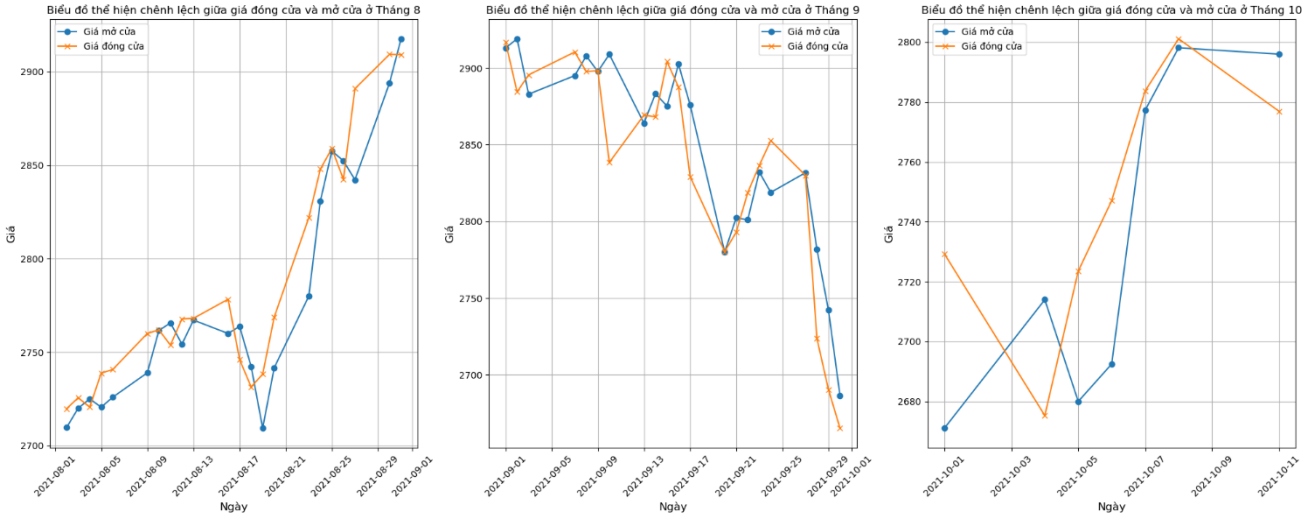
Khi đó sử dụng biểu đồ cột để trực quan khối lượng giao dịch trung bình của các thứ trong tuần như sau:




- **Tại sao sử dụng biểu đồ cột để trực quan khối lượng giao dịch trung bình của các thứ trong 3 tháng được xét?**
  - Biểu đồ cột rất phù hợp trong việc sử dụng để so sánh các giá trị với nhau. Ở đây, người dùng cần so sánh đâu là thứ có khối lượng giao dịch là lớn nhất. Chính vì thế, sử dụng biểu đồ cột có thể giúp người nhìn dễ dàng nhận ra giá trị nào là lớn nhất.
  - Đồng thời biểu đồ cột còn rất thích hợp cho những dữ liệu có phân loại, vì có thể sắp xếp các thuộc tính phân loại đó một cách dễ dàng theo thứ tự.
  - Mặt khác, biểu đồ cột còn giúp người dùng không chỉ dễ dàng nhận diện về sự chênh lệch của các thứ, mà còn nắm được xu hướng thay đổi của dữ liệu theo các thứ trong tuần.
- **Nhận xét và kết luận:**
  - Vào **các ngày thứ 6, khối lượng giao dịch trung bình cao hơn** những ngày bình thường. Điều này có thể xuất phát từ việc các nhà đầu tư thường tận dụng những ngày cuối tuần - đây là những thời gian mà họ (đặc biệt là những nhà đầu tư cá nhân) cảm thấy thoải mái khi tham gia vào thị trường sau một tuần bận rộn. Đồng thời, đây cũng là thời điểm vô cùng thích hợp để họ có thể tận dụng cơ hội, chuẩn bị cho một tuần giao dịch mới.

**2. Biến động giá cổ phiếu trong ngày của 3 tháng gần nhất có cho thấy sự đặc trưng nào tiêu biểu không? (Ví dụ giá mở cửa thường thấp hơn giá đóng cửa)?**

