MachineLearning-CAP5610

UNIVERSITY OF CENTRAL FLORIDA | DATA ANALYTICS

assignment-1

bhuvanesh jeevarathinam

2020

Task 1

Attributes

|  |  |  |
| --- | --- | --- |
| Attributes | Type 1 | Type 2 |
| Rating of an Amazon product by a person on a scale of 1 to 5 | Discrete | Ordinal |
| The Internet Speed | Continuous | Interval |
| Number of customers in a store. | Discrete | Nominal |
| your Student ID | Discrete | Nominal |
| Distance | Continuous | Ratio |
| your letter grade (A, B, C, D) | Discrete | Ordinal |
| The temperature in the campus | Continuous | Interval |

Task 2

Distance/Similarity Measures

**Proximity Measure for Shape:**

We can use *ratio of (Length / width) for* determining the shape of the boxes. For eg: Rectangular boxes will have a different ratio of length to width whereas the Square boxes will have a similar ratio. In the given problem., The ratio of boxes with vector scale (1,1) and (3,3) are 1, whereas the ratio of boxes with vector sale (1,2) and (3,6) is not 1 but it is ½. Hence from this proximity measure we can identify that the group shape each box belongs.

**Proximity Measure for Size:**

We can use *ratio of (Length\*width)* for determining the size of the boxes. Size of a box is usually represented with the are occupied by the box. Area is determined basically by the Product of Length and its width. Hence in the given vector the Product of the elements in the vector will give us the area/Size occupied by the box. When we compare the ratio of the area of different boxes, we can get the classification for similar box size. For e.g.: Box with vector (1,1) has area = 1 unit whereas Box with vector (3,3) has area of 9 units. Hence the ratio of them is not equal to 1 which proves that the area of the two boxes are not similar. Hence, they both belong to different groups.

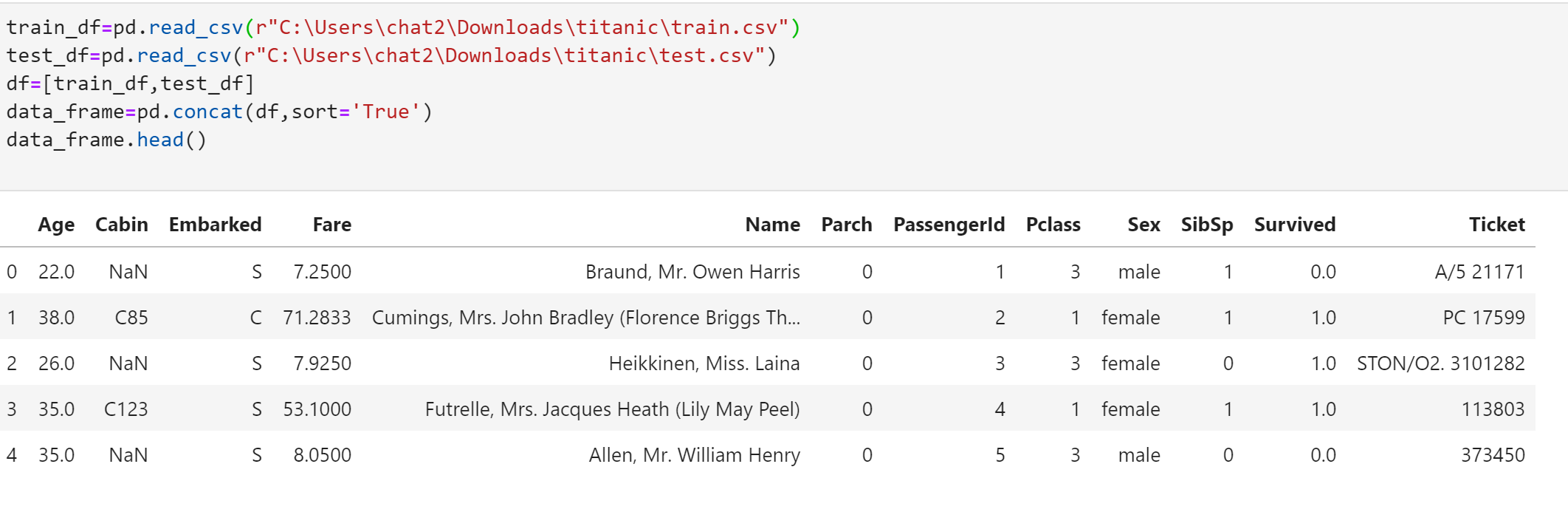
Task 3 - Data Preprocessing of Titanic – Part 1

I have pushed my code to a repository named CAP5610-MachineLearning. Kindly please access the code from the repo link provided.

