Airbnb NYC analysis program

This program includes four main parts: exploratory data analysis, data preprocessing data visualization and conclusions. This program runs in PyCharm.

1. Exploratory data analysis

df = pd.read\_csv('ABNYC.csv')  
print(df.head())  
print(df.shape)  
print(df.columns)

Result:

文本

描述已自动生成

Figure 1

This dataset has 48895 rows and 16 columns, all column names is showed as figure 1.

Check missing values, duplicates, and outliers.

# display all columns  
pd.set\_option('display.max\_columns', 16)  
print(df.isnull().sum())  
print(df.duplicated().sum())  
# cancel scientific notation  
pd.set\_option('display.float\_format', lambda x: '%.3f' % x)  
print(df.describe())  
print(df.info())

Result:

文本

描述已自动生成

Figure 2

As figure 2 showed us, missing values showed up in following columns (name, host\_name,

last\_ review, and review\_per\_month), and there are no duplicates in the dataset.

手机屏幕的截图

描述已自动生成

Figure 3

表格

中度可信度描述已自动生成

Figure 4

After checked statistical information of each column, price column has minimum value 0 and minimum\_nights column has maximum value 1250 which can be considered as unreasonable outliers.

1. Data preprocessing

2.1 missing value

# filling missing values  
airbnb = df.copy()  
  
airbnb['reviews\_per\_month'].fillna(0, inplace=True)  
airbnb['name'].fillna('Unknown', inplace=True)  
airbnb['host\_name'].fillna('Unknown', inplace=True)  
airbnb.drop('last\_review', 1, inplace=True)  
  
print(airbnb.isnull().sum())

Result:

文本

描述已自动生成

Figure 5

Missing values in review\_per\_month filled with 0 and missing values in name and host\_name filled with Unknown. Last\_review column is unnecessary for this analysis, so I decided to delete it.

* 1. outliers

文本

描述已自动生成