```
/* Generated Code (IMPORT) */
/* Source File: Walmart_Store_sales.csv */
/* Source Path: /folders/myshortcuts/myfolder */
/* Code generated on: 7/16/20, 1:31 AM */

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/folders/myshortcuts/myfolder/Walmart_Store_sales.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=WORK.IMPORT;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;


%web_open_table(WORK.IMPORT);

/* 1. Which store has maximum sales*/
PROC SQL;
    SELECT P.Store, P.Weekly_Sales as Max_Weekly_Sales, P.Date FROM IMPORT
        P WHERE Weekly_Sales=(SELECT Max(S.Weekly_Sales) FROM IMPORT as S
        WHERE S.Store=P.Store) GROUP BY P.Store, P.Weekly_Sales ORDER BY
        P.Weekly_Sales Desc;
QUIT;

/* Result : Store number 14 is having max Weekly sales value of 3818686.45 on 24 Dec 2010*/
/* 2. Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to st
/* First sorting data BY store */
PROC SORT data=IMPORT;
BY Store;
RUN;

/* Calculating the Standard deviation */
PROC MEANS Data=IMPORT std Max CV Maxdec=2;
Title 'Standard Deviation BY Store';
Class Store;
Var Weekly_Sales;
RUN;

/* 3. Which store/s has good quarterly growth rate in Q32012 */
/* Adding quarter column*/
DATA IMPORT_Sales_Q;
set IMPORT;
BY Store;
format Quarter yyq.;
Quarter=cats(yyq(year(date), qtr(Date)));
QuaterNum=qtr(Date);
Year=year(Date);
Format Date Date9.;
RUN;

/* Filter the data for 2012 years for finding growth rate quarter wise for all store */
PROC SQL;
Create table Walmart_Quarter_Growth as SELECT Store, sum(Weekly_Sales) as
Weekly_Sales, Quarter FROM IMPORT_Sales_Q WHERE year(Quarter)=2012
GROUP BY Store, Quarter;
QUIT;

/* Calculating the growth rate for all store w.r.t. each quarter*/
DATA Walmart_quarter_analysis (drop=LAG_Rate DIF_Rate);
set Walmart_Quarter_Growth;
BY Store Quarter;
LAG_Rate=LAG(weekly_sales);
DIF_Rate=DIF(weekly_sales);

IF FIRST.Store THEN
DO;
LAG_Rate=.;
DIF_Rate=.;
END;
GrowthRate=(DIF_Rate/LAG_Rate)*100;
RUN;

/* Finding the store with Max growth rate in Q3 for year 2012. */
PROC SQL;
SELECT G.Store, G.Quarter, G.GrowthRate as Max_GrowthRate FROM
Walmart_quarter_analysis G WHERE G.GrowthRate=(SELECT Max(S.GrowthRate) FROM
```

```
Walmart_quarter_analysis S WHERE qtr(S.Quarter)=3);
QUIT;

/* Result: For Store no. 7 growth rate in Q3 is highest with 13.33% */



/* 4 Some holidays have a negative impact on sales. Find out holidays which have higher sales than
the mean sales in non-holiday season for all stores together */
PROC MEANS Data=IMPORT mean Maxdec=2;
Title 'Mean across sales during Holiday/Non Holidays.';
class Holiday_Flag;
Var Weekly_Sales;
RUN;



/* Mean sales for non-holiday is - 1041256.38 while Mean sales for holiday is 1122887.89

/* To find the holidays where sales is greater than 1041256.38 */
PROC SQL;
Select * , count(*) as count from IMPORT where
Weekly_Sales > 1041256.38 and Holiday_Flag=1;
QUIT;


/* Result : There are 220 such records where sales are higher than the means sales in non-holiday seasonc for all stores
/* PROC contents data =IMPORT_sales_q; */
/* Statistical Model */
/* For Store 1 – Build prediction models to forecast demand */
/* Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010
(starting from the earliest date in order).
Hypothesize if CPI, unemployment, and fuel price have any impact on sales. */
/* Utilize variables like date and restructure dates as 1 for 5 Feb 2010
(starting from the earliest date in order) */
DATA WALMART_PRED_STORE1;
SET IMPORT;
BY Date;

/* data set have must be sorted by Date */
WHERE Store=1;

IF first.Date THEN
DO;
Counter+1;
END;
ELSE
Counter+1;
RUN;

/*Hypothesize if CPI, unemployment, and fuel price have any impact on sales. */
/* Cheking model with CPI independent variable */
PROC REG DATA=WALMART_PRED_STORE1;
MODEL WEEKLY_SALES=CPI;
RUN;


/* Result : Analysis
CPI variable has significant positive impact on dependent variable Weekly_sales where p-value is less
than 0.05. We also have joint significance of coefficient where p-value is less than 0.05
hence the variable has significantly related.
