```
/* Retail Analysis */

FILENAME READFILE '/home/u42880822/walmart_store_sales.csv';

PROC IMPORT DATAFILE = READFILE REPLACE

DBMS=CSV
OUT=RetailAnalysis;
GETNAMES=YES;

RUN;

PROC CONTENTS DATA=RetailAnalysis;
RUN;

/* Check the missing value */

PROC MEANS DATA=RetailAnalysis NMISS;
RUN;

/* Which store has maximum sales */
PROC MEANS DATA=RetailAnalysis MAX;
BY STORE;
VAR WEEKLY_SALES;
TITLE 'Maximum sales/Per Store';
RUN;

/*Store has maximum sales*/
PROC SQL;
CREATE TABLE TOTALSALES AS
SELECT STORE, SUM(WEEKLY_SALES) AS TOTAL_SALES_BY_STORE
FROM RetailAnalysis
GROUP BY STORE
ORDER BY TOTAL_SALES_BY_STORE DESC;
TITLE 'Store has maximum sales';
QUIT;

PROC PRINT DATA=TOTALSALES;
TITLE 'Store has maximum sales';
RUN;
```

```
PROC MEANS DATA=RetailAnalysis;
CLASS STORE;
VAR WEEKLY_SALES;
OUTPUT OUT=MEANANA(DROP= _TYPE_ _FREQ_ )  STD=STDBYSALES
MEAN=MEANBYSALES;
RUN;

/*store has maximum standard deviation*/
PROC SORT DATA=MEANANA;
TITLE 'Store has maximum standard deviation';
BY DESCENDING STDBYSALES;
WHERE STORE NE .;

RUN;


/* Find out the coefficient of mean to standard deviation */
PROC MEANS DATA=RetailAnalysis CV;
CLASS STORE;
VAR WEEKLY_SALES;
RUN;

/* Which store/s has good quarterly growth rate in Q3’2012 */

DATA RetailAnalysis_2012;
SET RetailAnalysis;
WHERE YEAR(DATE) =2012;
RUN;

/* Calculate growth rate */
DATA GROWTH;
FORMAT GROWTH_RATE PERCENT8.2;
SET RetailAnalysis_2012;
BY STORE DATE WEEKLY_SALES;
LAG_SALES=IFN(FIRST.STORE,0,LAG(WEEKLY_SALES));
IF LAG_SALES NE 0
THEN
GROWTH_RATE=(WEEKLY_SALES/LAG_SALES)-1;
RUN;
```

```
PROC PRINT DATA=GROWTH;
RUN;



PROC TIMESERIES DATA=GROWTH OUT=CLEAN_GROWTH;
BY STORE;
ID DATE INTERVAL=QTR ACCUMULATE=TOTAL;
VAR GROWTH_RATE;
RUN;

PROC PRINT DATA=CLEAN_GROWTH;
RUN;

/* From timeseries data filterd only Q3 observations */

DATA QTR_CLEAN_GROWTH;
SET CLEAN_GROWTH;
WHERE QTR(DATE)=3;
RUN;

PROC SORT DATA=QTR_CLEAN_GROWTH;
BY DESCENDING GROWTH_RATE;
RUN;

PROC PRINT DATA=QTR_CLEAN_GROWTH;
RUN;

DATA HOLIDAY;
SET RetailAnalysis;
WHERE HOLIDAY_FLAG=1;
RUN;

DATA NON_HOLIDAY;
SET RetailAnalysis;
WHERE HOLIDAY_FLAG=0;
RUN;

/* Calculate the mean weekly_sales of the non-holiday data */

PROC MEANS DATA=NON_HOLIDAY;
OUTPUT OUT=MEAN_NON_HOLIDAY;
VAR WEEKLY_SALES;
```

```
RUN;

PROC PRINT DATA=MEAN_NON_HOLIDAY;

RUN;

PROC MEANS DATA=HOLIDAY;
OUTPUT OUT=MEAN_HOLIDAY;
VAR WEEKLY_SALES;
RUN;

PROC SQL;
TITLE 'Higher Sales during Holidays';
CREATE TABLE HOLIDAY_SALES AS
SELECT STORE, WEEKLY_SALES, DATE,HOLIDAY_FLAG AS HOLIDAY,
CASE
WHEN WEEKLY_SALES >= 1041256.38 THEN 'HIGHER'
WHEN WEEKLY_SALES < 1041256.38 THEN 'LOWER'
END
AS HIGHER_SALES
FROM HOLIDAY;


QUIT;

DATA HIGHER_HOLIDAY_SALES;
SET HOLIDAY_SALES;
WHERE HIGHER_SALES='HIGHER';
DROP HIGHER_SALES;
RUN;


PROC PRINT DATA=HIGHER_HOLIDAY_SALES;
RUN;

/* Provide a monthly and semester view of sales in units and give
insights */

/* Monthly view of sales in units */

