MỤC LỤC

LỜI MỞ ĐẦU

Trong lĩnh vực kinh doanh, khai phá dữ liệu không chỉ là một phương tiện, mà là chìa khóa mở ra những cánh cửa mới của sự hiểu biết và quản lý. Dữ liệu không chỉ giúp chúng ta đánh giá hiệu suất và xu hướng thị trường, mà còn là nguồn động lực cho sự sáng tạo và đổi mới.

Khi chúng ta chìm đắm vào biển số liệu và thông tin, chúng ta như những nhà thám hiểm của thị trường, đặt ra những câu hỏi sâu sắc về hành vi khách hàng, xu hướng tiêu dùng, và cơ hội mới. Dữ liệu không chỉ là công cụ hỗ trợ quyết định mà còn là bản đồ chỉ dẫn cho chiến lược kinh doanh.

Qua việc khai phá dữ liệu, chúng ta có thể nhận biết những khu vực tiềm năng, tối ưu hóa chuỗi cung ứng, và tạo ra những sản phẩm dịch vụ phản ánh chính xác nhu cầu của thị trường. Dữ liệu giúp chúng ta hiểu rõ hơn về đối thủ cạnh tranh, tăng cường tính cạnh tranh và tạo ra những chiến lược phát triển bền vững.

Dựa vào những số liệu mà BTC cung cấp, nhóm Dataminds chúng em đã phân tích và đưa ra những dự đoán về bộ dữ liệu của BTC.

I. KHAI BÁO DỮ LIỆU.

Để bắt đầu cuộc phiêu lưu khai phá dữ liệu của mình, chúng ta không thể bỏ qua bước quan trọng của việc khai báo các thư viện, đặc biệt được thiết kế để làm sạch và phân tích dữ liệu.

Đầu tiên và quan trọng nhất, chúng ta cần tích hợp Pandas, một thư viện Python mạnh mẽ dành cho xử lý và phân tích dữ liệu có cấu trúc. Nhóm chúng em dùng thư viện Pandas để phục vụ cho việc đọc dữ liệu.

Tiếp theo, để làm sạch dữ liệu, chúng ta sử dụng thư viện NumPy, re, datetime, timedelta để xử lý dữ liệu.

Ngoài ra, để trực quan hóa dữ liệu, không thể bỏ qua thư viện Matplotlib – một thư viện giúp chúng ta tạo ra các biểu đồ và đồ thị một cách dễ dàng và thẩm mỹ.

```
# Khai báo thư viện
import pandas as pd # đọc dữ liệu

# xử lý dữ liệu
import numpy as np
import re
from datetime import datetime, timedelta

import matplotlib.pyplot as plt # vẽ biểu đồ
```

Bằng cách tích hợp những thư viện này, nhóm chúng em đã trang bị bộ công cụ mạnh mẽ, sẵn sàng chinh phục mọi thách thức trên hành trình khám phá dữ liệu.

II. ĐỌC DỮ LIỆU.

1. Đọc dữ liệu từ file Excel và gán giá trị vào biến data.

Trong quá trình khai thác dữ liệu trong kinh doanh, việc đọc dữ liệu từ file Excel và gán giá trị vào biến data đóng vai trò quan trọng để phân tích và hiểu rõ hơn về thông tin kinh doanh của doanh nghiệp. Bằng cách sử dụng các thư viện và công cụ phổ biến như Pandas trong Python.

2. Gán bảng vào biến tương ứng.

Với bộ dữ liệu được cung cấp, file Excel gồm 3 bảng dữ liệu: customer, ticket và film, đóng vai trò quan trọng trong quá trình phân tích dữ liệu kinh doanh của doanh nghiệp. Để thuận tiện cho việc quản lý và sử dụng thông tin từ mỗi bảng, chúng ta tiến hành đọc dữ liệu từ file Excel và gán chúng vào các biến tương ứng.

```
# Đọc dữ liệu từ file excel
data = pd.ExcelFile("data/DATA-SET-VÒNG-1-CUỘC-THI-DATA-GOT-TALENT-2023.xlsx")
customer = pd.read_excel(data, "customer")
ticket = pd.read_excel(data, "ticket")
film = pd.read_excel(data, "film")
```

Nhờ vào việc gán mỗi bảng vào biến tương ứng, chúng ta dễ dàng tiến hành các phân tích, truy vấn và tương tác với dữ liệu từng khía cạnh của doanh nghiệp một cách linh hoạt và hiệu quả. Điều này giúp tạo ra một cơ sở dữ liệu mạnh mẽ để nhóm hiểu rõ hơn về hành

vi của khách hàng, thông tin về vé và thông tin về các bộ phim, đồng thời hỗ trợ quyết định kinh doanh dựa trên những thông tin chi tiết và toàn diện.

