Stat 130

September 12, 2024

[18]:

```
import pandas as pd
url = "https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.
 ↪csv"
df = pd.read_csv(url)
df.isna().sum()
```

[18]:

```
Survived                   0
Pclass                     0
Name                       0
Sex                        0
Age                        0
Siblings/Spouses Aboard    0
Parents/Children Aboard    0
Fare                       0
dtype: int64
```

[19]:

```
# Number of rows and columns
print(df.shape)

# First few rows of the dataset
print(df.head())
```

```
(887, 8)
   Survived  Pclass                                               Name  \
0         0       3                             Mr. Owen Harris Braund
1         1       1  Mrs. John Bradley (Florence Briggs Thayer) Cum…
2         1       3                              Miss. Laina Heikkinen
3         1       1        Mrs. Jacques Heath (Lily May Peel) Futrelle
4         0       3                            Mr. William Henry Allen

      Sex   Age  Siblings/Spouses Aboard  Parents/Children Aboard     Fare
0    male  22.0                        1                        0   7.2500
1  female  38.0                        1                        0  71.2833
2  female  26.0                        0                        0   7.9250
3  female  35.0                        1                        0  53.1000
4    male  35.0                        0                        0   8.0500
```

[20]:

```python
import pandas as pd

# Load the dataset
url = "https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.
 ↪csv"
df = pd.read_csv(url)

# General summary for numerical columns
print("Summary statistics for numerical columns:")
print(df.describe())

# Summary for categorical columns
print("\nCounts for categorical columns:")
print(df['Pclass'].value_counts())
print(df['Sex'].value_counts())
print(df['Siblings/Spouses Aboard'].value_counts())
print(df['Parents/Children Aboard'].value_counts())

# Data types of each column
print("\nData types of each column:")
print(df.dtypes)
```

```
Summary statistics for numerical columns:
         Survived      Pclass         Age  Siblings/Spouses Aboard  \
count  887.000000  887.000000  887.000000               887.000000
mean     0.385569    2.305524   29.471443                 0.525366
std      0.487004    0.836662   14.121908                 1.104669
min      0.000000    1.000000    0.420000                 0.000000
25%      0.000000    2.000000   20.250000                 0.000000
50%      0.000000    3.000000   28.000000                 0.000000
75%      1.000000    3.000000   38.000000                 1.000000
max      1.000000    3.000000   80.000000                 8.000000

       Parents/Children Aboard       Fare
count               887.000000  887.00000
mean                  0.383315   32.30542
std                   0.807466   49.78204
min                   0.000000    0.00000
25%                   0.000000    7.92500
50%                   0.000000   14.45420
75%                   0.000000   31.13750
max                   6.000000  512.32920

Counts for categorical columns:
Pclass
3    487
1    216
```

```
2    184
Name: count, dtype: int64
Sex
male      573
female    314
Name: count, dtype: int64
Siblings/Spouses Aboard
0    604
1    209
2     28
4     18
3     16
8      7
5      5
Name: count, dtype: int64
Parents/Children Aboard
0    674
1    118
2     80
5      5
3      5
4      4
6      1
Name: count, dtype: int64

Data types of each column:
Survived                     int64
Pclass                       int64
Name                        object
Sex                         object
Age                        float64
Siblings/Spouses Aboard      int64
Parents/Children Aboard      int64
Fare                       float64
dtype: object
```

2.2 general definitions of the meaning Survived - Whether the passenger survived (0 = No, 1 = Yes) Pclass - Passenger class (1st, 2nd, or 3rd) Name - Name of the passenger Sex - Gender of the passenger Age - Age of the passenger Siblings/Spouses Aboard - Number of siblings or spouses aboard Parents/Children Aboard - Number of parents or children aboard Fare - Fare paid by the passenger

5, The difference between an “attribute” such as df. shape which does not end with ( ) and a “method”, such as df. describe() which does end with () Attribute’s definition is something that is a property of an object and it stores some kind of information about the object. And when using an attribute it does not require any ( ) since it is simole data stored with in the object. Method’s definition is a function that is associated with an object and it performes action or calculates based on the objects data. and methodmust need ( ) since they are functions that are needed to be excluded. The difference between these two is attribute stores information and no parentheses needed and on the other side method is a function that performs an operation and requires parentheses.

https://chatgpt.com/share/6e872396-74bf-44c7-bb98-a127fda61a49

https://chatgpt.com/share/a37648f2-b481-4451-9b44-67913864d6f5

https://chatgpt.com/share/006757d6-91f0-49aa-8dfb-420941240f

post lecture HW

Post Lecurtre HW 6 Cout The number of valid data point that exist in the column Mean The average of all data values It is operated by summing all value and then dividing by the number of non-null entries. STD a Mearure of how many amout of varitation or dispersion in a set of values. a higher standard divition means that the values are spread over a wider range. Min The smallest value in a set of data. 25% the values that are below 25% of the data. it repersents the first quartile. 50% the middle value of the data sets. and if the data set is evenly distributed, it divides the sataset into two equal halves. 75% The value below which 75%of the data falls ot represents the upper boundary of the interquartiles range Max The largest value in the data set.

https://chatgpt.com/share/66e34bff-04bc-8007-a8e2-cdae6313565c

7.1 When using df. dropna (subset=[’Age’l) The action of this is removing only the rows with missing values. the benefit of using df. dropna is preserves the rest of the datasets retaining valuble information in other columns However When using del dfl’Age’l The problem is it will losses all data in the “age” column which might be important for analysis.

7.2 for example when simplifying a data set by removing an irrelevent column ticket which does not contribute to the data set. Using del df [‘Ticket’] can completely remove the Ticket column however df. dropna() does not get the action done.

7.3 When appling del dfI’col’l before df. dropna() can ensure efficency and avoid unessary row removal and improve data quality by focesing more on the more important data sets.

7.4 For example Ticket is the useless column First i will use del dfl ’Ticket’l to remove the column and then i will use df. dropna() to remove rows with any missing values in the remaining columns.

8.1 groupby (“col1”) This groups the data based on the unique values in Coll [“col2”] specificies the column for which you want to calculate summary statistics describe(): calculates summeries of statistics.

8.2 df. describe()shows a higher level of overview of the missing data and df. groupby (“col1”) [“col2”]. describe() shows how data exists within each group.

8.3 When using chatGpt is way faster than doing a google search since chat gpt will tell you whats wrong about the code and youtube will give you links that might help you with your code or might not.

https://chatgpt.com/share/66e36ac1-79a0-8007-bbd2-8bldc78a1f10

Chat GPT 8 ABCDEFG https://chatgpt.com/share/66e36e2e-546c-8007-b63c-a78d1b47bcce

Yes