Assignment by Himanshu Shukla

Q.1 Estimate the probability that your vote will impact the election of your MP

in your constituency.

Ans. Probability of my vote impact will be = 1/ no.s of voters in my constituency for M.P. election.

Total no.s of voters in my constituency (Lucknow central) for M.P. election = 424848

Probability of my vote impact will be 2.353783000037661e-6

Q.2 Say whether the following is true or false and support your answer with a

proof: for any integer n, the number n^2+n+1 is odd.

Ans. It is true. Let suppose we are taking any even no. n= 2

n^2 = 4 and n^2+n = 6

n^2+n+1 will be 7 which is odd.

And Let take any odd no. n = 3

n^2 = 9 and n^2+n = 12

n^2+n+1 will be 13 which is odd.

Explanation:

If n is odd:
n2 is also odd
n2+n (sum of two odds) is even
n2+n+1 is odd

If n is even:
n2 is also even
n2+n (sum of two evens) is even
n2+n+1 is odd

Q3. Extract all those Customer ID wherever the customer has done IMPS Credit

followed by IMPS Debit within 10 minutes. Refer Below Data for answer.

1. Explain the Logical Approach

Data: Transaction Dump of a Bank on 1 Year

Column:

1. Customer ID: Identifier for a particular Customer
2. Transaction Type: Credit/ Debit
3. Transaction Status: Success/ Failed/ Others
4. Transaction Usecase: IMPS, UPI, etc.
5. Transaction Timestamp: Date + Time (e.g. 2022-01-01 12:35:22)
6. Customer Category: A/B/C1/C2

Ans. SELECT 'Customer ID',
'Transaction Timestamp' – lag ('Transaction Timestamp', 1) OVER (ORDER BY 'Customer ID', 'Transaction Timestamp') as diff
FROM data
HAVING diff < 10;

Explanation:

First of all, we will arrange data w.r.t customer ID and date time then we will find difference between date time from previous date time. Then select only those customer ID whose difference coming less than 10. I am assuming that the difference is coming in minutes if it is not in minutes then we have to convert it in minutes.

Q4. Extract all those Customer ID wherever the Customer Category is C1 or C2,

such that the latest transaction status of customer across any Customer

category should not be "Others". Refer Below Data for answer.

1. Explain the Logical Approach

2. Write an SQL Query

3. Write a Python Code using Pandas

Data: Transaction Dump of a Bank on 1 Year

Column:

1. Customer ID: Identifier for a particular Customer
2. Transaction Type: Credit/ Debit
3. Transaction Status: Success/ Failed/ Others
4. Transaction Usecase: IMPS, UPI, etc.
5. Transaction Timestamp: Date + Time
6. Customer Category: A/B/C1/C2

Ans. 1) we will apply filter in data wherever the Customer Category is C1 or C2 and Transaction Status should not be equal to Others

2) SELECT 'Customer ID'
FROM data
WHERE 'Customer Category' = 'C1' OR 'Customer Category' = 'C2' AND 'Transaction Status' != 'Others';

3) import pandas as pd
df = pd.read_csv("data")
df [(df["Customer Category"]=='C1') | df ["Customer Category"]=='C2') & df ['Transaction Status'] != 'Others')]

Q5. Consider the below Text File. Print all the words from the text such that output should contain any word only once. For example: “FTL” should be printed only once.

1. Explain the Logical Approach
2. Write the Python Code using basic datatypes
3. Write the Python Code using Natural Language Processing

Sample Content of Text File:

Historically, the aviation industry has taken a regulatory approach to fatigue prevention through the specification of flight and duty time limitations in a Flight Time Limitations (FTL) Scheme. This is done by limiting the number of hours aircrew can work and specifying the minimum rest time which is required before commencement of each flight duty period.

The purpose of an FRMS is to support the safe application of such FTL Schemes by recognising the need for aircrew be adequately rested before commencing and during flying duties by facilitating both proactive and reactive interventions in relation to the implementation of FTL Schemes

Ans. 1) convert this para into lower case and then split it and then find count of each word frequency in para.

2)
```
import pandas as pd
import re
para=”””Historically, the aviation industry has taken a regulatory approach to fatigue prevention through the specification of flight and duty time limitations in a Flight Time Limitations (FTL) Scheme. This is done by limiting the number of hours aircrew can work and specifying the minimum rest time which is required before commencement of each flight duty period.
The purpose of an FRMS is to support the safe application of such FTL Schemes by recognising the need for aircrew be adequately rested before commencing and during flying duties by facilitating both proactive and reactive interventions in relation to the implementation of FTL Schemes
”””
Para=para.lower()
x=re.split(';| |,|\n',para)
df=pd.Series(x).value_counts()
df[df.values<2]
```

3) I am not very good in NLP

Q6. Consider the data below:

1. What is the Mean of below Dataset 01, Dataset 02, Dataset 03. Explain the approach used to extract the Mean in each case.
2. What is the Median of below Dataset 01, Dataset 02, Dataset 03. Explain the approach used to extract the Median in each case.
3. Is there any relation between Mean & Median for below Dataset 01, Dataset 02, Dataset 03. Explain reason for your answer.

Dataset 01: 1, 3, 7, 5, 9, 11, 13

Dataset 02: 1, 5, 9, 13, 17, 21, 25

Dataset 03: 2, 4, 8, 9
Note: Do not Use Python or SQL to answer above Question.

Ans. 1) for finding mean we will count no. of elements in dataset and find the sum of all elements Then divide sum from no. of elements.
Dataset 1 mean: (1 + 3 + 7 + 5 + 9 + 11 + 13)/7= 7
Dataset 2 mean: (1+ 5+ 9+ 13+ 17+ 21+ 25)/7= 13
Dataset 3 mean: (2 + 4 + 8 + 9)/4= 5.7

2) for finding median we will count no. of elements in dataset and after arranging the elements in ascending order if odd no. is there then (count+1)/2th element is median or if even no. is there then mean of (count/2)th and (count+1/2)th is median.
Dataset 1 mean: (1 + 3 + 7 + 5 + 9 + 11 + 13) = 5
Dataset 2 mean: (1+ 5+ 9+ 13+ 17+ 21+ 25) = 13
Dataset 3 mean: (2 + 4 + 8 + 9)/4= 6

3) in dataset 1 median < mean i.e. it is rightly skewed data
In dataset2 mean = median that means it is normally distributed.
In dataset 3 mean < median i.e. Left skewed data.