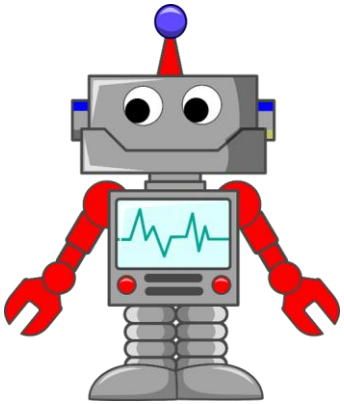
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

Data validation & EDA application

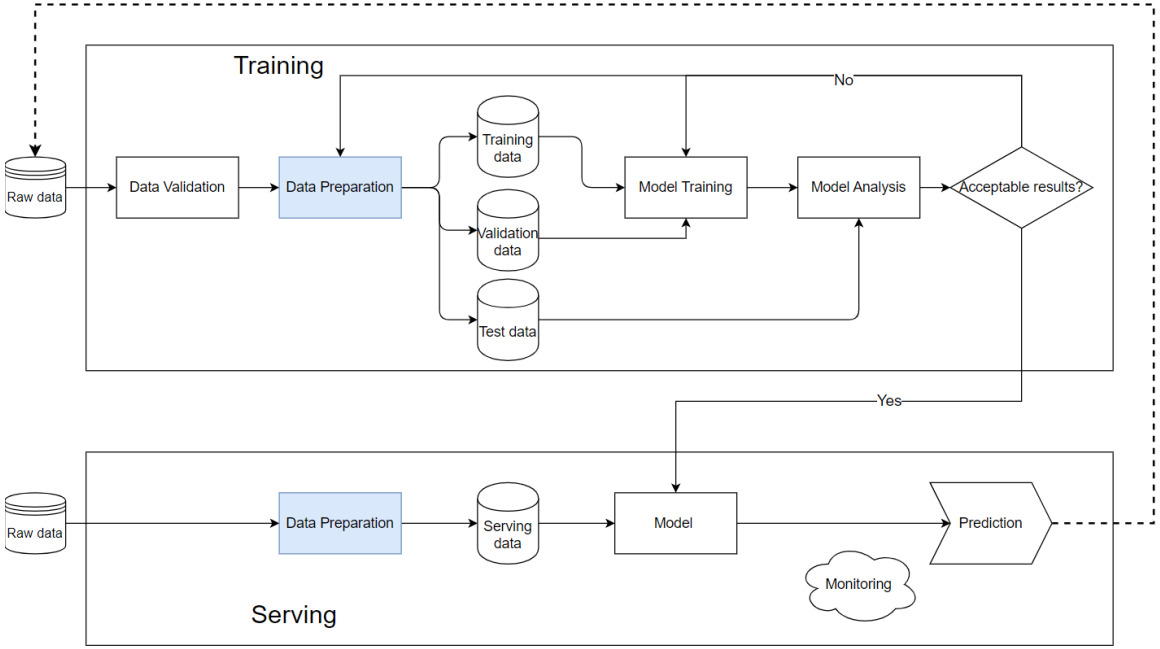


TS. Nguyễn Vinh Tiệp

NỘI DUNG

- Outlier detection
- EDA và ứng dụng EDA vào dữ liệu
- Automic EDA tools

ML Pipeline

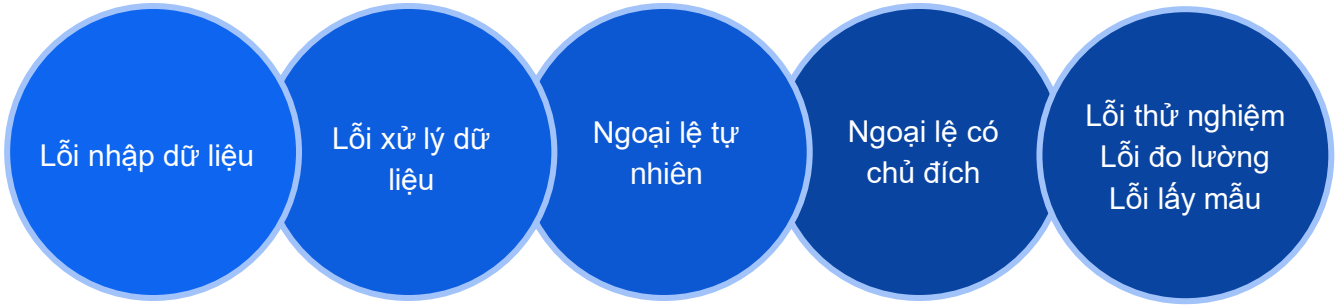


https://machinelearningcoban.com/tabml_book/ch_intro/pipeline.html

Data Validation

1. Phát hiện các ngoại lệ
2. Xử lý các ngoại lệ (ngoại bỏ, thay thế giá trị)

Ngoại lệ: Các kiểu ngoại lệ khác nhau



Lỗi của con người xảy ra trong quá trình thu thập, ghi hoặc nhập dữ liệu

Trích xuất dữ liệu từ nhiều nguồn (một số lỗi thao tác hoặc trích xuất)

Lỗi không phải do con người tạo ra

Điều này thường thấy trong các đo lường tự báo cáo liên quan đến dữ liệu nhạy cảm.