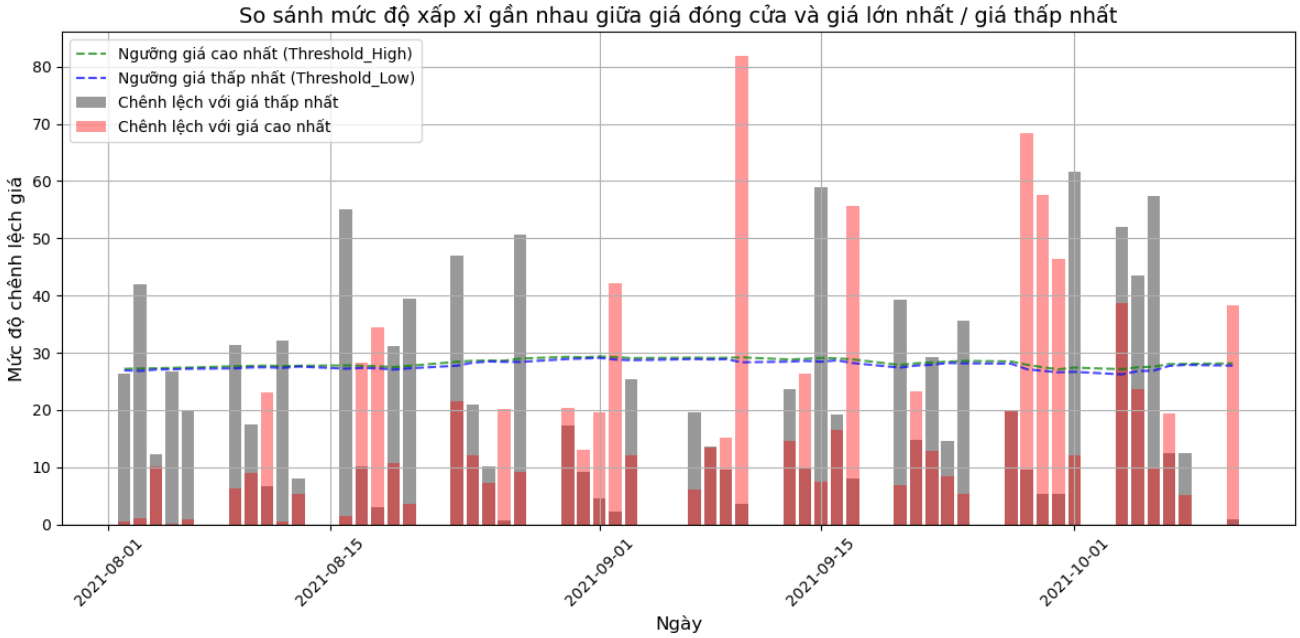
- Ta sử dụng biểu đồ đường có chia theo các khoảng thời gian để dễ dàng thể hiện biến động giữa giá mở cửa và giá đóng cửa của cổ phiếu trong ngày.




- **Tại sao sử dụng biểu đồ đường có chia theo khung thời gian lại phù hợp cho việc trực quan xu hướng của giá đóng cửa và giá mở cửa trong 3 tháng gần nhất?**
  - Dữ liệu được phân nhóm rõ ràng, chia theo từng tháng sẽ giúp người xem dễ dàng tập trung vào từng giai đoạn cụ thể, từ đó mà không bị rối khi xem toàn bộ nội dung trong ba tháng cùng một lúc.
  - Đồng thời, người dùng còn có thể xem được xu hướng của giá mở cửa và giá đóng cửa theo từng tháng và không chỉ so sánh được xu hướng đó trong 1 tháng, mà còn so sánh được xu hướng của các tháng với nhau.
- **Nhận xét và kết luận:**
  - Vào tháng 8, biểu đồ cho thấy có nhiều ngày có giá đóng cửa cao hơn giá mở cửa.
  - Vào tháng 9, xu hướng này lại không còn quá rõ ràng khi vẫn có một số ngày có giá mở cửa cao hơn giá đóng cửa.
  - Vào tháng 10, xu hướng lại đặc biệt rõ ràng khi hầu hết các ngày trong tháng có giá đóng cửa cao hơn giá mở cửa.
  - Nhìn nhận chung, trong ba tháng gần nhất, phần lớn các ngày (cụ thể là 31 ngày) có giá đóng cửa cao hơn giá mở cửa). Điều này cho thấy xu hướng của cổ phiếu trong 3 tháng gần nhất là tăng giá trong ngày. Đồng thời giá cổ phiếu tăng trong ngày còn thể hiện rằng những nhà đầu tư có nhiều kỳ vọng tích cực cho cổ phiếu của tập đoàn Alphabet Inc. (hay Google Inc.)

**3. Có khi nào giá đóng cửa xấp xỉ gần với giá cao nhất hoặc giá thấp nhất trong ba tháng gần nhất hay không? Nếu có, điều này mang ý nghĩa gì đối với doanh nghiệp?**

- Ta sử dụng biểu đồ cột chồng lồng ghép (clustered bar chart) để đánh giá mức độ chênh lệch giữa giá đóng cửa và giá cao nhất / giá thấp nhất trong 3 tháng gần nhất.