/* Convert walmart data into timeseries data */
PROC TIMESERIES DATA=RetailAnalysis OUT=MONTHLY_SALES;
BY STORE;
```

```
ID DATE INTERVAL=MONTH ACCUMULATE=TOTAL;
VAR WEEKLY_SALES HOLIDAY_FLAG TEMPERATURE FUEL_PRICE CPI UNEMPLOYMENT;
FORMAT WEEKLY_SALES DOLLAR16.2;
RUN;

PROC PRINT DATA=MONTHLY_SALES;

RUN;

/* Giving insights */

/* Checking the correlation */

PROC CORR DATA=MONTHLY_SALES;
RUN;

/* 1. Doing Comparison */

/* a) Bar Chart */
PROC SGPLOT DATA=MONTHLY_SALES;
HBAR STORE/RESPONSE=WEEKLY_SALES STAT=SUM
DATALABEL DATALABELATTRS=(WEIGHT=BOLD) ;
TITLE 'Total Views by Store';
RUN;

/* b) Clustered Bar Chart / Column Chart */

DATA STORE_MONTHLY_SALES;
SET MONTHLY_SALES;
MONTHNAME=PUT(DATE, MONNAME.);

RUN;

PROC SGPLOT DATA=STORE_MONTHLY_SALES;
VBAR STORE/RESPONSE=WEEKLY_SALES GROUP=MONTHNAME GROUPDISPLAY=CLUSTER
DATALABEL DATALABELATTRS=(WEIGHT=BOLD) DATASKIN=gloss;
YAXIS GRID;
TITLE 'Total View by monthly wise';
RUN;

/* Studying relationship */

/* a) Bubble Chart */
```

```
PROC SGPLOT DATA=MONTHLY_SALES;
BUBBLE Y=WEEKLY_SALES X=STORE SIZE=WEEKLY_SALES
/FILLATTRS=(COLOR=GREEN) DATALABEL=STORE;
RUN;

/* b) Scatter Plot for Relationship */

PROC SGPLOT DATA=MONTHLY_SALES;

TITILE 'Relationship of Store with Weekly_sales';
SCATTER X=STORE Y=WEEKLY_SALES /
MARKERATTRS=(SYMBOL=CIRCLEFILLED SIZE=10) DATALABEL=STORE;

RUN;

/* Studying Distribution */

/* a) Histogram */

PROC SGPLOT DATA=MONTHLY_SALES;
HISTOGRAM WEEKLY_SALES/FILLATTRS=(COLOR=YELLOW) SCALE=PROPORTION;
DENSITY WEEKLY_SALES;
RUN;

/* b) Scatter Plot */

PROC SGPLOT DATA=MONTHLY_SALES;

SCATTER X=DATE Y=WEEKLY_SALES/GROUP=STORE GROUPDISPLAY=CLUSTER
MARKERATTRS=(SYMBOL=CIRCLEFILLED SIZE=10);
RUN;

/* Composition */

/* a) Stacked Column Chart: */

PROC SGPLOT DATA=MONTHLY_SALES;

TITLE 'Weekly_sales by Store and date';
VBAR DATE/RESPONSE=WEEKLY_SALES GROUP=STORE STAT=PERCENT DATALABEL
GROUPORDER=DESCENDING;
XAXIS DISPLAY=(NOLABEL);
```

```
YAXIS GRID LABEL='Weekly_sales';
RUN;


/* For Store 1 – Build prediction models to forecast demand */

/* Store-1 data */

DATA STORE1;
TITLE 'STORE 1 DATA';
SET RetailAnalysis;
WHERE STORE=1;
RUN;

/* Convert store-1 data into timeseries data */

PROC TIMESERIES DATA=STORE1
OUT=TIME_STORE1;
BY STORE;
ID DATE INTERVAL=MONTH ACCUMULATE=TOTAL;
VAR WEEKLY_SALES HOLIDAY_FLAG TEMPERATURE FUEL_PRICE CPI UNEMPLOYMENT;
RUN;

PROC PRINT DATA=TIME_STORE1;
TITLE 'Convert store-1 data into timeseries data';
RUN;

/* Build Model */

/* Linear Regression_Hypothesize if CPI, unemployment, and fuel price
have any impact on sales.*/
PROC REG DATA=STORE1;
MODEL WEEKLY_SALES=CPI UNEMPLOYMENT FUEL_PRICE;
RUN;


/*Time series forecasting model_Forcast 6 month sales*/

PROC ARIMA DATA=TIME_STORE1
PLOTS=( FORECAST(FORECAST FORECASTONLY));
IDENTIFY VAR=WEEKLY_SALES(1);
ESTIMATE P=(1 2 3) Q=1 METHOD=ML;
FORECAST LEAD=6 BACK=0 ALPHA=0.05;
```

```
OUTLIER;
RUN;
```