Loại ngoại lệ khác

Làm sạch dữ liệu - Ngoại lệ

- Có hai loại ngoại lệ:
  - Các ngoại lệ đơn biến: các điểm dữ liệu **có giá trị nằm ngoài phạm vi giá trị dự kiến**
  - Các ngoại lệ đa biến: có các ngoại lệ **phụ thuộc** vào **mối tương quan** giữa **hai biến**

Làm sạch dữ liệu: Hướng tiếp cận đơn giản

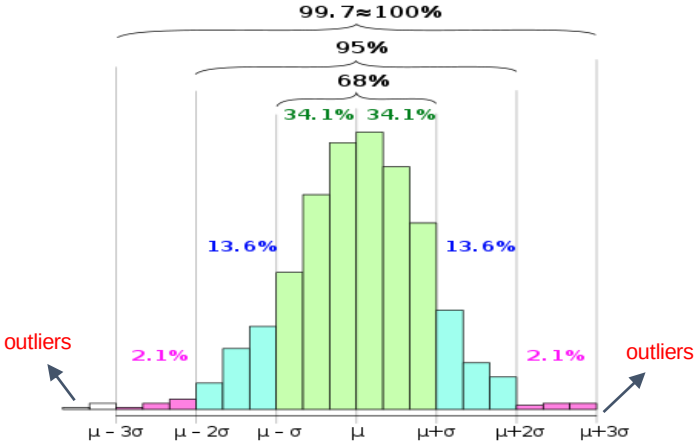
- Một số bước có thể được thực hiện để làm sạch dữ liệu:
  - Xóa các giá trị bị thiếu, trùng lặp, ngoại lệ và các hàng/cột không cần thiết
  - Lập lại chỉ mục và định dạng lại dữ liệu

```
# Drop rows with missing value
data.dropna(inplace = True)

# Remove duplicates
data.drop_duplicates()

# Drop unnecessary columns
data.drop(columns = [list cols], axis = 1)

# Drop/Filter unnecessary rows
data.drop([0, 1], inplace = True)
data[data[‘column_filter'] == ‘abc']
```

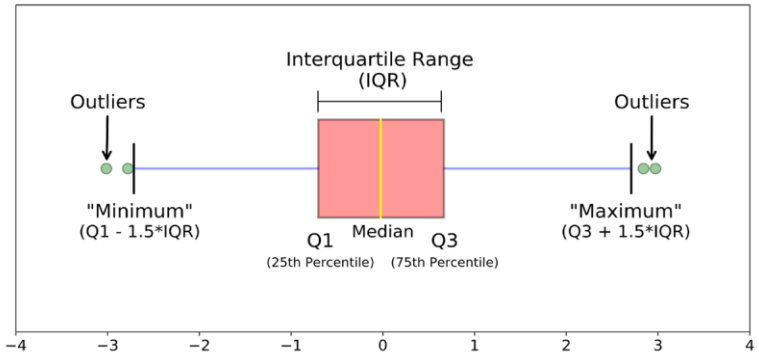
```
# Re indexing
data.set_index(‘column', inplace = True)
data.reset_index(drop = True)

# Re-formatting
data[‘column'] = data[‘column'].astype(int)

# Correcting inconsistent data
data[‘column'].replace(old_value, new_value, inplace =
True)
```

Dự đoán ngoại lệ như thế nào

- Phương pháp thống kê:
  - Phương pháp tính độ lệch chuẩn và trung bình: giới hạn để xác định các giá trị ngoại lệ (Gaussian hoặc Gaussian-like)
  - Phương pháp Interquartile Range (IQR): một thống kê tốt để tóm tắt mẫu dữ liệu phân phối không phải Gaussian



https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule



https://github.com/NaysanSaran/stats101/blob/master/

Tự động phát hiện ngoại lệ

- Tự động phát hiện ngoại lệ:
  - Yếu tố ngoại lệ cục bộ (Local Outlier Factor): Xác định các ngoại lệ là xác định vị trí của các mẫu ở xa mẫu khác
  - Rừng cách ly (Isolation Forest) : Thuật toán phát hiện bất thường dựa trên cây
  - Xác định hiệp phương sai tối thiểu (EllipticEnvelope): Tập dữ liệu theo phân phối chuẩn.
  - One-class SVM : Phát hiện ngoại lệ không được giám sát

Yếu tố ngoại lệ cục bộ - Local Outlier Factor (LOF)

- Thuật toán yếu tố ngoại lệ cục bộ có thể được chia thành bốn phần

Step 1

K-Distance and K-Neighbors

Nếu k nhỏ thì thuật toán trở nên nhạy cảm với nhiễu và nếu k lớn, nó có thể không nhận ra được các dị thường cục bộ.

Step 2

Reachability Distance

Biểu thị khoảng cách lớn nhất của hai điểm và khoảng cách-k của điểm thứ hai

Step 3

Local Reachability Density

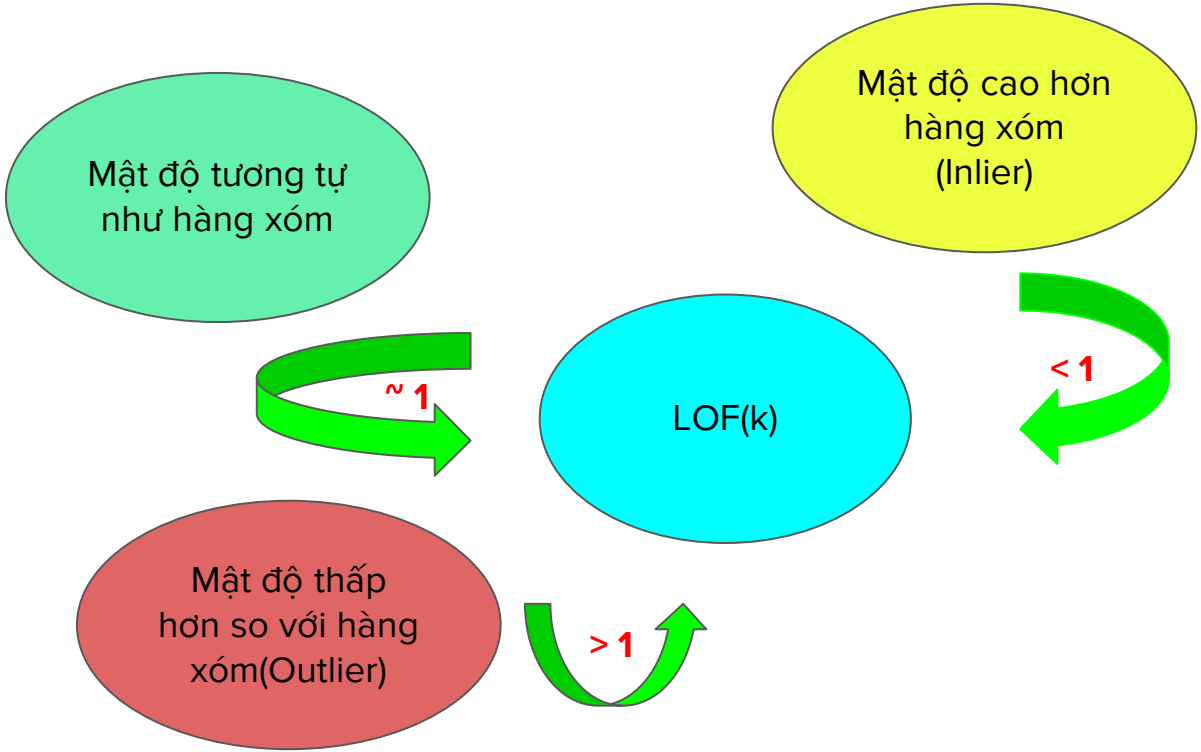
Đề cập đến việc chúng ta cần đi bao xa từ điểm hiện tại để đến điểm hoặc tập hợp điểm tiếp theo

