DATA SCIENCE PROJECT

Report

Members of the group :

Name	ID
Youssef Mahmoud Youssef	42010659
Rwan Malek	42010454
Anas Mohamed	42010069
Abdullh Rabea	42010168
Amira boctor klini	41910174

Introduction:

In this report, we will discuss how to make a code in Jupyter Notebook to preprocessing on pc games dataset . Preprocessing is an important step in data analysis as it helps to clean and prepare the data for further analysis. In this report, we will focus on the preprocessing steps of cleaning and removing duplicates.

Dataset:

The dataset used in this report is pc games dataset. The dataset contains information about pc games that include name of the game ,genre ,price, reviews and other important information about the game.

Preprocessing:

Preprocessing is an essential step in data analysis that involves cleaning and preparing the data for further analysis. In this report, we will discuss some common preprocessing techniques that we used in this project.

1- Handling Missing Values:

Missing values are a common issue in any dataset and can be caused by a variety of reasons such as data entry errors, sensor failures, or data loss during transmission.

Examples

```
df.isna().sum()
```

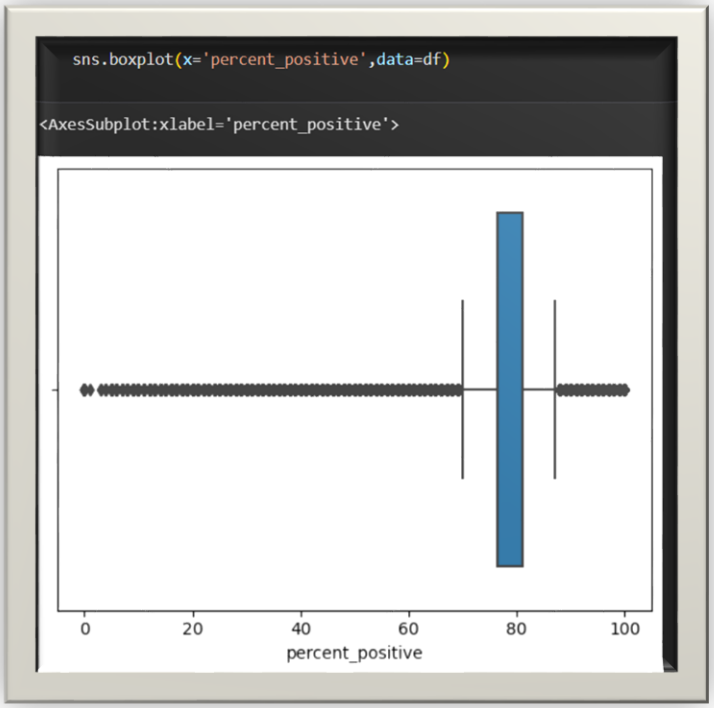
```
id                0
title             0
genres            0
price             0
overall_review    0
reviews           0
percent_positive  0
win_support       0
mac_support       0
lin_support       0
dtype: int64
```

