ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG QUỐC TẾ
VNU-INTERNATIONAL SCHOOL

PROJECT REPORT ON:

FORECASTING THE SALES OF A SUPERMAKET DURING FASTIVAL SEASON

SUBJECT:

DATAWAREHOUSING AND BUSINESS ANALYTICS

Contents

Executive summary

We are living in the era of the knowledge economy. All our activities want to be highly effective, it is necessary to have methods to get the necessary information and knowledge quickly and accurately.

The application of information technology in production and business practice has brought about great efficiency and benefits. Technology is increasingly developed and perfected to meet the increasing requirements of research, production management and professional practice. The scale of application from day-to-day business successes has progressed to meeting requirements at a higher level, helping operational managers not only know how the work is going, but also Knowing what to expect, that is, the information is analytical, and such systems face many technical limitations, especially as the size and complexity of the information environment increases, systems Traditionally built information systems do not satisfy users and information system managers.

The best solution that meets the ability to support the business while not affecting the operation of the business system is to build a data warehouse.

I. Introduction

One challenge of retail data modeling is the need to make decisions based on limited history. If Christmas comes but once a year, it's an opportunity to see how strategic decisions have impacted the bottom line.

In this report, we use historical sales data of 45 Walmart stores located in different regions. Each store contains multiple departments, and we had to forecast sales for each department in each store. Events marking the holidays were selected and included in the dataset. These drops are known to have an impact on sales, but predicting which departments are affected and to what extent is a challenge.

In addition, Walmart organizes several promotional events throughout the year. These drops come before prominent holidays, the four biggest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. Weeks that include these holidays are weighted five times more than non-holiday weeks. Part of the challenge this contest presents is modeling the impact of price drops during these weeks in the absence of adequate/ideal historical data.

stores.csv
This file contains anonymized information about the 45 stores, indicating the type and size of store. This is the historical training data, which covers 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

test.csv
This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

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

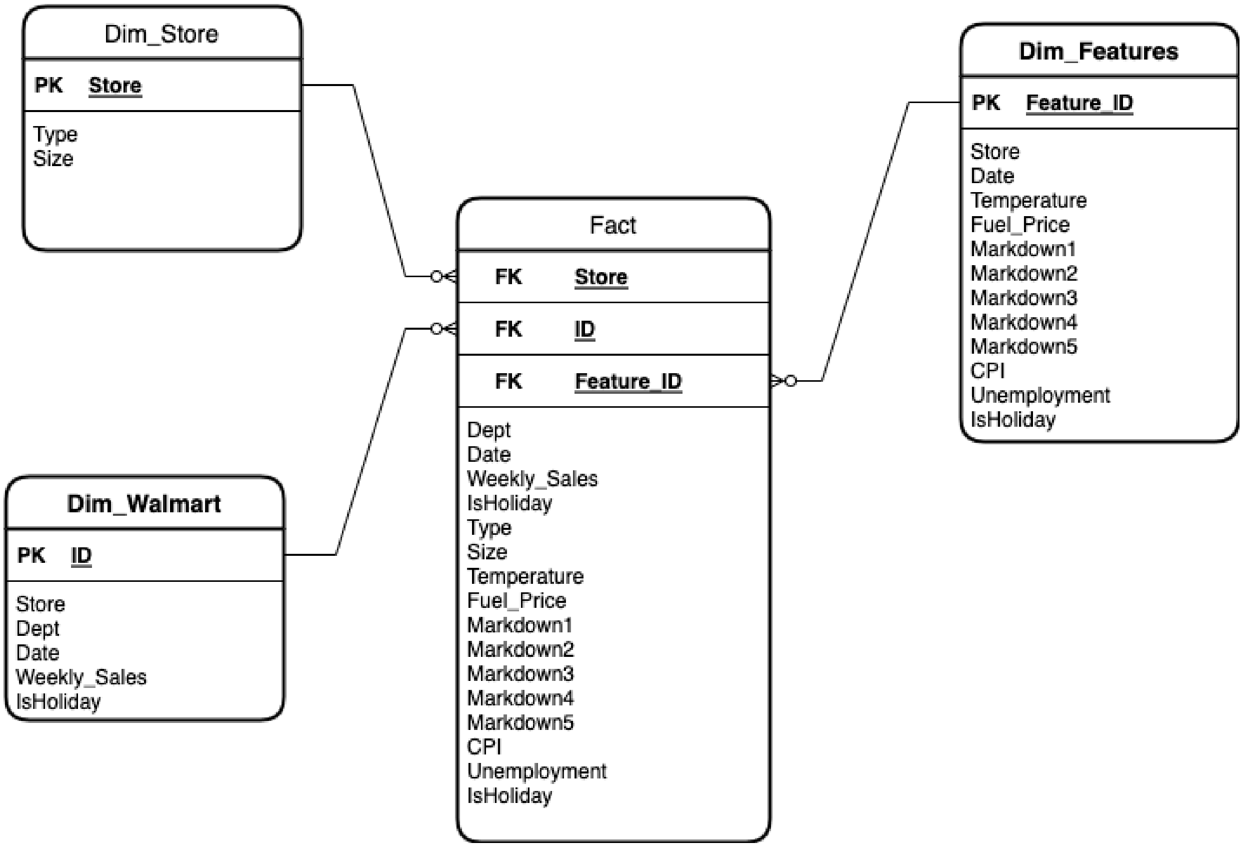
Asking questions:

1. Which store sells the most and least during the holiday?

2. Which time during the holiday has the highest and lowest total weekly sales and rank it?
3. Which time during the weekdays have the highest total weekly sales?
4. What type of store has highest sales revenue? Calculate in percentage?
5. Which department has the highest weekly sales?
6. Create table to see the weekly sale, size by Holiday and Type store
7. Determine the relationship between each column?
8. Which quarter has the highest average weekly sales?
9. Which type of store is the most popular?
10. Which store has the highest average weekly sales and Size?

II. Data Base Design

Dimensional Modeling - Star Schema Generation



We have 3 dimensional tables and 1 Fact table. Each primary key of dimensional table will connect with foreign key of Fact table. This means Fact table will contain dimension key column of dimension tables and other features to meet analytical needs.

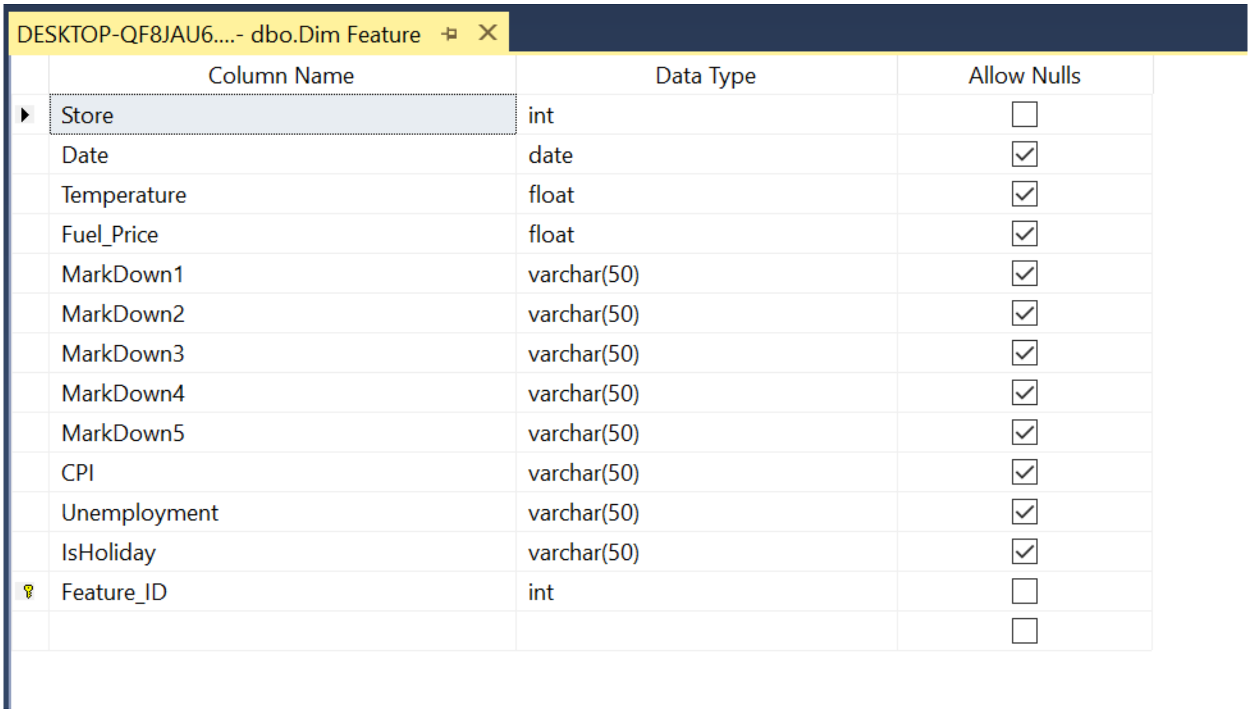
Build Data Base in SQL Server

Step 1: Create Database

In Object Explorer, right – click on Databases, select New Database, name this DB.

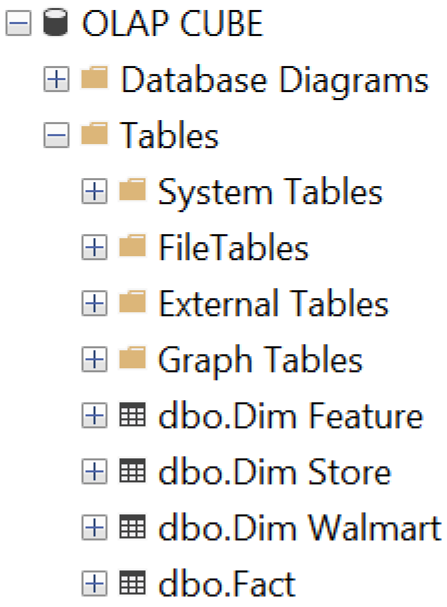
Step 2: Create Dimension Table

In Object Explorer, click on New Database that we just create and right – click on Tables, select New Table.



Column Name	Data Type	Allow Nulls
Store	int	☐
Date	date	☑
Temperature	float	☑
Fuel_Price	float	☑
MarkDown1	varchar(50)	☑
MarkDown2	varchar(50)	☑
MarkDown3	varchar(50)	☑
MarkDown4	varchar(50)	☑
MarkDown5	varchar(50)	☑
CPI	varchar(50)	☑
Unemployment	varchar(50)	☑
IsHoliday	varchar(50)	☑
Feature_ID	int	☐
		☐

We will create table like the figure above, and do the same with other Dimensions.

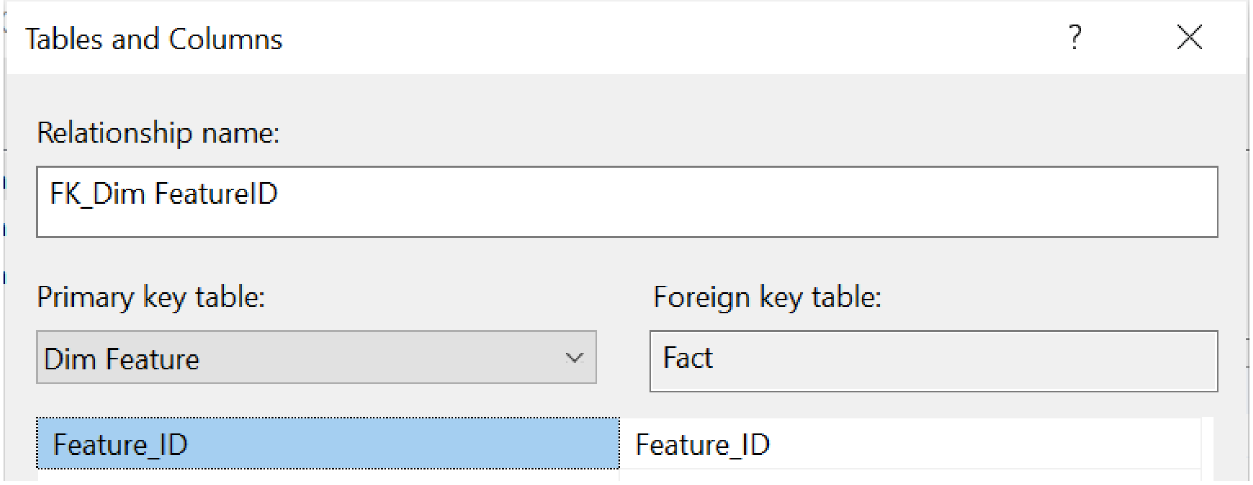


- OLAP CUBE
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - External Tables
    - Graph Tables
    - dbo.Dim Feature
    - dbo.Dim Store
    - dbo.Dim Walmart
    - dbo.Fact

After finished, we will have 4 tables like this figure.

Step 3: Create Relationship between Dimension tables and Fact Table

Right – click on dbo.Fact, select Design and right – click on random column then select Relationships. Then we set the relationship between dimension tables and Fact table by connect PK and FK like figue below:



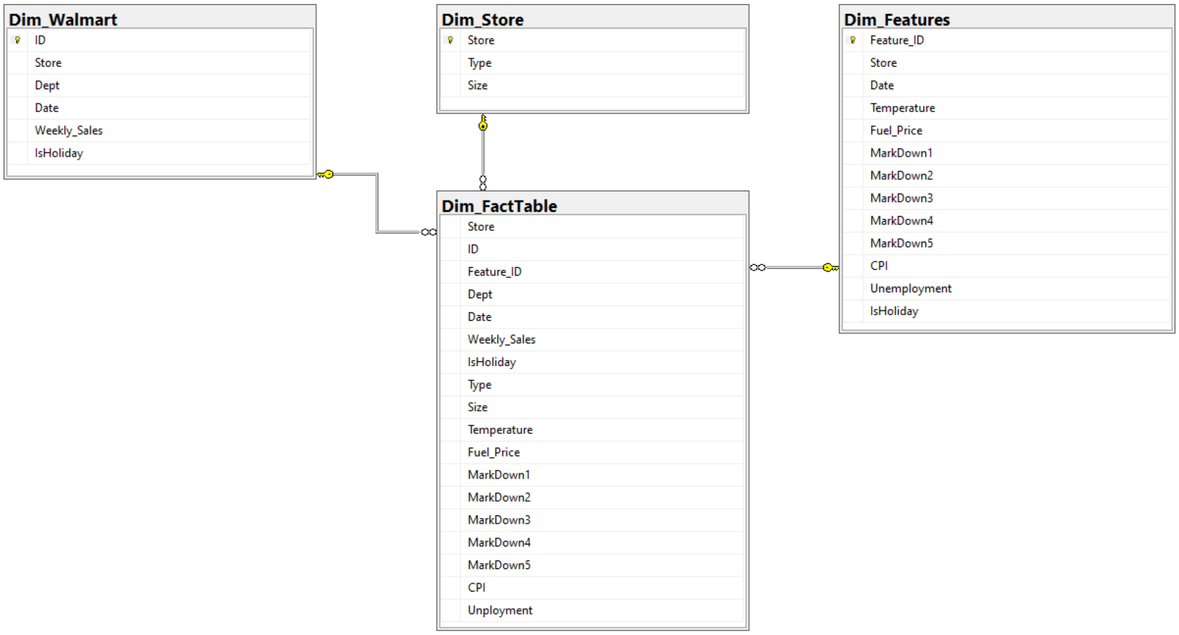
Tables and Columns

Relationship name:

FK_Dim FeatureID

Primary key table:	Foreign key table:
Dim Feature	Fact
Feature_ID	Feature_ID

After connect 3 dimension tables with fact table, we have 3 relationship. We can visual this by right – click on Database Diagrams then click New Database Diagram. The result below:

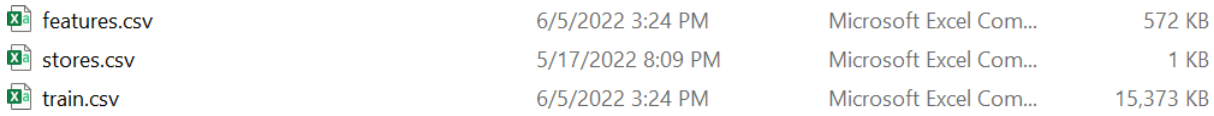


III. Building DataWareHouse

To build a Data Warehouse, we need to do 4 steps. In ETL process, we using SQL Server Integration Services Projects (SSIS)

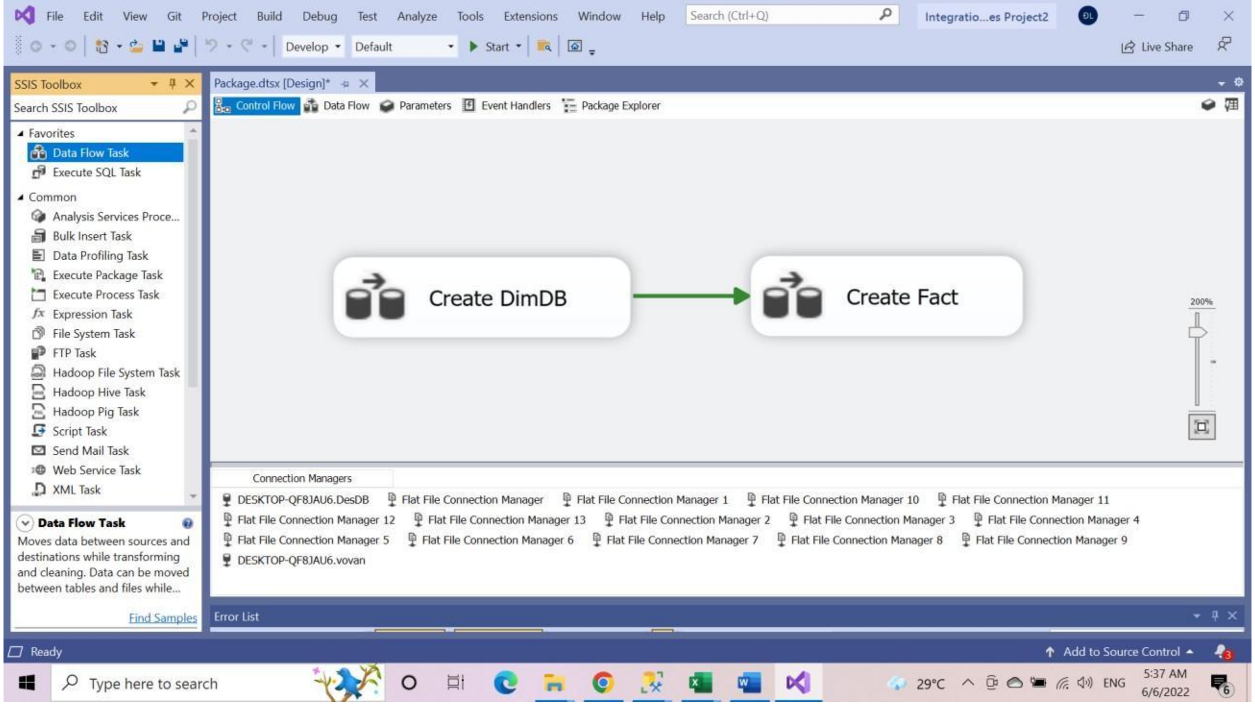
Step 1: Extract

We need to extract data from csv file. We have 3 .csv file in local computer. What we need to do is extract them to Dimension table.