Step 4

Local Outlier Factor Calculation

Mật độ khả năng tiếp cận cục bộ được tìm thấy được so sánh với mật độ khả năng tiếp cận cục bộ của k hàng xóm gần nhất

LOF: Làm thế nào để phát hiện các ngoại lệ?



Ví dụ

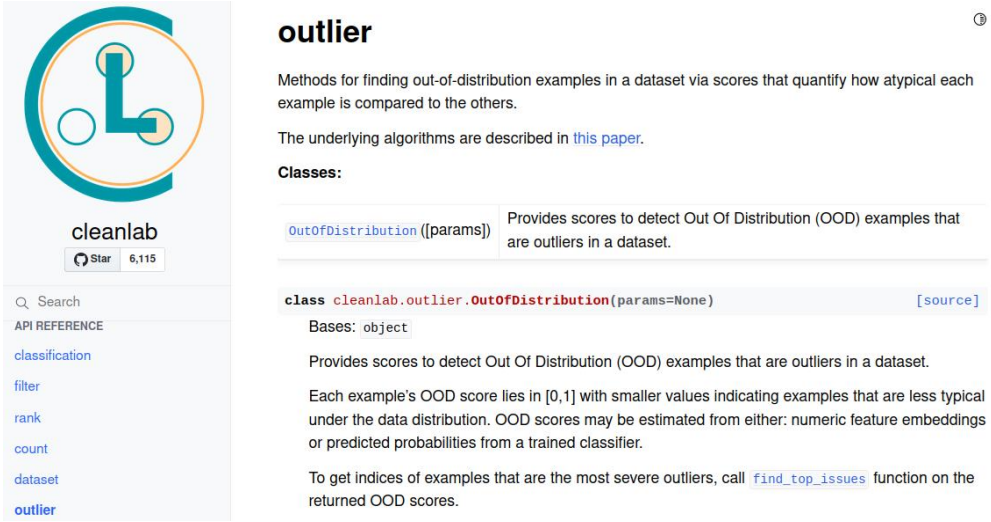
- Example
- LOF with sklearn

Cleanlab: Công cụ AI tập trung vào dữ liệu tiêu chuẩn

- github

Cleanlab: Tìm mẫu OOD (Out-Of-Distribution)

- Source code



Cleanlab: Tìm vấn đề về nhãn

- Phát hiện các vấn đề về dữ liệu: ngoại lệ, trùng lặp, lỗi nhãn,..

```
KNN = NearestNeighbors(metric='euclidean')
KNN.fit(X_processed.values)

knn_graph = KNN.kneighbors_graph(mode="distance")
```

```
data = {"X": X_processed.values, "y": labels}

lab = Datalab(data, label_name="y")
lab.find_issues(pred_probs=pred_probs, knn_graph=knn_graph)
```

```
Finding label issues ...
Finding outlier issues ...
Finding near_duplicate issues ...
Audit complete. 357 issues found in the dataset.
```

```
lab.report()
```

```
Here is a summary of the different kinds of issues found in the data:

    issue_type  num_issues
         label         294
       outlier          46
near_duplicate          17

Dataset Information: num_examples: 941, num_classes: 5


---------------------- label issues ----------------------

About this issue:
        Examples whose given label is estimated to be potentially incorrect
    (e.g. due to annotation error) are flagged as having label issues.
```

Data preparation

- Data fusion
- Data cleaning
- Data augmentation
- Data visualization
- Data splitting
- ...

