<https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting>

Task: ทำนาย colum Weekly\_Sales ในไฟล์ train.csv

ไฟล์ที่ให้ ประกอบไปด้วย

1. **stores.csv**

This file contains anonymized information about the 45 stores, indicating the type and size of store.

1. **train.csv**

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales - sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

1. **features.csv**

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

* Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
* Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
* Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
* Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

1. ไฟล์ test.csv จะส่งให้ในวันทดสอบโมเดล โดยไฟล์จะมีฟอร์แมทเหมือน train.csv แต่จะไม่มีข้อมูล weekly sales ให้

ในการทดลอง นิสิตควร preprocess data ในรูปแบบต่างๆ เช่น

1. แทนที่จะมองว่า Is\_Holiday = Yes or No อาจแยกประเภท Holiday ไปเลยโดยดูจากวันที่
2. จะจัดการกับข้อมูล MarkDown ที่มีไม่ครบอย่างไร

* ตัดออก
* ลองทำ Store clustering โดยใช้ assumption ว่า store ในกลุ่มเดียวกันน่าจะใช้ markdown คล้ายกัน (หรือแบ่ง store ตาม type-size-region)
* ใช้วิธีการดู markdown ปีติดๆ กัน (ก่อนหรือหลัง แล้วแต่ availability) ช่วงเวลาเดียวกัน ร้านค้าเดียวกัน น่าจะใช้ Markdown แบบเดียวกัน
* ใช้วิธี plot graph เพื่อช่วยในการ interpolate MarkDown
* ฯลฯ