**13장 연습문제**

1826074 오현진

2021/12/08

[EDA 예제]

소스코드 :

# EDA는 데이터를 시각화 및 변환하여 직관적이고 체계적인 방법으로 탐구하는 방법 및 활동을 의미

# EDA의 과정

# 1. 분석의 목적 확인

# 2. 어떠한 변수가 존재하는지를 확인(변수의 이름, 설명 등)

# 3. 데이터를 전체적으로 살펴보며 데이터에 문제가 없는지 확인하고, head와 tail부분을 확인

# 4. 이상한 수치나 NULL(NaN)과 같은 데이터의 존재유무 확인

# 5. 속성간의 관계를 분석하기

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

# 이용한 데이터 : sales.csv

# 1. 분석의 목적 : 한 브랜드의 2013.01 ~ 2015.10 까지 판매량 조사

# 2. 변수 소개 : date -> 날짜, shop\_id -> 매장코드, item\_id -> 상품명,

# item\_price -> 상품가격, item\_cnt\_day -> 해당상품의 판매량

data = pd.read\_csv('sales.csv')

df = pd.DataFrame(data)

# 3. 데이터를 전체적으로 살펴보기 (head, tail부분을 확인한다.)

# 분석 방향에 대한 고찰 : a. 2013.01.05의 경우, 판매량이 -1인 것으로 보아, 환불건이 있었던 것 같다.

# b. item\_price의 경우, 판매 금액의 단위가 필요할 것 같다.

#. c. 매장코드와 상품코드는 관계자만 알 수 있는 정보로 판단된다.

#. d. 1년이 다 지나지도 않았는데 100만건이 넘는 판매정보가 있는 것 보니, 장사가 잘 되나?

print(df.head(), '\n')

print(df.tail(), '\n')

# 4. 이상치나 NULL(NaN)과 같은 데터의 존재 유무 확인

# 적절한 통계 요약지표 확인

print(df.describe(), '\n')

# item\_price에 대한 속성값 확인

print(df['item\_price'].describe(), '\n')

# 판매량이 음수인 날 확인 : 100만건 이상의 판매 데이터중, 반품이 들어온 건수는 2000여 건 뿐임.

print(df.loc[df['item\_cnt\_day'] < 0, :], '\n')

# 가격이 1000보다 작은 것 확인 : 100만건 이상의 판매 데이터중, 1000보다 작은 판매 건수는 거의 900만건임.

print(df.loc[df['item\_price'] < 1000, :], '\n')

# 시각화 도구 사용

# 가격과 판매량의 분포 : 가격이 싸다고 해서 무조건 판매량이 높은게 아니라, 적당한 가격에 형성된 상품이 판매량이 높음.

sns.set(style='darkgrid')

sns.boxplot(x='item\_price', y='item\_cnt\_day', data=data)

sns.pairplot(df)