Làm sạch dữ liệu – thiếu giá trị

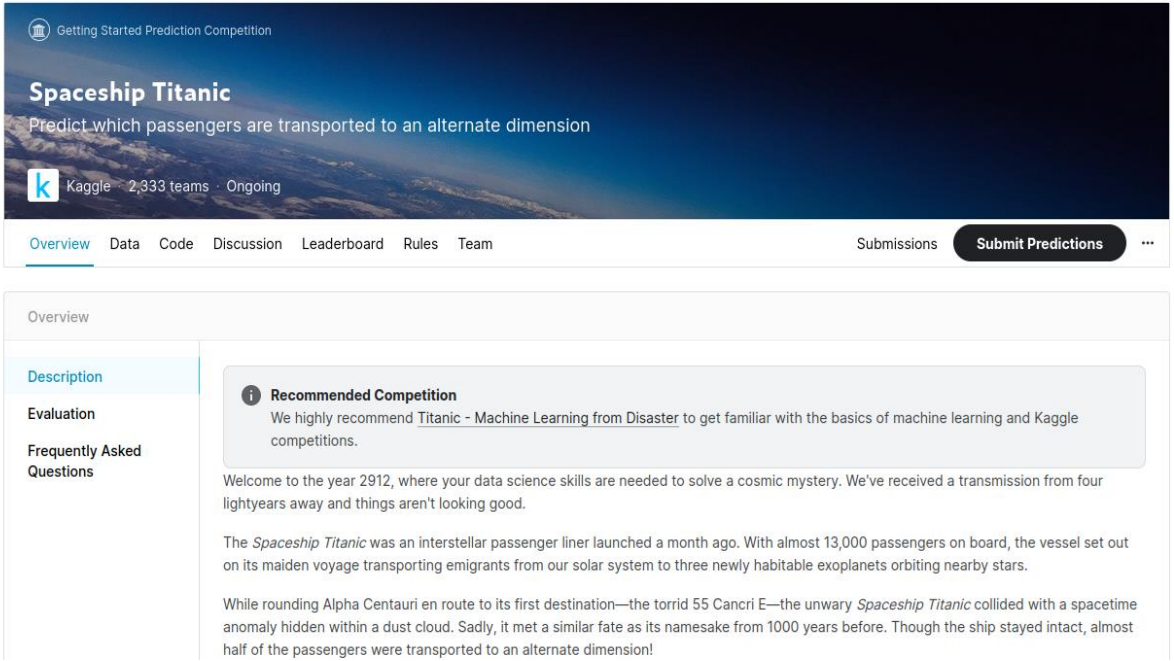
- Bộ dữ liệu chứa các **giá trị bị thiếu**, thường được mã hóa dưới dạng trống, NaN,..
- Phương pháp đơn giản:
  - Bỏ các cột có tỷ lệ thiếu giá trị cao (chẳng hạn như 80%)
  - Tự động điền giá trị còn thiếu
- Phương pháp khác

EDA

- Phân tích các đặc trưng chính của dữ liệu
- Mô tả bằng số liệu, biểu đồ thống kê hoặc trực quan hóa dữ liệu

EDA: Ví dụ về dữ liệu

- *Dataset link*



Data loading

Tải dữ liệu và in ra màn hình các dữ liệu đầu tiên

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Các loại dữ liệu

Có 2 loại dữ liệu chính:

- **Numerical** data: dữ liệu dạng số
- **Categorical** data: dữ liệu dạng phân loại