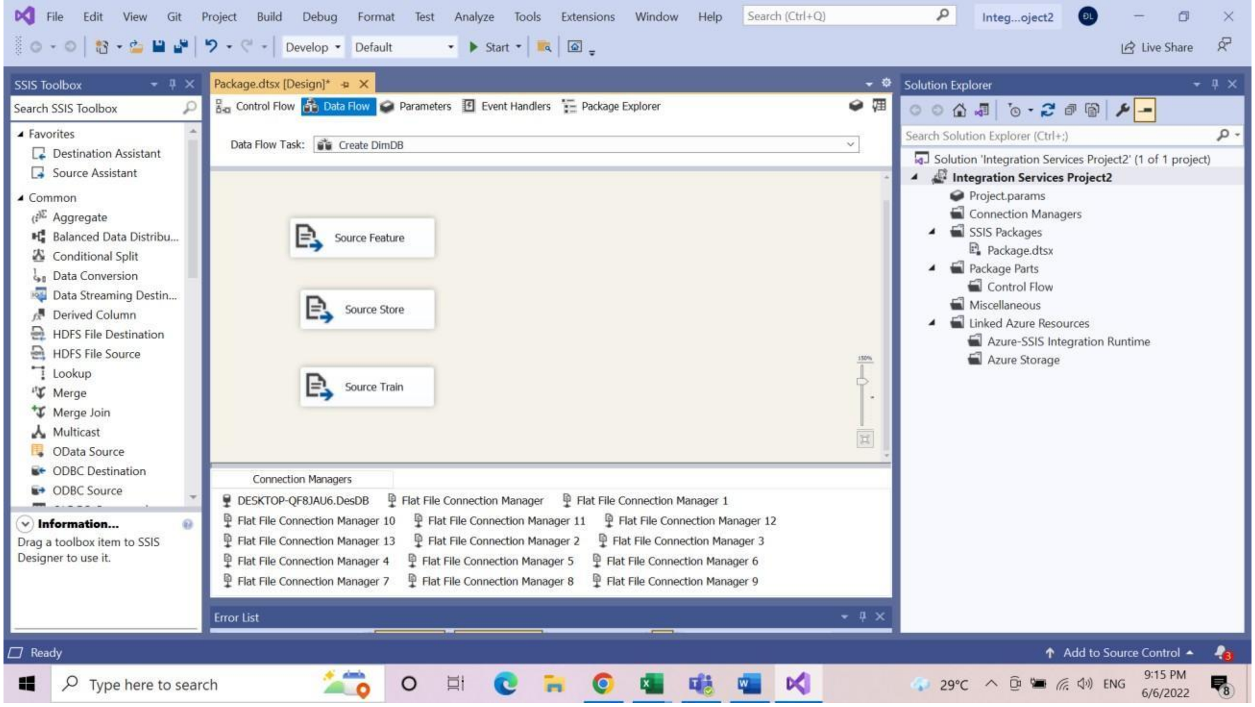


features.csv	6/5/2022 3:24 PM	Microsoft Excel Com...	572 KB
stores.csv	5/17/2022 8:09 PM	Microsoft Excel Com...	1 KB
train.csv	6/5/2022 3:24 PM	Microsoft Excel Com...	15,373 KB

Firstly, create new SSIS project, to create data flow to extract data, we must to create control flow. In SSIS Toolbox, drag Data Flow Task and drop into control flow then clip to it. We create two Data Flow Task one for Dimension Table and other for Fact Table.


Package.dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
SSIS Toolbox
Favorites
Data Flow Task
Execute SQL Task
Common
Create DimDB
Create Fact
Connection Managers
DESKTOP-QF8JAU6.DesDB
Flat File Connection Manager
Flat File Connection Manager 1
Flat File Connection Manager 10
Flat File Connection Manager 11
Flat File Connection Manager 12
Flat File Connection Manager 13
Flat File Connection Manager 2
Flat File Connection Manager 3
Flat File Connection Manager 4
Flat File Connection Manager 5
Flat File Connection Manager 6
Flat File Connection Manager 7
Flat File Connection Manager 8
Flat File Connection Manager 9
DESKTOP-QF8JAU6.vovan
Data Flow Task
Moves data between sources and destinations while transforming and cleaning. Data can be moved between tables and files while...
Find Samples
Error List
Ready

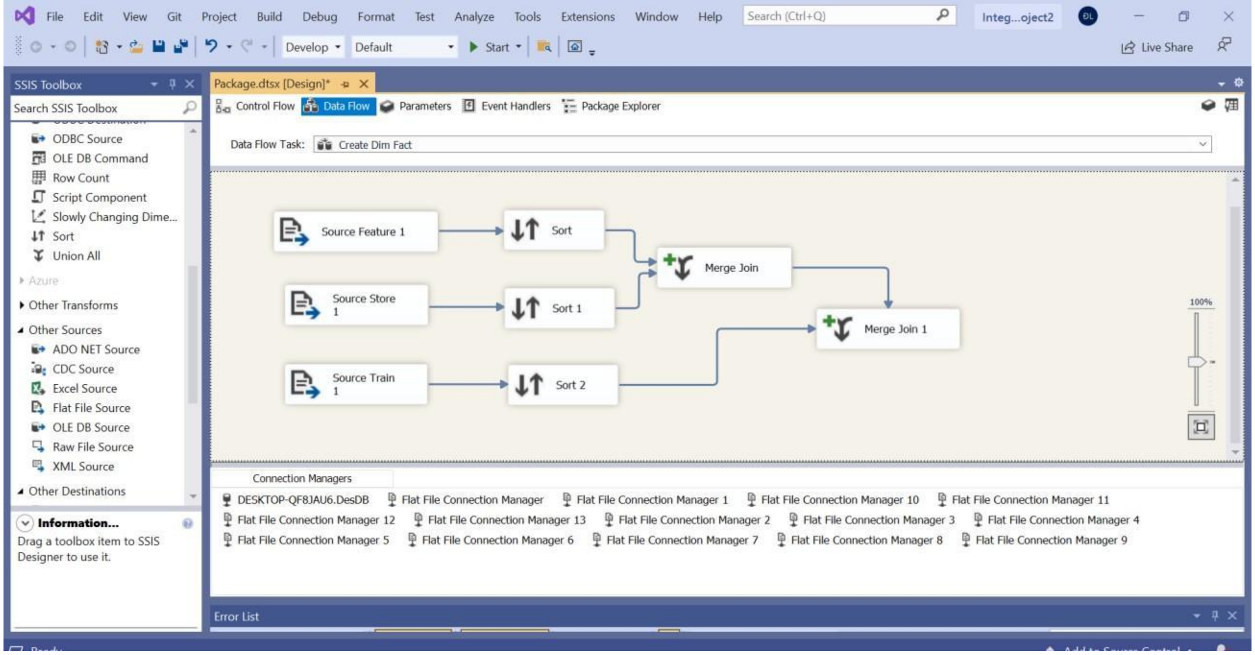
The first is the flow to extract dimension tables to SQL server. Double click to Data Flow Task, we will automatically move to Data Flow. Using Flat File Source in SSIS Toolbox, we read three csv file



We do the same as Create Fact Data Flow.

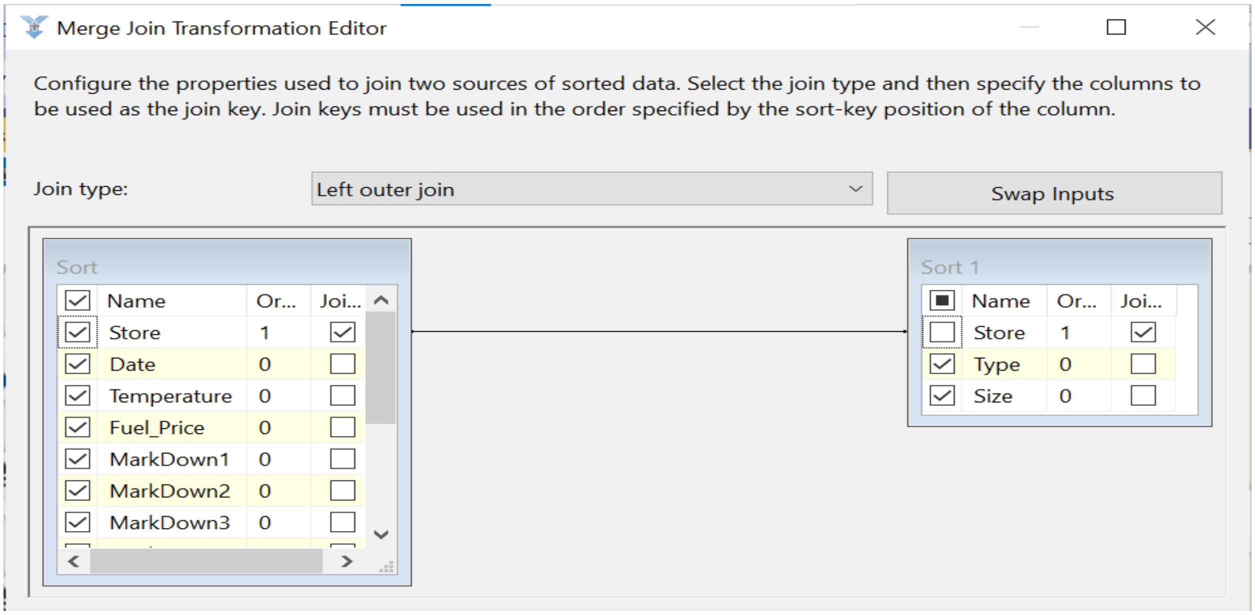
Step 2: Transform

In Transforming step, we will do it in Create Fact Data Flow.



Data Flow Task: Create Dim Fact

The 3 source data that we extracted in previous step will be sorted by Store. Then, the first sort and the second sort will join together by Merge Join. Right – click to Merge Join to modify type of join is left join. n and join on Store column.



Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

Join type: Left outer join

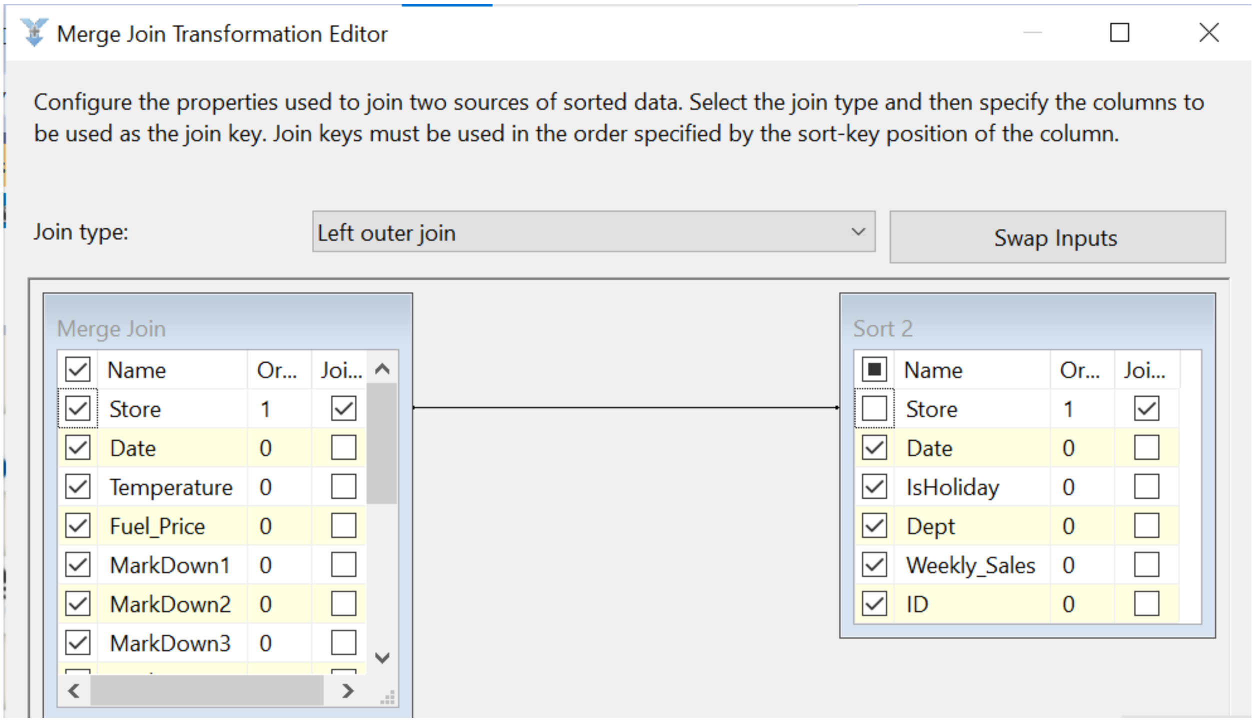
Sort

☑	Name	Or...	Joi...
☑	Store	1	☑
☑	Date	0	□
☑	Temperature	0	□
☑	Fuel_Price	0	□
☑	MarkDown1	0	□
☑	MarkDown2	0	□
☑	MarkDown3	0	□

Sort 1

■	Name	Or...	Joi...
□	Store	1	☑
☑	Type	0	□
☑	Size	0	□

Next, we still using left outer join. Merged the previous Merge join and the third sort.