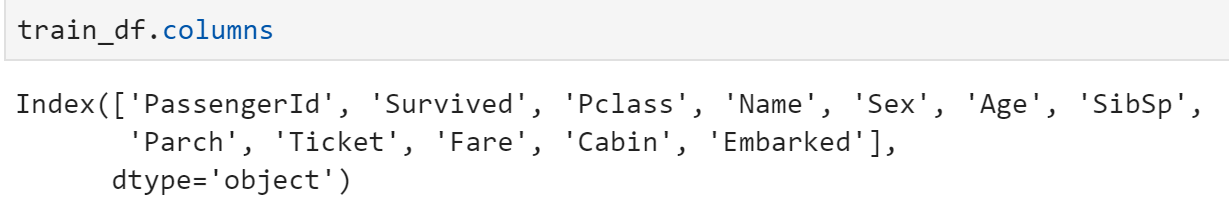
Codelink:

https://github.com/bhuvaneshkj/CAP5610-MachineLearning-Assignment.git

Subtask 1: Analyze by describing data



Q1: Which features are available in the dataset?



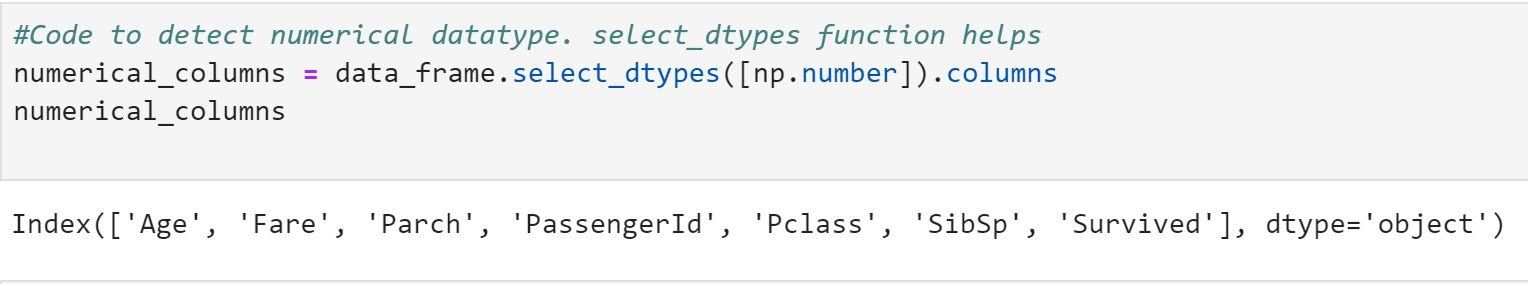
Features which are available in dataset are PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked respectively.

Q2: Which features are categorical?



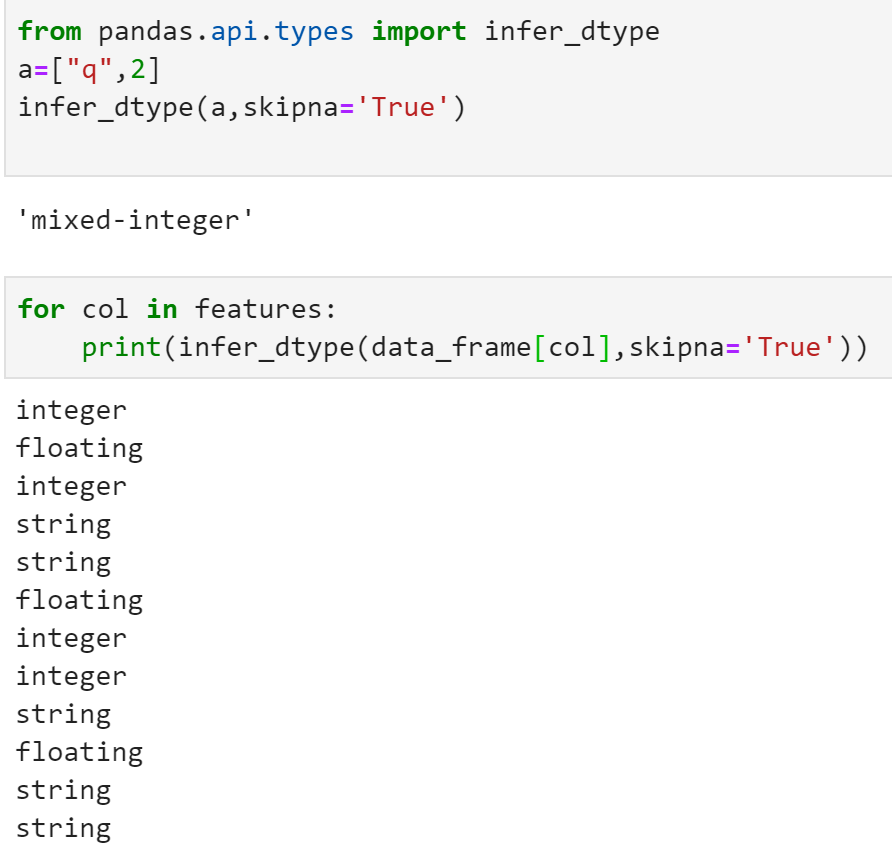
From the Dataset we can find that each feature has its respective unique value counts. From the observation we can see that Features like Survived, Pclass, Sex, SibSp, Parch, Embarked etc has unique value count of less than 10. For a feature to be considered Categorical it should have a count which can be in a countable and measure level. Features like Cabin have 186 unique value which cannot be considered as categorical as 186 categories of cabins cannot exist.

