HW1\_Assignment

UCF | DATA ANALYTICS

Aravind Kumar Sankara Narayanan

ar914633

2020

**Task 1 Attributes :**

**Classify the following attributes as binary, discrete, or continuous. Further classify the attributes as nominal, ordinal, interval, ratio.**

(a) Rating of an Amazon product by a person on a scale of 1 to 5 – **Discrete , Ordinal**

(b) The Internet Speed - **continuous,ratio**

(c) Number of customers in a store. – **discrete,ratio**

(d) your Student ID – **discrete,nominal**

(e) Distance –  **Continuous,ratio**

(f) your letter grade (A, B, C, D) – **Discrete , Ordinal**

(g) The temperature in the campus – **Continuous,Interval**

**Task 2 Distance/Similarity Measures**

**Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. Briefly explain your choice.**

**• Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)? Justify your answer.**

Considering the shapes (length-width-ratio) , I can infer that correlation will be the best proximity measure to group boxes based on shapes. The answer can be justified with two inferences

Inference 1 :

Considering the length and width of first rectangle having dimensions of (length, width) = (2,1) and the length and width of bottom left rectangle having dimensions of (length, width) = (6,3), there is a correlation in increase of length and width 3 times in the second rectangle than the first rectangle. Hence rectangle 1(2,1) and rectangle 2 (6,3) can be grouped using correlation proximity measure

Inference 2:

Incase of the first square with sides (a = 1) , and the second square with sides (a=3) , there is correlation in increase of sides 3 times in the second square than the first. Hence square 1 with sides a=1, and square 2 with sides a=3 can be grouped using correlation proximity measure

Hence from both the inferences we come to a conclusion that correlation is the best proximity measure to group boxes based on their shapes.

**• Which proximity measure would you use to group the boxes based on their size? Justify your answer.**

Considering the size of the boxes, the boxes can be grouped using Euclidian Distance Proximity measure. To brief here, lets consider the rectangle 1 having dimensions (2,1) as length and width respectively, the area of the rectangle is 2, and the square with side a =1 , has an area of 1. So considering their respective areas, these rectangle and squares can be grouped together based on their sizes.

Similarly consider the rectangle having dimensions (6,3) as length and width respectively, the area of the rectangle is 18, and the square with side a = 3 has an area of 9. So considering their respective areas , these rectangle and squares with areas more likely to be closer than the other boxes, hence these can be grouped together based on their sizes.

**Task 3 Data Preprocessing of Titanic**

**Subtask 1: Analyze by describing data**

**Q1: Which features are available in the dataset?**

Available Features: Age, Sex, class, Ticket Number, Embarked,Fare,Cabin,PassengerID

New Features created: Age Group, Adult (Yes/No)

**Q2: Which features are categorical?**

Available Features : Sex,Sibsp,Parch, PassengerID

New Features : Age Group, Adult

**Q3: Which features are numerical?**

Age, Pclass, Fare

**Q4: Which features are mixed data types?**

Ticket

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

Age, Cabin

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

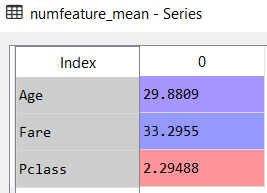
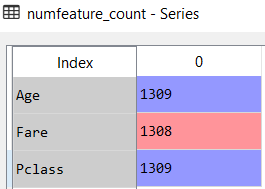
|  |  |
| --- | --- |
| Feature | Type |
| Age | Float |
| Class | Integer |
| Sex | Strings |
| Ticket Number | Strings |
| Age Group | Strings |
| Adult(Yes/No) | Integer |
| Embarked | Strings |

**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?**

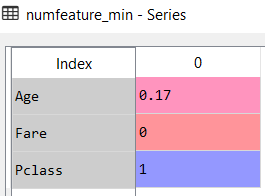
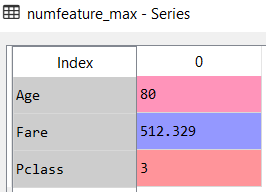
Given :