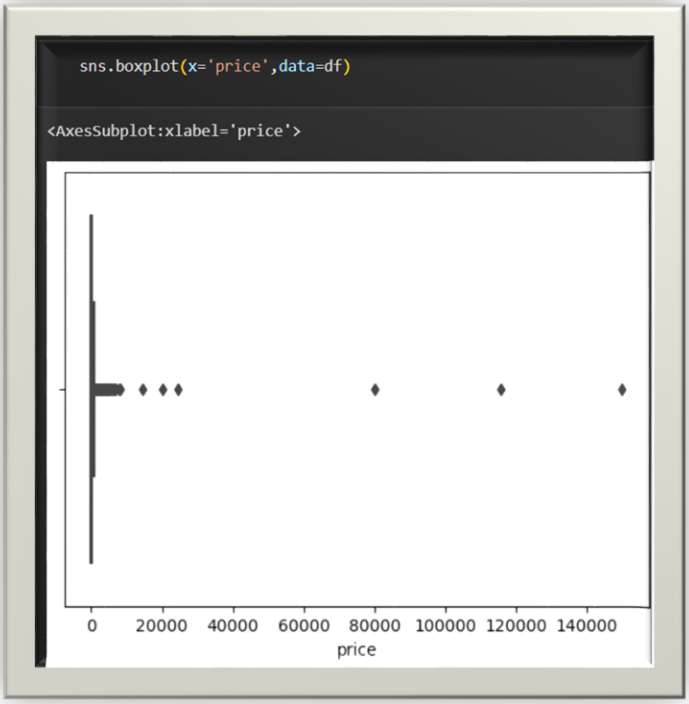
```
df.info()
```

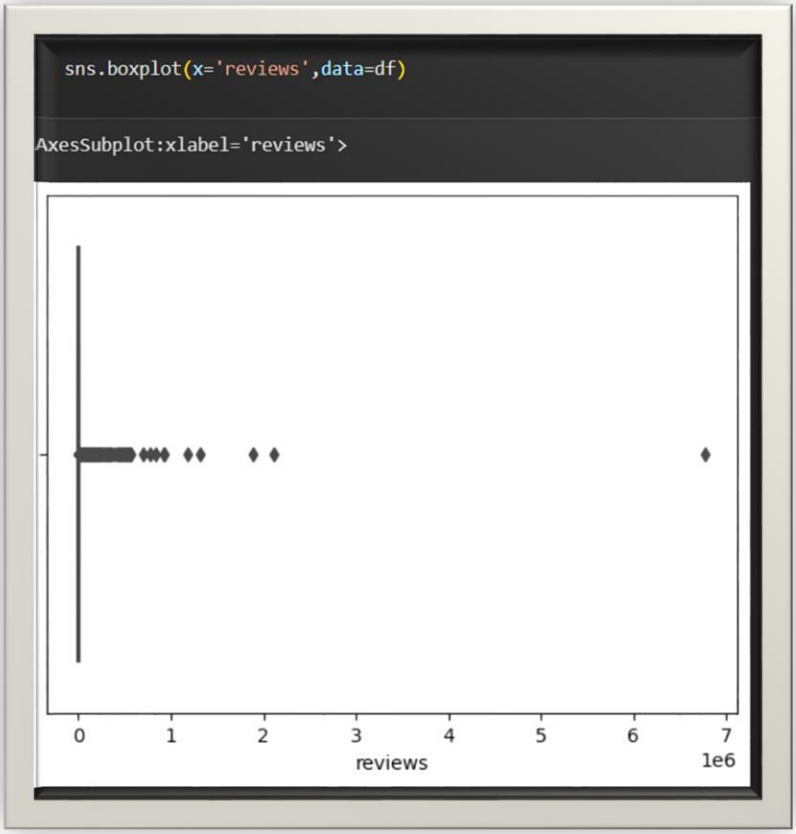
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 73340 entries, 0 to 73344
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                73340 non-null  object
 1   title             73340 non-null  object
 2   genres            73340 non-null  object
 3   price             73340 non-null  float64
 4   overall_review    73340 non-null  object
 5   reviews           73340 non-null  float64
 6   percent_positive  73340 non-null  float64
 7   win_support       73340 non-null  object
 8   mac_support       73340 non-null  object
 9   lin_support       73340 non-null  object
dtypes: float64(3), object(7)
memory usage: 6.2+ MB
```

3-**handling the outliers:**
We didn't do much in the outliers because it seems to be important in that dataset so decide to leave it as it is

-example







Data visualization :

is the process of representing data in a graphical or visual format. It is an essential component of data analysis as it allows us to communicate complex data insights in a clear and concise manner. In this response, I will discuss some common types of data visualizations and how they can be used to communicate insights from a dataset.

