# Tài liệu Hướng dẫn Mã nguồn

Tài liệu này cung cấp mô tả chi tiết về quy trình phân tích dữ liệu được thực hiện trong dự án phân tích tình hình kinh doanh tại siêu thị.

Link gihub: https://github.com/Maianh2510/Project--Superstore-Business-Analysis.git

Script này thực hiện các bước tải dữ liệu, làm sạch, xử lý và trực quan hóa để hỗ trợ phân tích trên Power BI.

## 1. Nhập thư viện cần thiết

Script sử dụng các thư viện sau để xử lý và trực quan hóa dữ liệu:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

## 2. Tải dữ liệu

Tập dữ liệu được tải từ tệp Excel bằng pandas. Script hiển thị 5 dòng đầu tiên và thông tin chung về tập dữ liệu.

df = pd.read\_excel('/content/Superstore.xlsx')

df.head()

df.info()

## 3. Làm sạch dữ liệu

Các bước làm sạch dữ liệu bao gồm kiểm tra trùng lặp, chuyển đổi kiểu dữ liệu.

# Kiểm tra dữ liệu trùng lặp

df.duplicated()

# Chuyển đổi các cột sang kiểu string

columns\_string = ['Customer\_ID','Customer\_Name','Postal\_Code','Product\_ID','Product\_Name']

df[columns\_string] = df[columns\_string].astype('string')

# Kiểm tra lại các cột string

df[columns\_string ].dtypes

# Kiểm tra dữ liệu có bao nhiêu unique

df['Ship\_Mode'].unique()  
df['Segment'].unique()

df['Country'].unique()

df['City'].unique()

df['State'].unique()

df['Region'].unique()

df['Category'].unique()

df['Sub\_Category'].unique()

# Chuyển đổi các cột danh mục sang kiểu category

columns\_category = ['Ship\_Mode','Segment','Country','City','State','Region','Category','Sub\_Category']

df[columns\_category] = df[columns\_category].astype('category')

# Kiểm tra lại các cột category

df[columns\_category].dtypes

# Chuyển đổi cột ngày tháng sang datetime

df['Order\_Date'] = pd.to\_datetime(df['Order\_Date'])

df['Order\_Date'].head()

df['Ship\_Date'] = pd.to\_datetime(df['Ship\_Date'])

df['Ship\_Date'].head()

# Kiểm tra lại data

df.info()

## 4. Xử lý giá trị khuyết

Các giá trị bị thiếu trong cột 'Ship\_Mode' và 'Country' được điền bằng mode.

# Kiểm tra giá trị bị thiếu

df.isnull().sum()

# Điền giá trị bị thiếu bằng mode

df['Ship\_Mode'] = df['Ship\_Mode'].fillna(df['Ship\_Mode'].mode()[0])

df['Country'] = df['Country'].fillna(df['Country'].mode()[0])

# kiểm tra mô tả thống kê dữ liệu

df.describe()

## 5. Xử lý giá trị ngoại lai

Sử dụng biểu đồ hộp để phát hiện các điểm ngoại lai trong cột 'Sales' và 'Profit'. Áp dụng phương pháp IQR để loại bỏ.

plt.figure(figsize=(8,5))

sns.boxplot(data=df[['Sales']])

plt.title("Boxplot for Sales ")

plt.show()

plt.figure(figsize=(8,5))

sns.boxplot(data=df[['Profit']])

plt.title("Boxplot for Profit ")

plt.show()

# Hàm loại bỏ giá trị ngoại lai

def rm\_outliner(col\_name):

    Q1 = df[col\_name].quantile(0.25)

    Q3 = df[col\_name].quantile(0.75)

    IQR = Q3 - Q1

    lower\_bound = Q1 - 1.5 \* IQR

    upper\_bound = Q3 + 1.5 \* IQR

    df\_result = df[(df[col\_name] >= lower\_bound) & (df[col\_name] <= upper\_bound)]

    return df\_result

df = rm\_outliner(col\_name = 'Sales')

df = rm\_outliner(col\_name = 'Profit')

## 6. Phân tích tương quan

Ma trận tương quan được tạo để đánh giá mối quan hệ giữa các biến số, sau đó trực quan hóa bằng heatmap.

numberical\_columns =  ['Sales','Quantity','Discount','Profit']

numberical\_columns

coreclation\_matrix = df[numberical\_columns].corr()

coreclation\_matrix

sns.heatmap(coreclation\_matrix,annot=True, cmap='coolwarm')

plt.title('sales correlation matrix')

plt.show()

## 7. Xuất dữ liệu

Dữ liệu sau khi làm sạch và xử lý được xuất ra tệp CSV để trực quan hóa trên Power BI.

df.to\_csv("sales.csv", index=False)