Q3: Which features are numerical?



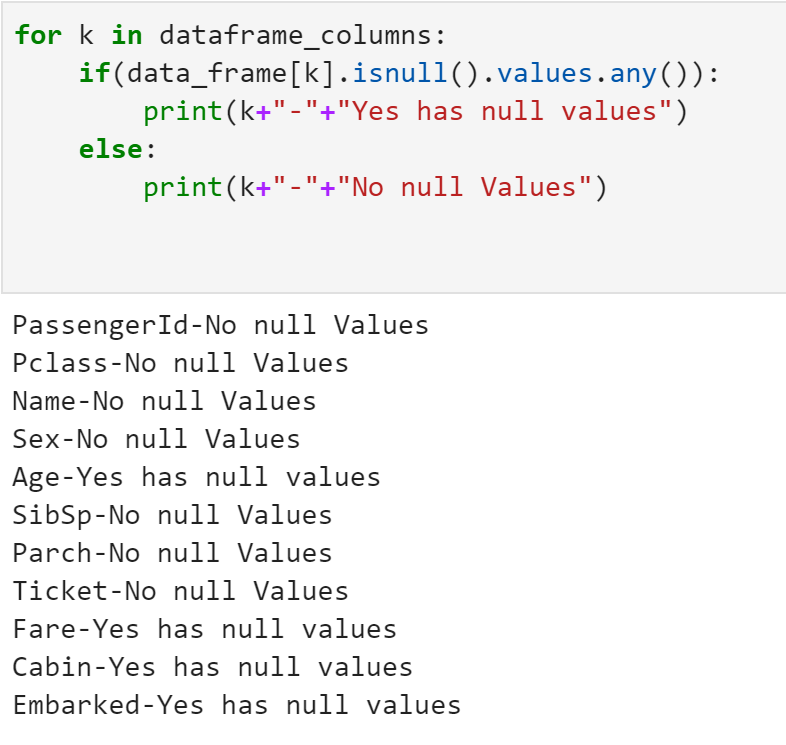
Features such as PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare etc are numerical columns.

Q4: Which features are mixed data types?



We tend to find the mixed data types in columns of dataset by using a function in Pandas Api which checks for it. Here we see that none of the features have mixed datatype.

Q5: Which features contain blank, null or empty values?



We can infer from code that features like Age, Fare, Cabin, Embarked etc. has null values.

Q6: What are the data types (e.g., integer, floats or stringsfor various features?



Features and its Dataypes Listed:

Passenger – Integer

Survived – float

Pclass – Integer

Name – Object

Sex – Object

Age – Float

SibSp - Int

Parch – Int

Ticket – Object

Fare – Float

Cabin – Object

Embarked – Object

Q7: To understand what is the distribution of numerical feature values across the samples, please list the properties (count, mean, std, min, 25% percentile, 50%percentile, 75% percentile, max) of numerical features?



The above tabular description shows us the Statistics of the Titanic Training Dataset.

Q8: To understand what is the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please the properties (count, unique, top, freq) of categorical features?

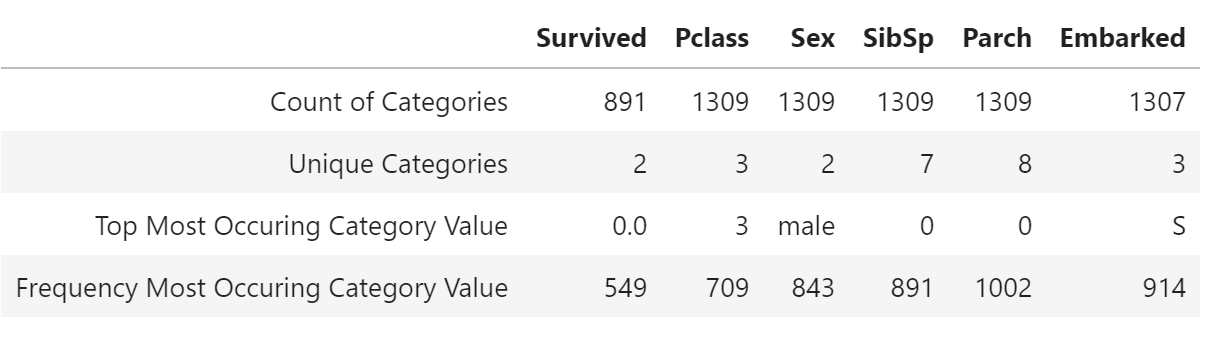
To find:

