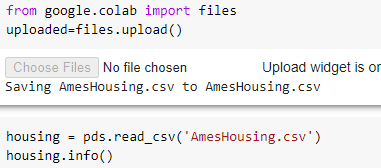
1. **Đọc dữ liệu**

Đầu tiên ta upload file hoặc gán đương link file vào lệnh Read()

Sử dụng phương thức read(size) để lấy về dữ liệu có kích thước bằng size. Nếu để trống tham số này thì nó sẽ đọc hết file hoặc nếu file quá lớn thì nó sẽ đọc đến khi giới hạn của bộ nhớ cho phép.

* Đọc file ở local host:



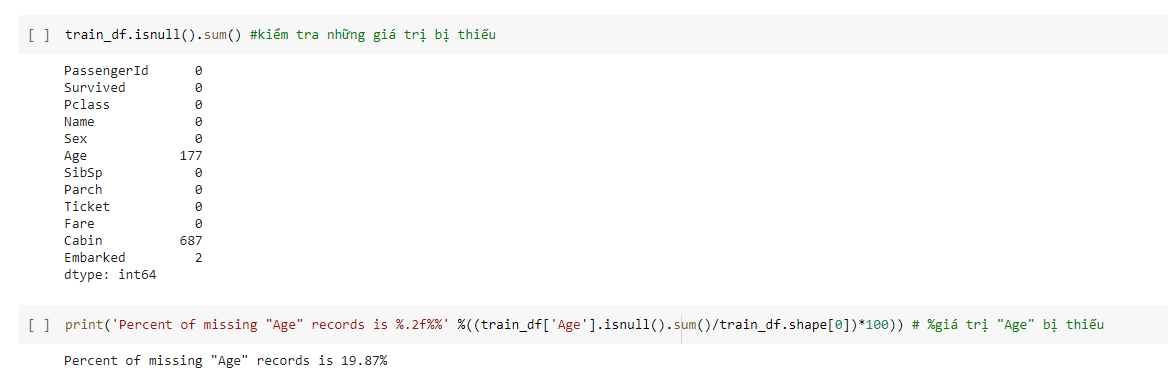
* Sử dụng data ở một driver khác

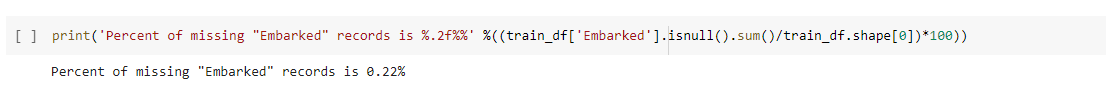
     train\_df = pd.read\_csv(‘URL’)

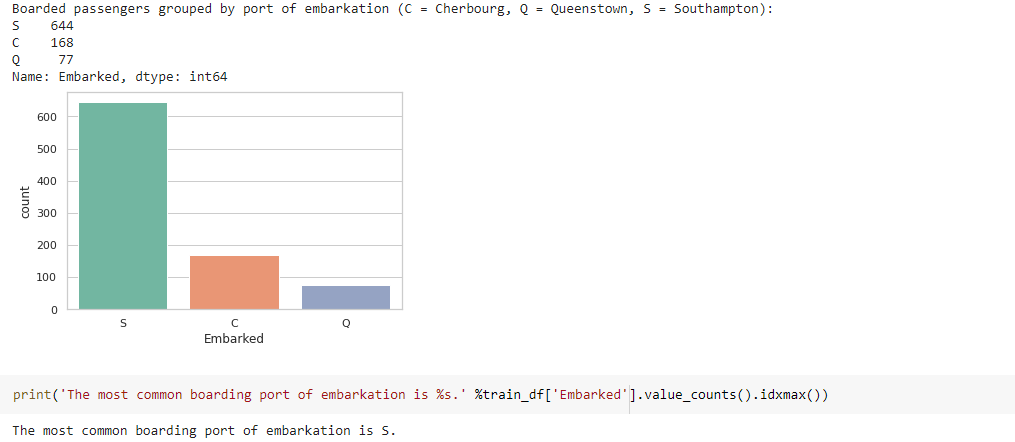
     test\_df = pd.read\_csv(‘URL’)

