**TUTORIAL: IOWA LIQUOR SALES ANALYSIS**

ARSHITA DIXIT

ABHISHEK SHAH

KAUSTUBH PADHYA

**OBJECTIVE**

Liquor markets in the whole world see a great amount of trends in its sales which can be analyzed using Big Data.

In this tutorial, we will make an analysis on the sales of liquor in the state of Iowa using Hive, Pig and Tableau.

**INTRODUCTION**

In this tutorial, we will first use excel to clean and filter the downloaded dataset and upload it to the local using FileZilla FTP.

Then we will learn how to use BigInsight to:

* Download and upload **csv** files.
* Create Hive tables and queries for the analysis of ‘liquor sales’ data.
* Create Pig queries to analyze the same and store the data.

Finally,

* Use Tableau for the visualization of the analysis made.

**PREREQUISITES**

Everything we need to go through the scripts and queries is already provisioned with the cluster.

For cleaning and filtering the data you need **Microsoft Excel 2010**.

We must FileZilla FTP to upload the dataset to the local of Hadoop.

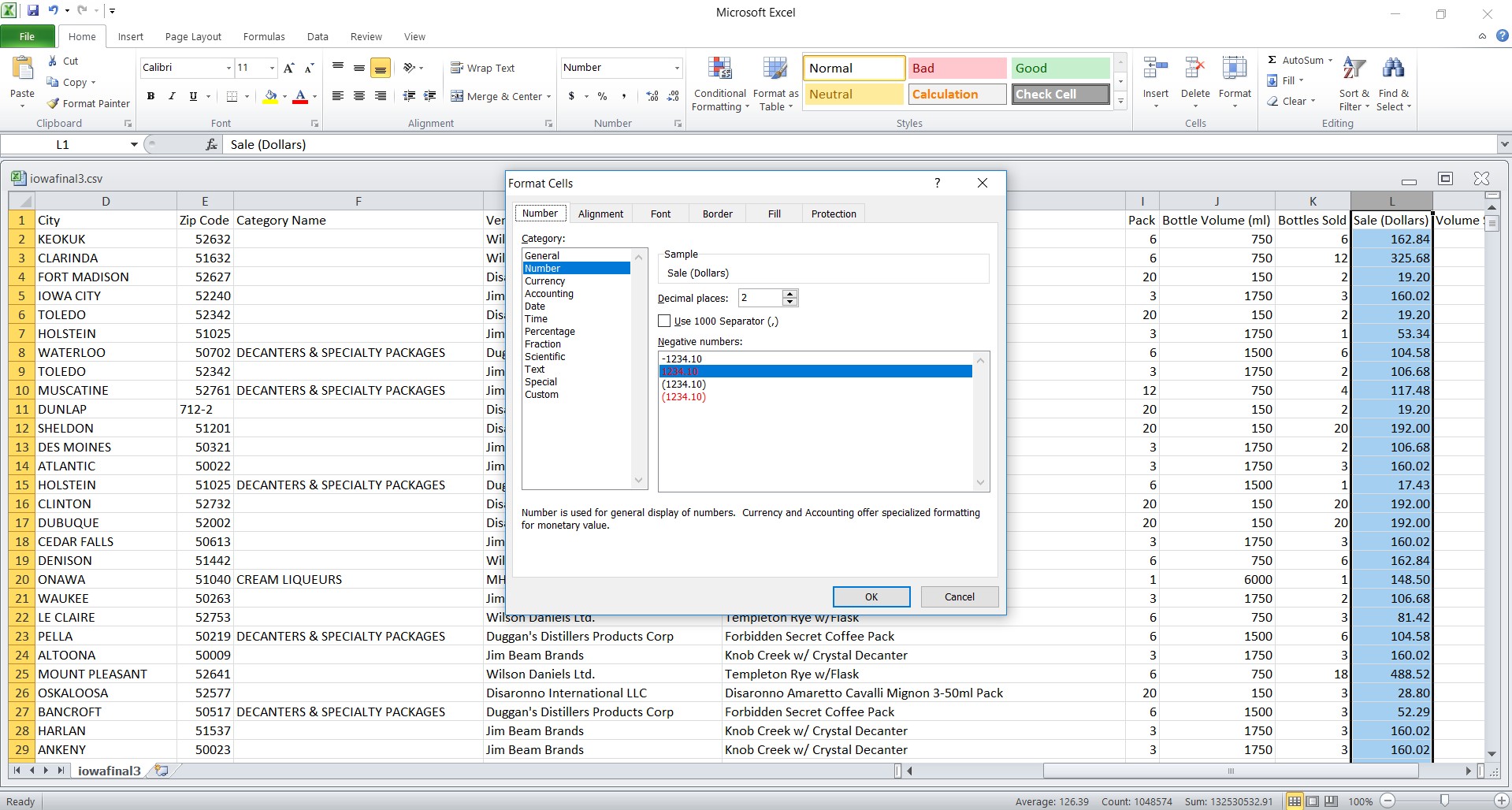
To export the analyzed data to Tableau for the visualizations, you must have **Tableau 10.1** installed.