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

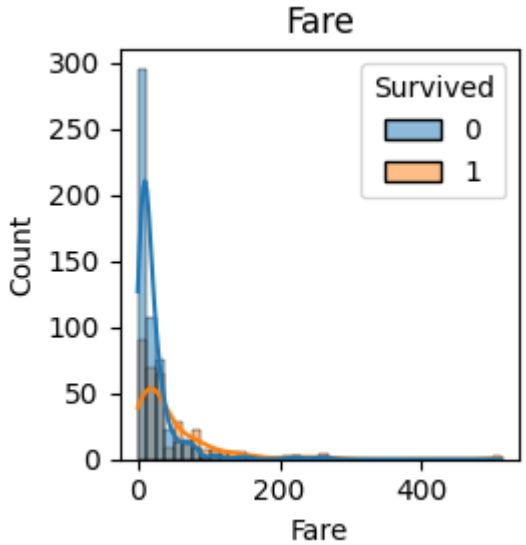
Kiểm tra các đặc trưng thống kê

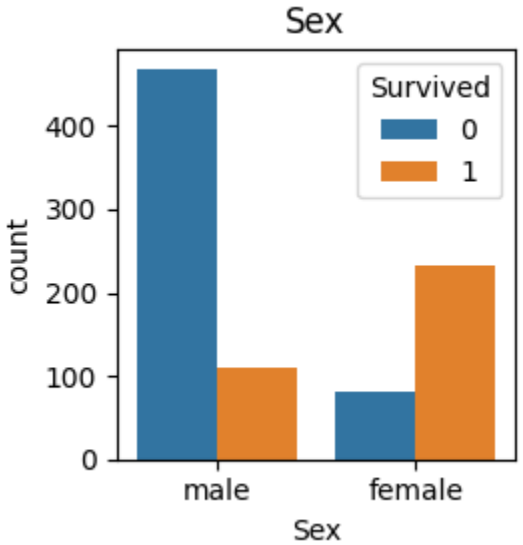
```
1  df_train.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Đồ thị thống kê

Sử dụng thư viện **matplotlib** hoặc **seaborn** để vẽ đồ thị





Các kiểu phân tích dữ liệu bằng cách thăm dò

- Có ba loại EDA chính
  - Phân tích đơn biến
  - Phân tích hai biến
  - Phân tích đa biến

Phân tích đơn biến

- Phân tích dữ liệu của chỉ một biến (đặc trưng/cột)
- Phân tích đơn biến không dùng biểu đồ
  - Trung tâm dữ liệu (Central Tendency): đề cập đến giá trị nằm ở vị trí trung tâm hoặc khu vực giữa của dữ liệu (các tham số ước lượng: trung bình, trung vị, mode).
  - Phạm vi: Sự **khác biệt** giữa giá trị **tối đa** và **tối thiểu** trong dữ liệu
  - Phương sai và độ lệch chuẩn

Phân tích đơn biến không dùng biểu đồ

- Tính trung bình, Độ lệch chuẩn, giá trị nhỏ nhất, lớn nhất và phân vị thứ nhất (first quartile), phân vị thứ ba (third quartile), trung vị của các cột số trong tập dữ liệu.

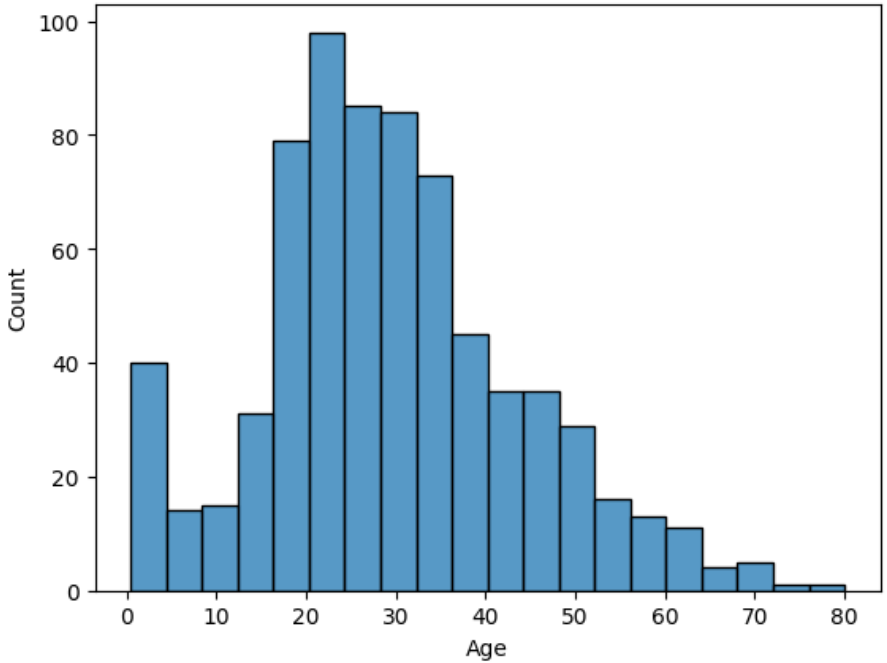
```
1  df_train[['Survived','Pclass','Age','SibSp','Fare']].describe()
```

	Survived	Pclass	Age	SibSp	Fare
count	891.000000	891.000000	714.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	32.204208
std	0.486592	0.836071	14.526497	1.102743	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	512.329200

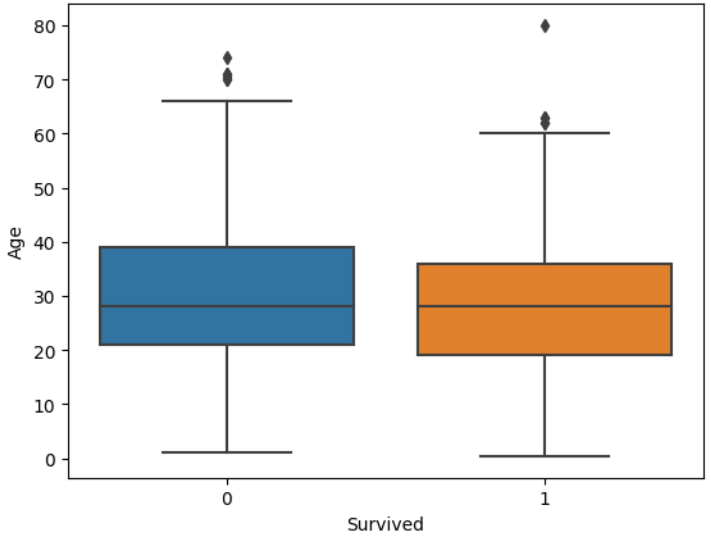
Phân tích đơn biến

- Phân tích đơn biến sử dụng biểu đồ
