Data Scientist Assessments

**Question 1 (3 sections)**

* 1. Please list the different types of join statements and briefly explain the operation they represent.  
      Inner Join:  
      JOIN: Return matching records from both tables.  
      Outer Join:  
      LEFT JOIN: Return all records from the left table, plus matching records from the right table.  
      RIGHT JOIN: Return all records from the right table, plus matching records from the left table.  
      FULL OUTER JOIN: Return all records from both tables.

1.2) Answer the below questions on different data query statements by listing the results.

* 1. E.g: Table A

|  |  |  |
| --- | --- | --- |
| columnA | columnB | columnC |
| 1 | abc | 123 |
| 2 | def | 456 |
| 3 | ghi | 789 |

Table B

|  |  |  |
| --- | --- | --- |
| columnD | columnE | columnF |
| 2 | jkl | 999 |
| 3 | mno | 888 |
| 4 | pqr | 777 |
| 5 | stu | 555 |

What is the result of table A inner join table B on columnA=columnD? (list all columns from both tables)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| columnA | columnB | columnC | columnD | columnE | columnF |
| 2 | def | 456 | 2 | jkl | 999 |
| 3 | ghi | 789 | 3 | mno | 888 |

What is the result of table A left outer join table B on columnA=columnD?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| columnA | columnB | columnC | columnD | columnE | columnF |
| 1 | abc | 123 | null | null | null |
| 2 | def | 456 | 2 | jkl | 999 |
| 3 | ghi | 789 | 3 | mno | 888 |

What is the result of table A right outer join table B on columnA=columnD?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| columnA | columnB | columnC | columnD | columnE | columnF |
| 2 | def | 456 | 2 | jkl | 999 |
| 3 | ghi | 789 | 3 | mno | 888 |
| null | null | null | 4 | pqr | 777 |
| null | null | null | 5 | stu | 555 |

What is the result of table A union/append table B? (if columnD is renamed as columnA, columnE is renamed as columnB and columnF is renamed as columnC)

|  |  |  |
| --- | --- | --- |
| columnA | columnB | columnC |
| 1 | abc | 123 |
| 2 | def | 456 |
| 3 | ghi | 789 |
| 2 | jkl | 999 |
| 3 | mno | 888 |
| 4 | pqr | 777 |
| 5 | stu | 555 |

What happens to table A if you Transpose it?

I do not think there will be any difference because it has the same number of rows and columns.

1.3) Explain what is the difference between relational vs non-relational database

|  |  |
| --- | --- |
| Relational DB | Non-Relational DB |
| Data organised into tables with fixed schema of rows and columns | Uses various data models like key-value pairs, documents, graphs, column family |
| Relationships between tables established through primary and foreign keys | More flexible schema with dynamic structure to allow different data types in the same collection |
| ACID (atomicity, consistency, isolation, durability) to ensure data integrity and transactional reliability | BASE (basically available, soft state, eventually consistent) over strict ACID |
| Suitable for structured data with defined relationships and complex queries | Suitable for unstructured or semi-structured data, high data volumes, and applications requiring horizontal scalability |

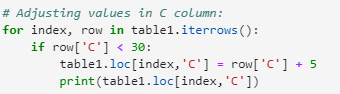
**Question 2 (4 sections)**

Refer to ‘table1’ to answer Question 2.1 to Question 2.4.

table1 :

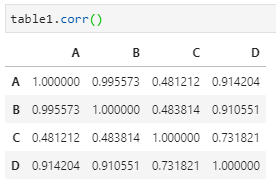


2.1) For the following iterating code, what is the possible output that will be printed?



30.6  
17.6  
27.3  
32.5  
28.8

2.2) Correlation between variables:



A, B, C are the independent variables while D is the dependent variable in this case study. What are your insights on the correlation values obtained between them?

Please provide short and precise answers.

We can observe several correlations among the independent variables:

A-B: strong positive

A-C: weak positive

A-B: weak positive

We can also see several correlations among the independent variables and the dependent variable:

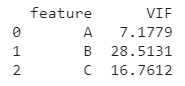
A-D: strong positive

B-D: strong positive

C-D: moderate positive

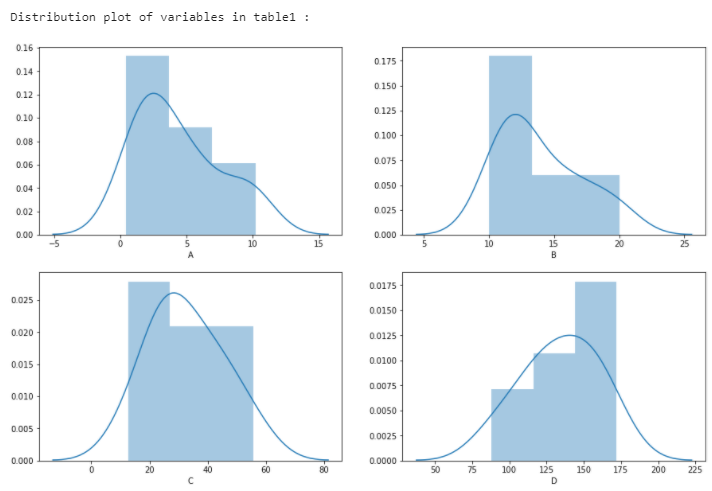
Thus, from here we can see that A and B somehow have the same contributions towards D. It is possible to experiment on the possibility of removing one of them. At the same time the contribution of C can be reassessed to ensure that it is really needed in modelling.

2.3) The following VIF values indicating variables multicollinearity are identified from the dataset. What is your suggestion to handle this occurrence? What’s the reason for your suggestion?



The multicollinearity of different variables need to be addressed. This can be done by removing variables via stepwise elimination or PCA. These two approaches are suggested because they can help to remove redundant variable/combine several dimensions into one. They can also help to reduce the computational cost. However, advices from experts need to be considered as well.

2.4) The distribution plot of the variables are as follows:



For variable A, the business team would like to have a categorical variable to represent this variable, instead of the original numerical variable.

By referring to the distribution plot above, what is the suitable classification or ranges that you can suggest for conversion from numerical to categorical classes?

I would suggest cut-off point as follow:

X <= 7: 0

X > 7: 1

This is because we can see that there is a sudden jump in value (no value between 6 and 9) and there are two peaks in the distribution curve. However, clustering analysis can be performed prior to making any decision.

**Question 3:**

In a case study under the Logistic department, the department would like to have a prediction model to predict the stop duration period for every truck’s stop point.

Based on the brainstorming with the Logistic department personnel, several factors are identified as impacting a truck’s stop duration period.

The factors include the type of products delivered, the delivery area, the combination of drivers (Driver 1 and Driver 2), and the dimension of the products. In addition, the personnel gave feedback that the total weight based on quantity per product delivered also impacts the stop duration time in two ways - the total weight of products being delivered, and the total weight of containers being returned.

Based on this problem statement, please structure the possible independent variables and the dependent variable in this case study.

|  |  |  |
| --- | --- | --- |
| No | Independent variables | Dependent variables |
| 1 | productType | stopDuration |
| 2 | driverCombination |  |
| 3 | deliveryArea |  |
| 4 | quantityPerProduct |  |
| 5 | *weightDelivered* |  |
| 6 | *weightReturned* |  |
| 7 |  |  |
| 8 |  |  |

\* Note that the number of rows provided does not indicate maximum or minimum variables required. You may add in or delete rows / columns as you find appropriate.