Project 1

<Task>

1. The answers about goals - Why do these anomalies happen?
2. ~~Change the bootstrap sample size into 440~~
3. ~~Which plot? - to check the separability clearly~~
4. Evaluation metric ?
5. ~~Which plot? - Relation Between Anomalies and Meta-Data (Histogram)~~
6. ~~Combine the EDA part (from Evita and Jing)~~
7. ~~Change the way to get the mean - in bootstapping~~

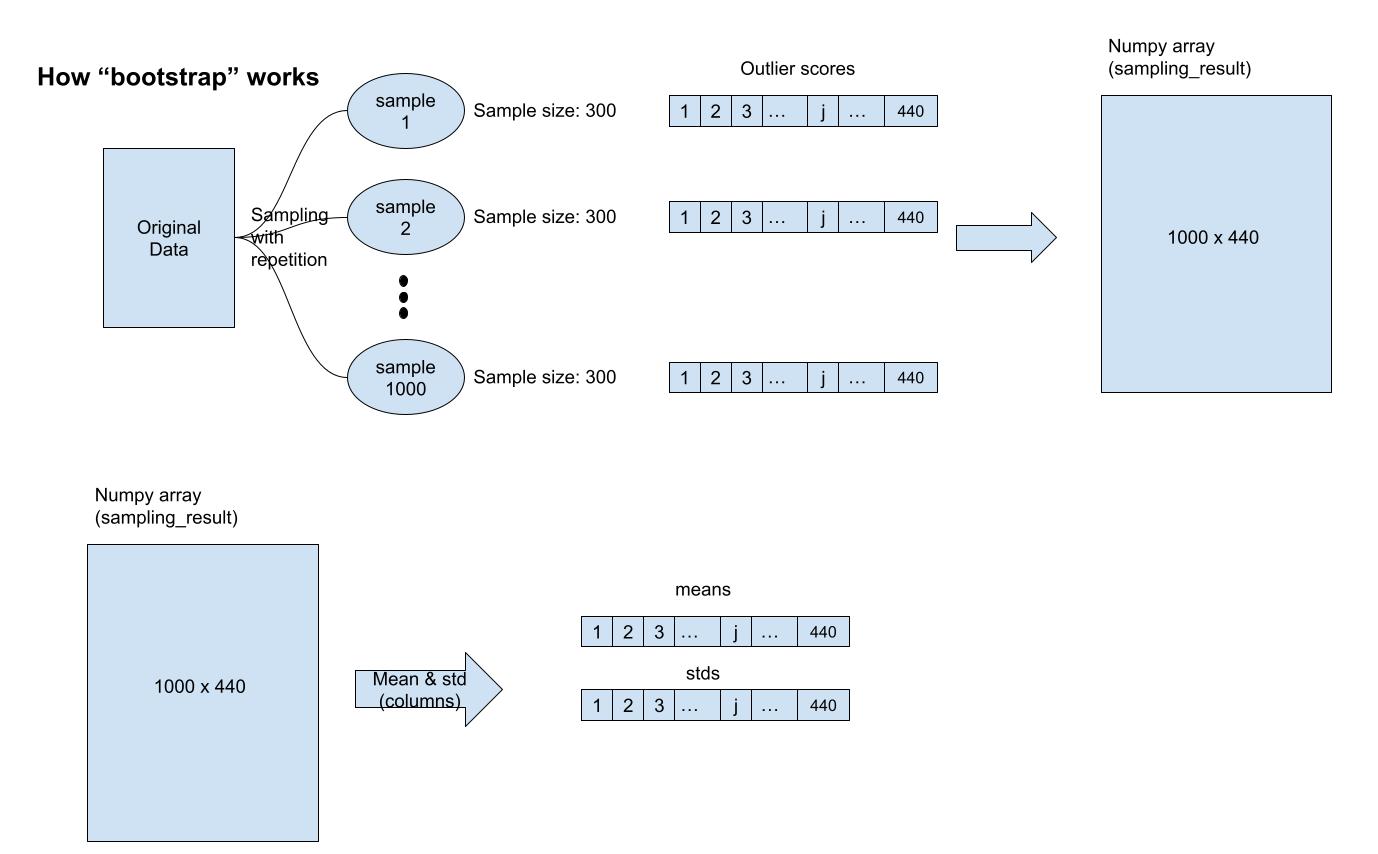
# <Data>

1. The annual spending on different product categories by a collection of wholesale customers.
2. Features (8): Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicassen
3. Instances (440) : wholesale customers.
4. Goals
   1. Find instances with anomalous spending <- anomaly prediction
   2. Find instances of similar behavior <- cluster prediction
   3. Why do these anomalies happen?
   4. Make it reproducible <- bootstrapping

# Loading the data, Preprocessing, Initial Data Analysis

* 1. **Loading the data**
     1. Using the package “ucimlrepo”
     2. Data
        1. X: Channel, Fresh, Milk, Grocery, Frozen, Detergents\_Paper, Delicassen
        2. y: Region
        3. data: X + y
  2. **Preprocessing**
     1. Check the missing values
     2. Drop the metadata (Channel, Region)
        1. data -> X
     3. Convert the data into numpy array
        1. X -> X\_arr
  3. **Initial Data Analysis**
     1. Pairplot on the original data
     2. Pairplot on the non-linear transformation data
        1. Θ = 1
        2. Θ = 10 (because it looks more like Gaussian dist.)
        3. Θ = 100
     3. Non-linear transformation (Θ = 10)

# Detecting Anomalies

* 1. **Hard min (the nearest neighbor) [Eq(1)]**
     1. Calculate the outlier scores for each instance using hard min function
     2. List the top 10 instances by its outlier score
  2. **Soft min [Eq(2)]**
     1. Calculate the outlier scores for each instance using soft min function
     2. List the top 10 instances by its outlier score (gamma == 1)
  3. **Bootstrapping with softmin**
     1. To find the proper gamma
     2. Resampling with repetition
     3. The num. of sampling: 1000
     4. Sample size: 300
     5. Plotting the results of Bootstrapping according to its gamma
        1. Gamma == 0.001
        2. Gamma == 0.1
        3. Gamma == 1
        4. Gamma == 1.5
        5. Gamma == 3
        6. Gamma == 5
        7. Gamma == 7.5 (the best separability between anomaly and non-anomaly)
        8. Gamma == 10
        9. Gamma == 100
        10. Gamma == 200
     6. Select the right gamma
        1. Visual intuition
        2. Evaluation metric
           1. Clustering metric ex) Silhouette Score:
     7. List the top 10 instances by its outlier score (gamma == 7.5)

# Getting Insights into Anomalies

* 1. **Relation Between Anomalies and Meta-Data (Histogram)**
     1. Based on “Channel”
     2. Based on “Region”
  2. **Identifying Input Features that Drive Anomaly**
     1. Calculating RK with gamma == 7.5 **[Eq(3)]**
        1. RK: a 2D array of a collection of the contributions of data point k to each instance (440 x 440)
        2. example) How much does the wholesale customer k contribute to the outlier score of the wholesale customer 1?
     2. Calculating RI with gamma == 7.5 **[Eq(4)]**
        1. RI: an array of a collection of the contributions of input feature i to each instance (440 x 8)
        2. example) How much does Fresh feature contribute to the outlier score of the wholesale customer 1?
     3. Embed RI into 2D using T-SNE
     4. Check some instances’ RI with barplot