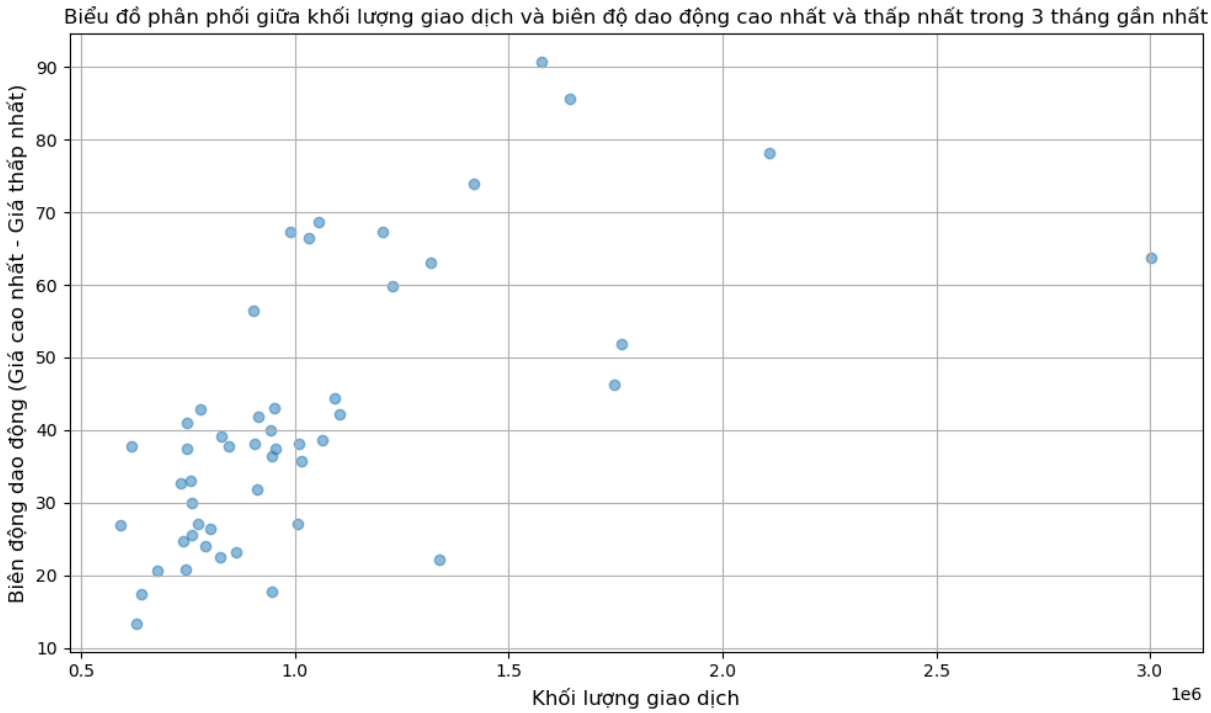



- **Giải thích biểu đồ:**
  - Trong biểu đồ, có 2 ngưỡng giá cần quan tâm, đó là ngưỡng giá cao nhất (là giá cao nhất được biểu diễn ở ngưỡng 1%) và ngưỡng giá thấp nhất (là giá thấp nhất được biểu diễn ở ngưỡng 1%). Tuy nhiên, do 2 đường khá sát nhau, nên ta sẽ xem 2 đường là một đường ngưỡng chung.
  - Vùng có màu xám đen thể hiện các mức độ chênh lệch theo ngày của giá đóng cửa và giá thấp nhất.
  - Vùng có màu đỏ nhạt thể hiện các mức độ chênh lệch theo ngày của giá đóng cửa và giá cao nhất.
  - Vùng màu đỏ đậm xuất hiện do có sự giao nhau giữa vùng màu đỏ nhạt và vùng màu xám đen (tức là khi chênh lệch của giá đóng cửa với giá thấp nhất trùng với chênh lệch của giá đóng cửa với giá cao nhất).
  - Nếu mức chênh lệch của giá đóng cửa với giá thấp nhất trong ngày (cột màu xám đen) nhỏ hơn đường ngưỡng, ngày đó có giá đóng cửa gần với giá thấp nhất trong ngày.
  - Nếu mức chênh lệch của giá đóng cửa với giá cao nhất trong ngày (cột màu đỏ) nhỏ hơn đường ngưỡng, ngày đó có giá đóng cửa gần với giá cao nhất trong ngày.