Merge Join Transformation Editor

Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

Join type: Left outer join

Merge Join

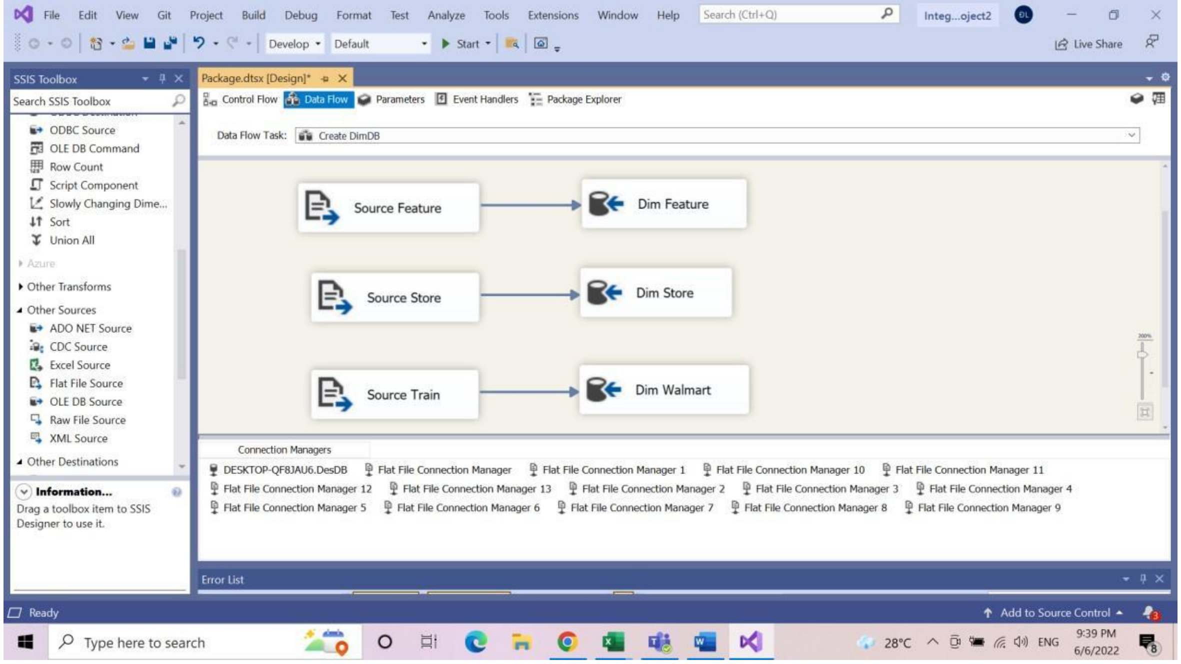
☑	Name	Or...	Joi...
☑	Store	1	☑
☑	Date	0	□
☑	Temperature	0	□
☑	Fuel_Price	0	□
☑	MarkDown1	0	□
☑	MarkDown2	0	□
☑	MarkDown3	0	□

Sort 2

■	Name	Or...	Joi...
□	Store	1	☑
☑	Date	0	□
☑	IsHoliday	0	□
☑	Dept	0	□
☑	Weekly_Sales	0	□
☑	ID	0	□

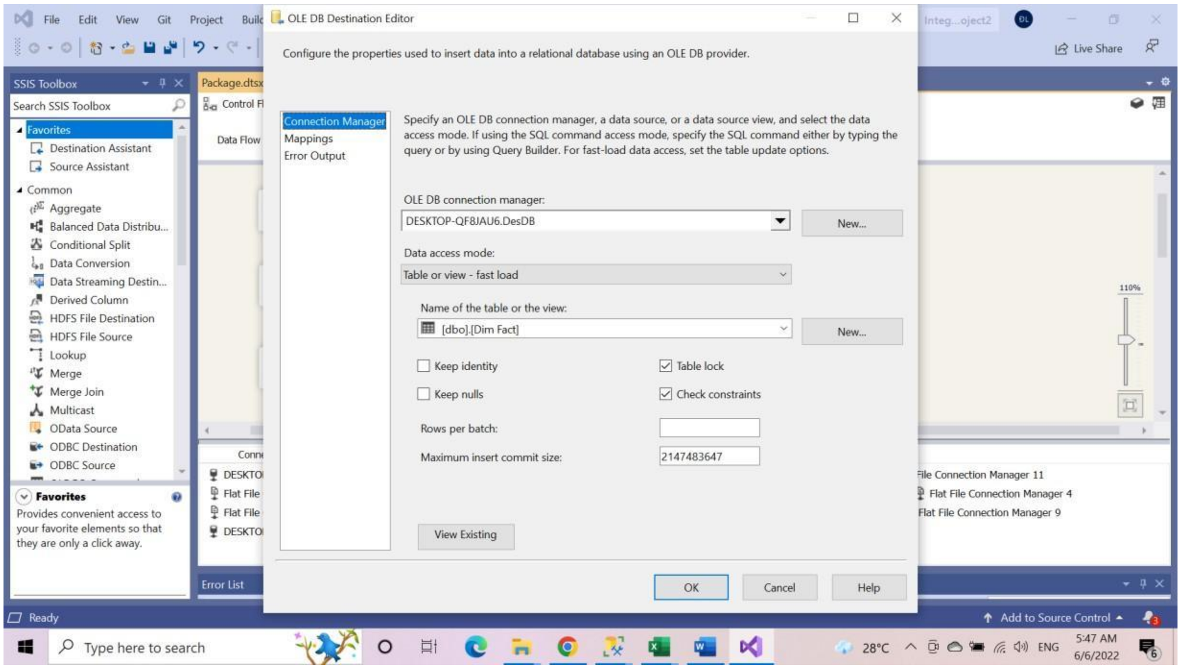
Step 3: Loading

The final step in ETL is Loading. In Loading, we will link it to SQL Server by using OLE DB Destination.



Data Flow Task: Create DimDB

After to create a destination, double click on OLE DB Destination and select your SQL Server name and name of table destination and do the same with other table The example is in figure below:



OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager: DESKTOP-QF8JAU6.DesDB

Data access mode: Table or view - fast load

Name of the table or the view: [dbo].[Dim Fact]

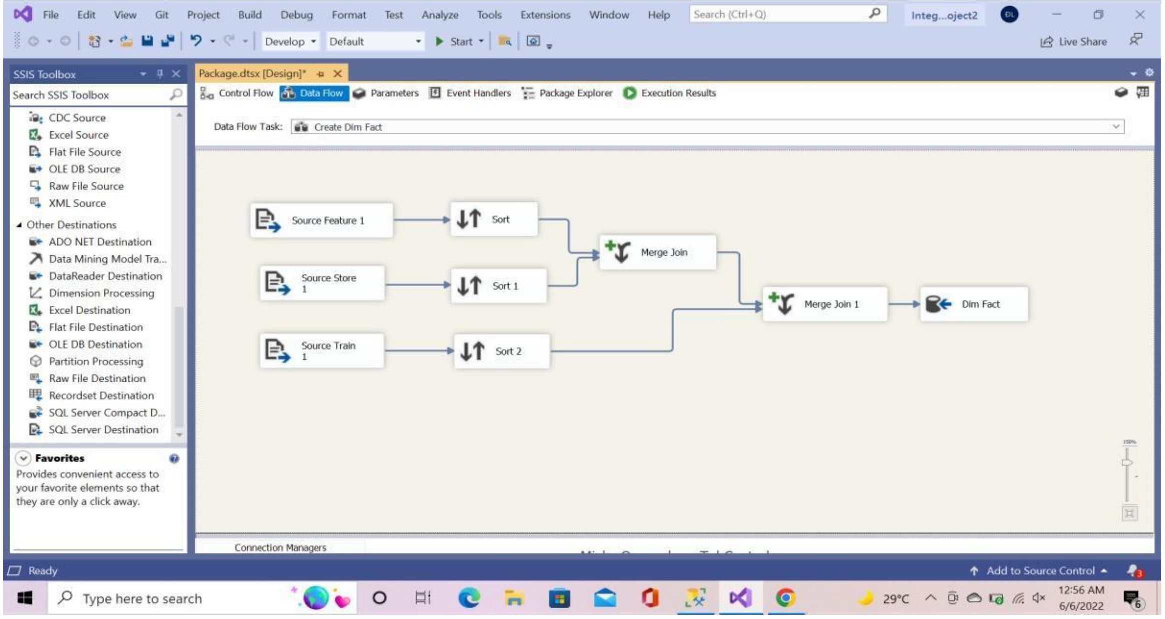
☐ Keep identity ☑ Table lock

☐ Keep nulls ☑ Check constraints

Rows per batch:

Maximum insert commit size: 2147483647

Next, we will move to create Fact Table, we do the same in Control Flow section, in Data Flow, we will move 3 table by using left join. However, before doing merged, we need to sort data. After merge join finish, linked it to Fact table.



Data Flow Task: Create Dim Fact

Finally, finished all task, data has been loaded to DataWareHouse. Preparing for the next step.

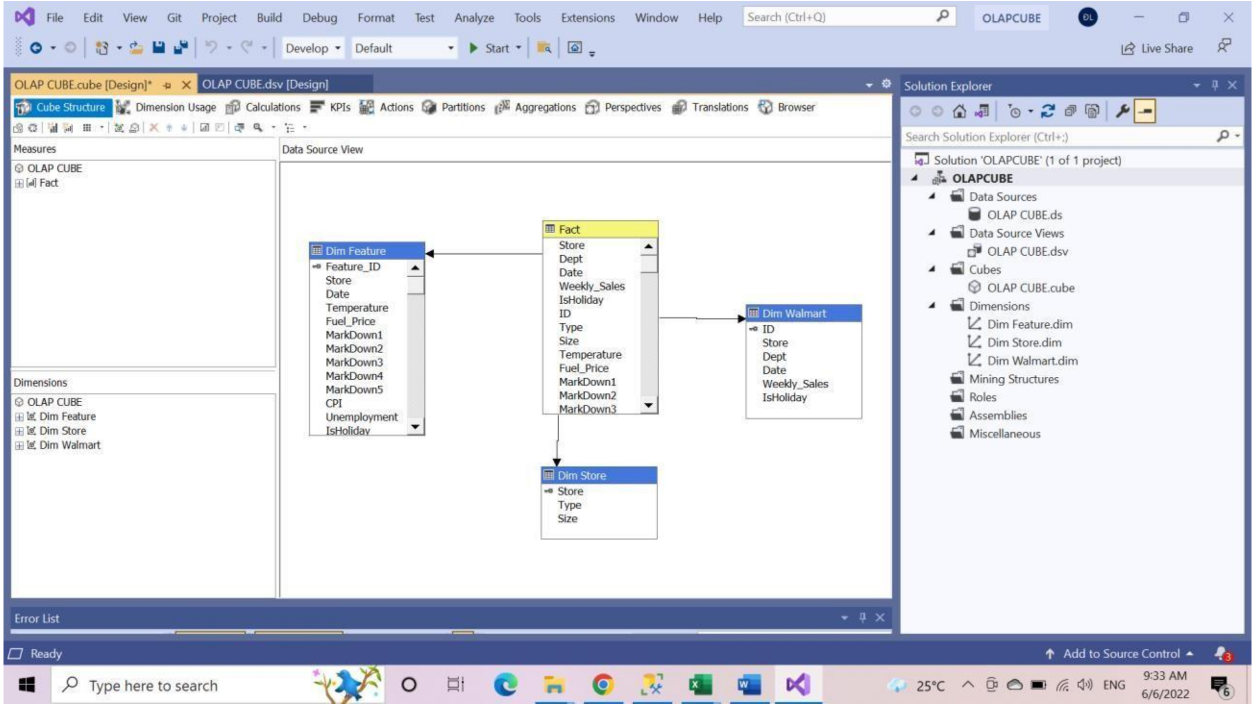
Step 4: Build the OLAP (On-line transactional processing) Cube and analysis (SSAS – Analysis Services Multidimensional and Data Mining Project)

Firstly, create new project and select Analysis Services Multidimensional and Data Mining Project. In Solution Explorer, right – click on Data Source, select New Data Source to connect with data source.

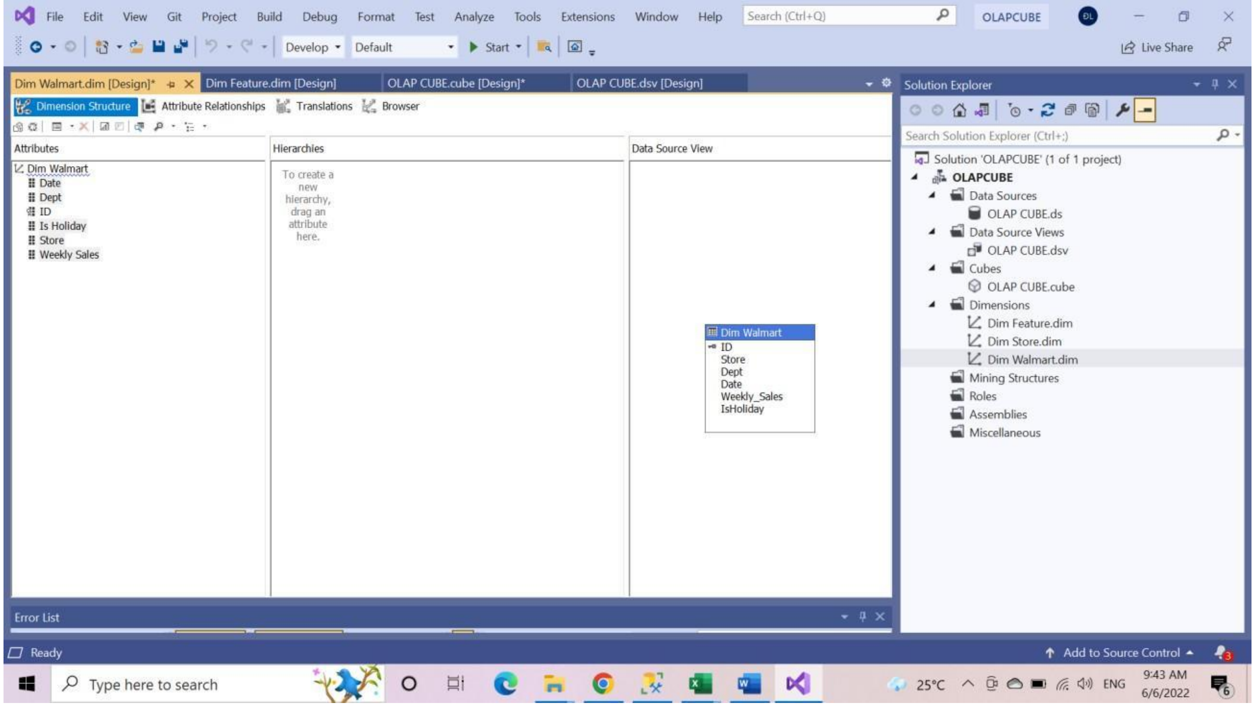
Secondly, right – click on Data Source Views and click New Data Sources, select relationship table to make a view.

Thirdly, right – click on Cubes and click New Cube, select the Fact table in Measure Group Tables then select all Measure, next select all Dimension tables then finish.

Here is our cube:

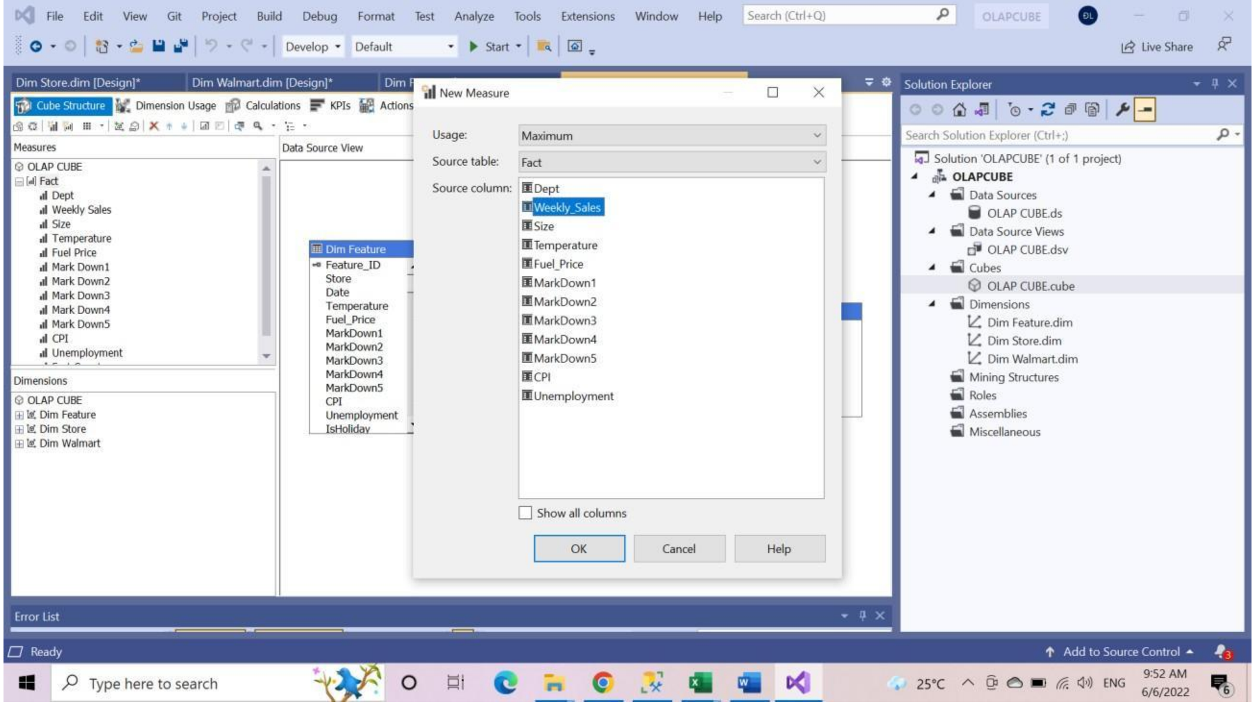


Fourth, select attributes by click dimension tables in Dimensions.



Fifth, adding measures for analysis purpose.

Right-click on Fact, click new measure and select measure that want to add.