Numerical Features available in dataset : Age, Pclass, Fare

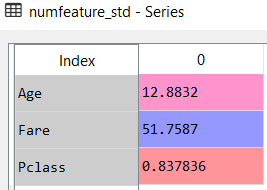
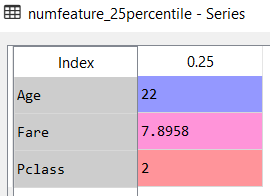
Mean : Count :

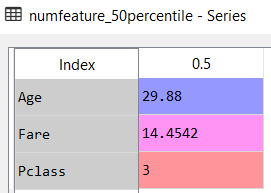
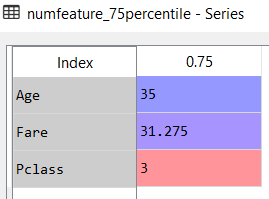
Min : Max :

Standard Deviation : 25 Percentile

50 Percentile : 75 Percentile :

**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?**

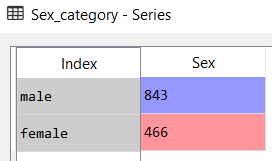
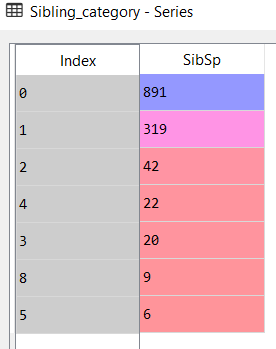
Given :

Categorical Features available in Data set : Sex,Sibsp,Parch, PassengerID

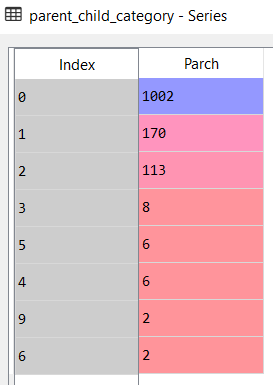
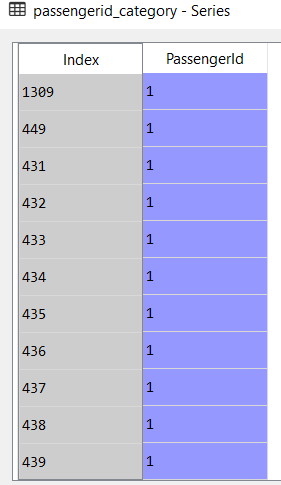
New Features generated using Feature extraction : Age Group, Adult

**i)Count** :

Sex : SibSp :

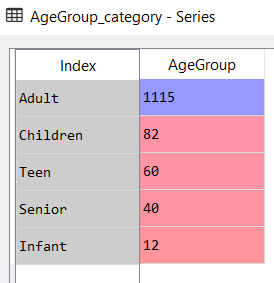
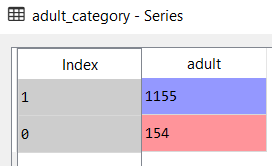
 

Parch : Passenger ID :

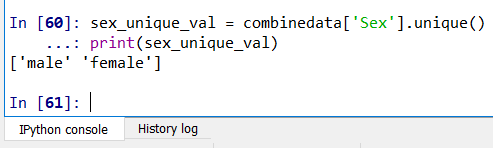
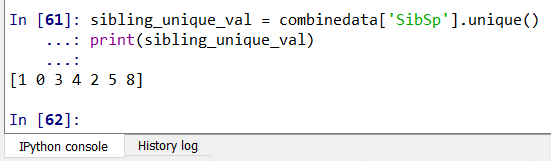
[Note : Since all passenger\_Id are unique, hence it returns all values in the column PassengerId]

AgeGroup : Adult :

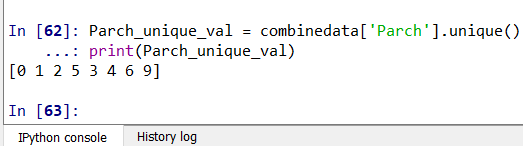
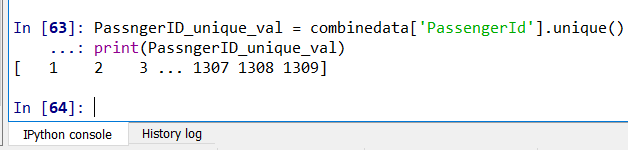
**ii)Unique :**