- **Tại sao sử dụng biểu đồ cột chồng lồng ghép trong việc đánh giá mức độ chênh lệch giữa giá đóng cửa và giá cao nhất/ giá thấp nhất trong 3 tháng gần nhất là phù hợp?**
  - Biểu đồ cột chồng lồng ghép cho phép người dùng nhận diện cụ thể hai mức chênh lệch khác nhau giữa giá đóng cửa với giá thấp nhất và giá đóng cửa với giá cao nhất. Nhờ vậy, người dùng có thể dễ dàng nhận biết giá đóng cửa gần với giá nào hơn mà không bị nhầm lẫn.
  - Không chỉ vậy, người dùng còn có thể dễ dàng so sánh được các xu hướng của mức chênh lệch theo các ngày liên tiếp nhau.

- **Nhận xét và kết luận:**
  - Ta có thể thấy, số ngày có giá đóng cửa gần giá cao nhất là cao hơn số ngày có giá đóng cửa gần với giá thấp nhất (cụ thể là số ngày có giá đóng cửa gần giá cao nhất là 40, và số ngày có giá đóng cửa gần với giá thấp nhất là 34 ngày).
  - Điều này cho thấy các nhà đầu tư có niềm tin cao đối với cổ phiếu và doanh nghiệp, đây là một dấu hiệu tích cực về triển vọng kinh doanh của doanh nghiệp. Mặt khác, giá đóng cửa gần giá cao nhất có thể cho thấy sự ổn định trong kỳ vọng cũa doanh nghiệp do ít biến động tiêu cực trong ngày.
  - Tuy nhiên, số ngày có giá đóng cửa gần với giá thấp nhất cũng khá cao. Điều này cho thấy có vài ngày, giá cổ phiếu thiếu sự ổn định. Việc này có thể xuất phát từ việc quá trình được xét (năm 2021) là thời gian chịu nhiều tác động tiêu cực từ đại dịch COVID-19, với sự suy giảm kinh tế toàn cầu. Chính vì thế, tâm lý của các nhà đầu tư có sự thay đổi liên tục (một số nhà đầu tư, đặc biệt là các nhà đầu tư nhỏ lẻ, chọn cách bán cổ phiếu để thu vốn, đảm bảo quyền lợi của bản thân).
  - Vì thế, sự chênh lệch giữa giá đóng cửa và giá thấp nhất / giá cao nhất chưa được chênh lệch nhiều.

4. **Mối quan hệ giữa khối lượng giao dịch và biên độ dao động (giá cao nhất - giá thấp nhất) hằng ngày trong 3 tháng gần nhất là gì?**
    - Ta sử dụng biểu đồ phân tán để minh họa mối quan hệ giữa khối lượng giao dịch và biên độ dao động (giá cao nhất - giá thấp nhất) hằng ngày trong 3 tháng gần nhất.




- **Tại sao sử dụng biểu đồ phân tán trong việc thể hiện mối quan hệ giữa khối lượng giao dịch và biên độ dao động (cao nhất - thấp nhất) trong 3 tháng gần nhất là phù hợp?**
    - Biểu đồ phân tán giúp người dùng nhìn thấy mối quan hệ tuyến tính giữa khối lượng giao dịch và biên độ dao động giá (khi khối lượng giao dịch tăng, biên độ dao động giá cũng tăng).
    - Đồng thời, người dùng có thể quan sát được sự phân bố của các cụm điểm quan hệ, các điểm gần nhau cho thấy phần lớn dữ liệu sẽ tập trung ở một khu vực nhất định.
    - Mặt khác, người dùng còn có thể so sánh mối quan hệ này của các ngày giao dịch với nhau và đưa ra nhận xét rằng ngày nào có giao dịch cao hơn, ngày nào có giá biến động cao hơn.
- **Nhận xét và kết luận:**
    - Phân bố dữ liệu của cụm điểm tập trung ở biên độ dao động từ 20 đến 50 và khối lượng giao dịch từ 0.5 đến 1.5 triệu cổ phiếu.
    - Với hệ số tương quan là *0.63* (mối tương quan dương trung bình), ta có thể thấy khi khối lượng giao dịch tăng, biên độ dao động giá cũng có xu hướng tăng theo, nhưng

không quá chặt chẽ. Chính vì thế, ta có thể nhận định rằng, ngoài khối lượng giao dịch, biên độ dao động giá có thể bị ảnh hưởng bởi các yếu tố bên ngoài khác dẫn đến sự chênh lệch của giá cao nhất và giá thấp nhất.

