Starting the test	Wednesday, January 3, 2024, 4:39 p.m
State	Completed
Completion of the test	Wednesday, January 3, 2024, 6:40 p.m
Trial length	2 hours

Information

PART Relational Databases

This part consists of five tasks. It is recommended to work with **Microsoft SQL Management Studio** and **Server** , but you can use other relational database management software as well. For the PowerBI use PowerBI desktop software.

Task Overview:

1. **AdventureWorks database** (1 task, 5 points) **.**
2. **AdventureWorksDW database** (2 tasks, 6 points) **.**
3. **PowerBI** (2 tasks, 9 points).

Important Note:
Careful reading and understanding of each task's instructions are crucial for successful completion. Pay special attention to the specific requirements and guidelines provided within the tasks.

Information

You will need for this part the AdventureWorks database. It is deployed on your local MS SQL Server (Server address= "localhost"). If you cannot access the database, it is possible to download a database backup from Microsoft (use version 2012+):

https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms

For better orientation in database schema, you can use the following document, but there can be minor differences according to the version of your database:

AdventureWorks database schema

Task **1**

Done Number of points out of 5

What is the name of best salesman in **AdventureWorks** database according to total revenue of sales (Sales.SalesOrderHeader) and what is the value? Write First name, Last name and total sum of all orders. Paste the SQL code.

Linda

Mitchell

11695019.06

```
SELECT Top 1 pp.FirstName,pp.LastName,
ROUND(SUM(soh.TotalDue), 2) AS TotalSales FROM Person.Person pp
JOIN Sales.SalesOrderHeader soh ON soh.SalesPersonID = pp.BusinessEntityID
GROUP BY pp.FirstName,pp.LastName
ORDER BY TotalSales DESC;
```

Task **2**

Done Number of points from 1

What type of data warehouse schema is in AdventureWorksDW database?

Answer: Snowflake Schema

Task **3**

Not answered Number of points out of 5

Use the **AdventureWorksDW** to answer the question **.** What customer has the highest number of sales in the 4th quarter of all years? Write the last name of the customer, the number and paste the full SQL query.

Information

Download the PowerBI file from this link:
https://elearning.jcu.cz/pluginfile.php/779043/question/questiontext/835399/6/2569203/Power%20BI%20test%20-%20MAID.pbix

Task **4**

Done Number of points out of 3

What is the total income of "John Che"? Write the number and upload a screenshot (PrintScr). The number must be directly visible in the screenshot.

4087512

Screenshot (2).png

Task **5**

Done Number of points out of 6

What is the average salary of employees in the Czech Republic? Write the number and upload a screenshot (PrintScr). The number must be directly visible in the screenshot.

4054.25

Screenshot (4).png

Information

PART NoSQL

This part consists of five tasks. It is highly recommended to work with **RavenDB** , but you can use other document databases as well (specify in answers if you use a different database).

Task Overview:

1. **Northwind database** (3 tasks, 10 points) **.**
2. **Major League Baseball database** (2 Tasks, 10 points) **.**

Important Note:
Careful reading and understanding of each task's instructions are crucial for successful completion. Pay special attention to the specific requirements and guidelines provided within the tasks.

Information

Create a new non-sharded **document database** and import **Northwind** sample data (Northwind example database).

About Northwind: https://ravendb.net/docs/article-page/6.0/csharp/start/about-examples

---

Task **6**

Done Number of points out of 2

Create a query that returns a list of **all employees and their superiors** (for example king - Buchanan, Suyama - Buchanan, etc.).
* Last names are sufficient.
Write the full query.

```
{'_id': 'Peacock', 'names_player': ['employees/2-A']}
{'_id': 'Suyama', 'names_player': ['employees/5-A']}
{'_id': 'Callahan', 'names_player': ['employees/2-A']}
{'_id': 'King', 'names_player': ['employees/5-A']}
{'_id': 'Davolio', 'names_player': ['employees/2-A']}
{'_id': 'Fuller', 'names_player': ['']}
{'_id': 'Leverling', 'names_player': ['employees/2-A']}
{'_id': 'Dodsworth', 'names_player': ['employees/5-A']}
{'_id': 'Buchanan', 'names_player': ['employees/2-A']}
```

Task **7**

Not answered Number of points out of 3

Create an index that will allow to search **orders by the last name of the employee who handled the order** .
(For example the index for query:
from index '...' where Employee == 'Buchanan')
Write the full index syntax.

Task **8**

Not answered Number of points out of 5

How many pieces of **Maxilaku** were sold in total?

