**Understanding the business problem**

Proper understanding of the problem helps a Data scientist to develop adequate solution

. Once the problem is well understood, appropriate data and information are gathered.

**Data gathering**

Data gathering is the process of retrieving data from various sources to be used in our DS process.

**Data processing**

Data processing is the process of converting the gathered data into: easily readable formats, or easily process-able formats.

**Data analysis**

Data analysis is the process of analyzing data sets to summarize their main characteristics.

**Data visualization**

Data visualization is the graphical/pictorial representation of the information and data that we have. This allows us to observe trends/patterns in our data set. Taking a look at a large set of numbers in tables may not be that helpful in aiding understanding of inherent trends.

**Data cleaning**

Data cleaning is the process of removing unwanted or inaccurate records from a table or a dataset.

**Creating and testing a model**

A model is a mathematical construct that finds patterns when we feed raw data into it. Models are developed for specific use cases. In order words, a model built for a specific problem mayn't provide adequate solution for other problems. The model enables us to classify the input or make prediction about future trends of things etc.

**Python**

Supports multiple paradigms (Functional, structured, & procedural programming. Most applied language for Data Science tasks today. Free and Open Source, Highly Readable, Memory Management, Clean Visual Layout, Support OOPS, High Performance. Support for several Libraries required for DS tasks. Could be supported by several tools.

**R**

It is an open-source programming language developed in 1991 (by Ross Ihaka and Robert Gentleman). For statistically oriented DS tasks, R is the perfect language. Hence, it is very popular among statisticians. With over 10k packages in the open-source repository of CRAN, R caters for all statistical applications. Can handle complex linear algebra. Other than just statistical analysis, R can also be used for building neural networks models.

**2.**

**Direct download of data file (or files) manually**

That is, such data maybe obtained from: Internal Repository (private): Mostly owned and managed by a specific company. External repository (publicly available): Can be freely accessed

**Query data from a database**

This involves writing simple/complex queries with the aid of a programming language to access data from a database of interest. A typical database query language is the SQL (Structured Query Language) that allows access to the data. It is important to know that queries (SQL commands) are meant to operate on structured data, mostly organized in tabular forms. The tables that are contain data and are related in a way that records can be easily retrieved or stored or modified.

**Query an API (usually web-based)**

This involves writing queries with the aid of a programming language to access data via an API. Such queries are often issued via HTTP Request. That is, The vast majority of automated data queries will run via HTTP requests. HTTP GET is the most common method, but there are also PUT, POST, DELETE methods change some state on the server.

**Scrap data from a webpage**

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.

**Acquisition of data by oneself**

Another means of obtaining data for Data Science solution would be to collected the data based on designed experimental protocol. Though this approach is mostly not considered in the field of Data Science because the data gotten from such means are usually not large in volume.

**CSV** (comma separate value) files

Refers to any delimited text file (not always separated by commas)

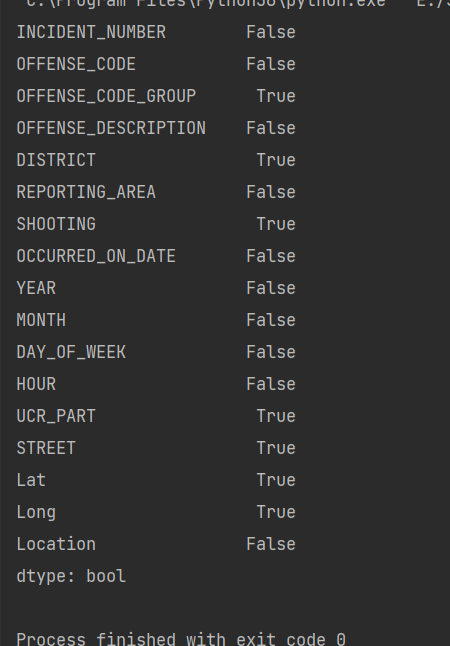
**JSON** (Javascript object notation) files and strings

JSON originated as a way of encapsulating Javascript objects. A number of different data types can be represented