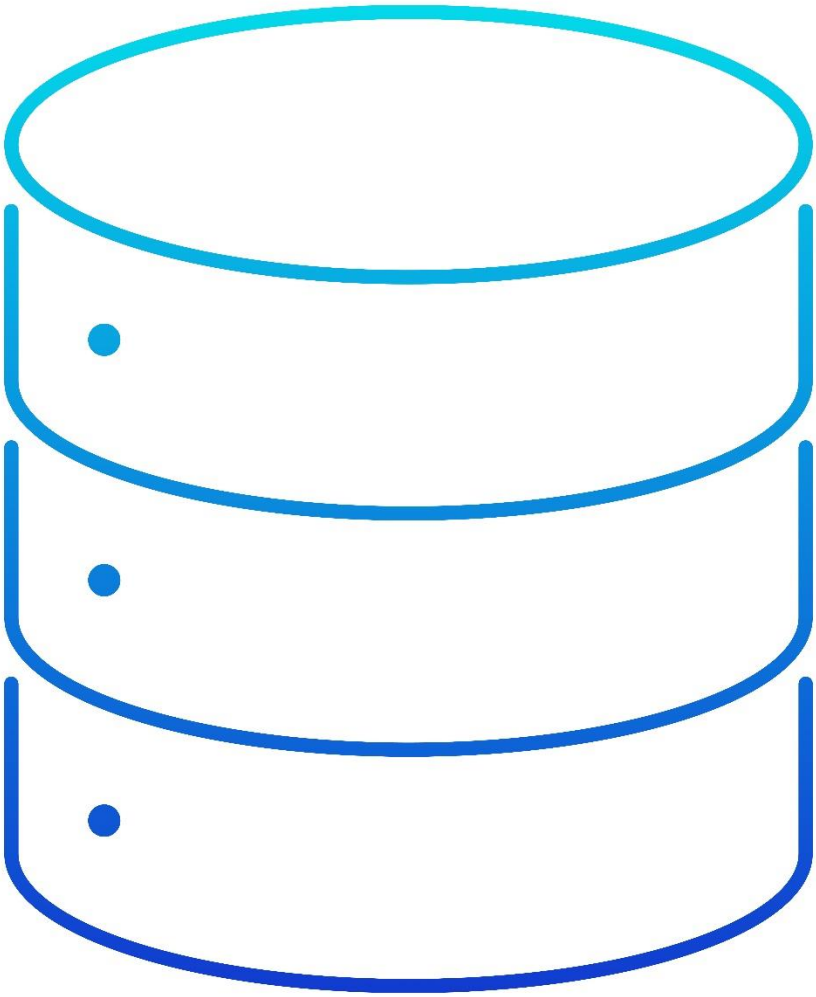
- Khi khối lượng giao dịch tăng, giá cả cũng có xu hướng biến động nhiều hơn, doanh nghiệp có thể tận dụng thông tin này để dự đoán và chuẩn bị cho các giai đoạn biến động lớn khi khối lượng giao dịch tăng cao. Đồng thời sự tăng của khối lượng giao dịch kèm theo biên độ dao động lớn có thể phản ánh tâm lý biến động của các nhà đầu tư, từ đó, doanh nghiệp có thể theo dõi để nhận biết các xu hướng, điều chỉnh các chính sách, truyền thông để tăng thêm thu hút từ các nhà đầu tư.

## KẾT LUẬN TỔNG QUAN VỀ KHOẢNG THỜI GIAN NGẮN HẠN (3 THÁNG GẦN NHẤT) CỦA CỔ PHIẾU GOOGLE:

Với 3 tháng gần nhất (tháng 8/2021, 9/2021, 10/2021):

- Xu hướng khối lượng giao dịch trung bình cao nhất nằm ở thứ 6 cuối tuần. Điều này cho thấy nhà đầu tư có xu hướng gia tăng hoạt động giao dịch trước khi thị trường đóng cửa cuối tuần. Có thể một phần các nhà đầu tư khá lo ngại về những thông tin bất ngờ vào cuối tuần của doanh nghiệp.
- Biến động cổ phiếu Google cho thấy xu hướng khá tích cực trong 3 tháng gần nhất, với giá đóng cửa thường cao hơn giá mở cửa. Điều này cho thấy cổ phiếu Google có sức hút và giá trị vững chắc trên thị trường.
- Mặt khác, giá đóng cửa có xu hướng xấp xỉ giá cao nhất trong ngày. Điều này cho thấy các nhà đầu tư thường kỳ vọng cao vào cổ phiếu Google và giữ vững nhu cầu đến cuối ngày giao dịch. Tuy nhiên, vẫn có nhiều ngày giá đóng cửa có xu hướng xấp xỉ giá thấp nhất trong ngày. Có thể điều này xuất phát từ tâm lý khá biến động của các nhà đầu tư khi thế giới đang có nhiều vấn đề về khủng hoảng kinh tế trong và sau đại dịch COVID-19.
- Bên cạnh đó, khi khối lượng giao dịch tăng, biên độ dao động của giá cao nhất và thấp nhất cũng có xu hướng thăng theo. Donah nghiệp có thể sử dụng những thông tin này để dự đoán các giai đoạn thị trường có thể xuất hiện cơ hội và rủi ro lớn, từ đó tối ưu hóa chiến lược tài chính của bản thân.