*/
/* Cheking model with Unemployment independent variable */
PROC REG DATA=WALMART_PRED_STORE1;
MODEL WEEKLY_SALES=Unemployment;
RUN;



/* Result : Analysis
For unemployment the the p-value is significantly very high. Hence Unemployment is not significantly
related with weekly_sales */
/* Cheking model with Fuel_Price independent variable */
PROC REG DATA=WALMART_PRED_STORE1;
MODEL WEEKLY_SALES=Fuel_Price;
RUN;



/* Result : Analysis
For fuel_price the the p-value is significantly very high. Hence fuel_price is not significantly
related with weekly_sales */
```

```
/* Cheking model combine with all three independent variable that includes CPI, Unemployment and CPI */
PROC REG DATA=WALMART_PRED_STORE1;
MODEL WEEKLY_SALES=CPI Unemployment Fuel_Price;
RUN;


/* Result : Analysis
While combining all three variable it shows relationship is significant and less than critical value of 0.05
*/
/* 5. Time series forecasting model – */
/* Hypothesize if the data is fit for time series analysis – check for white noise probability test */
PROC SORT data=IMPORT;
by store Date;
RUN;

ODS GRAPHICS ON;

PROC TIMESERIES DATA=IMPORT out=WALMART_TS_HYPOTHESIS1;
TITLE 'PREDICTING SALES USING DATE';
by Store Date;
Id Date Interval=WEEK Accumulate=total;
Var Weekly_Sales;
RUN;

/* Before we start looking into Arima model lets first try to plot graph*/
PROC sgplot DATA=IMPORT;
scatter y=Weekly_Sales x=Date;
RUN;

QUIT;


/* STEP 1: ARIMA Identification */
/* Descriptive Statistics */
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales nlag=24;
RUN;
QUIT;



/* Result :
As P-value less for lags which is less than 0.05 so the P-value is significant  hence the variables
are highly correlated between lags.
By observing the Weekly_Sales plot the mean, variance are not constant and from ACF plot as
As p-value is significant the data is not random and hence not stationary. This also conclude that
it does not have White Noise.
*/
/* Testing with First order differencing for stationarity */
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(1);
RUN;


/* Result : By looking at ACF plot there is steep fall and p-value is significant. Hence we conclude the first order
provides stationary statistics.*/
/* Testing with Second order differencing */
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(2);
RUN;




/* Result: After checking second order differencing it has found that first order is better so we can proceed with first
differencing. */
/* STEP 2: Estimation and Diagnostic Check */
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(1);
Estimate p=1;
RUN;



/*Result: ITs not normal distributed and high difference. After the dignostic check it has found that the
p-value is less than 0.05 hence residuals are co-related and so this model is not suitable to use */
/* Will use another model with MA term lets start with q=1 */
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(1);
Estimate p=1 q=1;
RUN;
```

```
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(1);
Estimate p=2 q=1;
RUN;
```

```
PROC ARIMA DATA=IMPORT;
Identify var=Weekly_Sales(1);
Estimate p=2 q=2;
RUN;

/* From above 3 model ARIMA 1,1,1 is better than other models so will choose it to forecast*/
```

```
/* 6. Build ARIMA model to forecast 6 months i.e., input utilize only till April 2012. */
/* Predict next 6 months i.e., June to Oct 2012. Check for MAPE. */
/* 6.1 First only use the holiday and then use only non-holiday observation for data till April 2012*/
/* Holiday obeservation */
DATA IMPORT_Holiday;
set IMPORT;
BY Store;
Format Date Date9.;
Holiday_Flag=0;
where Date <='30Apr2012'd;
RUN;

/*Non Holiday observation */
DATA IMPORT_NonHoliday;
set IMPORT;
BY Store;
Format Date Date9.;
Holiday_Flag=1;
where Date <='30Apr2012'd;
RUN;
```

```
/* Forecasting for Holiday observation */
PROC ARIMA DATA=IMPORT_Holiday;
Identify var=Weekly_Sales(1);
Estimate p=1 q=1;
forecast lead=6 interval=Month id=Date out=forecast_result;
RUN;
```