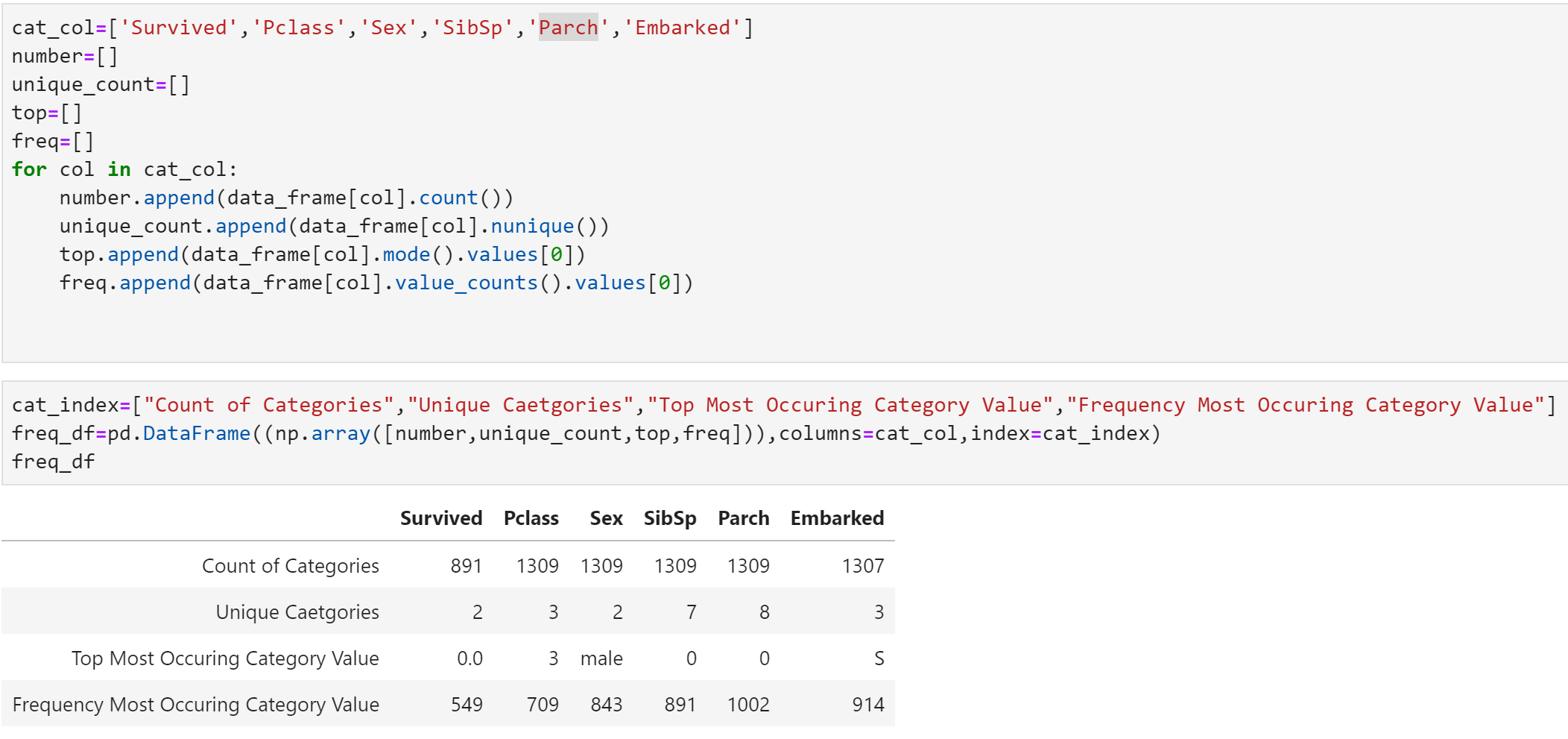
Number of Categorical Values: count() is used.

Number of Unique Categories – nunique() is used

Maximum/Top Occurring – mode() is used

Frequency of Top Occurring – value\_counts().values[0] is used





Codelink :

https://github.com/bhuvaneshkj/CAP5610-MachineLearning-Assignment.git