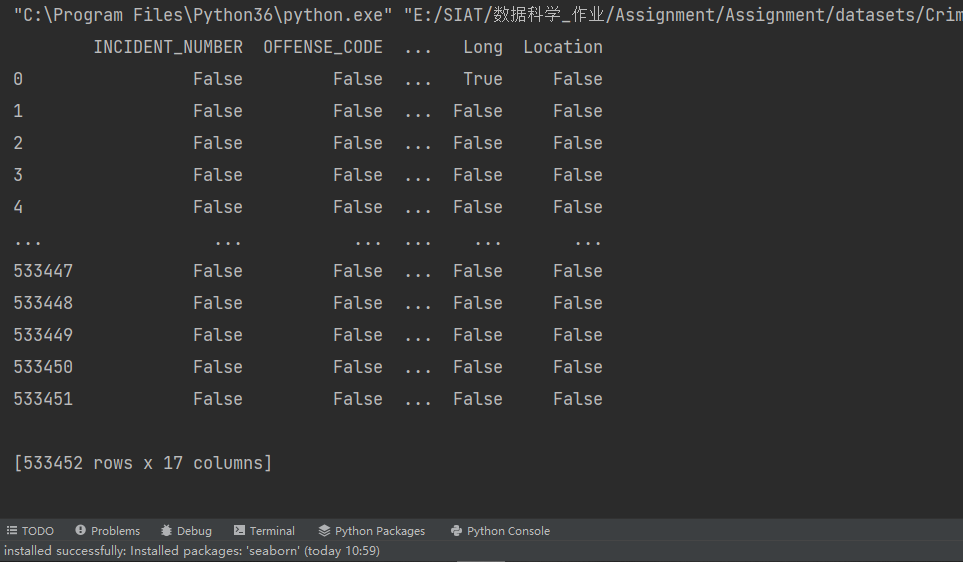
**HTML/XML** (hypertext markup language / extensible markup language) files

3.

