Data Preparation: Merging, Cleaning, and Transforming the Data

We put together all the three .csv files (train.csv, features.csv and stores.csv). All the ‘NA’ values in the table features.csv was changed to NULL. The ‘TRUE’ and ‘FALSE’ values for the ISHOLIDAY attribute were changed to binary values 1 and 0 respectively. For analytical purposes and visualization, the variables TEMPERATURE, CPI, UNEMPLOYMENT, FUEL\_PRICE and WEEKLY\_SALES were all categorized into ‘Low’, ‘Medium’ and ‘High’.

**SQL queries as below:**

CREATE TABLE Store\_Features

AS

SELECT \*

FROM Stores JOIN Features USING(Store);

CREATE TABLE Walmart\_Train

AS

SELECT \*

FROM Train JOIN Store\_Features USING(Store, Week, IsHoliday);

ALTER TABLE walmart\_train ADD sales\_class varchar(100);

UPDATE WALMART\_TRAIN SET sales\_class = (CASE WHEN Weekly\_Sales <= 0 THEN 'Negative'

WHEN Weekly\_Sales > 0 AND Weekly\_Sales <= 25000 THEN 'Low'

WHEN Weekly\_Sales > 25000 and Weekly\_Sales <= 100000 THEN 'Medium'

WHEN Weekly\_Sales > 100000 then 'High'

ELSE NULL

END);

ALTER TABLE walmart\_train ADD temp\_class varchar(100);

UPDATE WALMART\_TRAIN SET temp\_class = (CASE WHEN Temperature <= 50 THEN 'Cold'

WHEN Temperature > 50 AND Temperature < 80 THEN 'Comfortable'

WHEN Temperature >= 80 then 'Hot'

ELSE NULL

END);

ALTER TABLE walmart\_train ADD fuel\_class varchar(100);

UPDATE WALMART\_TRAIN SET fuel\_class = (CASE WHEN Fuel\_Price <= 2.8 THEN 'low'

WHEN Fuel\_Price > 2.8 AND Fuel\_Price < 3.5 THEN 'Medium’

WHEN Fuel\_Price >= 3.5 then 'High'

ELSE NULL

END);

ALTER TABLE walmart\_train ADD unemploy\_class varchar(100);

UPDATE WALMART\_TRAIN SET unemploy\_class = (CASE WHEN Unemployment <= 5 THEN 'low'

WHEN Unemployment > 5 AND Unemployment < 7 THEN 'Medium'

WHEN Unemployment >= 7 then 'High'

ELSE NULL

END);

ALTER TABLE walmart\_train ADD cpi\_class varchar(100);

UPDATE WALMART\_TRAIN SET cpi\_class = (CASE WHEN CPI <= 160 THEN 'low'

WHEN CPI > 160 AND CPI < 200 THEN 'Medium'

WHEN CPI >= 200 then 'High'

ELSE NULL

END);

ALTER TABLE walmart\_train ADD size\_class varchar(100);

UPDATE WALMART\_TRAIN SET size\_class = (CASE WHEN Size <= 100000 THEN 'low'

WHEN Size > 100000 AND Size < 200000 THEN 'Medium'

WHEN Size >= 200000 then 'High'

ELSE NULL

END);