Usage: Maximum

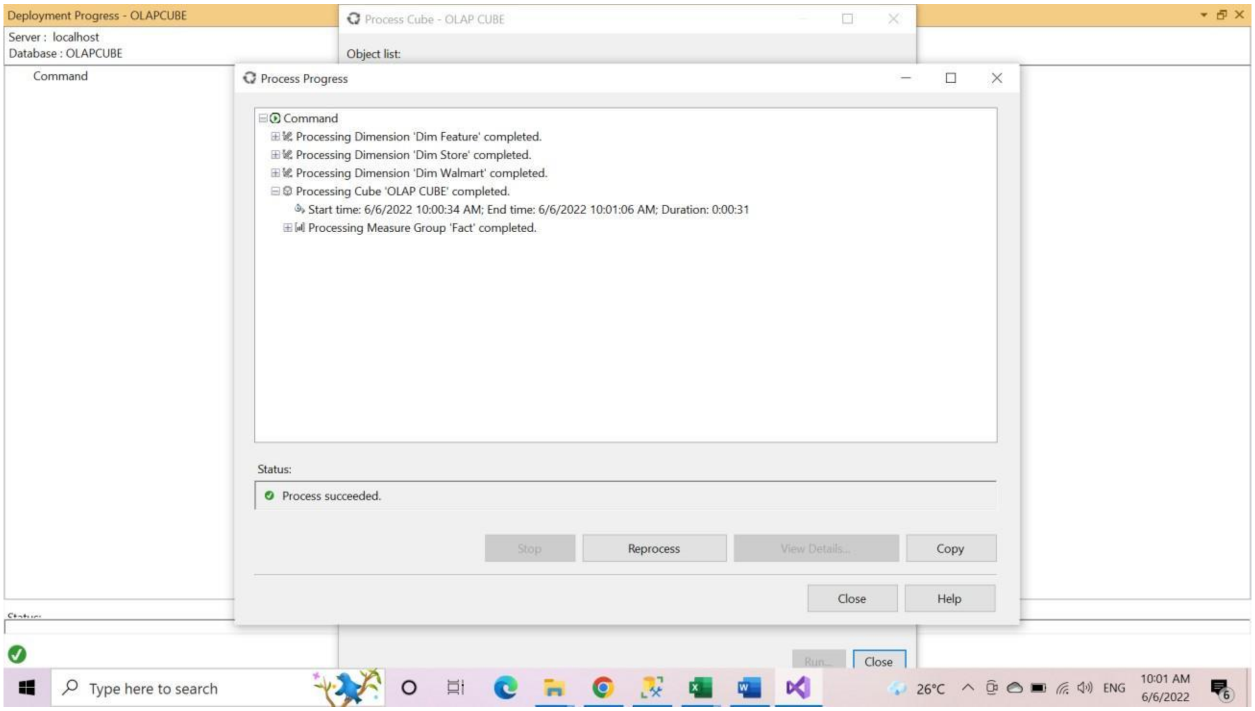
Source table: Fact

Source column: Weekly_Sales

□ Show all columns

Final step is process and deploy cube to Brower.

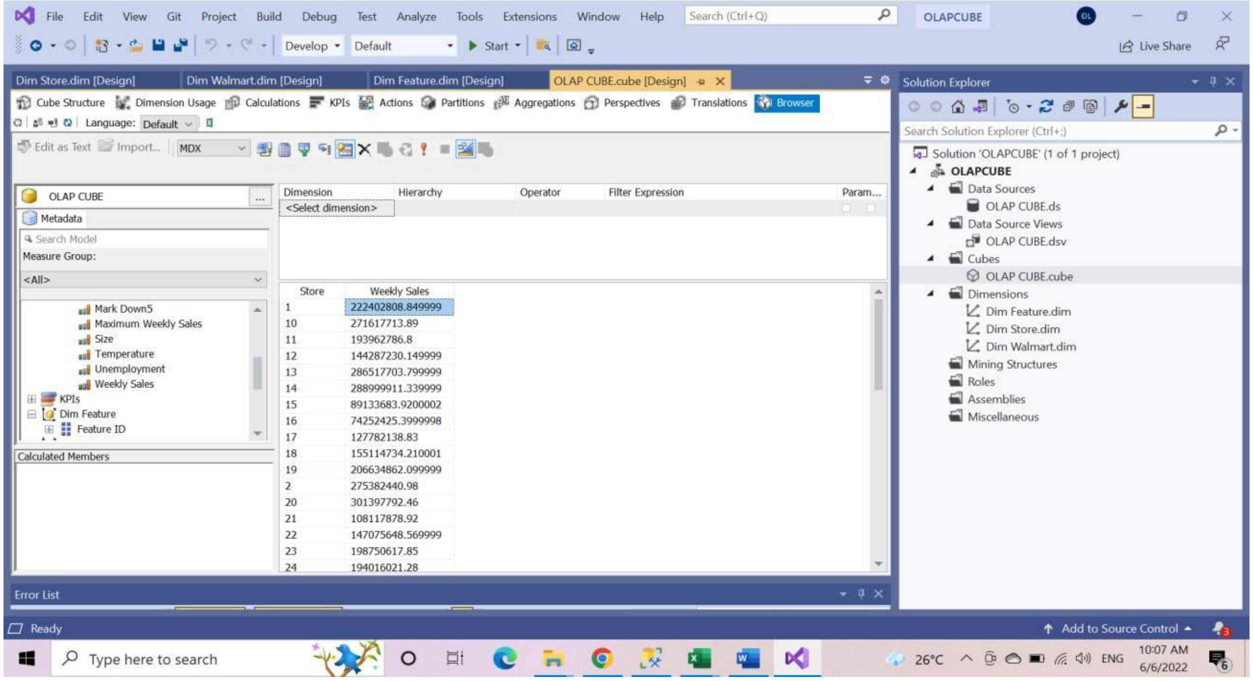
Right – click to the cube that we have already create and click Process



```
Command
  Processing Dimension 'Dim Feature' completed.
  Processing Dimension 'Dim Store' completed.
  Processing Dimension 'Dim Walmart' completed.
  Processing Cube 'OLAP CUBE' completed.
    Start time: 6/6/2022 10:00:34 AM; End time: 6/6/2022 10:01:06 AM; Duration: 0:00:31
    Processing Measure Group 'Fact' completed.
```

Status: Process succeeded.

When deploy successful, click to Browser to do same simple query. For example, we calculate sum of weekly sales in each store.

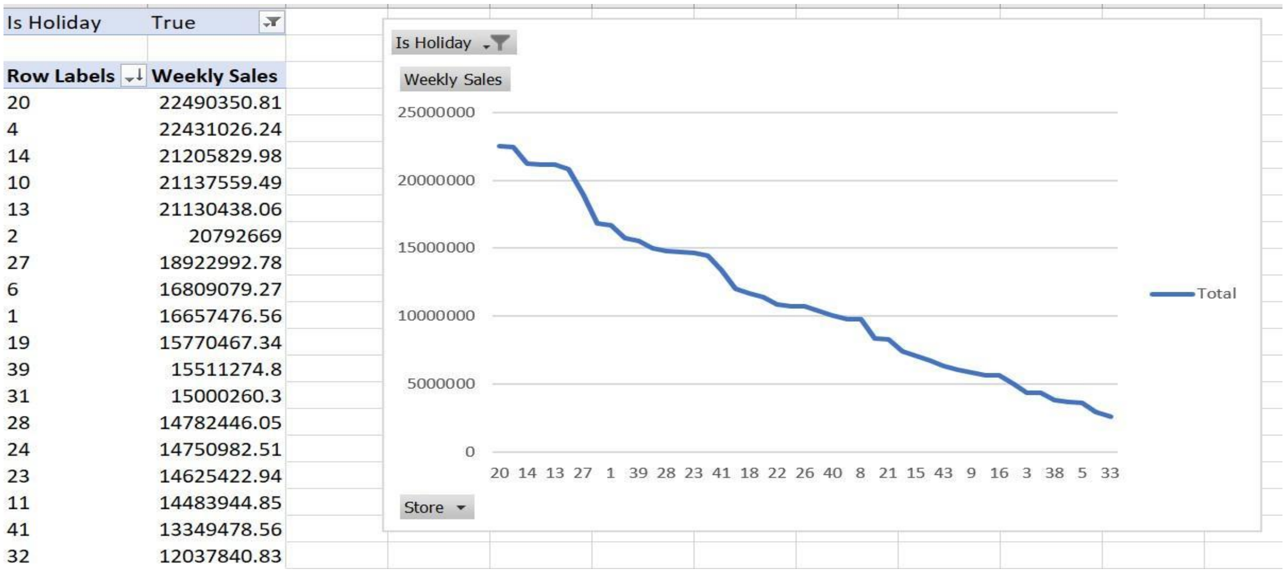


Store	Weekly Sales
1	222402808.849999
10	271617713.89
11	193962786.8
12	144287230.149999
13	286517703.799999
14	288999911.339999
15	89133683.9200002
16	74252425.3999998
17	127782138.83
18	155114734.210001
19	206634862.099999
2	275382440.98
20	301397792.46
21	108117878.92
22	147075648.569999
23	198750617.85
24	194016021.28

Answering the following questions:

However, most of time we will work with pivot table in excel. Just click on Excel symbol in Language near with Default. We will use excel pivot-table for answering the question.

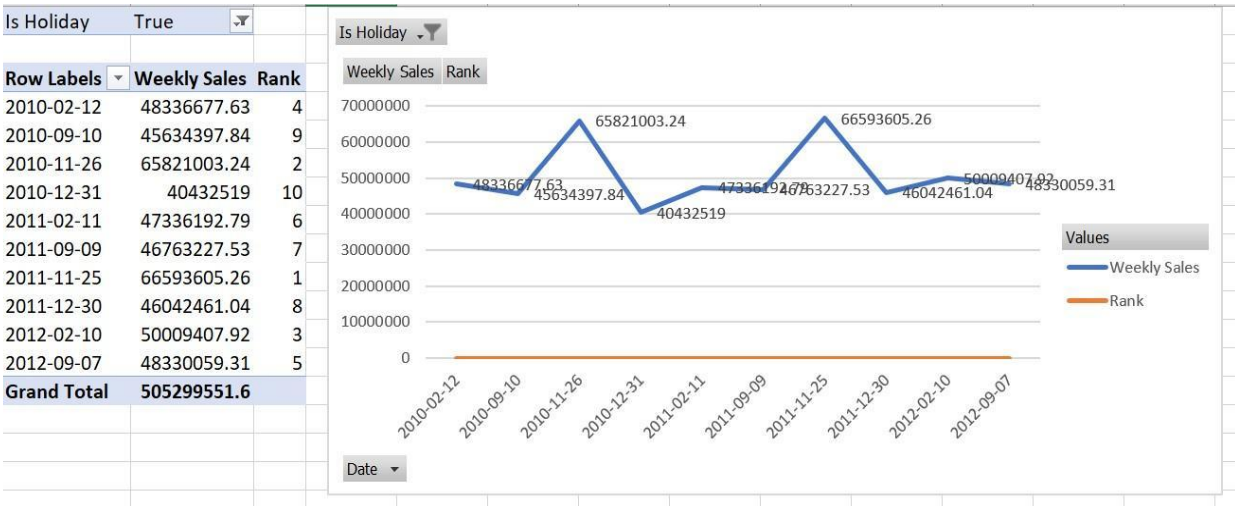
1. **Which store sells the most and least during the holiday?**



Is Holiday	True
Row Labels	**Weekly Sales**
20	22490350.81
4	22431026.24
14	21205829.98
10	21137559.49
13	21130438.06
2	20792669
27	18922992.78
6	16809079.27
1	16657476.56
19	15770467.34
39	15511274.8
31	15000260.3
28	14782446.05
24	14750982.51
23	14625422.94
11	14483944.85
41	13349478.56
32	12037840.83

Insight: Store 20 have the highest total sum of weekly sales during the holiday while Store 33 have the lowest total sum of weekly sales during the holiday.

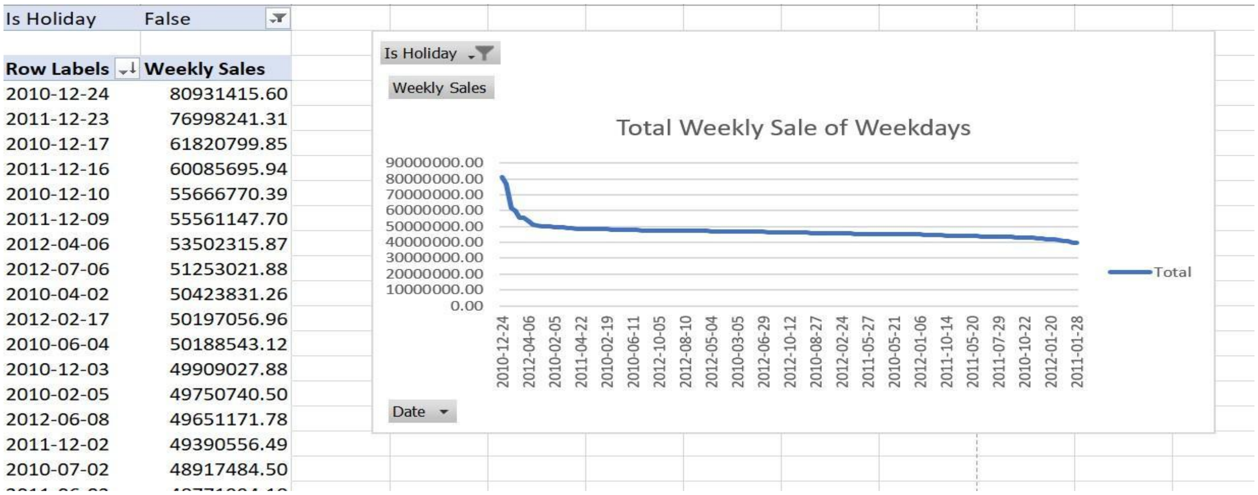
2. **Which time during the holiday has the highest and lowest total weekly sales and rank it?**



Is Holiday	True	
Row Labels	**Weekly Sales**	**Rank**
2010-02-12	48336677.63	4
2010-09-10	45634397.84	9
2010-11-26	65821003.24	2
2010-12-31	40432519	10
2011-02-11	47336192.79	6
2011-09-09	46763227.53	7
2011-11-25	66593605.26	1
2011-12-30	46042461.04	8
2012-02-10	50009407.92	3
2012-09-07	48330059.31	5
Grand Total	**505299551.6**	

Insight: 25/11/2011 have the highest weekly sale, the following is 26/11/2010, the third is 10/2/2012. So we can conclude that the last quarter of year will have the highest weekly sale in overall.

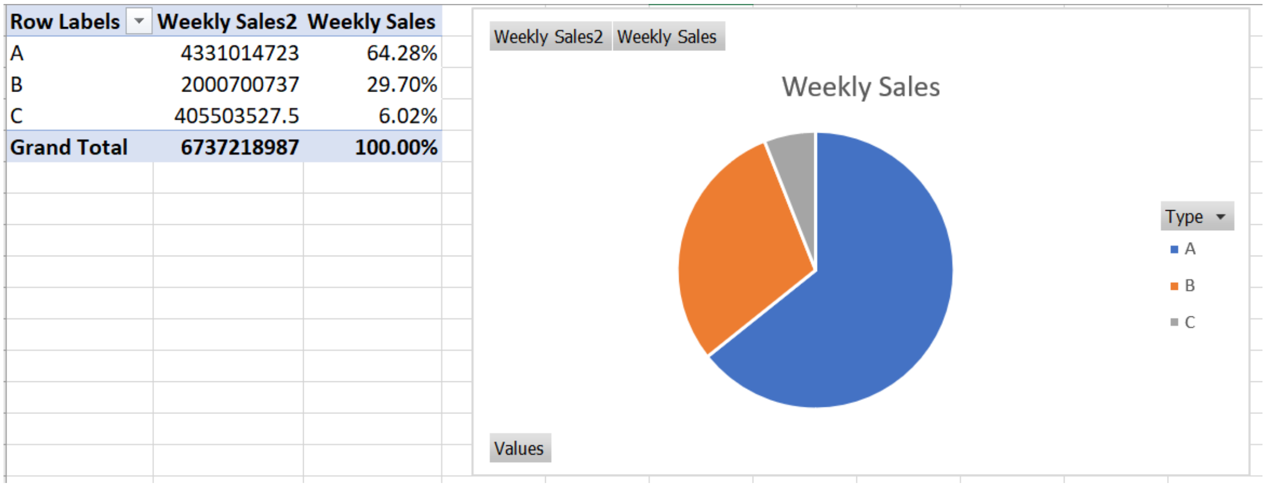
3. **Which time during the weekdays have the highest total weekly sales?**



Is Holiday	False
Row Labels	**Weekly Sales**
2010-12-24	80931415.60
2011-12-23	76998241.31
2010-12-17	61820799.85
2011-12-16	60085695.94
2010-12-10	55666770.39
2011-12-09	55561147.70
2012-04-06	53502315.87
2012-07-06	51253021.88
2010-04-02	50423831.26
2012-02-17	50197056.96
2010-06-04	50188543.12
2010-12-03	49909027.88
2010-02-05	49750740.50
2012-06-08	49651171.78
2011-12-02	49390556.49
2010-07-02	48917484.50

Insight: In the normal week (not in holiday week), 24/12/2010 has the highest weekly sales. The following is 23/12/2011. As we can see most of highest weekly sales come from December or the last quarter of the year.