## **Section 4:** *Xây dựng mô hình và đánh giá.*



# I. Giới thiệu mô hình và thuật toán sử dụng:

**Vì sao hồi quy tuyến tính (OLS) phù hợp để tính toán và dự đoán giá chứng khoán?**

Hồi quy tuyến tính dựa trên mô hình hóa mối quan hệ giữa một biến mục tiêu (giá chứng khoán, trong trường hợp này là Adj Close) và một hoặc nhiều biến độc lập (các đặc trưng như Open, High, Low, Volume).

Ordinary Least Squares (OLS) là phương pháp phổ biến nhất để ước lượng tham số của hồi quy tuyến tính, vì nó tối ưu bằng cách giảm thiểu tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế.

Đây là lý do tại sao hồi quy tuyến tính OLS có thể phù hợp cho dự đoán giá chứng khoán:

1. ***Hồi quy tuyến tính phù hợp nếu dữ liệu có mối quan hệ gần như tuyến tính:***

a. Giá chứng khoán thường có mối quan hệ gần tuyến tính giữa các đặc trưng:

**Ví dụ:** giá đóng cửa (Close) thường liên quan đến giá mở cửa (Open) và giá cao nhất (High) trong ngày.

b. Các đặc trưng như Volume (khối lượng giao dịch) có thể đóng vai trò như một yếu tố bổ sung, nhưng không quá phức tạp. Nếu mối quan hệ trong dữ liệu là tuyến tính (hoặc gần tuyến tính), hồi quy OLS là một công cụ tốt để dự đoán.

2. ***Đơn giản, dễ hiểu, và dễ giải thích:***

Hồi quy tuyến tính cung cấp một mô hình dễ giải thích:

Hệ số (coefficients) cho biết mỗi biến đầu vào ảnh hưởng đến biến mục tiêu như thế nào.

Điều này rất quan trọng trong tài chính, nơi các nhà đầu tư thường muốn hiểu lý do đằng sau dự đoán.

**Ví dụ:** Nếu Volume tăng 10%, hồi quy có thể cho thấy giá Adj Close tăng một giá trị nhất định.

3. ***Nhanh chóng và hiệu quả trên tập dữ liệu nhỏ:***

Hồi quy tuyến tính rất hiệu quả khi xử lý tập dữ liệu có số lượng dòng và đặc trưng vừa phải, như trong ví dụ của bạn.

Trong khi các mô hình phức tạp hơn như Random Forest hay LSTM cần nhiều tài nguyên tính toán hơn, hồi quy tuyến tính có thể cho ra kết quả nhanh chóng mà không cần nhiều tùy chỉnh.

4. ***Không đòi hỏi quá nhiều điều kiện về dữ liệu:***

Miễn là không có sự đa cộng tuyến nghiêm trọng (multicollinearity) giữa các đặc trưng và các biến độc lập không tương quan quá cao, hồi quy tuyến tính có thể hoạt động tốt.

Với các tập dữ liệu nhỏ, như trong ví dụ của bạn, việc đảm bảo dữ liệu không có outliers hoặc phân phối cực đoan là đủ để hồi quy OLS hoạt động.

Hồi quy tuyến tính OLS cũng có một số hạn chế:

1. ***Giả định tuyến tính tuyệt đối:***

   OLS giả định rằng quan hệ giữa các biến độc lập và biến mục tiêu là tuyến tính, trong khi giá cổ phiếu có thể chịu ảnh hưởng bởi các yếu tố phi tuyến (tin tức, tâm lý thị trường, v.v.).

   Nếu dữ liệu không thực sự tuyến tính, mô hình sẽ không đạt hiệu quả cao.

2. ***Nhạy cảm với ngoại lai (Outliers):***

   Các giá trị ngoại lai trong dữ liệu (ví dụ: khối lượng giao dịch bất thường hoặc biến động giá cực lớn) có thể làm lệch kết quả.

3. ***Không nắm bắt được sự phụ thuộc thời gian:***

   Hồi quy tuyến tính không trực tiếp xử lý tính chất chuỗi thời gian của dữ liệu. Nó xem xét từng quan sát một cách độc lập, nên không tận dụng được các mối quan hệ giữa các điểm thời gian khác nhau.

## II. Cài đặt mô hình và đánh giá mô hình:

Bước 1: Vì cột Date không có giá trị sử dụng trong mô hình này – vì mô hình này không sử dụng yếu tố thời gian. Do đó ta sẽ xóa cột này đi.