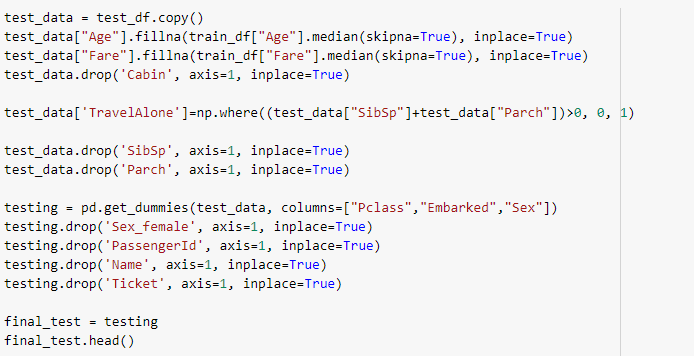
1. **Xử lý missing**

Đầu tiên chúng ta sử dụng lệnh data.isnull().sum() để xem dữ liệu bị missing. Sau đó tính phần trăm missing của dữ liệu để có thể thay thế bằng mean hoặc median còn nếu găp dữ liệu phân loại thì chọn dữ liệu xuất hiện nhiều nhất  để thay thế các dữ liệu  missing.









1. **Summary dữ liệu dùng các đại lượng thống kê mô tả**

Chúng ta sẽ sử dụng lệnh Describe () để xem các só liệu thống kê

* Count: là dữ liệu mẫu hoặc dữ liệu tổng
* Mean là trung bình
* Std là độ lệnh chuẩn là độ tương đối của điểm trên đồ thị
* Min: điểm nhỏ nhất
* 25% và 75%: Là tứ  vị ví dụ vị lớp 100 người bạn đứng 10 thì bạn hơn 89 người khác là bạn hơn 89%
* 50%:là trung vị của dữ liệu

Dữ liệu được chia làm 4 phần bằng nhau 25% để có thể dễ dàng tập trung phân tích vào các khoản dữ liệu mà mình muốn phân tích

Max là điểm cao nhất. Các điểm ngoài Min và Max là outlier

1. **Trực quan dữ liệu**

* Biến Định lượng thì  dùng biểu đồ histogram:

Histogram là biểu đồ áp dụng trên một biến liên tục nhằm tìm ra phân phối tần suất trong những khoảng giá trị được xác định trước của một biến.

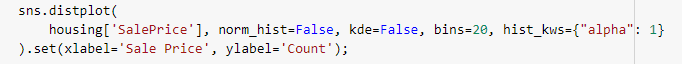
Có 2 cách tạo biểu đồ histogram theo các khoảng giá trị đó là:

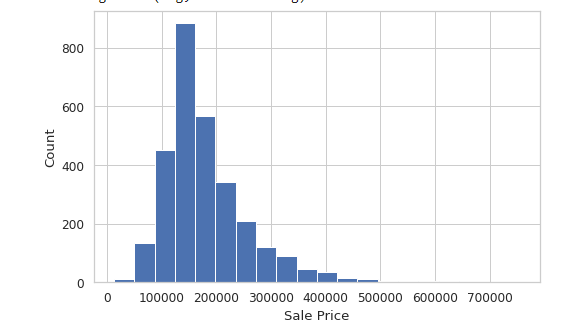
• Phân chia các khoảng giá trị có độ dài bằng nhau và độ dài được tính toán từ số lượng bins khai báo.