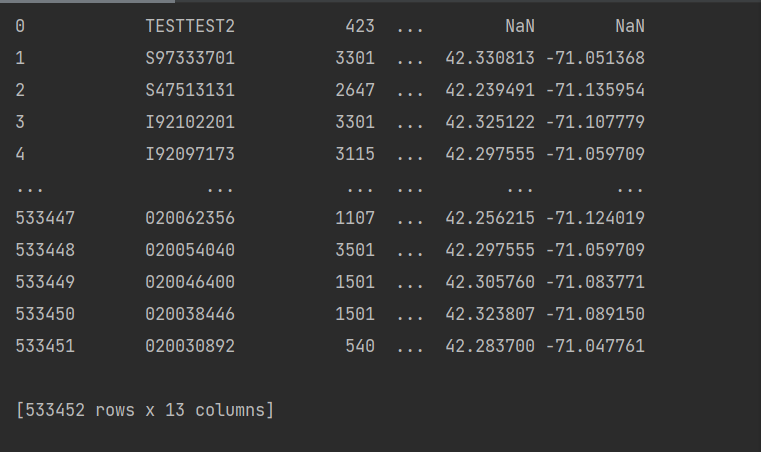
Null column



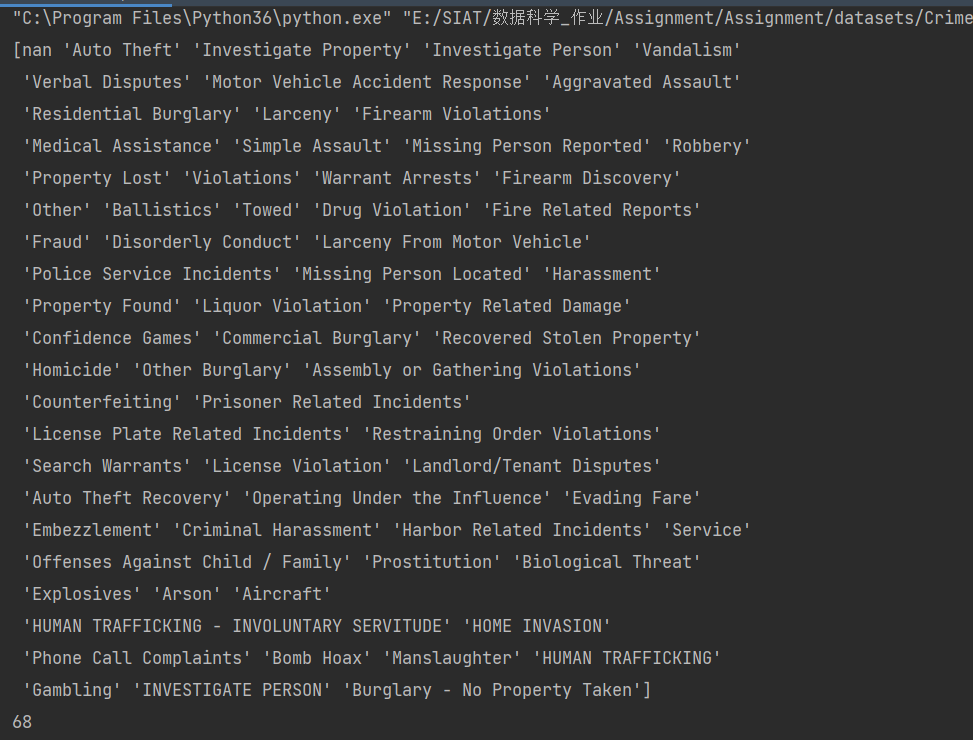
Missing values

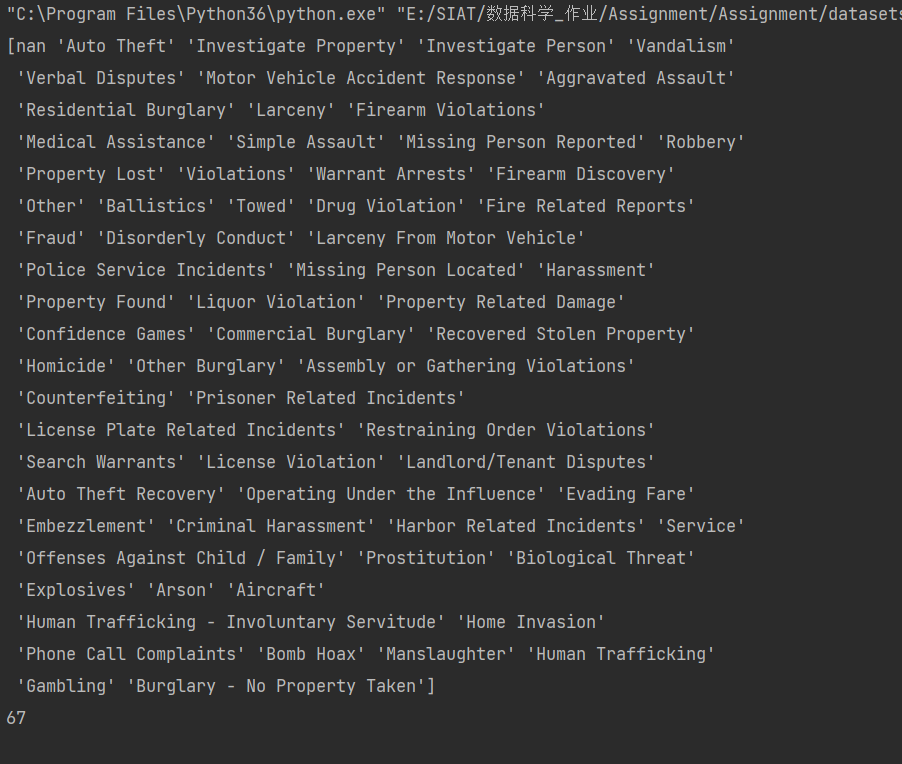


Duplication rows



Misspellings





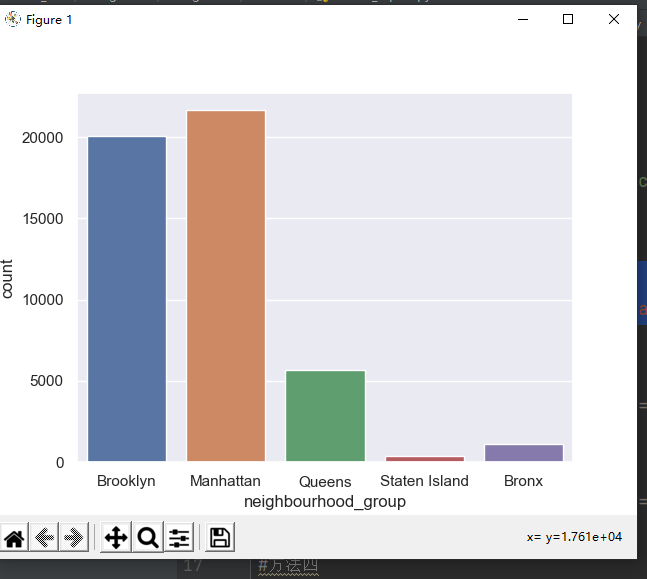
4.