III. MÔ TẢ DỮ LIỆU.

- Dữ liệu ở bảng.

Trước hết, chúng ta thực hiện một tổng quan về dữ liệu có trong các bảng để có cái nhìn tổng quan về cấu trúc và nội dung của chúng. Điều này đặc biệt quan trọng để xây dựng một cơ sở hiểu biết sâu sắc và toàn diện về thông tin doanh nghiệp.

Bằng cách sử dụng thư viện như Pandas trong Python để hiển thị và khám phá dữ liệu trong mỗi bảng. Chúng ta cần xem qua về dữ liệu bảng, bao gồm:

- Kích thước bảng
- Dữ liệu bảng đó 10 dòng đầu
- Dữ liệu bảng đó 10 dòng cuối
- Thông tin về các cột có trong bảng (tên cột, kiểu dữ liệu của mỗi cột, số giá trị không bị trống của mỗi cột, từ đó suy ra số giá trị trống của mỗi cột)

Tiếp đến, nhóm đã thực hiện việc xem bảng với 2 bảng còn lại.

⇨ Việc này giúp chúng ta hiểu rõ hơn về tính đầy đủ và chất lượng của dữ liệu.

Sau khi xem qua các bảng, ta có những chú thích như sau:

- Bảng `customer`:
  - Xuất hiện dòng có giá trị ở cột `customerid` có một dấu chấm ở cuối khác so với các dòng khác
  - Những dòng có giá trị ở cột `job` là "teenager" thì sẽ bị thiếu giá trị ở cột `industry`
- Bảng `ticket`:
  - Không phải mỗi vé trong bảng là một lần đặt
  - Các vé có cùng lần đặt sẽ có cùng giá trị ở cột `orderid`
  - Có 4 vé trong bảng bị thiếu dữ liệu (Khách hàng đặt vé trên website)

Qua bước này, chúng ta có thể xác định các thách thức có thể xuất hiện và xác định cách tiếp cận phù hợp để xử lý chúng. Ngoài ra, việc nhìn chung vào dữ liệu từ các bảng cũng giúp xây dựng một cái nhìn tổng thể, tạo nền tảng cho quá trình phân tích chi tiết hơn trong việc đưa ra quyết định kinh doanh.

IV. LÀM SẠCH VÀ XỬ LÝ DỮ LIỆU.

1. Chuẩn hóa tên cột theo một tiêu chuẩn.

Trước khi bắt đầu quá trình phân tích dữ liệu từ các bảng customer, ticket và film, bước quan trọng là chuẩn hóa lại tên cột theo một tiêu chuẩn chung. Việc này giúp tạo ra được sự đồng nhất và dễ quản lý trong quá trình xử lý dữ liệu, giúp tăng khả năng hiệu quả khi thực hiện các thao tác, truy vấn và phân tích dữ liệu sau này.

Chuẩn hóa tên cột không chỉ là bước quan trọng để tạo ra một cơ sở dữ liệu có tổ chức mạnh mẽ mà còn giúp tăng tính nhất quán và dễ bảo trì. Việc này làm cho quá trình làm việc với dữ liệu trở nên thuận lợi hơn, giúp định rõ từng thành phần và giảm rủi ro gặp lỗi trong quá trình xử lý dữ liệu.

2. Sửa lại dữ liệu Customerid.

Sau bước chuẩn hóa tên cột, tiếp theo chúng ta cần tập trung vào việc sửa lại dữ liệu Customerid của khách hàng để loại bỏ các dấu chấm thừa. Việc này là quan trọng để đảm bảo tính nhất quán và chính xác của dữ liệu. Bằng cách sử dụng các công cụ xử lý chuỗi trong ngôn ngữ lập trình, chúng ta có thể dễ dàng loại bỏ dấu chấm không mong muốn từ Customerid.