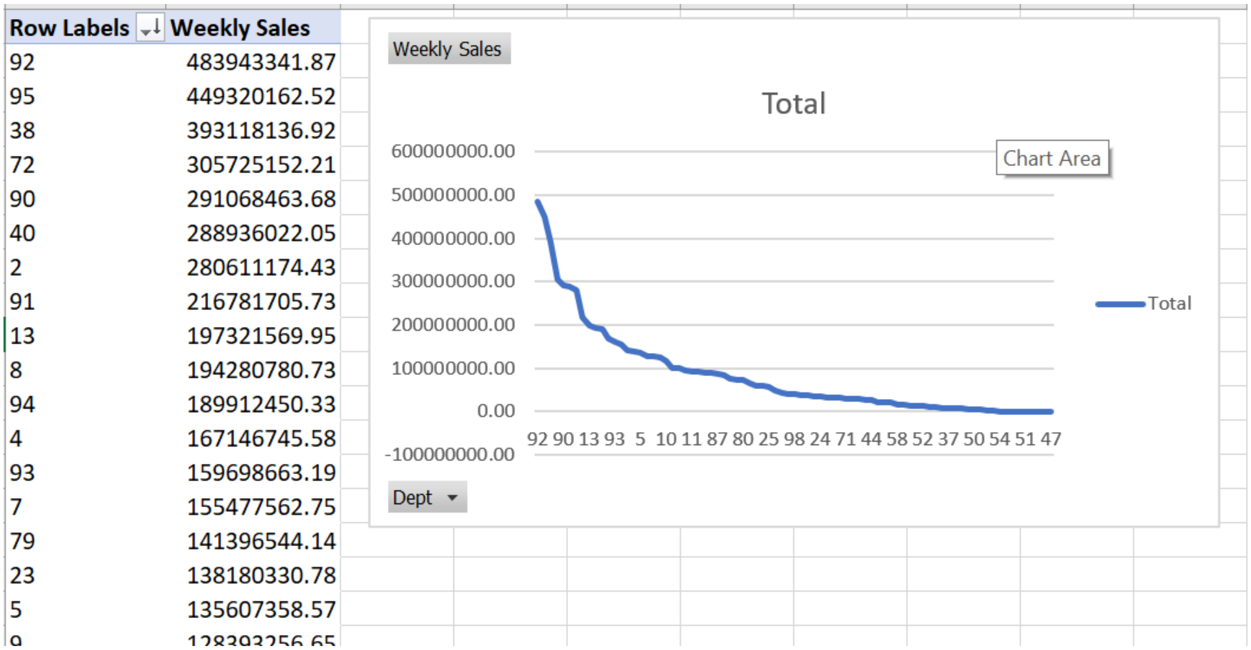
4. **What type of store has highest sales revenue? Calculate in percentage?**



Row Labels	Weekly Sales2	Weekly Sales
A	4331014723	64.28%
B	2000700737	29.70%
C	405503527.5	6.02%
Grand Total	**6737218987**	**100.00%**

Insight: Store A has the highest sales revenue, accounting for 64.28% of total sales.

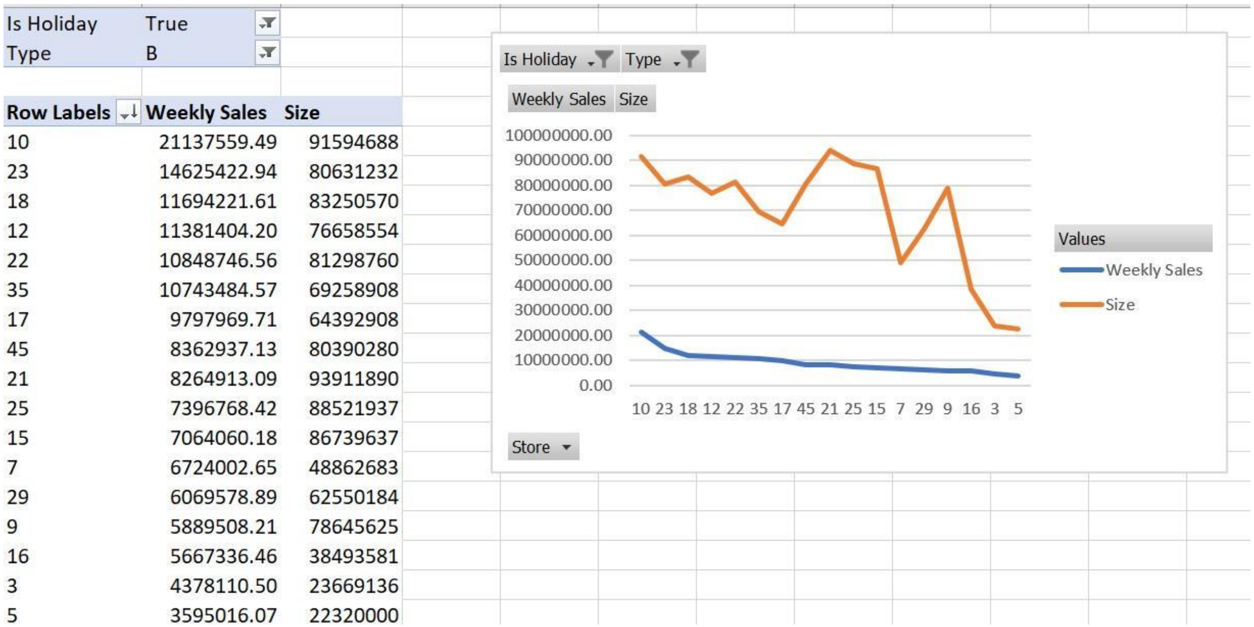
5. **Which department has the highest weekly sales?**



Row Labels	Weekly Sales
92	483943341.87
95	449320162.52
38	393118136.92
72	305725152.21
90	291068463.68
40	288936022.05
2	280611174.43
91	216781705.73
13	197321569.95
8	194280780.73
94	189912450.33
4	167146745.58
93	159698663.19
7	155477562.75
79	141396544.14
23	138180330.78
5	135607358.57
9	128393256.65

Insight: Department 92 has the highest sales revenue.

6. Create table to see the weekly sale, size by Holiday and Type store



Is Holiday	True
Type	B

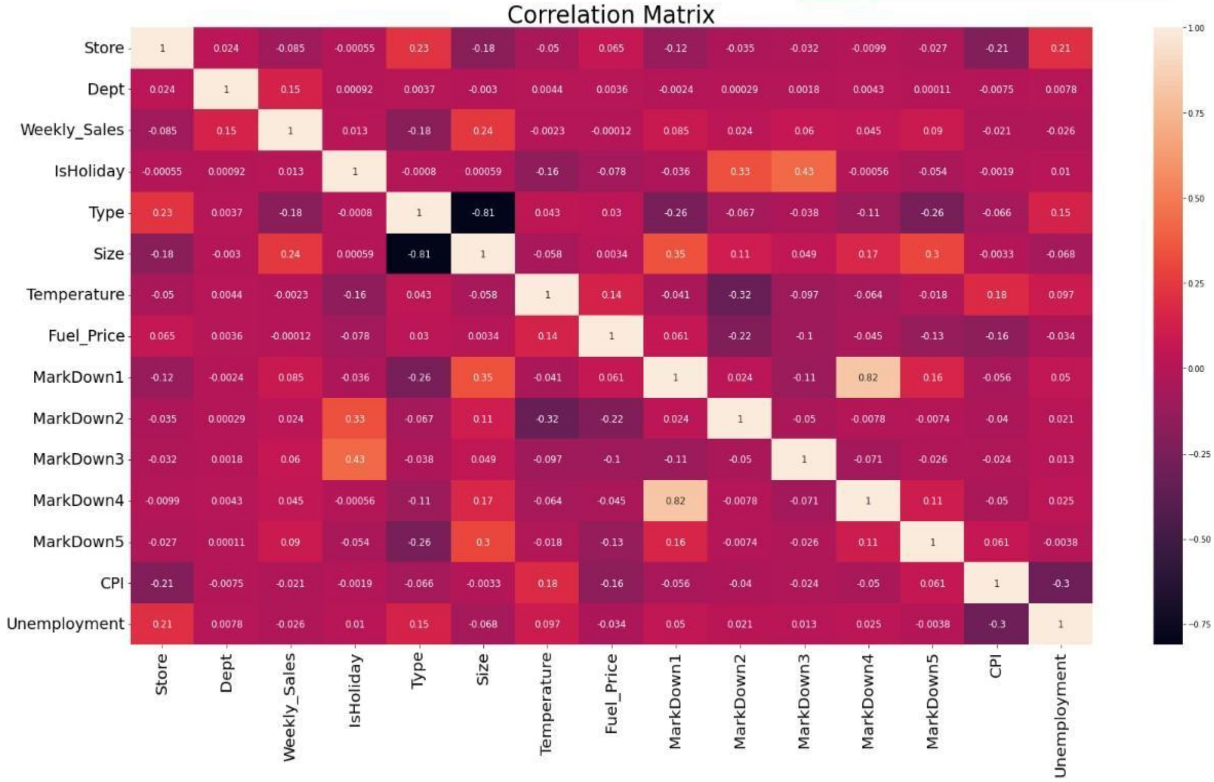
Row Labels	Weekly Sales	Size
10	21137559.49	91594688
23	14625422.94	80631232
18	11694221.61	83250570
12	11381404.20	76658554
22	10848746.56	81298760
35	10743484.57	69258908
17	9797969.71	64392908
45	8362937.13	80390280
21	8264913.09	93911890
25	7396768.42	88521937
15	7064060.18	86739637
7	6724002.65	48862683
29	6069578.89	62550184
9	5889508.21	78645625
16	5667336.46	38493581
3	4378110.50	23669136
5	3595016.07	22320000

Insight: When the size decrease, the weekly sales also decrease for all of store types.

IV. Data Mining

Tableau for Visualization

7. Determine the relationship between each column?

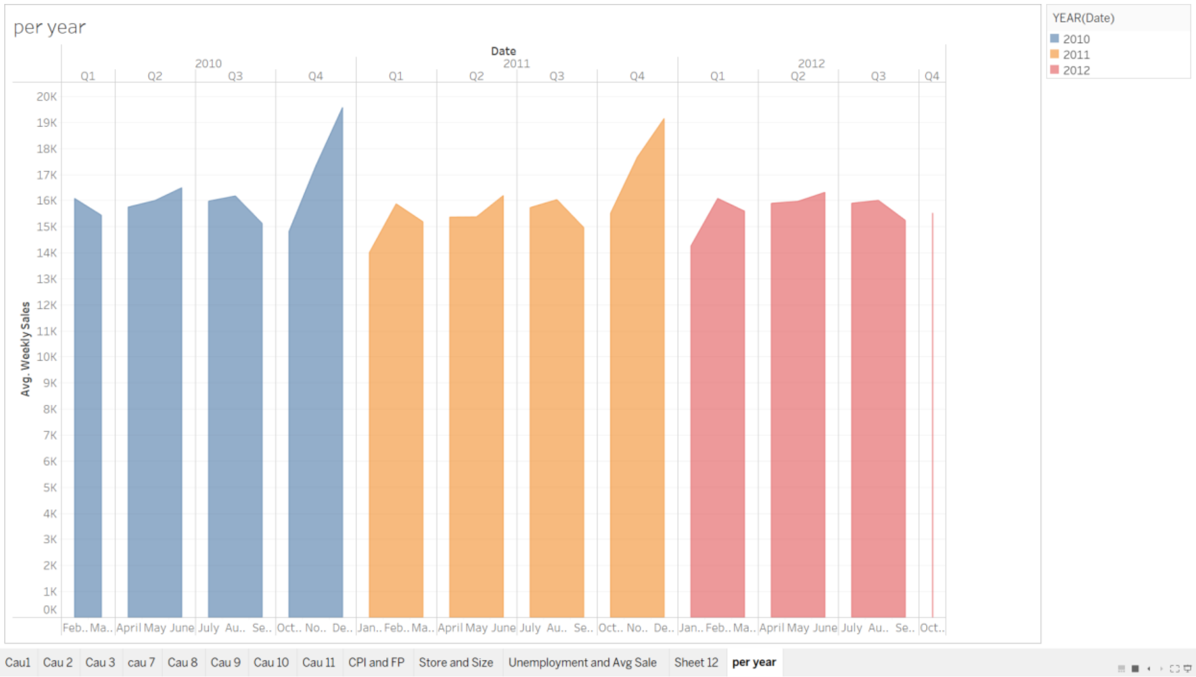


As the correlation between the Size and the Weekly_Sales is 0.24. It is a quite strong positive relationship. It can be concluded that when the Size increases, the Weekly_Sales also increases.

When the Size increases, MarkDown also increases.

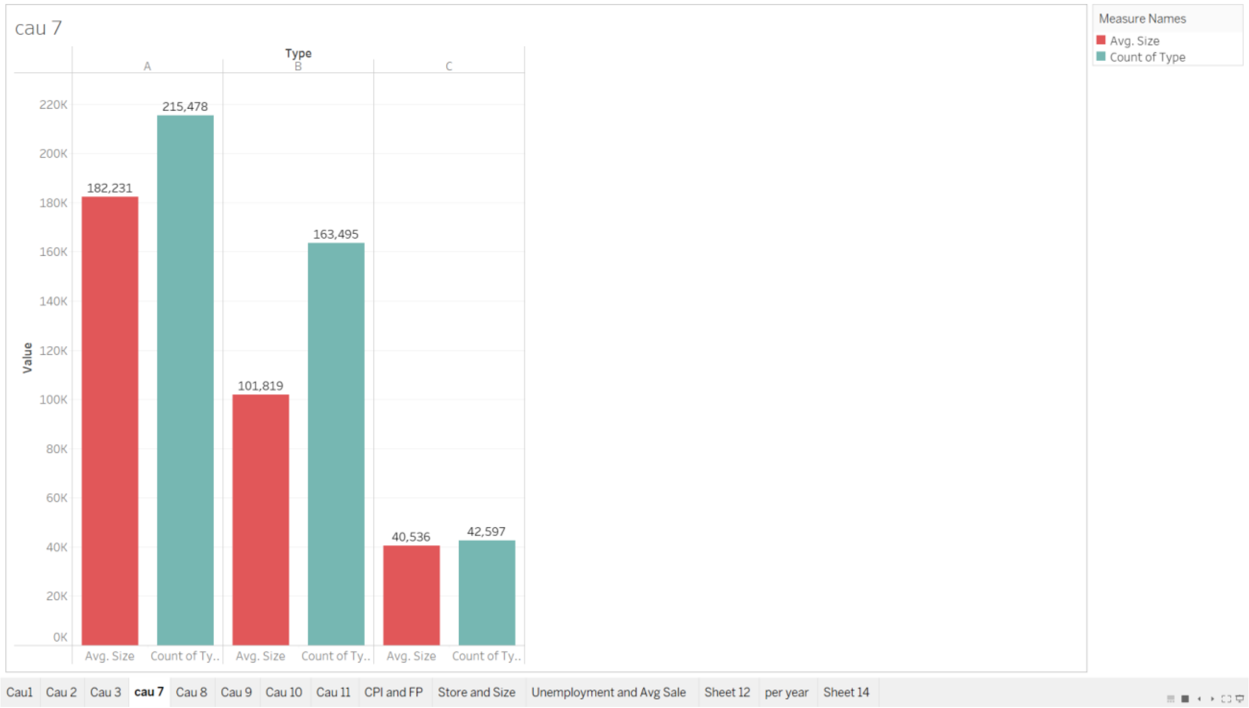
As the correlation between CPI and Unemployment is -0.3. It is a strongly negative relationship. It can be concluded that when the CPI increases, Unemployment decreases.