Answer:

Information

Create new non-sharded **document database** and **import data from files MLB** (Major League Baseball). Mind the metadata/headers (Body attributes should be nested).

https://elearning.jcu.cz/pluginfile.php/779043/question/questiontext/835399/14/2569769/mlb_players.csv

https://elearning.jcu.cz/pluginfile.php/779043/question/questiontext/835399/14/2569769/mlb_teams_2012.csv

Task **9**

Done Number of points out of 4

Replace team abbreviation in players with proper team name (full team name).

Write the update/patch script.
Attach export of the database (database dump).

```
collection.update_many({'Team':'CIN'},{'$set':{'Team':'Cortina Innovation Network'}})
```

Task **10**

Done Number of points out of 6

Which team has the shortest/smallest players on average (**team height average**)?
Write the answer (team name, average height) + full query/index.

```
{'_id': 'CWS', 'min_avgh': 74.63636363636364}
```

In [113]:

```
names=collection.aggregate([{'$group':{'_id':'$Team','min_avgh':{'$avg':'$Body - Height(inches)'}}},{'$sort ':{'avgweight':1}},
{'$limit':1}])
for name in names:
  pprint.pprint(name)
```

Information

PART DATA SCIENCE

This part comprises three distinct tasks. You will engage in two data processing tasks, each involving a different data file, and one task focused on assessing your basic knowledge.

Task Overview:

1. **Data Processing Tasks (2 Tasks):** You are expected to proficiently use libraries and methods from `pandas`, `numpy`,`PCA` (Principal Component Analysis), and clustering. Pay close attention to the instructions provided in each task to ensure accurate and complete responses. `KMeans`
2. **Basic Knowledge Question (1 Task):** This will test your fundamental understanding of the concepts covered in this course.

Submission Guidelines:

- At the conclusion of this part, submit your code files (either or format) that you used to solve the tasks. `.py .ipynb`
- Please ensure that your submissions are original. All code will be thoroughly checked for plagiarism.

Important Note:

Careful reading and understanding of each task's instructions are crucial for successful completion. Pay special attention to the specific requirements and guidelines provided within the tasks.

Task **11**

Done Number of points out of 9

You are provided with data (test_data.xlsx) from a retail shop that consists of 8 columns **(Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country)** .
You are supposed to analyze data in indexes between (190000, 194000) ! Do not forget to clean the data of NaNs, as failing to do so may render your answers invalid!
Fill in your answers to the text below, as if it were your report from the data.

The amount of **unique invoices** in the provided dataset for indexes (190000, 194000) is equal to [202] (integer).

The country with the **most orders** is [United Kingdom] (string) with the amount of [3015] (integer).

The country with the **least orders** is [Belgium] (string) with the amount of [0] (integer).

The **highest total price per item** in one order was [2174] (float) in order with id [507845] (integer).

The **second highest spending customer** with id [14911] (integer) ordered and total of [4] (integer) times.

Hints:
Total price per item is calculated as price times quantity.
9 out of 10 data scientists do not recommend .xlsx as a format to work with.
Methods to consider: dropna, group_by, sort_values

Task **12**

Done *Number of points out of 3*

How is the Gap Statistic used in determining the number of clusters?

Select one or more options:

- ☐ *and. By comparing within-cluster dispersion to that of null reference distribution.*
- ☐ *b. By measuring the compactness of clusters.*
- ☐ *C. By evaluating the mean distance between data points.*
- ☐ *d. By computing the total variance in the dataset.*

Task **13**

Done *Number of points out of 8*

You are provided with data (data_test_clustering_A1.csv) that consists of 8 features for each row.
Analysis of the data within the index range (47924, 48878) .
All floats are rounded to 3 digits!
Complete the report below with your findings from the data analysis.

Based on 2D PCA visualization of data between indexes (47924, 48878) , we can identify [3] *(integer) distinct clusters.*

We randomly selected two rows (indexes) within this range for further analysis. The selected indexes are (311, 502) .

*For index 311, the **most significant** PCA component value is* [2.935799] *(float).*

*For index 502, the **second most significant** PCA component value is* [2.230968] *(float).*

The Euclidean distance between the PCA components of indices (311, 502) is [4.228] *(float).*

*After applying K-Means clustering with **random_state=0** to the specified data segment, it was found that the features from index 311 belong to cluster label* [1] *(integer), and features from index 502 belong to cluster label* [1] *(integer).*

Hints:

- *Using a different random_state for K-Means clustering will result in incorrect answers.*
- *Consider using methods like PCA, KMeans, Euclidean distance, fit_transform, and fit_predict in your analysis.*

*Task **14***

Done *Not rated*

Submit your .py or .ipynb file which you've used during the test.

Your test will be considered invalid without propper code submission!

Test_dataAnalysis.ipynb

Clustering_of_features_with_K_means_and_PCA.ipynb