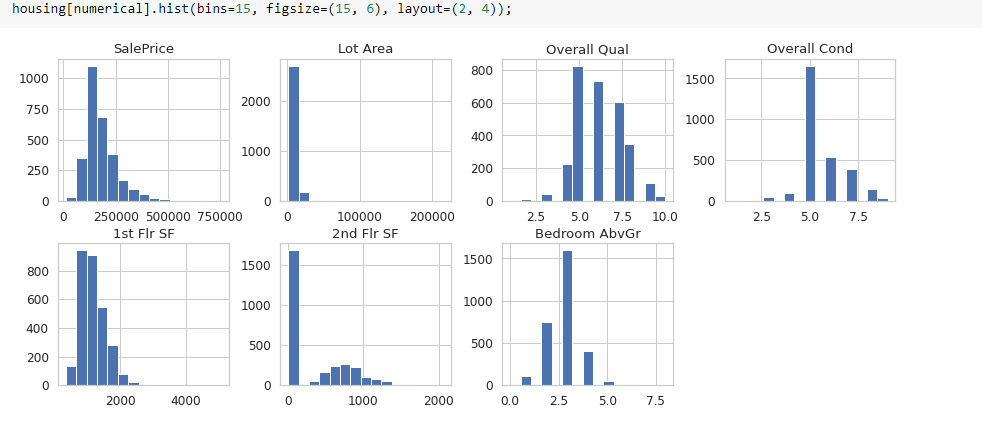
• Tự định nghĩa các khoảng giá trị dựa trên bins\_edge là các đầu mút của khoảng.

Plt.hist() dùng hàm này để vẽ biễu đồ histogram trong python

• Value distribution (histogram): Cho biết cách các giá trị trong cột của bạn được phân phối. Thanh càng cao, càng nhiều giá trị rơi vào một phạm vi.

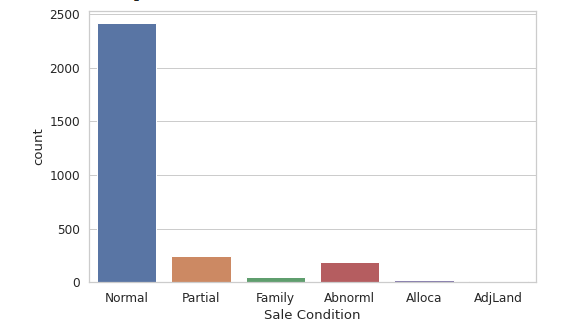






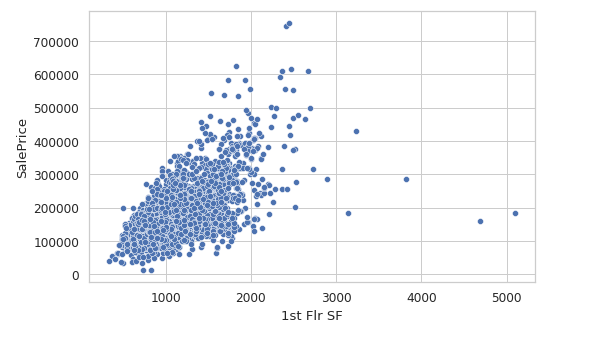
* Biểu đồ column hoặc biểu đồ bar để thể hiện các biến phân loại Sử dụng Countplot() dùng để biểu diễn các biến phân loại theo 2 trục là trục tung và trục hoành Sử dụng hàm sns.countplot() để vẽ biểu diễn biến phân loại.





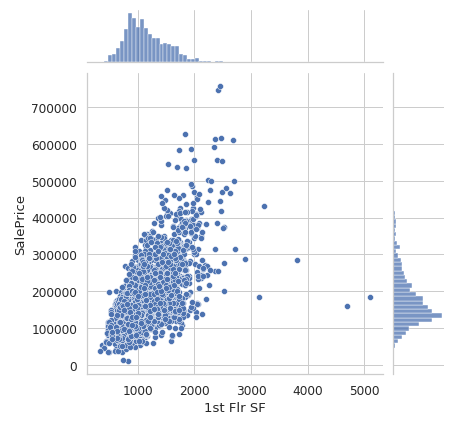
* Sự tương quan giữa 2 biến định lượng: Sử dụng scatterplot() để biển diễn sự quan hệ giữa 2 biến định lượng.





Ngoài ra còn có thể sử dụng biểu đồ phức tạp:





* Sự tương quan giữa 2 biến định lượng và biến phân loại Sử dụng boxplot() để biểu diễn sự tương quan giữa 2 biến định lượng và phân loại

Ở boxplot() biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max).