8. Which quarter has the highest average weekly sales?



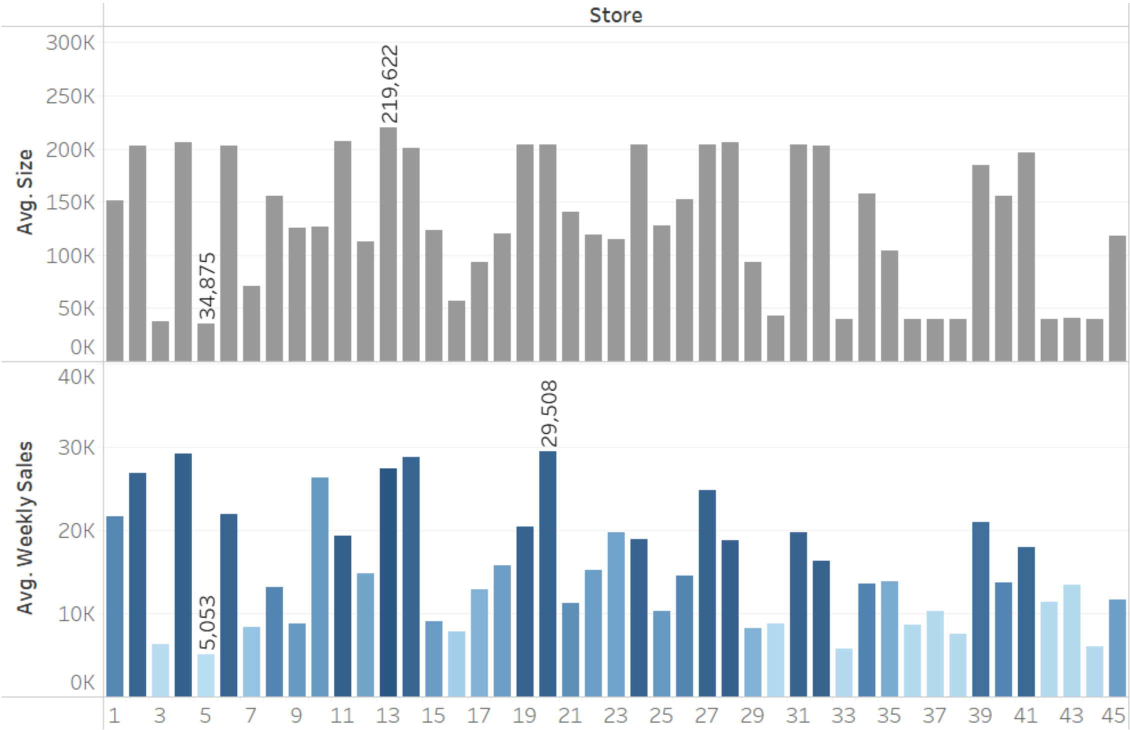
Insight: For both year 2010 and 2011 the average weekly sales highest in the fourth quarter. However, in 2012 we havent have data of the last quarter but we quite sure that the fourth quarter of 2012 is also high. In conclusion, Highest average sales at the last quarter.

9. Which type of store is the most popular?



Insight: Type A have the highest number of store and also have the highest average size.

10. Which store has the highest average weekly sales and Size?



Insight: We can see that stores have large size will have higher weekly sale than stores have small size. Store 20 has the highest average weekly sales and Store 13 has largest size.

Data Mining – SSIS

After analysis data to get the insight, now we can move to data mining. We use Analysis Services Multidimensional and Data Mining Project.

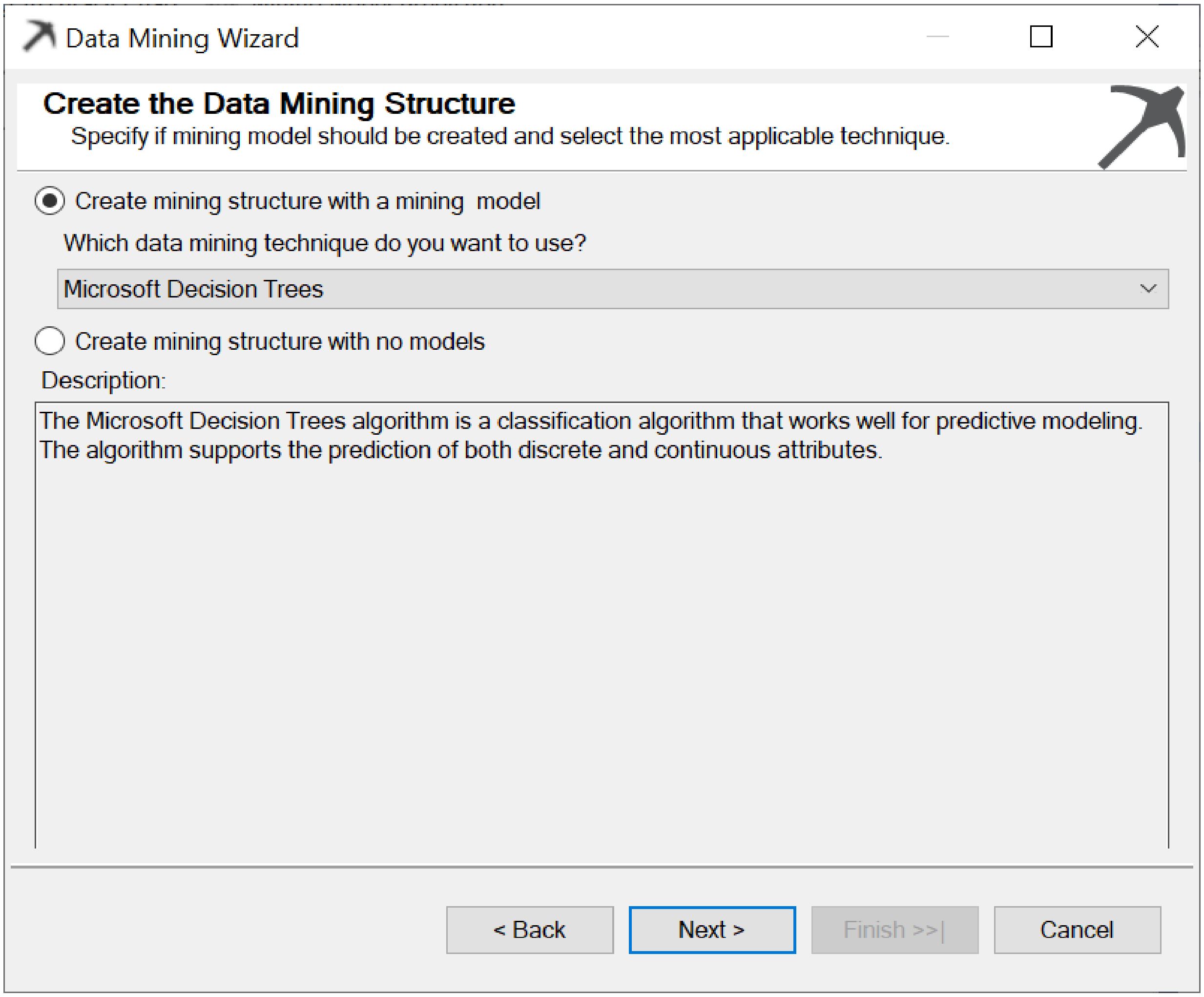
- **Microsoft Decision Trees:**

Step 1: Add Data Source

Step 2: Add Data Source View but select only one table for training model

Step 3: Right – Click on Mining Structures then select New Mining Strcuture.

Step 4: Select Algorithm



Create the Data Mining Structure

Specify if mining model should be created and select the most applicable technique.

● Create mining structure with a mining model

Which data mining technique do you want to use?

Microsoft Decision Trees

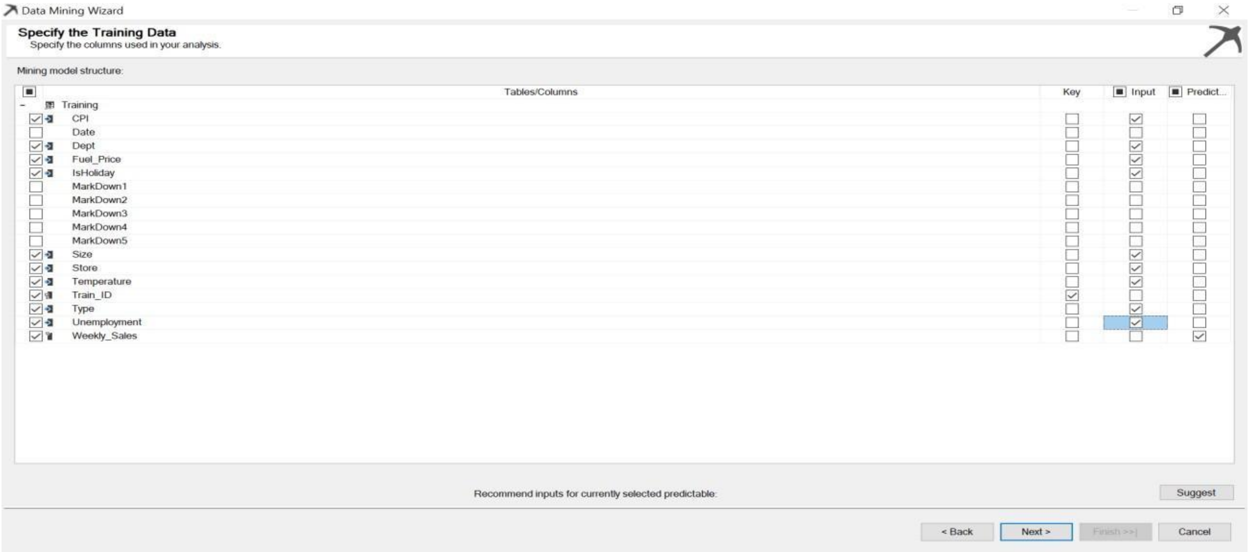
○ Create mining structure with no models

Description:

The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling. The algorithm supports the prediction of both discrete and continuous attributes.

In this window, we select Microsoft Decision Trees.

Step 5: Select input, output feature



Specify the Training Data

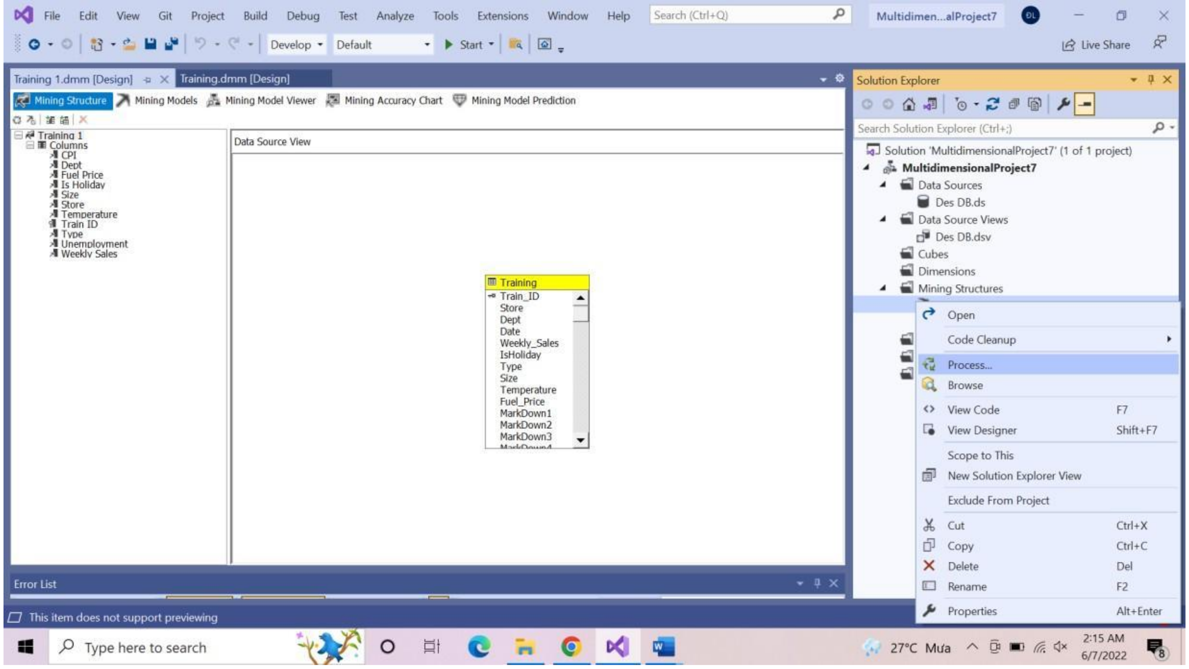
Specify the columns used in your analysis.

Mining model structure:

■	Tables/Columns	Key	■ Input	■ Predict...
−	Training			
☑	CPI	□	☑	□
□	Date	□	□	□
☑	Dept	□	☑	□
☑	Fuel_Price	□	☑	□
☑	IsHoliday	□	☑	□
□	MarkDown1	□	□	□
□	MarkDown2	□	□	□
□	MarkDown3	□	□	□
□	MarkDown4	□	□	□
□	MarkDown5	□	□	□
☑	Size	□	☑	□
☑	Store	□	☑	□
☑	Temperature	□	☑	□
☑	Train_ID	☑	□	□
☑	Type	□	☑	□
☑	Unemployment	□	☑	□
☑	Weekly_Sales	□	□	☑

Recommend inputs for currently selected predictable:

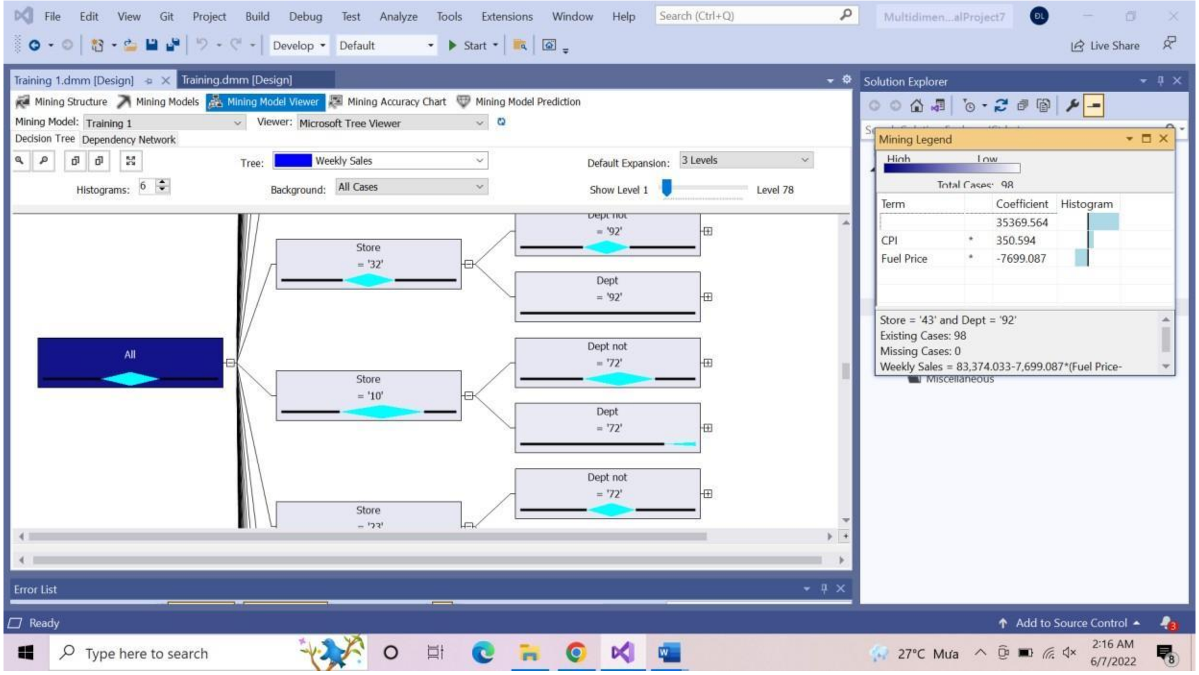
Step 6: Process and Deployment



Click into Process for training model.

Step 7: Mining Model Viewer

When running finish, click on Mining Model Viewer, we will see as the figure below


Mining Model: Training 1
Viewer: Microsoft Tree Viewer
Tree: Weekly Sales
Default Expansion: 3 Levels
Histograms: 6
Background: All Cases
Show Level 1 Level 78

Mining Legend

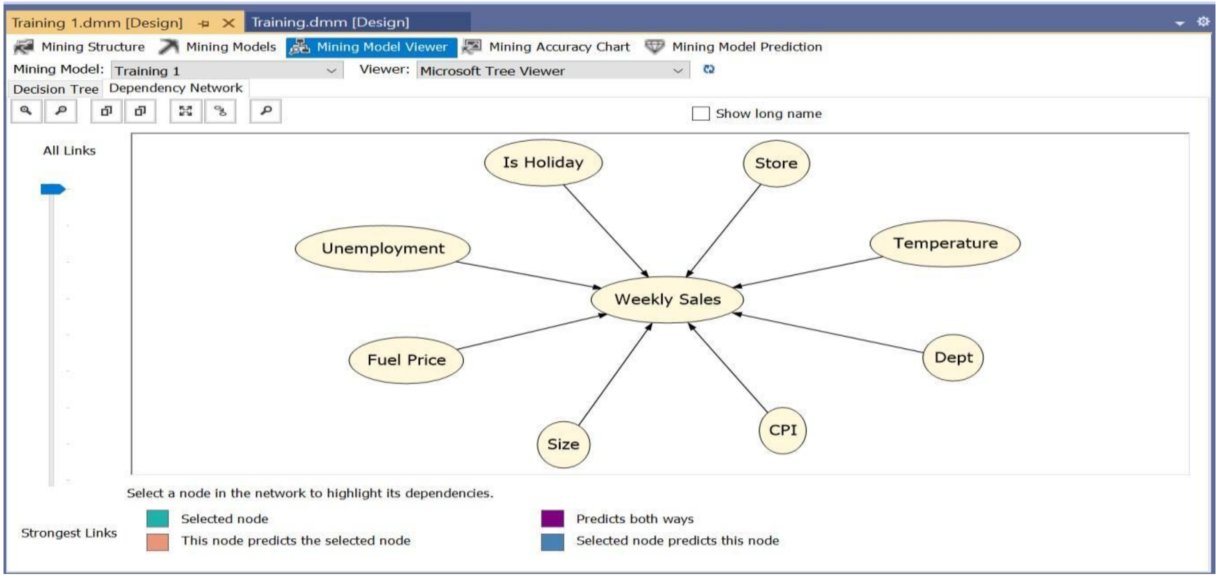
Total Cases: 98

Term		Coefficient
		35369.564
CPI	*	350.594
Fuel Price	*	-7699.087

Store = '43' and Dept = '92'
Existing Cases: 98
Missing Cases: 0
Weekly Sales = 83,374.033-7,699.087*(Fuel Price-

As Store has the most effect to the output values so the Store is the root note the following is Dept. If you want to see another leave click plus symbol near to the Dept note.