Bước 2: Sử dụng cột Adj Close như một tập dữ liệu mục tiêu – giá trị hướng đến của các yếu tốt khác trong mô hình.

Bước 3: Chia tách dữ liệu thành 2 tập khác nhau. Trong đó 80% dùng để huấn luyện và 20% dùng để kiểm tra và đánh giá mô hình. Lấy mẫu thông qua các giá trị random trong tập huấn luyện, ta sẽ cố định thông qua một seed có giá trị là 123.

Bước 4: Sử dụng thư viện scikitlearn để thực thi thuật toán linear regression. Và tiến hành huấn luyện.

Bước 5: Sau khi huấn luyện xong ta có được một mô hình với các tham số chuẩn chỉ, ta sẽ sử dụng mô hình này để tiến hành đánh giá kết quả. Một số giá trị cần đánh giá như:

1. R-squared: Đo lường tỷ lệ biến động của giá trị mục tiêu (y) được giải thích bởi mô hình.
2. Mean Absolute Error: Trung bình sai số tuyệt đối giữa giá trị thực và giá trị dự đoán.
3. Mean Squared Error: Trung bình bình phương sai số, nhạy cảm với sai lệch lớn.

4. Root Mean Squared Error: Căn bậc hai của MSE, đo lường sai số trung bình trong cùng đơn vị với giá trị thực.

**Kết quả thực hiện:**

```
R-squared (R²): 1.0
Mean Absolute Error (MAE): 3.21208080709699e-13
Mean Squared Error (MSE): 3.121070321707217e-25
Root Mean Squared Error (RMSE): 5.586654026971079e-13
```

**Nhận xét kết quả:**

1. $R^2 = 1.0$ cho thấy mô hình giải thích hoàn toàn biến động trong dữ liệu (một cách hoàn hảo).
2. MAE = 3.21208807096990e-13 (~0) cho thấy sai số dự đoán gần như không đáng kể. Cho thấy mô hình dự đoán chính xác từng giá trị của tập kiểm tra.
3. MSE = 3.121070321707217e-25 (~0) cho thấy sai số dự đoán rất nhỏ.
4. RMSE = 5.586654026971079e-13 (~0) chỉ ra rằng sai số trung bình là không đáng kể.

Bước 6: Tính toán hệ số hồi quy.

```
Coefficients:
   Feature   Coefficient
0     High -4.398789e-14
1      Low  4.356189e-14
2     Open  1.286118e-16
3    Close  1.000000e+00
4   Volume  2.710505e-20
```

**Đánh giá tổng quan:**

1. *Tác động của các đặc trưng:*
    - Chỉ có **Close** ảnh hưởng mạnh đến biến mục tiêu (Adj Close) với hệ số bằng 1.
    - Các đặc trưng khác (High, Low, Open, Volume) có hệ số rất nhỏ, gần bằng 0, và dường như không có tác động đáng kể.
2. *Mối quan hệ tuyến tính hoàn hảo:*
    - Giá trị Close **chi phối hoàn toàn** kết quả dự đoán, mốt số nguên nhân:
        - Adj Close thực tế bằng hoặc gần bằng Close trong dữ liệu ban đầu.

      **Ví dụ:** Trong dữ liệu bạn cung cấp, nếu Adj Close chỉ khác Close do một điều chỉnh nhỏ, mô hình sẽ tập trung hoàn toàn vào Close.

3. *Dữ liệu có vấn đề:*
    - Nếu Adj Close gần như là một bản sao của Close, việc huấn luyện mô hình hồi quy tuyến tính sẽ dẫn đến hiện tượng các đặc trưng khác như High, Low, Open, Volume bị bỏ qua.
    - Đây có thể là vấn đề dữ liệu (sự dư thừa giữa các đặc trưng) hoặc một tình huống quá đơn giản.
4. *Volume không có ý nghĩa:*
    - Khối lượng giao dịch (Volume) không có tác động đến kết quả dự đoán, nhưng điều này có thể do:
        - Volume không liên quan trực tiếp đến giá cổ phiếu trong dữ liệu mà ta cung cấp.
        - Cần xem xét thêm dữ liệu phức tạp hơn, hoặc tạo thêm các đặc trưng dựa trên Volume (ví dụ: phần trăm thay đổi khối lượng).

Điều trên và các chỉ số sai số đã chứng tỏ mô hình của chúng ta đang quá mức khớp (Overfitting) – hay cột Adj Close và Close gần như tương tự nhau. Dẫn đến kết quả sẽ bị sai lệch và thiên vị cho close.

- **Kết luận:**

Chúng ta cần phải thay đổi mô hình này.

