

Title: Multi-Data Preprocessing in MySQL

Authors: (Group5)DBer

姓名	系級	學號
林祐辰	經濟所碩二	r08323046
張燭郁	資料科學碩一	r09946031
曾尹均	資工所碩一	r09922080
簡丞珮	資工所碩一	r09922155

Objectives:

自動化資料分析中多種常見的 data preprocessing 流程

Motivation:

希望可以簡化常見的 data preprocessing 流程

Description:

預計完成以下幾點功能：

1. 自動計算Z分數
2. 計算不同變數間的 covariance
3. Sampling data and plot distribution (hist or CDF) (maybe plus Normal CDF)

4. Outlier detection
5. Detect inconsistent data
6. Group by any of (C1, C2, C3 if there are missing and all of them are keys)
7. Typo detection: Outlier detection from count(C1)
8. Select variable with enough variation(CV, Coefficient of Variation)
9. Time Series: lag column
10. Fill missing values

Outcome:

在透過我們實作的工具的幫助下，在使用 Data 之前可以快速分析內容、進行前處理，而不需要重複執行一樣的指令。例如，以指令

```
EXEC [dbo].[RemoveOutLier] 'Outlier_model', [A table]
```

即可以移除大量 table 中的 outlier、計算不同變數間的 covariance 等，不需要從外部的程式操作。

Implementation:

使用 stored procedure，將 description 所述的幾項功能包在 SQL 中，我們可能會將 python 或其他程式語言嵌入 stored procedure 內，來完成較複雜的演算法實作，完成後會將指定的 table 改成前處理後的樣子，或是回傳使用者有興趣的統計量。

Reference:

- <https://docs.microsoft.com/zh-tw/sql/machine-learning/tutorials/python-taxi-classification-on-deploy-model?view=sql-server-ver15>