Step 8: Dependence Network



Mining Model: Training 1

Viewer: Microsoft Tree Viewer

Select a node in the network to highlight its dependencies.

🟩 Selected node

🟧 This node predicts the selected node

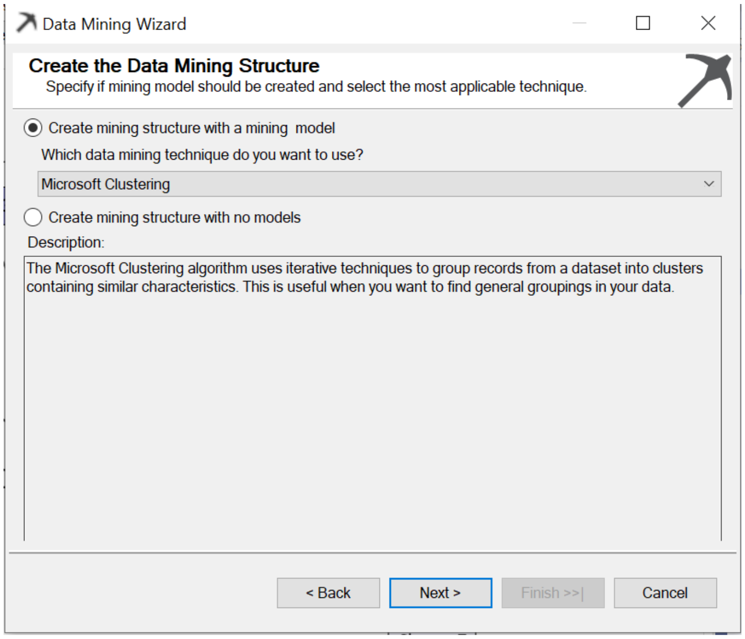
🟪 Predicts both ways

🟦 Selected node predicts this node

This network show the dependence of output column with input column

- Clustering:

Do the same task with Decision Tree, just difference in chose algorithm



Create the Data Mining Structure

Specify if mining model should be created and select the most applicable technique.

● Create mining structure with a mining model

Which data mining technique do you want to use?

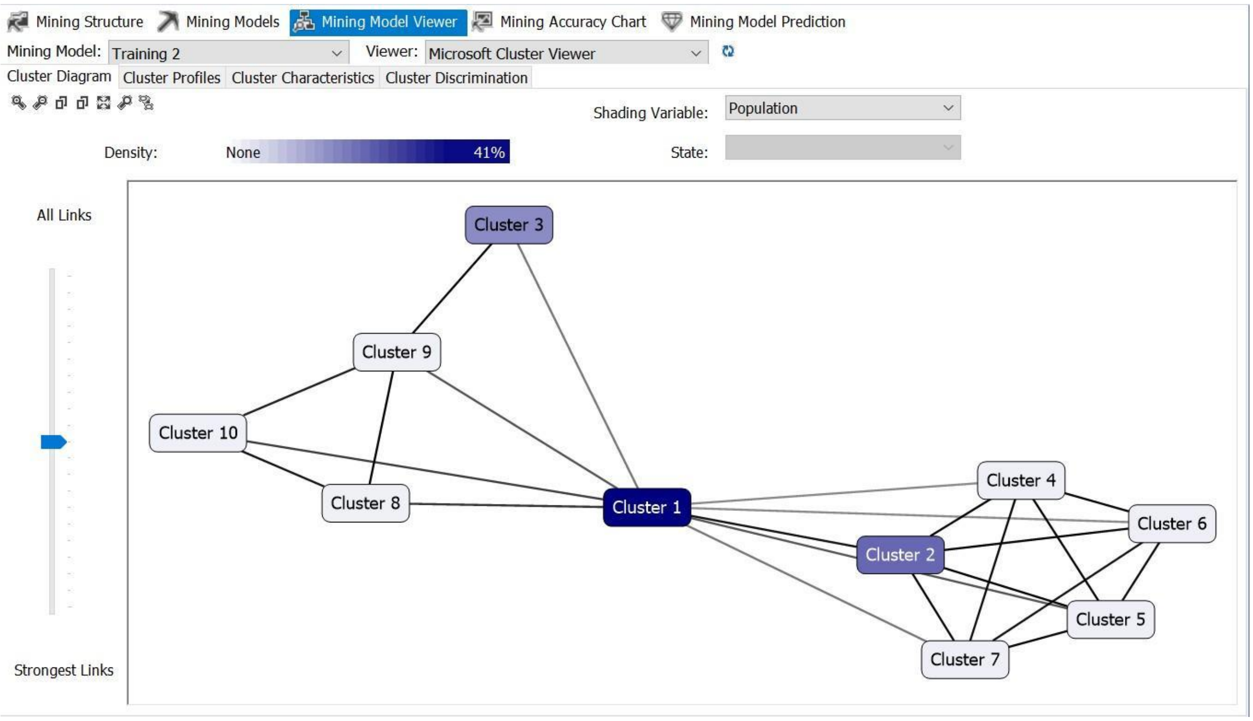
Microsoft Clustering

○ Create mining structure with no models

Description:

The Microsoft Clustering algorithm uses iterative techniques to group records from a dataset into clusters containing similar characteristics. This is useful when you want to find general groupings in your data.

Cluster Diagram:



Mining Model: Training 2

Viewer: Microsoft Cluster Viewer

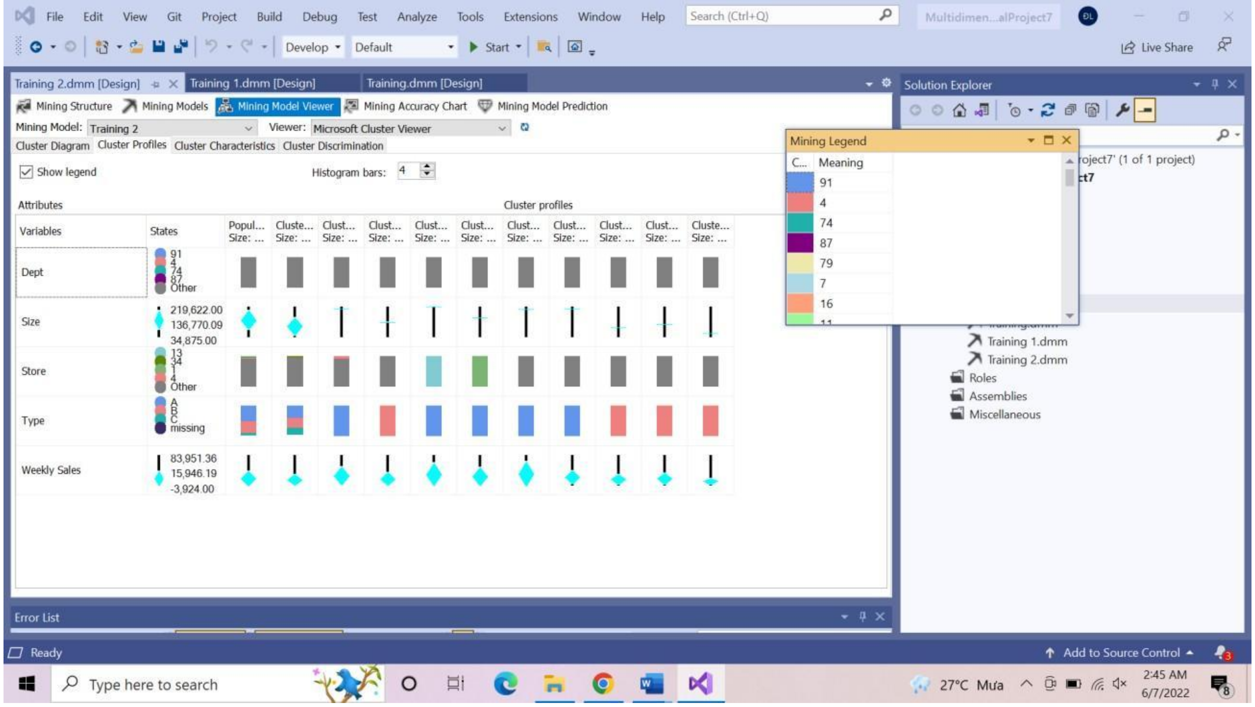
Shading Variable: Population

Density: None 41%

State:

We have 10 cluster from dataset.

Cluster Profiles:



Mining Model: Training 2

Viewer: Microsoft Cluster Viewer

☑ Show legend

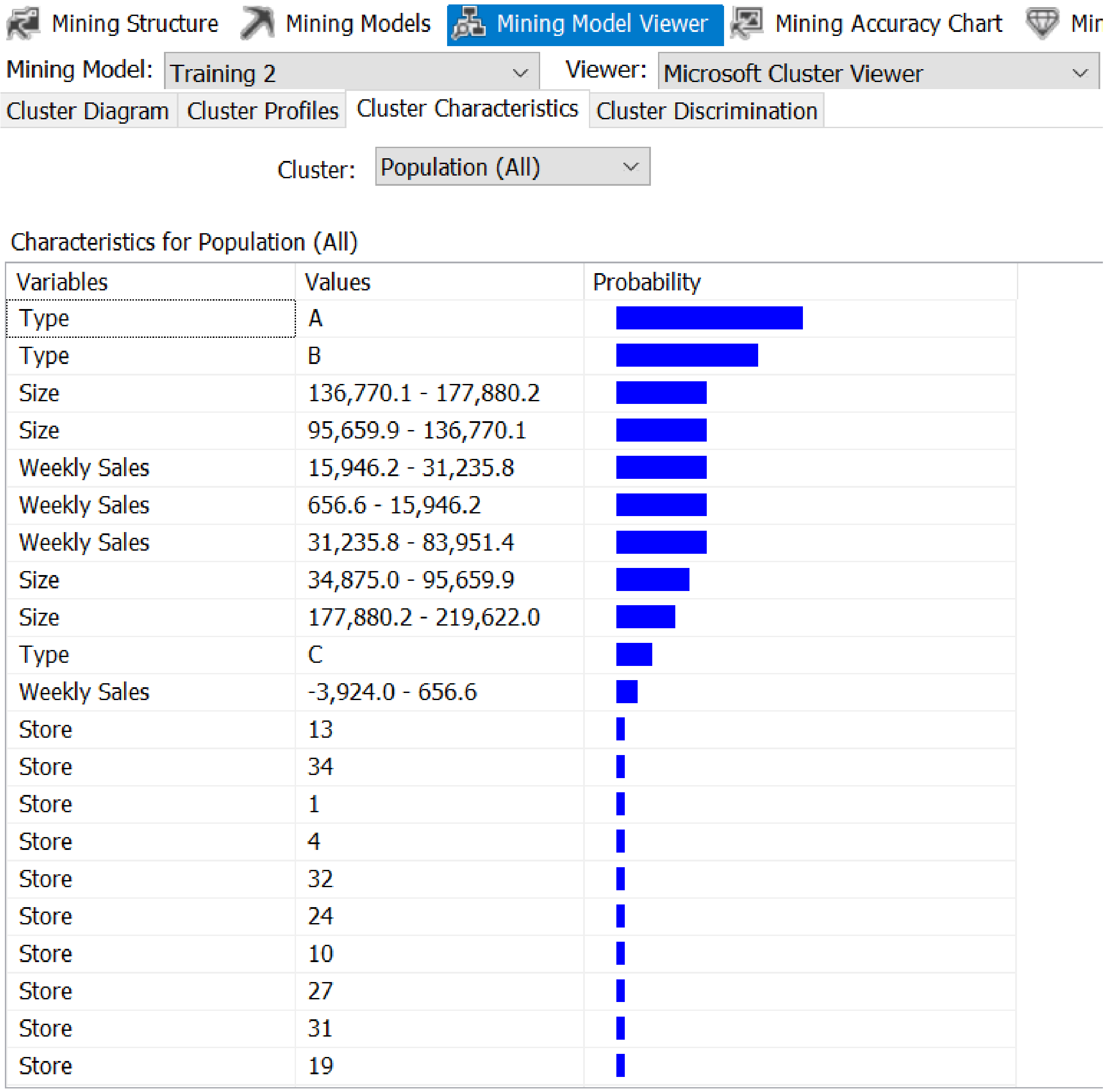
Histogram bars: 4

Mining Legend

C...	Meaning
🟦	91
🟥	4
🟩	74
🟪	87
🟨	79
🟦	7
🟧	16

We will have an overview the model. The distribution of a discrete variable is displayed as a colored bar with the maximum number of bars shown in the Histogram bar list. Continuous attributes are shown with a diamond histogram, representing the mean and standard deviation within each cluster.

Cluster Characteristics Tab:



Mining Model: Training 2

Viewer: Microsoft Cluster Viewer

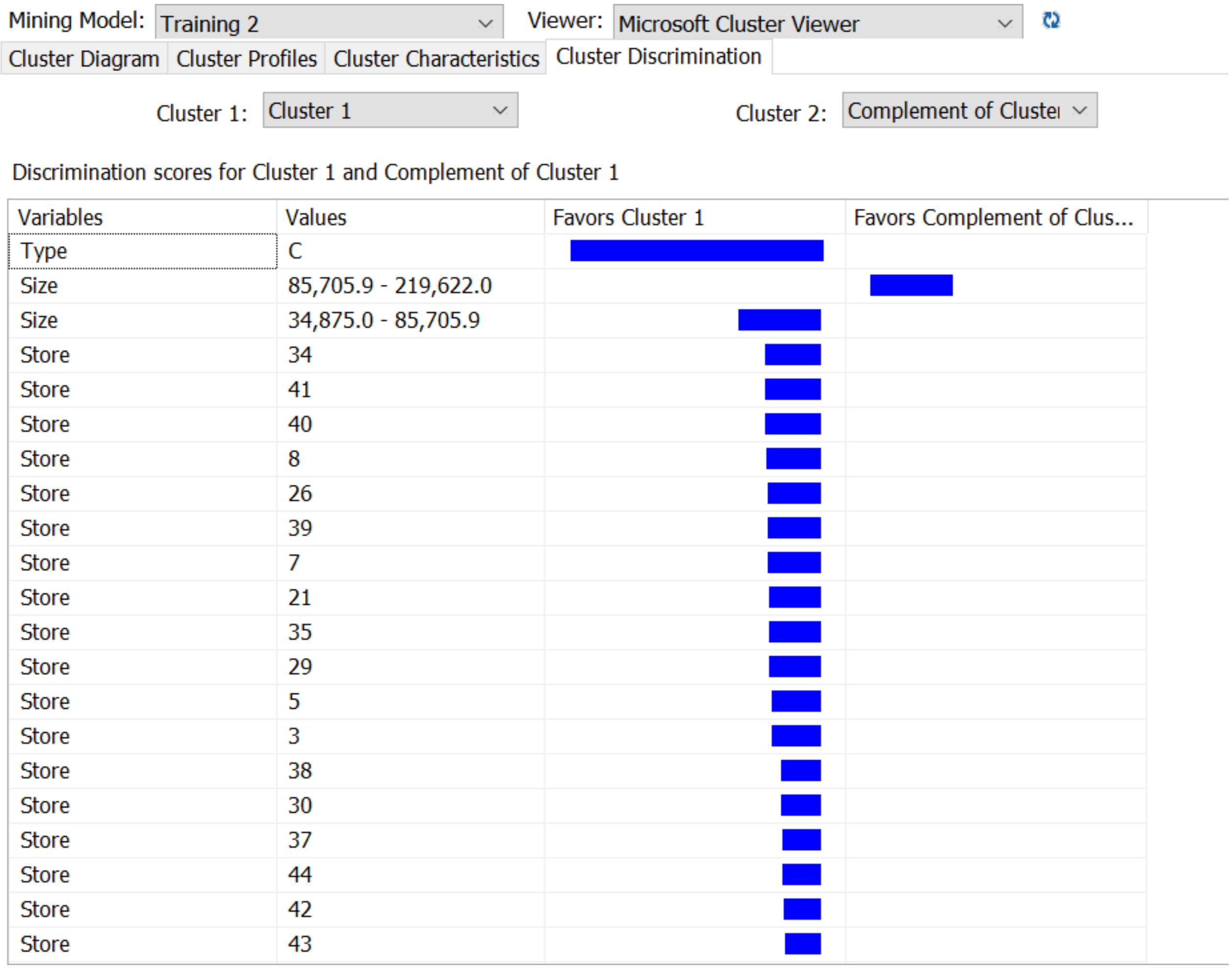
Cluster: Population (All)

