BÀI THỰC HÀNH 7

PHÂN TÍCH TRỰC QUAN DỮ LIỆU

1. Sử dụng bộ dữ liệu Cars. Link download

https://github.com/hohuongthien/Python-programing-for-Data-Analysis/blob/main/cars_dataset.zip

Sinh viên thực hiện các bước như bên dưới, sau đó thay đổi các tham số, các thuộc tính để so sánh sự khác nhau

```
#Import một số thư viện cần thiết

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

sns.set(color_codes=True)


# Load bộ dữ liệu

df = pd.read_csv("/data.csv")

df.head(5)

df.tail(5)


# Kiểm tra kiểu dữ liệu của các thuộc tính

df.dtypes


#Xem thống kê của bộ dữ liệu

df. describe()
```

```
# Loại bỏ những thuộc tính dư thừ
df = df.drop(['Engine Fuel Type', 'Market Category', 'Vehicle Style', 'Popularity', 'Number of Doors', 'Vehicle Size'], axis=1)
df.head(5)

#Đổi tên thuộc tính
df = df.rename(columns={"Engine HP": "HP", "Engine Cylinders": "Cylinders", "Transmission Type": "Transmission", "Driven_Wheels": "Drive Mode","highway MPG": "MPG-H", "city mpg": "MPG-C", "MSRP": "Price" })
df.head(5)

# Kiểm tra số dòng, số cột của bộ dữ liệu
df.shape

# Kiểm tra số lượng dòng dữ liệu bị trùng
duplicate_rows_df = df[df.duplicated()]
print("Số lượng dòng bị trùng: ", duplicate_rows_df.shape)

# Đếm số dòng dữ liệu
df.count()

# Xóa các dòng dữ liệu bị trùng
df = df.drop_duplicates()
df.head(5)

# Sau khi xóa các dòng dữ liệu trùng, kiểm tra lại số lượng dòng dữ liệu
df.count()
```

```
# Kiểm tra các giá tri null
print(df.isnull().sum())

# Xóa cá giá trị null
df = df.dropna()
df.count()

# Kiểm tra lại giá trị null của các thuộc tính
print(df.isnull().sum())

# Kiểm tra giá trị outliers của thuộc tính Cylinders
sns.boxplot(x=df['Cylinders'])

# Vẽ biểu đồ histogram
df.Make.value_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title("Số lượng xe theo hãng sản xuất")
plt.ylabel('Số lượng xe')
plt.xlabel('Hãng xe')

# Vẽ biểu đồ heatmaps
plt.figure(figsize=(10,5))
c= df.corr()
sns.heatmap(c,cmap="BrBG",annot=True)
c
```

```
# Vẽ biểu đồ scatterplot giữa 2 thuộc tính HP và Price
fig, ax = plt.subplots(figsize=(10,6))
ax.scatter(df['HP'], df['Price'])
ax.set_xlabel('HP')
ax.set_ylabel('Price')
plt.show()
```

2. Thực hiện các bước phân tích dữ liệu trực quan trên 2 bộ dữ liệu sau:

https://github.com/hohuongthien/Python-programing-for-Data-Analysis/blob/main/AmesHousing.zip

https://github.com/hohuongthien/Python-programing-for-Data-Analysis/blob/main/California%20Housing%20Prices.zip