  - Biểu đồ tần suất (Histogram): Biểu đồ dạng thanh trong đó **tần số của dữ liệu** được biểu thị bằng các thanh hình chữ nhật
  - Biểu đồ mật độ: giống như một phiên bản **mượt mà hơn** của **biểu đồ tần suất**
  - Box-plot: Ở đây thông tin được thể hiện dưới dạng các hộp (**giá trị nhỏ nhất, phân vị thứ nhất (first quartile), trung vị (median), phân vị thứ ba (third quartile), giá trị lớn nhất**)

Histogram & Box plot

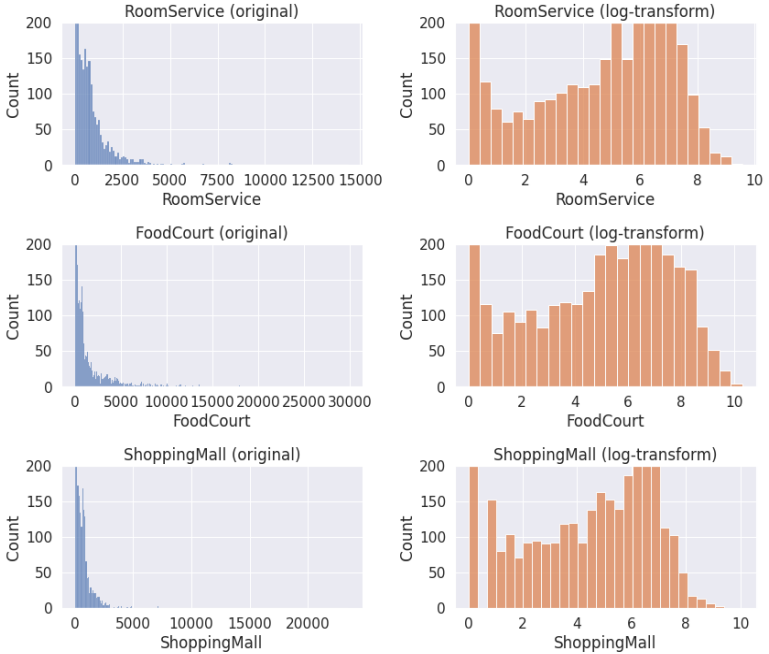


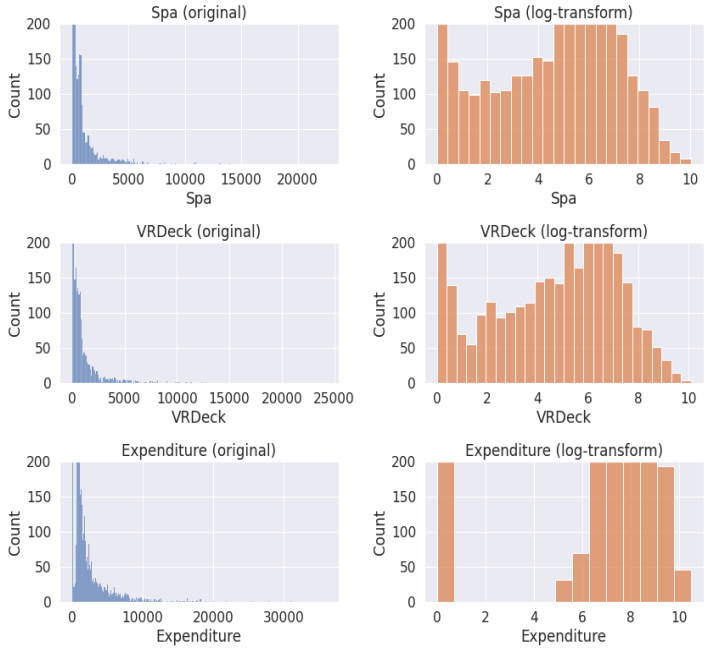
Biểu đồ tần suất về tuổi



Box-plot của tuổi

Ví dụ về phân tích đơn biến





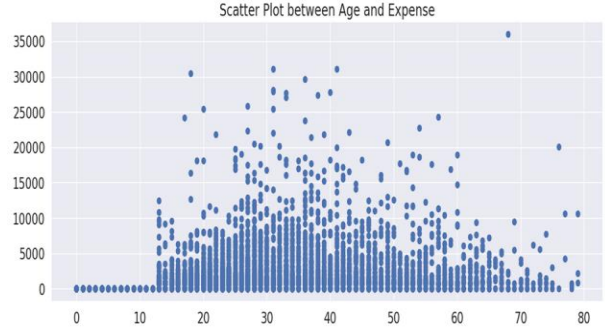
Phân tích hai biến

- Phân tích hai biến: xác định xem có tồn tại mối liên hệ thống kê giữa hai biến hay không
- Có ba loại chính:
  - Phân tích dạng số-số
  - Phân tích dạng số-phân loại
  - Phân tích dạng phân loại-phân loại

Phân tích dạng số - số

- Khi cả hai biến được so sánh đều có dữ liệu số
- Một số phương pháp trực quan có thể được sử dụng:
  - Biểu đồ phân tán (Scatter plot): được sử dụng để thể hiện mọi điểm dữ liệu trong biểu đồ
  - Biểu đồ cặp (Pair plot)
  - Ma trận tương quan (Correlation matrix)

Ma trận tương quan kết hợp bản đồ nhiệt



Biểu đồ phân tán giữa tuổi và chi tiêu

Phân tích dạng số - phân loại

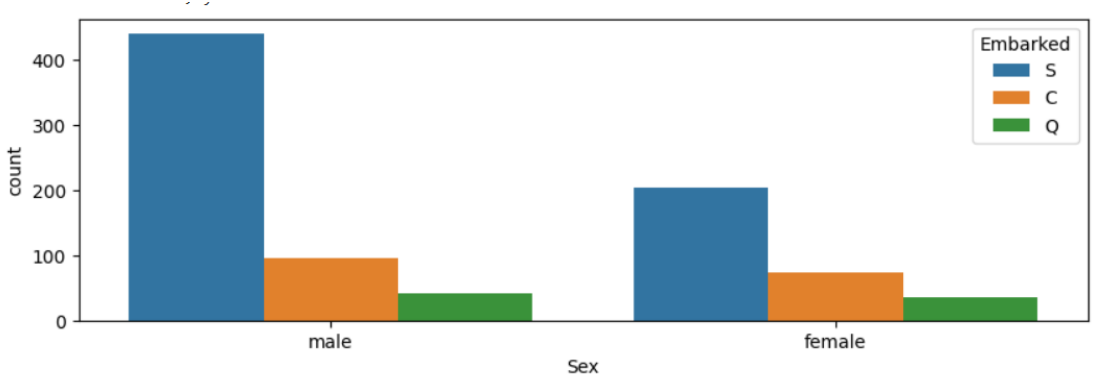
- Khi một biến có kiểu số và biến khác là biến phân loại
- Bạn có thể nhóm lại để sắp xếp dữ liệu thành các nhóm tương tự. Các hàng có cùng giá trị trong một cột cụ thể sẽ được sắp xếp thành một nhóm với nhau