Sex : SibSp :

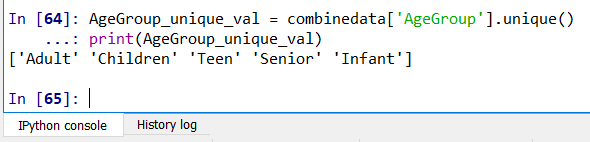
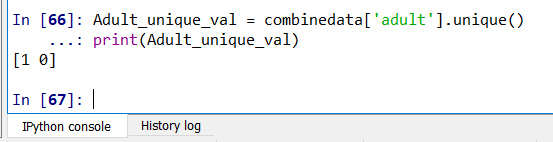
Answer : ‘male’, ‘female’ Answer : 0,1,2,3,4,5,8

Parch : PassengerId

Answer : [0,1,2,3,4,5,6,9] Answer : 1,2,3,…..,1307,1308,1309

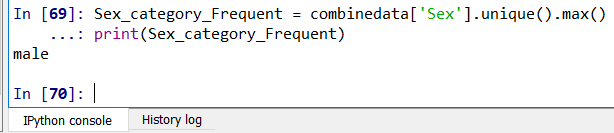
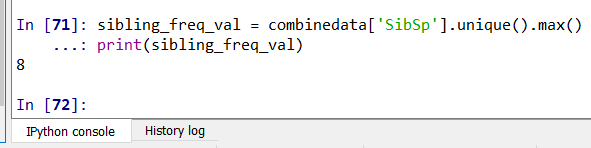
AgeGroup : Adult :

Answer: ‘Adult’,’Children’,’Teen’,’Senior’,’Infant’ Answer : 0 , 1 [0- No, 1- Yes]

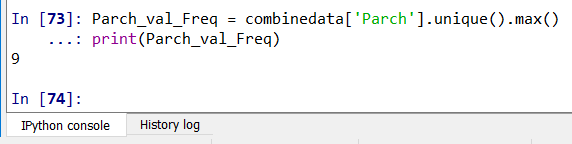
**iii)Top :**

Sex : Sibsp :

Answer : male Answer : 8

Parch :

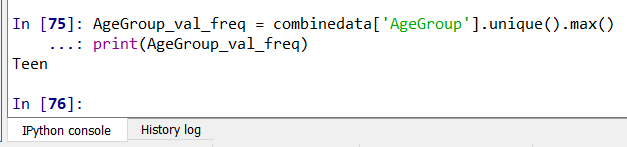
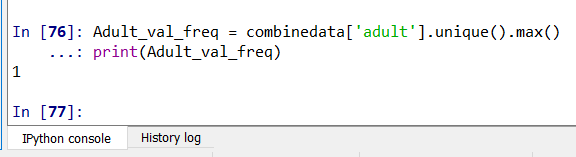


Answer : 9

PassengerId :

In passenger Id all values are unique with frequency 1, hence all the categorical value in passenger Id is equal

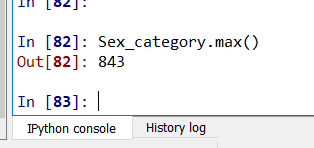
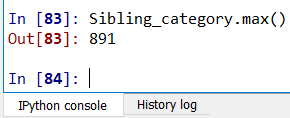
AgeGroup : Adult(Yes/No) :

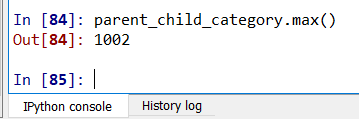
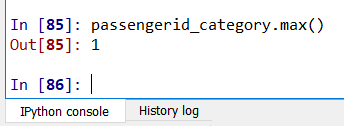
Answer : Teen Answer : 1 [Adults frequency is higher]

**iv)Freq :**

Sex : Sbsp:

Parch : PassengerId

AgeGroup: Adult (Yes/No)