Việc sửa lại dữ liệu customerid không chỉ giúp giảm nguy cơ xảy ra lỗi trong quá trình xử lý, mà còn cung cấp một cơ sở dữ liệu chắc chắn và dễ sử dụng cho các phân tích và đánh giá kinh doanh tiếp theo. Điều này đóng vai trò quan trọng trong việc xây dựng một hệ thống dữ liệu chất lượng và đáng tin cậy.

3. Biến đổi dữ liệu của cột dob thành định dạng ngày sinh và thêm cột tuổi ứng với ngày sinh của khách hàng.

Trong quá trình xử lý dữ liệu, chúng ta sẽ thực hiện bước biến đổi cột dob (ngày sinh) để đưa nó về định dạng ngày/tháng/năm. Điều này giúp tạo ra một cột mới chính xác và dễ đọc, làm việc với thông tin ngày sinh trở nên thuận tiện hơn.

Sau khi hoàn thành bước này, ta thêm một cột tuổi cho mỗi khách hàng, dựa trên thông tin ngày sinh đã được biến đổi. Bằng cách sử dụng tính toán giữa ngày hiện tại và ngày sinh, ta có thể tạo ra một cột tuổi có giá trị chính xác, cung cấp thông tin quan trọng về độ tuổi của từng khách hàng.

Quá trình biến đổi dữ liệu của cột dob và thêm cột tuổi không chỉ giúp làm sạch và cải thiện độ chính xác của dữ liệu, mà còn mang lại thông tin hữu ích về độ tuổi của khách hàng, hỗ trợ trong việc phân tích và hiểu rõ hơn về đặc điểm khách hàng theo nhóm tuổi.

4. Tính số lượng khách hàng theo độ tuổi.

Sau khi đã thêm cột tuổi vào dữ liệu, chúng ta tiếp tục quá trình phân tích bằng cách tính số lượng khách hàng theo độ tuổi. Tuy nhiên, trong quá trình này, chúng ta phát hiện ra rằng có một số khách hàng có độ tuổi là con số âm.

Để giải quyết vấn đề này, chúng ta cần tiến hành xử lý trước dữ liệu. Có thể xác định nguyên nhân của độ tuổi âm, có thể do nhập liệu không chính xác hoặc có sự lỗi trong dữ liệu gốc. Sau đó, ta có thể áp dụng các biện pháp xử lý như đặt giá trị tuyệt đối, loại bỏ các giá trị không hợp lý, hoặc thậm chí tìm kiếm nguồn gốc của lỗi để sửa chữa dữ liệu. Việc xử lý độ tuổi âm giúp đảm bảo tính chính xác của dữ liệu và làm cho quá trình phân tích theo độ tuổi trở nên đáng tin cậy hơn, đồng thời cung cấp cái nhìn toàn diện hơn về sự phân bố độ tuổi của khách hàng.

⇨ Sau khi đã xử lý và loại bỏ những khách hàng có giá trị tuổi là con số âm, chúng ta tiếp tục quá trình phân tích bằng cách tính số lượng khách hàng theo độ tuổi và biểu diễn thông tin này thông qua một biểu đồ.

Bằng cách sử dụng các thư viện đồ họa Matplotlib, ta tạo ra biểu đồ cột thể hiện phân bố khách hàng theo độ tuổi. Biểu đồ này sẽ giúp chúng ta dễ dàng nhận diện các nhóm độ tuổi phổ biến, đồng thời hiển thị sự phân bố một cách trực quan.

⇨ Sau khi quan sát biểu đồ ta nhận thấy rằng độ tuổi của khách hàng tập trung từ 24 đến 30 (ngoại trừ 25)

5. Tính số lượng khách hàng nam và nữ.

Để tính số lượng khách hàng nam và nữ từ dữ liệu, chúng ta sử dụng các thư viện xử lý dữ liệu như Pandas trong Python. Ta tạo các điều kiện để lọc dữ liệu theo giới tính và sau đó đếm số lượng khách hàng nam và nữ.

⇨ Từ kết quả tính toán, chúng ta nhận thấy rằng số lượng khách hàng nữ trong dữ liệu lớn hơn số lượng khách hàng nam khoảng 350 người. Đây là một thông tin quan trọng và có thể mang lại cái nhìn sâu sắc hơn về sự phân bố giới tính trong tập dữ liệu.