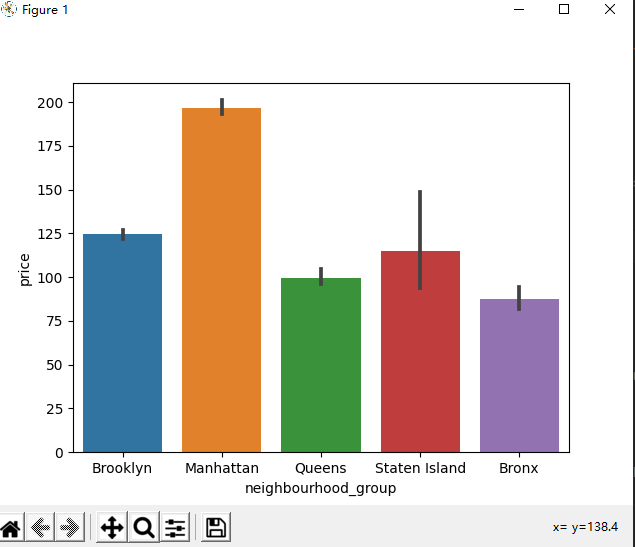
**To view changes over time easily via a visual aid rather that plain data**

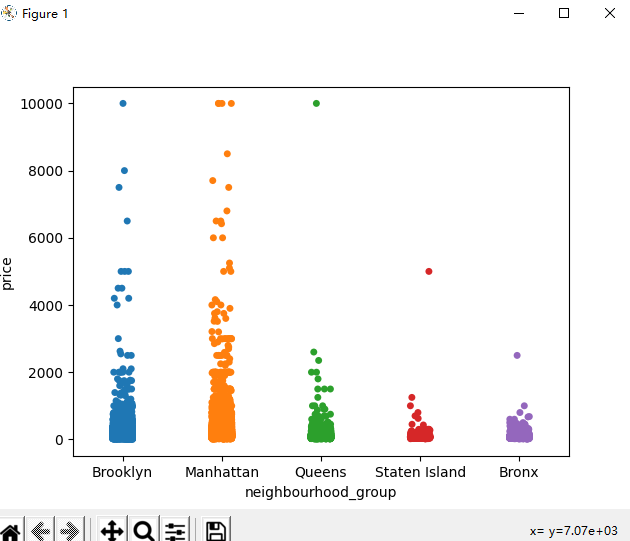
**To discover correlations among two or more variables seamlessly**

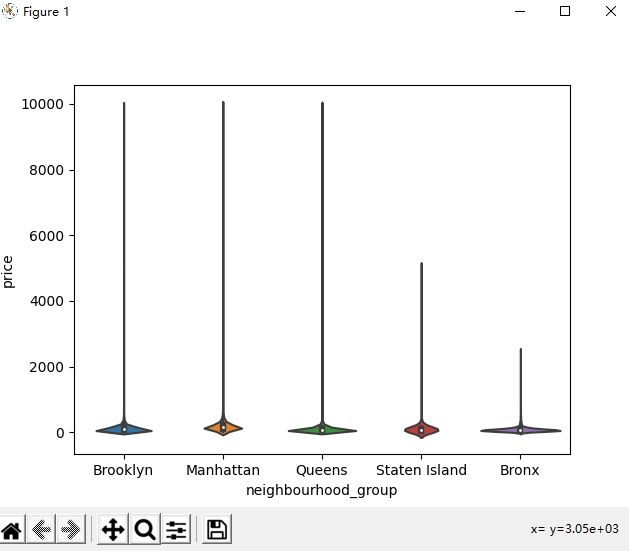
**To simplify complex information into user-friendly formats**

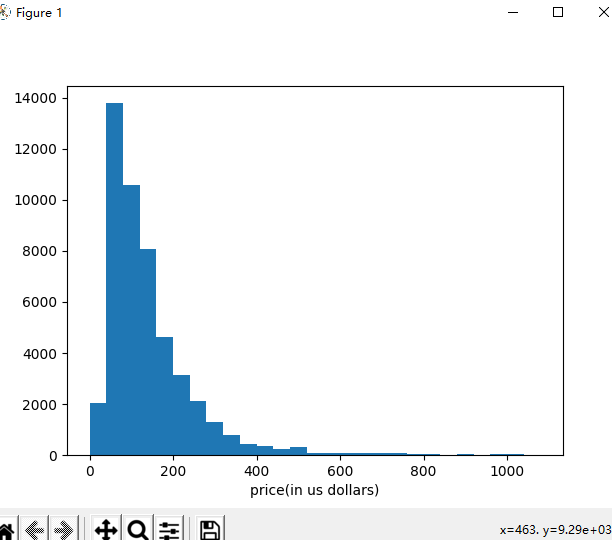
**To tell a better story with a bunch of pictures over time**