Characteristics for Population (All)

Variables	Values	Probability
Type	A	
Type	B	
Size	136,770.1 - 177,880.2	
Size	95,659.9 - 136,770.1	
Weekly Sales	15,946.2 - 31,235.8	
Weekly Sales	656.6 - 15,946.2	
Weekly Sales	31,235.8 - 83,951.4	
Size	34,875.0 - 95,659.9	
Size	177,880.2 - 219,622.0	
Type	C	
Weekly Sales	-3,924.0 - 656.6	
Store	13	
Store	34	
Store	1	
Store	4	
Store	32	
Store	24	
Store	10	
Store	27	
Store	31	
Store	19	

Check carefully features create a cluster

Cluster Discrimination Tab:



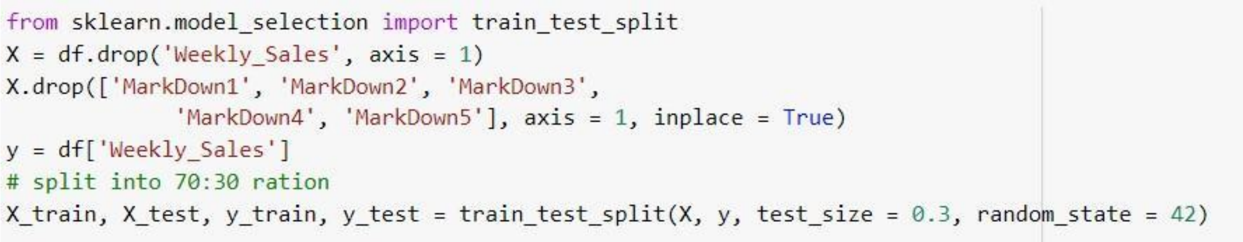
Discrimination scores for Cluster 1 and Complement of Cluster 1

Variables	Values	Favors Cluster 1	Favors Complement of Clus...
Type	C	█	
Size	85,705.9 - 219,622.0		█
Size	34,875.0 - 85,705.9	█	
Store	34	█	
Store	41	█	
Store	40	█	
Store	8	█	
Store	26	█	
Store	39	█	
Store	7	█	
Store	21	█	
Store	35	█	
Store	29	█	
Store	5	█	
Store	3	█	
Store	38	█	
Store	30	█	
Store	37	█	
Store	44	█	
Store	42	█	
Store	43	█	

Discover the characteristics that distinguish one phrase from another. After you select two clusters, one from the Cluster 1 list and the other from the Cluster 2 list, the viewer computes the difference between clusters and displays a list of attributes that distinguish the clusters the most.

Data Mining – Python

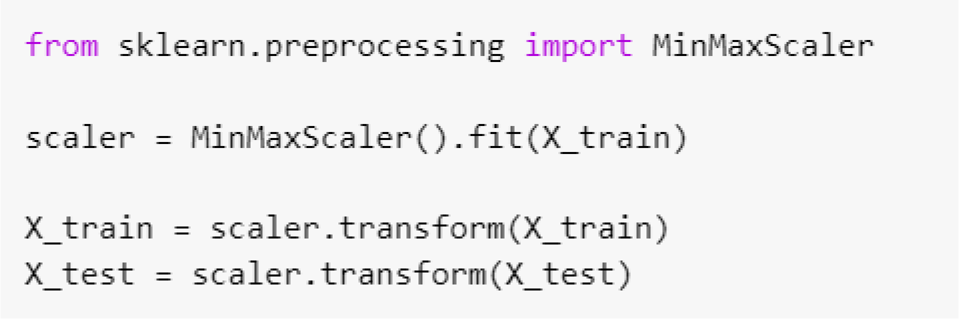
- **Feature Selection and Train Test Split**



```
from sklearn.model_selection import train_test_split
X = df.drop('Weekly_Sales', axis = 1)
X.drop(['MarkDown1', 'MarkDown2', 'MarkDown3',
             'MarkDown4', 'MarkDown5'], axis = 1, inplace = True)
y = df['Weekly_Sales']
# split into 70:30 ration
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
```

We will drop unneeded columns for input and split data 70% for training, 30% for test. The target column is Weekly Sales.

- **Data scaling:**



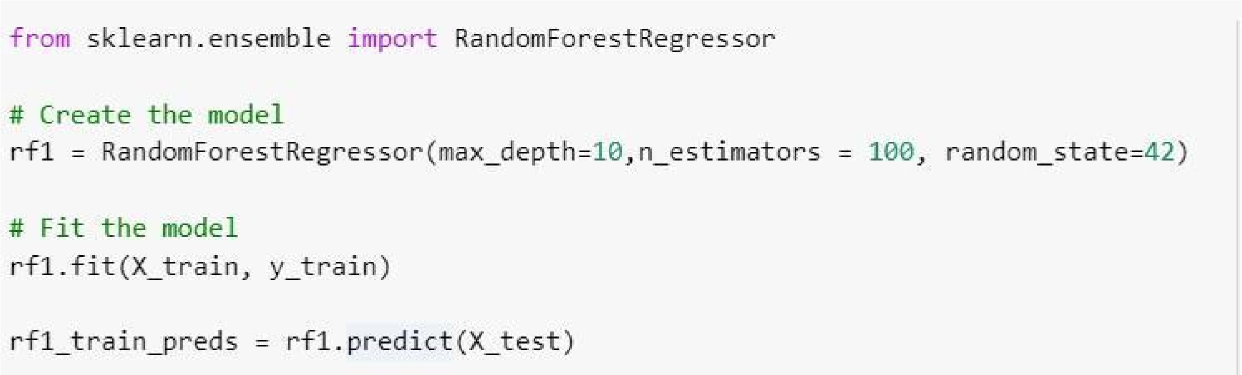
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler().fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

We should scale training data into the same range by MinMaxScaler method to get the better performance.

- **Training model:**

The output of the problem is continuous values so our problem is regressor.

- **Random Forest:**



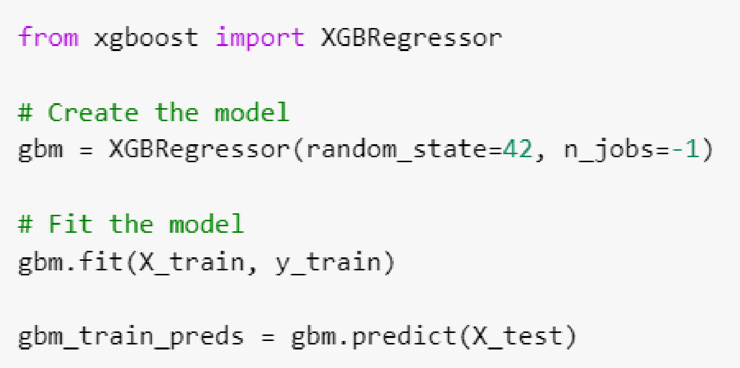
```
from sklearn.ensemble import RandomForestRegressor

# Create the model
rf1 = RandomForestRegressor(max_depth=10,n_estimators = 100, random_state=42)

# Fit the model
rf1.fit(X_train, y_train)

rf1_train_preds = rf1.predict(X_test)
```

We will select the max depth is 10, n_estimators is 100, set random_state is 42. Fit model with training data and predict in testing set.

- **XGBRegressor**

The second model is XGB Regressor. XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling. Shortly after its development and initial release, XG Boost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions.

Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XG Boost can be used directly for regression predictive modeling.

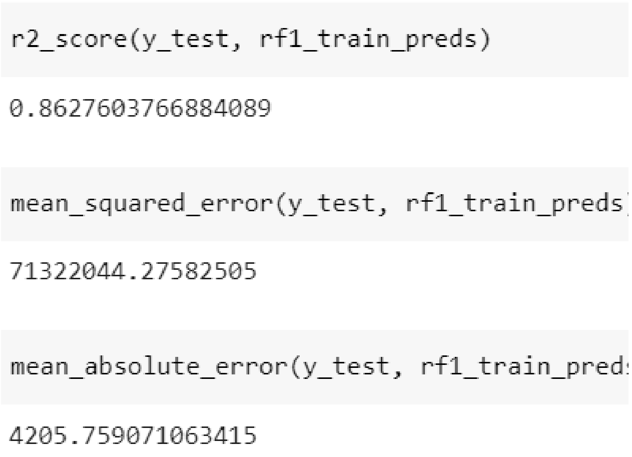


```
from xgboost import XGBRegressor

# Create the model
gbm = XGBRegressor(random_state=42, n_jobs=-1)

# Fit the model
gbm.fit(X_train, y_train)

gbm_train_preds = gbm.predict(X_test)
```

Doing the same with Random Forest. We will call the model, fit it with training dataset and predict in test set.

- **Model Evaluation**

- **Random Forest**



```
r2_score(y_test, rf1_train_preds)
```

```
0.8627603766884089
```

```
mean_squared_error(y_test, rf1_train_preds
```

```
71322044.27582505
```

```
mean_absolute_error(y_test, rf1_train_pred
```

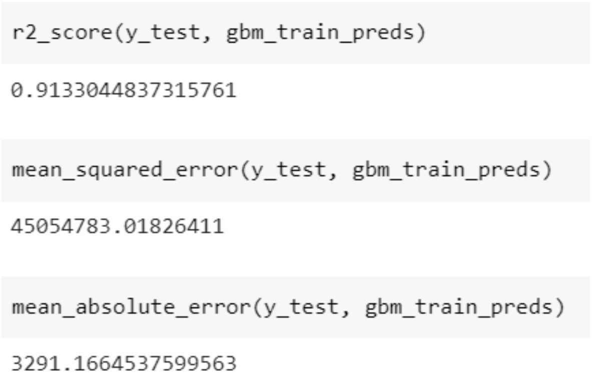
```
4205.759071063415
```

Firstly, we will calculate R^2 score. We get 86.276%.

Secondly, we will calculate mean squared error. We get 71322044.27582505

Thirdly, we get 4205.759071063415 in mean absolute error

- **XGBRegressor**



```
r2_score(y_test, gbm_train_preds)
```

```
0.9133044837315761
```

```
mean_squared_error(y_test, gbm_train_preds)
```

```
45054783.01826411
```

```
mean_absolute_error(y_test, gbm_train_preds)
```

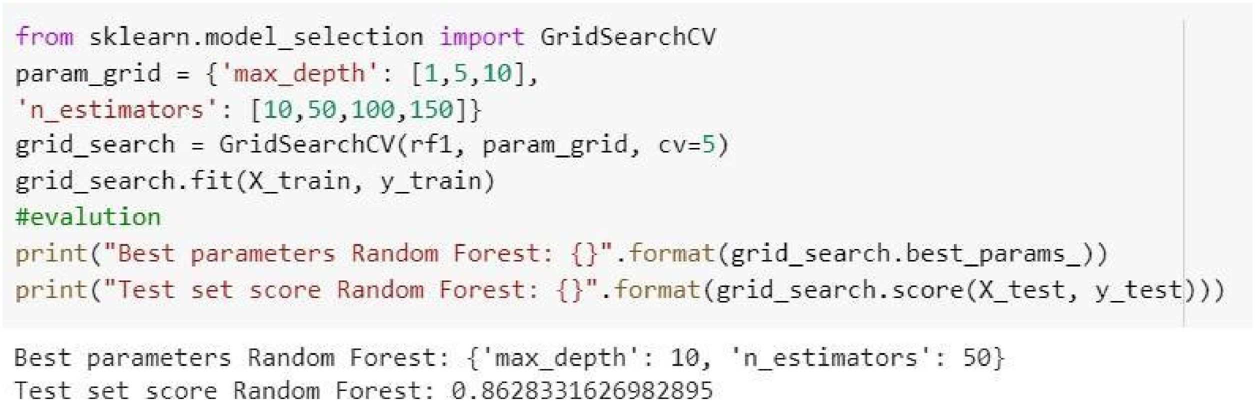
```
3291.1664537599563
```

R^2 Score is 91.33%

MSE score is 45054783.01826411

MAE score is 3291.166453759956

- **Fine Tuning**
- **Random Forest:**

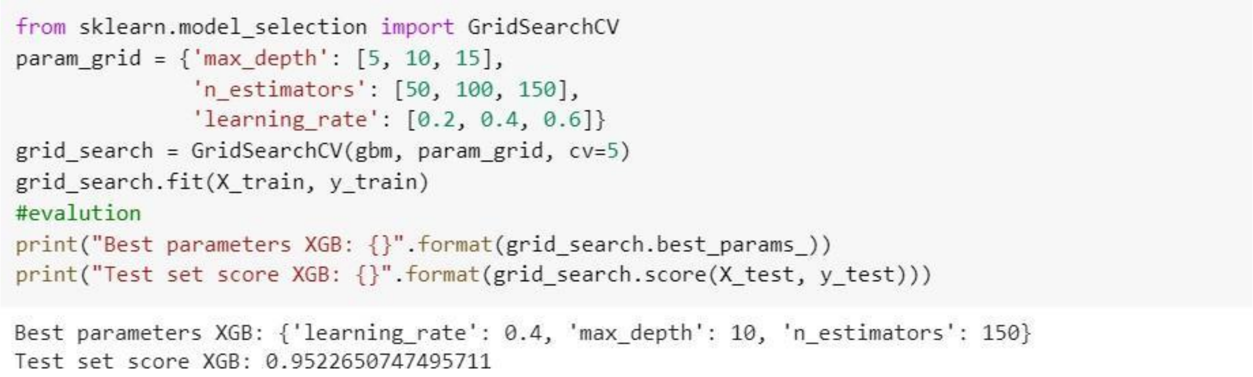


```
from sklearn.model_selection import GridSearchCV
param_grid = {'max_depth': [1,5,10],
'n_estimators': [10,50,100,150]}
grid_search = GridSearchCV(rf1, param_grid, cv=5)
grid_search.fit(X_train, y_train)
#evalution
print("Best parameters Random Forest: {}".format(grid_search.best_params_))
print("Test set score Random Forest: {}".format(grid_search.score(X_test, y_test)))
```

```
Best parameters Random Forest: {'max_depth': 10, 'n_estimators': 50}
Test set score Random Forest: 0.8628331626982895
```

Which max_depth is 10 and n_estimators is 50, we reach the highest score is 86.28%

- **XGBRegressor**



```
from sklearn.model_selection import GridSearchCV
param_grid = {'max_depth': [5, 10, 15],
              'n_estimators': [50, 100, 150],
              'learning_rate': [0.2, 0.4, 0.6]}
grid_search = GridSearchCV(gbm, param_grid, cv=5)
grid_search.fit(X_train, y_train)
#evalution
print("Best parameters XGB: {}".format(grid_search.best_params_))
print("Test set score XGB: {}".format(grid_search.score(X_test, y_test)))
```

```
Best parameters XGB: {'learning_rate': 0.4, 'max_depth': 10, 'n_estimators': 150}
Test set score XGB: 0.9522650747495711
```

After searching parameters, we get the best parameter for XGBRegressor with learning rate is 0.4, max depth is 10 and number of estimators is 150. We get the score for test set is 0.95, that is very very high score for regressor problem.

In conclusion, we see that XGBoost is the best algorithm for our problem with highest performance.

V. References:

[1] Zina A. S. Abdullah, Taleb A. S. Obaid, Design and Implementation of Educational Data Warehouse Using OLAP: International Journal of Computer Science and Network, Volume 5, Issue 5, October 2016

Available at:
http://eprints.rclis.org/30201/1/Design-and-Implementation-of-Educational-Data-Warehouse-Using-OLAP.pdf

[2] Jens Otto Sørensen, Karl Alnor: Creating a Data Warehouse using SQL Server: Department of Information Sciences The Aarhus School of Business, Denmark Available at:
http://ceur-ws.org/Vol-19/paper10.pdf

[3] R. Kimball, The Data Warehouse Toolkit. New York: Wiley Computer Publishing, 1996

[4] Rajeev Kaula, Business Rules for Data Warehouse: World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:3, No:11, 2009 Available at:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.866.4304&rep=rep1&type=pdf