```
df_train[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare'\
    ,'Embarked']].groupby(['Survived', 'Embarked']).mean()
```

Survived	Embarked	Pclass	Age	SibSp	Parch	Fare
0	C	2.200000	33.666667	0.253333	0.253333	35.443335
	Q	2.936170	30.325000	0.510638	0.276596	13.335904
	S	2.545667	30.203966	0.611241	0.348946	20.743987
1	C	1.634409	28.973671	0.494624	0.451613	79.720926
	Q	2.866667	22.500000	0.300000	0.000000	13.182227
	S	1.967742	28.113184	0.493088	0.539171	39.547081

Phân tích dạng phân loại – phân loại

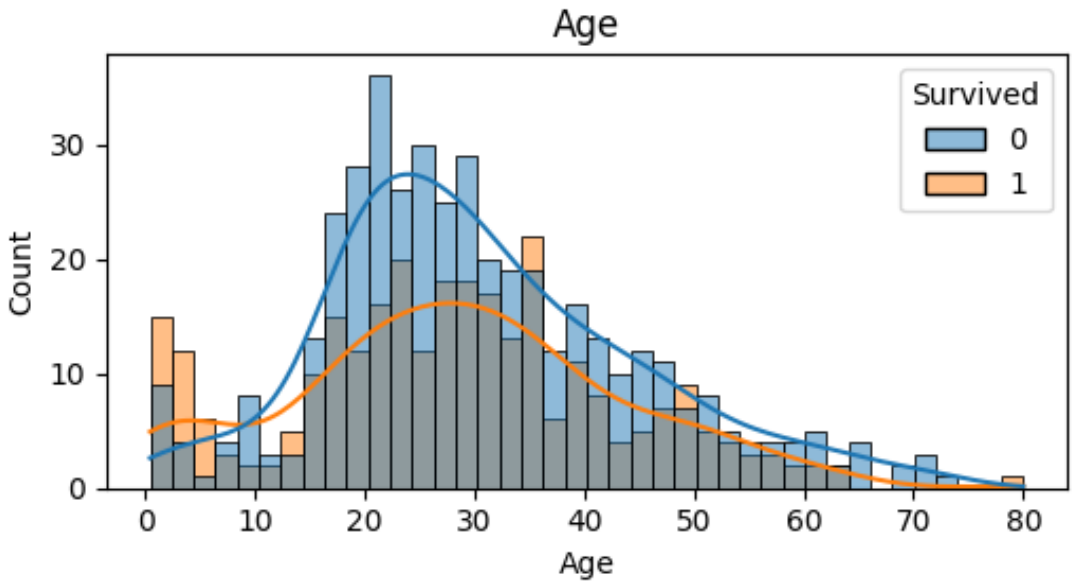
- Khi cả hai biến đều có tính phân loại
- Một số phương pháp trực quan có thể được sử dụng:
  - Biểu đồ cột xếp chồng (Stacked Bar Chart hay Segmented Bar Chart)
  - Biểu đồ cột tạo cụm (Cluster Bar Chart)



Phân bố của giới tính và cảng khởi hành

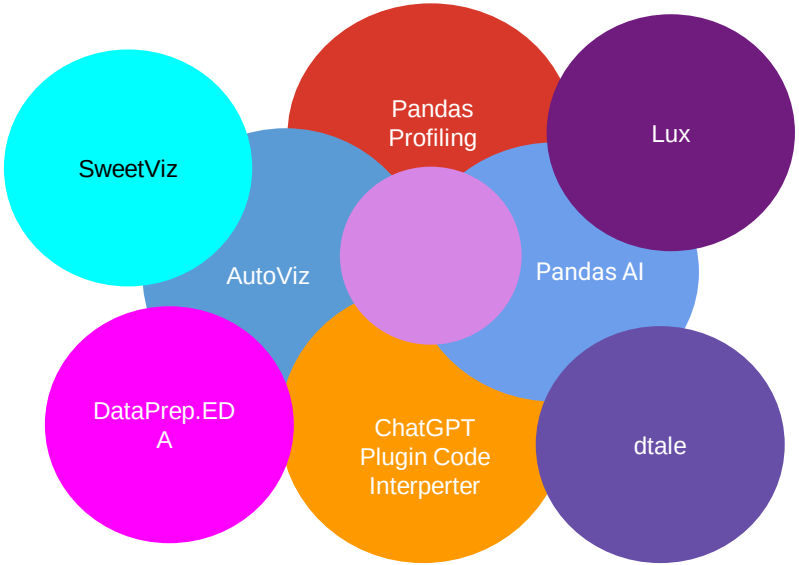
EDA- Age vs Survived