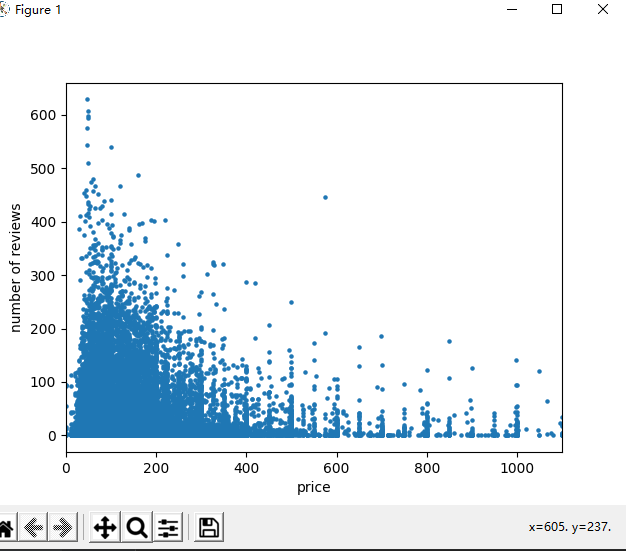












5.探索性数据分析（EDA）是数据科学项目的一个重要方面。您需要对上一节课中交给您的“FIFA世界杯”数据集进行简明而翔实的描述，并执行以下EDA任务：（a）找到数据的汇总统计数据，并对数据集中任何（单个）变量的统计数据进行简要说明，（b）找到平均值、中值，以及变量“合格球队”的模式，（c）为变量“获胜者”生成直方图图，以了解一个国家在1930年至2014年期间赢得世界杯的次数，（d）使用“热图”绘图，（e）绘制“目标得分”、“合格团队”、“匹配显示”变量的成对绘图，并总结绘图中的信息。注意：使用EDA任务结果和相应代码的屏幕截图清楚地报告您的结果。

a.

Year is which year hold this world-cup

Country is in where hold this world-cup

Winner is in this year world-cup which country is the champion

Runners-up is which country get the second

Third is which country get the third

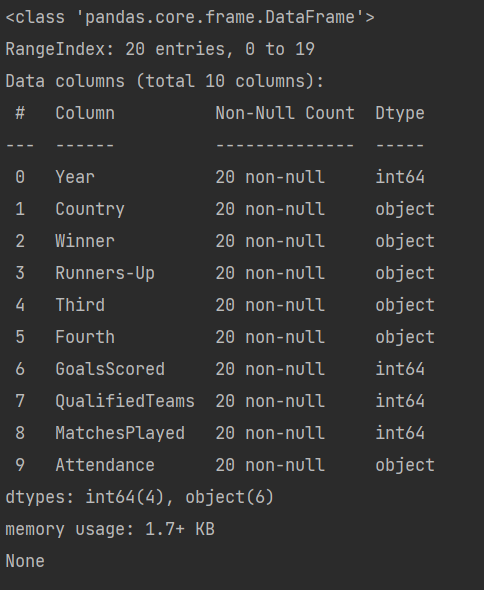
Fourth is which country get the fourth

GoalsScored is the whole score in this year

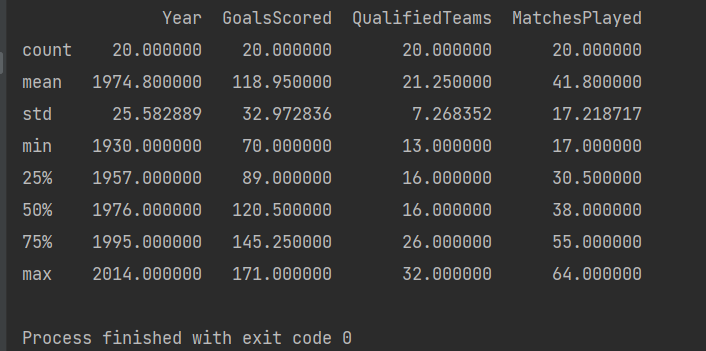
QualifiedTreams is how much team take part in this game

MathsPlayed is how many games in this year

Attendance is how many people come



b.



Mean 21.25

Median 16.0

Mode 16

c.