Sự chênh lệch này có thể là một yếu tố quan trọng trong các chiến lược kinh doanh và tiếp thị, vì nó có thể tạo ra cơ hội để tập trung vào nhu cầu và mong muốn cụ thể của đối tượng khách hàng nữ. Các chiến lược quảng cáo, sản phẩm hay dịch vụ có thể được tối ưu hóa dựa trên sự hiểu biết này về sự chênh lệch giới tính trong khách hàng.

6. Tính số lượng khách hàng theo nghề nghiệp và biểu diễn bằng biểu đồ tròn

Tiếp theo trong quá trình phân tích, chúng ta tính toán số lượng khách hàng theo nghề nghiệp để có cái nhìn rõ ràng về sự đa dạng của đối tượng khách hàng. Bằng cách sử dụng Pandas để lọc và đếm số lượng khách hàng theo từng nghề nghiệp, chúng ta có thể tạo ra một bảng dữ liệu thống kê.

Để biểu diễn thông tin này một cách trực quan, ta có thể sử dụng một biểu đồ tròn. Biểu đồ tròn sẽ thể hiện phần trăm mỗi nhóm nghề nghiệp so với tổng số khách hàng. Điều này giúp chúng ta nhanh chóng nhận thức được những nghề nghiệp nào có đóng góp lớn và những nghề nghiệp nào chiếm tỷ lệ thấp trong khách hàng.

⇨ Sau khi quan sát biểu đồ phân phối nghề nghiệp của khách hàng, chúng ta nhận thấy một hiện tượng đáng chú ý: hơn 50% khách hàng thuộc nhóm sinh viên (students) và thanh thiếu niên (teenagers). Đây là một khám phá quan trọng, làm cho chúng ta hiểu rõ hơn về đặc điểm độ tuổi của đối tượng khách hàng, đặc biệt là trong khoảng từ 24 đến 30 (ngoại trừ 25).

Để tận dụng thế mạnh của nhóm đối tượng này, chúng ta có thể đưa ra chiến lược kinh doanh hướng đến việc thêm những đợt giảm giá đặc biệt cho học sinh và sinh viên, việc này không chỉ thu hút họ đến rạp chiếu phim mà còn tạo điều kiện thuận lợi để tăng cường xuất chiếu của những bộ phim đang hot trend, đặc biệt được ưa chuộng trong cộng đồng giới trẻ.

7. Tính số lượng vé được đặt trên Website.

Để đo lường sự tương tác và hoạt động trực tuyến, chúng ta tính số lượng vé được đặt trên website. Bằng cách sử dụng các công cụ phân tích web và theo dõi giao dịch trực tuyến, ta có thể thu thập thông tin về số lượng vé mà khách hàng đã đặt qua website.

Bằng cách theo dõi và phân tích số lượng vé đặt trên website, chúng ta có thể đánh giá hiệu suất của các chiến lược quảng cáo trực tuyến và các ưu đãi đặc biệt mà doanh nghiệp có thể cung cấp.

⇨ Nhìn vào dữ liệu, chúng ta nhận thấy rằng số lượng vé được đặt trên website ít hơn so với số lượng vé mua trực tiếp tại quầy. Điều này có thể chỉ ra rằng một số khách hàng vẫn chưa có biết đến hoặc ít sử dụng phương pháp đặt vé trực tuyến.

⇨ Để khuyến khích việc đặt vé trên website, chúng ta có thể đưa ra những đợt giảm giá đặc biệt dành cho khách hàng thực hiện giao dịch trực tuyến.

8. Tính tổng số lượng vé theo thời gian và lấy ra 10 mốc thời gina được khách hàng đặt nhiều nhất.

Tiếp theo trong quá trình phân tích, chúng ta tính tổng số lượng vé theo thời gian để xác định các khoảng thời gian nào được khách hàng ưa chuộng nhất.

Sau khi thực hiện tính toán, chúng ta lựa chọn 10 mốc thời gian mà khách hàng đặt vé nhiều nhất. Kết quả cho thấy, khoảng thời gian từ 19h đến 21h30 là thời điểm mà khách hàng tập trung đặt vé nhiều nhất.