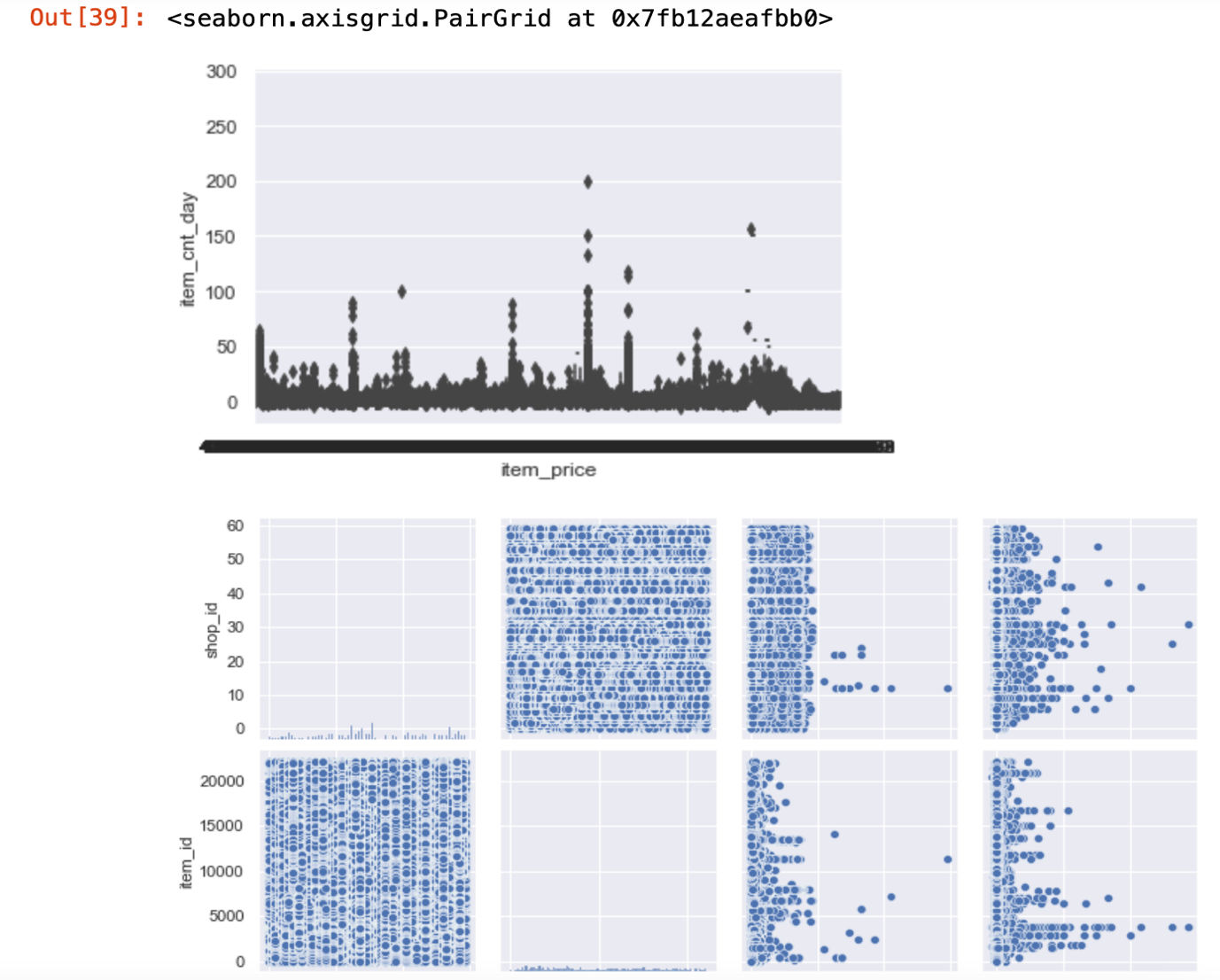
실행결과 :

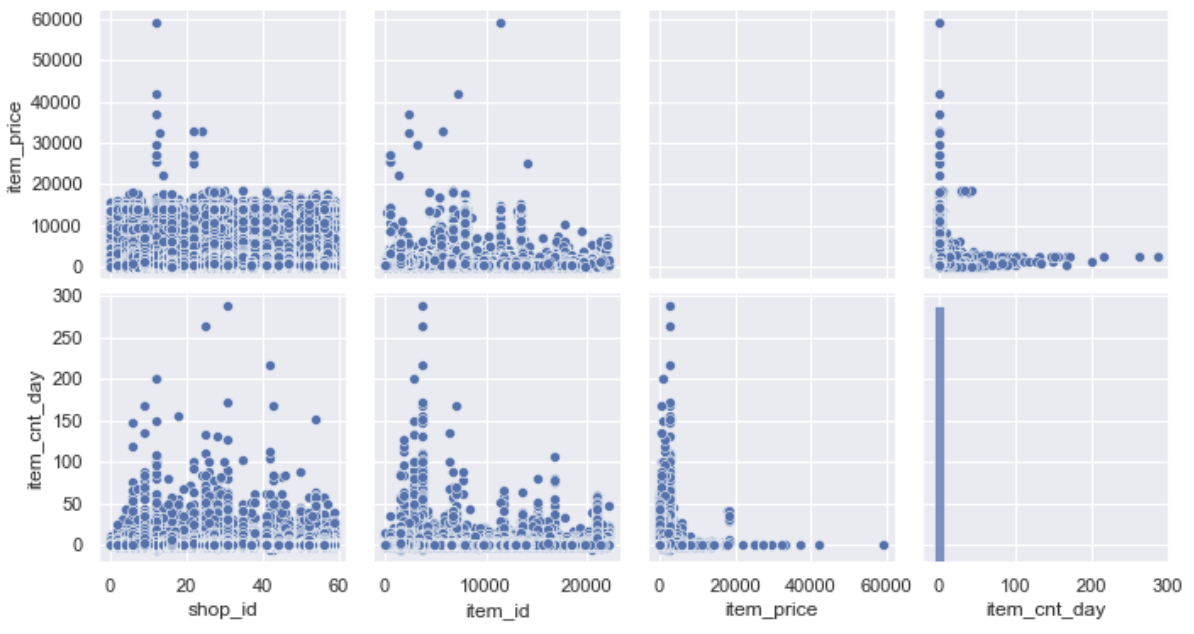
테이블이(가) 표시된 사진

자동 생성된 설명

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명





추가설명 : 과제를 sales.csv 데이터를 사용하여 진행하였습니다. 각 EDA 과정에 대한 설명은 코드 내 주석으로 작성해두었습니다.