CISC 520-2021/Fall Data Engineering & Mining

Assignment #2

Name: <insert name here>

Start date 28 September, due 5 October.

In this assignment, you need to clean `Groceries data`. You can either download from Groceries data or directly from Course Canvas.

The file includes {InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country} coulums.

Part 1. (50 points) Clean your data: In this process, you should include:

- Fill in missing values.
- Identify outliers and smooth out noisy data.
- Correct inconsistent data

```
import pandas as pd
import numpy as np
```

```
# Loading the Data
data = pd.read_excel('/content/sample_data/Online Retail.xlsx')
print(data.shape)
```

```
(541909, 8)
```

```
# print columns
print(data.columns)
```

```
Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
```

```
def clean_data(data):
  # 1. Fill in missing values.

  # 2. Identify outliers and smooth out noisy data.

  # 3. Correct inconsistent data

  return(data)
```

Part 2. (50 points)

Suppose a market shopping data warehouse consists of *four dimensions*: `customer`, `date`, `product`, and `store`, and *two measures*: `count`, and `avg sales`, where `avg sales` stores the real sales in dollar at the lowest level but the corresponding average sales at other levels.

1. (20 points) Draw a snowflake schema diagram (sketch it, do not have to mark every possible level, and make your implicit assumptions on the levels of a dimension when you draw it).
2. (20 points) Starting with the base cuboid [*customer, date, product, store*], what specific OLAP operations (e.g., roll-up student to department (level)) that one should perform in order to list the average sales of each cosmetic product since January 2005 ?
3. (10 points) If each dimension has 5 levels (excluding all), such as *`store-city-state-region-country`*, how many cuboids does this cube contain (including base and apex cuboids)?