```
sns.histplot(df_train, x='Age', hue='Survived', bins=40, kde=True)
```



Một số công cụ Automatic EDA

- Comparison notebook



Ưu điểm của Automatic EDA tools

01	Hiệu quả	• Phân tích dữ liệu nhanh chóng và tạo ra cái nhìn sâu sắc • Cho phép Nhà khoa học dữ liệu dành ít thời gian hơn cho việc khám phá sơ bộ
02	Tính toàn diện	• Quét qua mọi thành phần trong bộ dữ liệu • Góc nhìn toàn diện và đảm bảo rằng không thành phần nào bị bỏ sót.
03	Sự tiện lợi	• Tóm tắt và trực quan hóa mà không yêu cầu mã hóa rõ ràng cho từng cái
04	Kiểm tra chất lượng dữ liệu	• Kiểm tra chất lượng dữ liệu: thiếu giá trị, ngoại lệ, không nhất quán
05	Hiểu rõ hơn về dữ liệu	• Giúp hiểu được cấu trúc cơ bản của dữ liệu, mối tương quan và xu hướng trong dữ liệu.

Pandas Profiling

Overview

Overview | Warnings 14 | Reproduction

Dataset statistics

Number of variables	12
Number of observations	891
Missing cells	866
Missing cells (%)	8.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	83.7 KiB
Average record size in memory	96.1 B

Variable types

Numeric	5
Categorical	7

Pandas profiling

Tổng quan

Overview

Overview | Warnings 14 | Reproduction

Warnings

`Name` has a high cardinality: 891 distinct values	High cardinality
`Ticket` has a high cardinality: 681 distinct values	High cardinality
`Cabin` has a high cardinality: 147 distinct values	High cardinality
`Age` has 177 (19.9%) missing values	Missing
`Cabin` has 687 (77.1%) missing values	Missing
`PassengerId` is uniformly distributed	Uniform
`Name` is uniformly distributed	Uniform
`Ticket` is uniformly distributed	Uniform
`Cabin` is uniformly distributed	Uniform
`PassengerId` has unique values	Unique
`Name` has unique values	Unique
`SibSp` has 608 (68.2%) zeros	Zeros
`Parch` has 678 (76.1%) zeros	Zeros
`Fare` has 15 (1.7%) zeros	Zeros

Pandas profiling

Thiếu giá trị

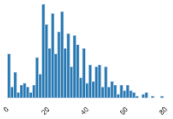
Age

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct	88
Distinct (%)	12.3%
Missing	177
Missing (%)	19.9%
Infinite	0
Infinite (%)	0.0%

Mean	29.69911765
Minimum	0.42
Maximum	80
Zeros	0
Zeros (%)	0.0%
Memory size	7.1 KiB



Toggle details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

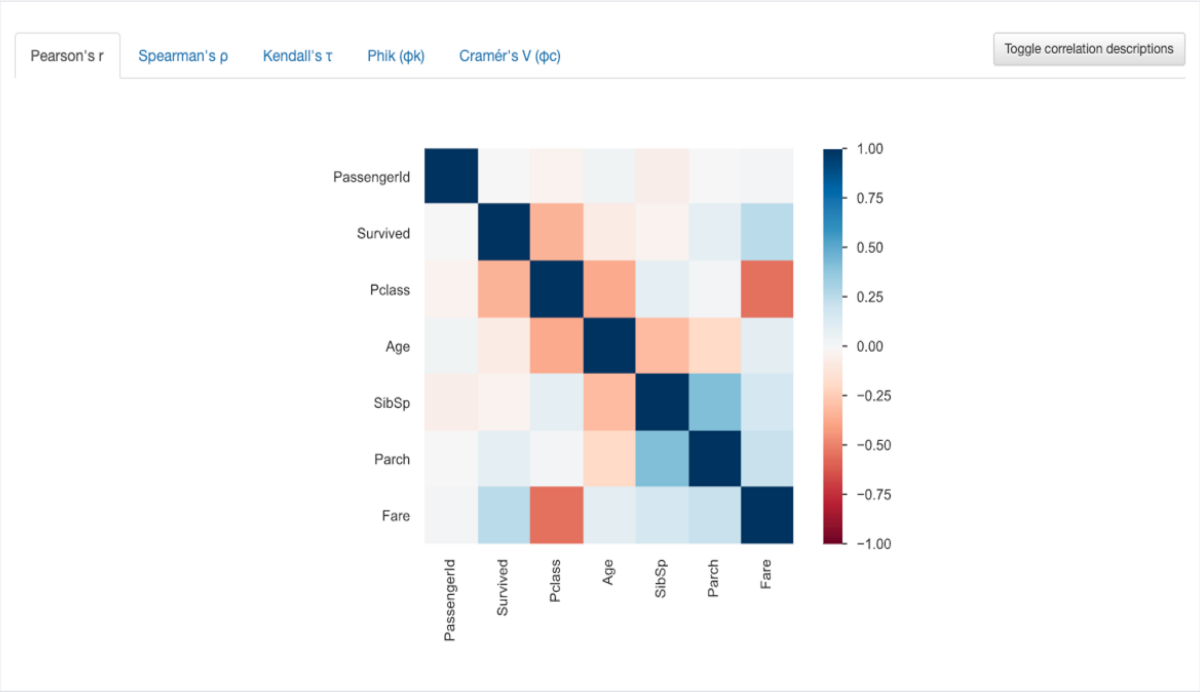
Minimum	0.42
5-th percentile	4
Q1	20.125
median	28
Q3	38
95-th percentile	56
Maximum	80
Range	79.58
Interquartile range (IQR)	17.875

Descriptive statistics

Standard deviation	14.52649733
Coefficient of variation (CV)	0.4891221855
Kurtosis	0.1782741536
Mean	29.69911765
Median Absolute Deviation (MAD)	9
Skewness	0.3891077823
Sum	21205.17
Variance	211.0191247
Monotocity	Not monotonic

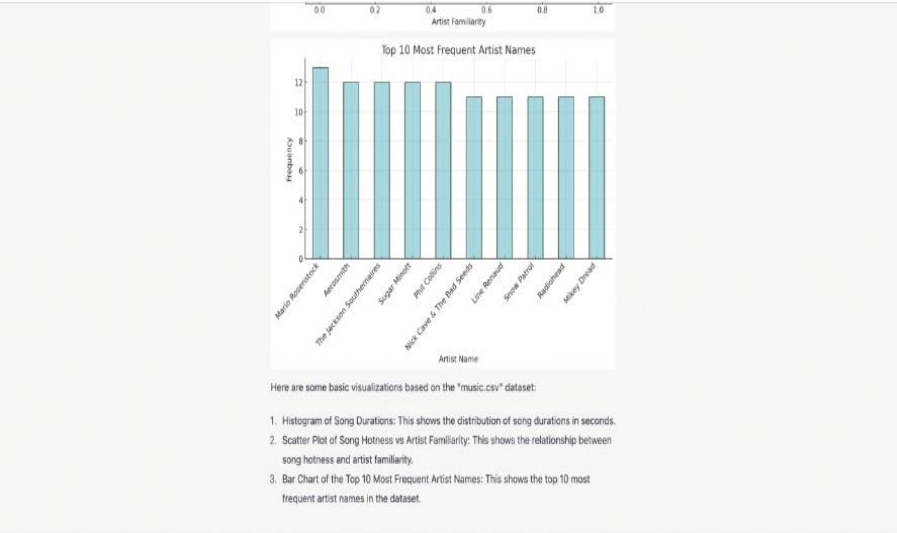
Pandas profiling

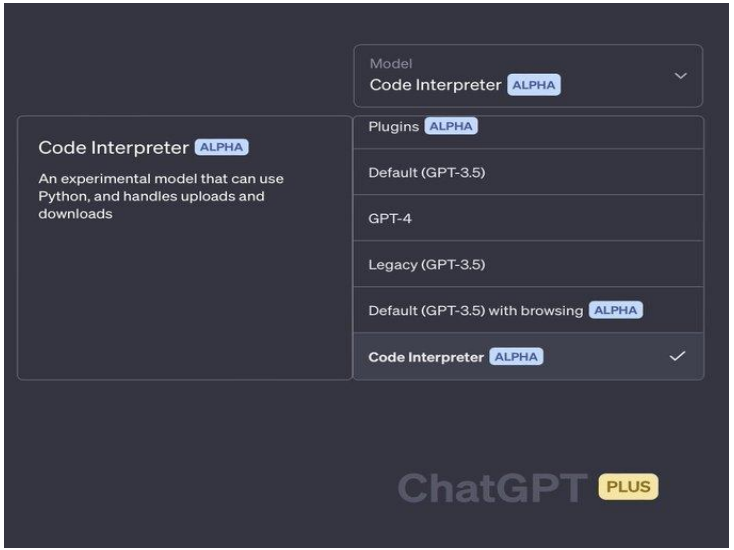
Sự tương quan



Pandas profiling
Trình thông dịch

- Thực hiện phân tích và trực quan hóa dữ liệu





Pandas AI

- Pandas AI là một thư viện Python bổ sung các khả năng AI tổng quát cho Pandas
- Làm cho Pandas có thể trò chuyện, cho phép bạn đặt câu hỏi về dữ liệu của mình
- Github: link

Installation

```
pip install pandasai
```

HỎI ĐÁP