- **Giải pháp thay đổi:**

Không sử dụng cột Close, thay vào đó sẽ sử dụng cột Adj Close – vì đây là giá đóng cửa cuối cùng trong ngày – để làm giá trị mục tiêu cho mô hình. Khi đó kết quả nhận được sẽ là:

```
R-squared (R²): 0.9999306541744295
Mean Absolute Error (MAE): 2.7730290079803472
Mean Squared Error (MSE): 22.010669878785006
Root Mean Squared Error (RMSE): 4.69155303484731
```

**Nhận xét:**

1. $R^2 = 0.99993$ (~1) cho thấy mô hình giải thích được **99.99% biến động** của giá trị mục tiêu (Adj Close) dựa trên các đặc trưng đầu vào.
2. Giá trị MAE = 2.77 cho biết, trung bình, dự đoán của mô hình chênh lệch khoảng **2.77 đơn vị** so với giá trị thực.

3. MSE = 22.01 cho thấy mức độ sai số tổng thể nhỏ và phân phối sai số không có giá trị lớn bất thường.
4. RMSE = 4.69 cho biết sai số trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng **4.69 đơn vị**.

| | Feature | Coefficient |
|---|---|---|
| 0 | High | 7.585631e-01 |
| 1 | Low | 7.955970e-01 |
| 2 | Open | -5.534607e-01 |
| 3 | Volume | 5.542495e-09 |

**Nhận xét:**

1. **High: 0.7585631**
   - Khi giá High tăng thêm 1 đơn vị, giá trị dự đoán của Adj Close sẽ tăng trung bình **0.76 đơn vị** (với giả định các yếu tố khác không đổi).
   - Đây là một hệ số khá cao, cho thấy giá cao nhất trong ngày có ảnh hưởng đáng kể đến giá đóng cửa điều chỉnh (Adj Close).
2. **Low: 0.7955970**
   - Khi giá Low tăng thêm 1 đơn vị, giá trị dự đoán của Adj Close sẽ tăng trung bình **0.80 đơn vị** (với các yếu tố khác không đổi).
   - Hệ số này cao hơn một chút so với High, điều này cho thấy giá thấp nhất trong ngày cũng có ảnh hưởng mạnh đến giá điều chỉnh.
3. **Open: -0.5534607**
   - Khi giá Open tăng thêm 1 đơn vị, giá trị dự đoán của Adj Close sẽ giảm trung bình **0.55 đơn vị**.
   - Hệ số âm chỉ ra rằng giá mở cửa có tác động ngược chiều với giá Adj Close trong dữ liệu này. Đây có thể phản ánh rằng giá cổ phiếu trong ngày có xu hướng giảm khi giá mở cửa cao.
4. **Volume: 5.542495e-09 (~0)**
   - Khối lượng giao dịch (Volume) có hệ số rất nhỏ ($10^{-9}$), điều này chỉ ra rằng khối lượng giao dịch hầu như không ảnh hưởng đến giá Adj Close.
   - Trong trường hợp này, khối lượng giao dịch có thể không cung cấp nhiều thông tin hữu ích để dự đoán giá cổ phiếu.

## III. Một số hạn chế của mô hình:

1. **Quá khớp (Overfitting):**

- $R^2$ gần như hoàn hảo (0.99993) có thể chỉ ra rằng mô hình ghi nhớ dữ liệu thay vì học được mối quan hệ tổng quát.
- Một số lý do:
    - Tập dữ liệu huấn luyện và kiểm tra không đủ lớn hoặc không đủ đa dạng.
    - Một số đặc trưng có tương quan rất cao, dẫn đến đa cộng tuyến (multicollinearity).

2. **Thiếu khả năng tổng quát hóa:**

- Nếu tập dữ liệu kiểm tra không đại diện cho dữ liệu thực tế (chẳng hạn, nó có cấu trúc đơn giản hoặc rất tương tự dữ liệu huấn luyện), mô hình có thể hoạt động kém trên dữ liệu mới.

3. **Tầm quan trọng thấp của Volume:**

- Mặc dù Volume thường được xem là yếu tố quan trọng trong tài chính, hệ số nhỏ của nó trong mô hình này chỉ ra rằng nó không đóng góp nhiều thông tin. Điều này có thể do:
    - Volume không ảnh hưởng trong dữ liệu hiện tại.
    - Có thể vẫn cần phải tạo thêm một số đặc trưng từ volume để làm nổi bật lên tác động của nó. Ví dụ như thêm cột % change in Volume – tỉ lệ thay đổi khối lượng giao dịch, trung bình động của volume – Moving Average, khối lượng giao dịch chuẩn hóa, …