Thông qua việc xác định được khoảng thời gian này, chúng ta có thể tối ưu hóa chiến lược tiếp thị và quảng cáo, chẳng hạn như đặt các đợt giảm giá, sự kiện đặc biệt hoặc quảng cáo chính sách đặt vé trước đóng cửa để kích thích việc mua vé trong khoảng thời gian này.

9. Tính khoảng thời gian vé được bán.

Tiếp theo, chúng ta tính khoảng thời gian mà vé được bán để hiểu rõ về chu kỳ bán vé và xác định các tháng có hoạt động mua bán sôi động nhất. Kết quả cho thấy rằng tất cả các vé đều được bán trong tháng 5 năm 2019.

Qua thông tin này, chúng ta có thể tập trung chiến lược tiếp thị và quảng cáo vào tháng 5, tận dụng những chiến lược hiệu quả đã thấy trong quá khứ và tạo ra các sự kiện hoặc chương trình khuyến mãi đặc biệt để tối ưu hóa việc bán vé trong khoảng thời gian này.

10. Thêm cột thứ vào bảng ticket từ ngày được bán của vé đó.

Để phân tích thêm về biểu đồ thời gian, chúng ta thêm một cột mới là "thứ" vào bảng ticket, dựa trên ngày mà vé được bán. Từ đó ta dễ dàng nhận biết và phân tích xu hướng theo ngày trong tuần, giúp hiểu rõ hơn về thói quen mua vé của khách hàng vào các ngày cụ thể.

Sau khi thêm cột "thứ", chúng ta có thể thực hiện phân tích thêm về mức độ hoạt động mua bán vé trên từng ngày trong tuần, giúp tối ưu hóa các chiến lược tiếp thị và quảng cáo theo đúng thời điểm khách hàng thích hợp nhất.

11. Tính tổng số lượng và doanh thu theo ngày trong tháng và biểu diễn bằng biểu đồ.

Tiếp theo, chúng ta tính tổng số lượng vé và doanh thu theo từng ngày trong tháng để hiểu rõ hơn về sự biến động và tương quan giữa các chỉ số này.

Kết quả cho thấy rằng có sự tỉ lệ thuận giữa số lượng vé và doanh thu trong suốt tháng. Đặc biệt, ngày 03/05/2019 được nhận diện là ngày có số lượng vé và doanh thu thấp nhất, trong khi ngày 05/05/2019 và 11/05/2019 lại là những ngày có số lượng vé và doanh thu cao nhất.

Thông qua việc nhận diện các ngày có sự biến động lớn, chúng ta có thể tập trung chiến lược tiếp thị, quảng cáo hoặc chương trình khuyến mãi đặc biệt vào những ngày này để tối ưu hóa doanh thu và tăng cường hoạt động mua bán vé.

12. Tính số lượng và doanh thu theo ngày trong tuần và biểu diễn bằng biểu đồ.

Tiếp theo, chúng ta tính tổng số lượng vé và doanh thu theo từng ngày trong tuần để xác định sự biến động và tương quan giữa các chỉ số này. Sau khi tính toán số lượng và doanh thu theo ngày trong tuần và biểu diễn thông tin này qua biểu đồ, chúng ta nhận thấy một số điều quan trọng về mối liên kết giữa các chỉ số này.

Cụ thể, có sự tỉ lệ thuận giữa số lượng vé và doanh thu theo ngày trong tuần. Thứ 5 và chủ nhật được xác định là những ngày có số lượng vé và doanh thu cao nhất, trong khi thứ 3 lại là ngày có số lượng vé và doanh thu thấp nhất.

Dựa trên thông tin này, chúng ta có thể đề xuất một chiến lược tăng thêm xuất chiếu vào khoảng thời gian từ 19h đến 21h30 những ngày chủ nhật để tối ưu hóa doanh thu, thông qua việc áp dụng chiến lược giảm giá, tổ chức sự kiện đặc biệt hoặc quảng cáo đặc trưng cho các buổi chiếu phim trong khoảng thời gian này.

13. Xuất dữ liệu ra các file tương ứng.

Cuối cùng, để lưu giữ kết quả phân tích và số liệu đã tính toán, chúng ta quyết định xuất dữ liệu ra các file tương ứng