1-Heat Maps:
Heat maps are a visualization technique used to represent large amounts of data in a condensed and visual format. They are useful for identifying patterns or correlations between different variables in a dataset. Heat maps can be simple or complex, depending on the number of variablesand the level of detail required.

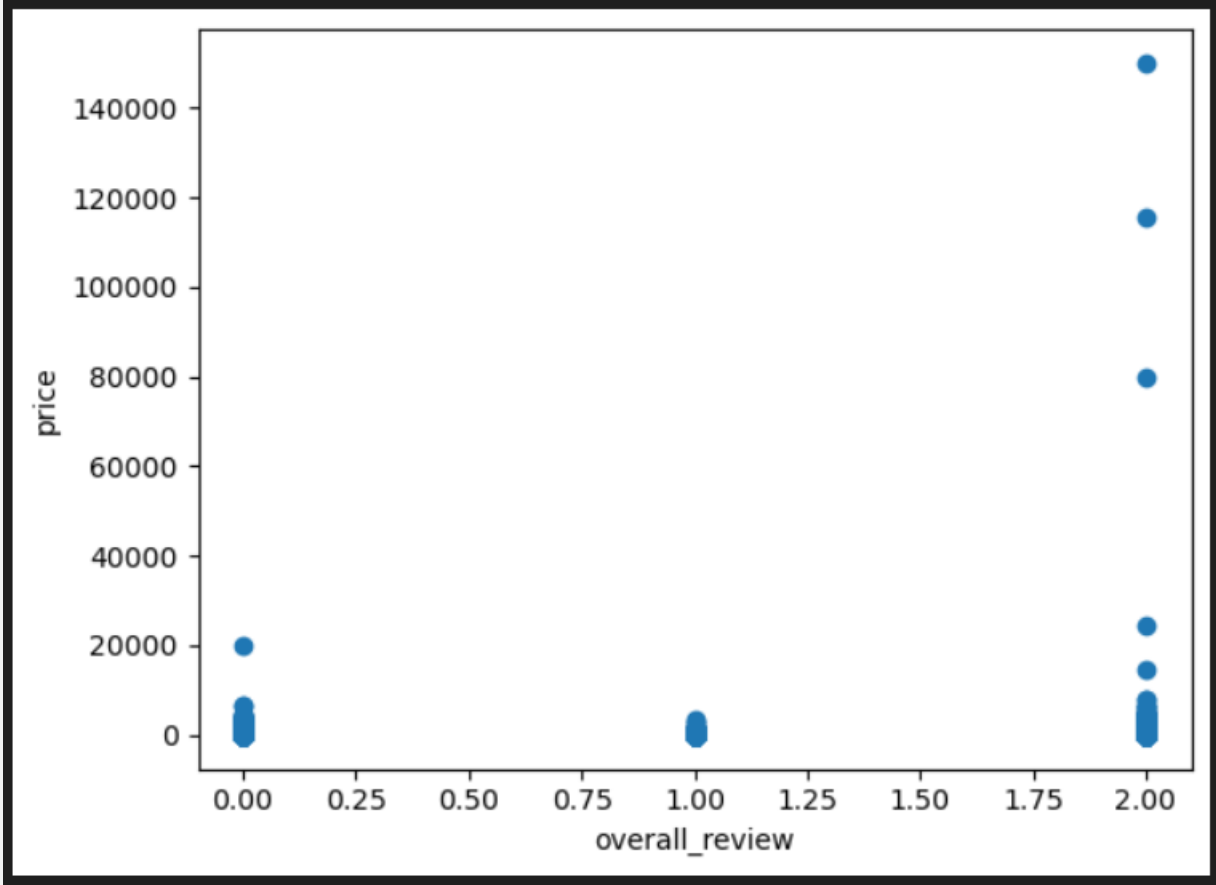
```
: sns.heatmap(df.corr(),annot=True)
: <AxesSubplot:>
```

2- Scatter Plots:

Scatter plots are a popular visualization technique used to represent the relationship between two continuous variables in a dataset. They are useful for identifying patterns or correlations between different variables in a dataset. Scatter plots can be simple or complex, depending on the number of variables and the level of detail required.

```
import matplotlib.pyplot as plt
plt.scatter(x,y)
plt.xlabel('overall_review')
plt.ylabel('price')
```

```
Text(0, 0.5, 'price')
```



Power PI:

Power BI is a business analytics service that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their reports and dashboards. It allows users to connect to different data sources, transform data, and create interactive visualizations, reports, and dashboards.