**Download the Dataset on your PC using the link given below**

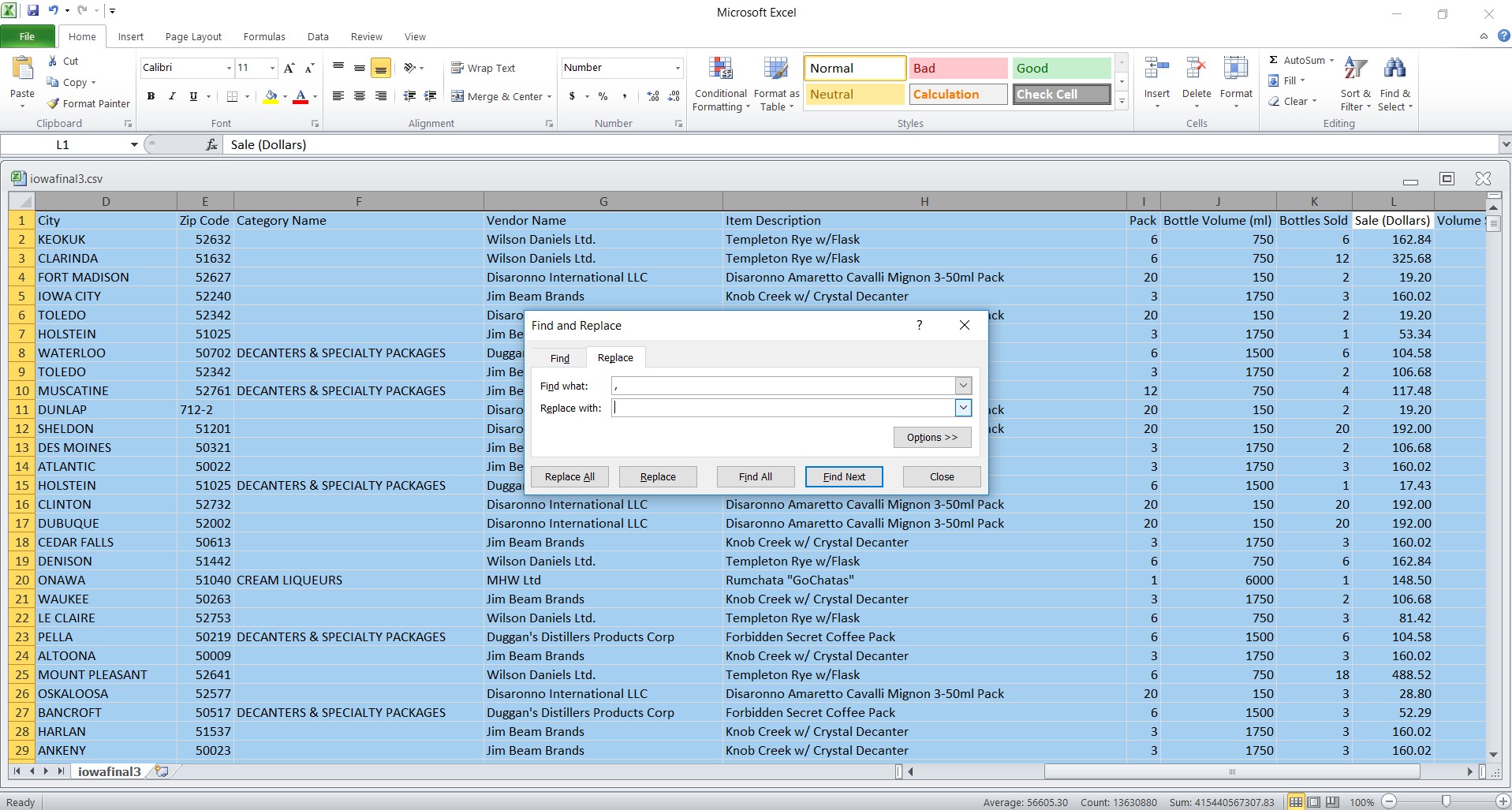
<https://data.iowa.gov/api/views/m3tr-qhgy/rows.csv?accessType=DOWNLOAD&bom=true>

**Cleaning using Excel**

* Open the downloaded file using Excel.
* Select the Sales (in Dollars) column and press **Ctrl+1**. Now you will get the Format Cells dialogue box as shown in the figure. Choose the second number system to get rid of the ‘$’ sign which makes computations easy in hive.



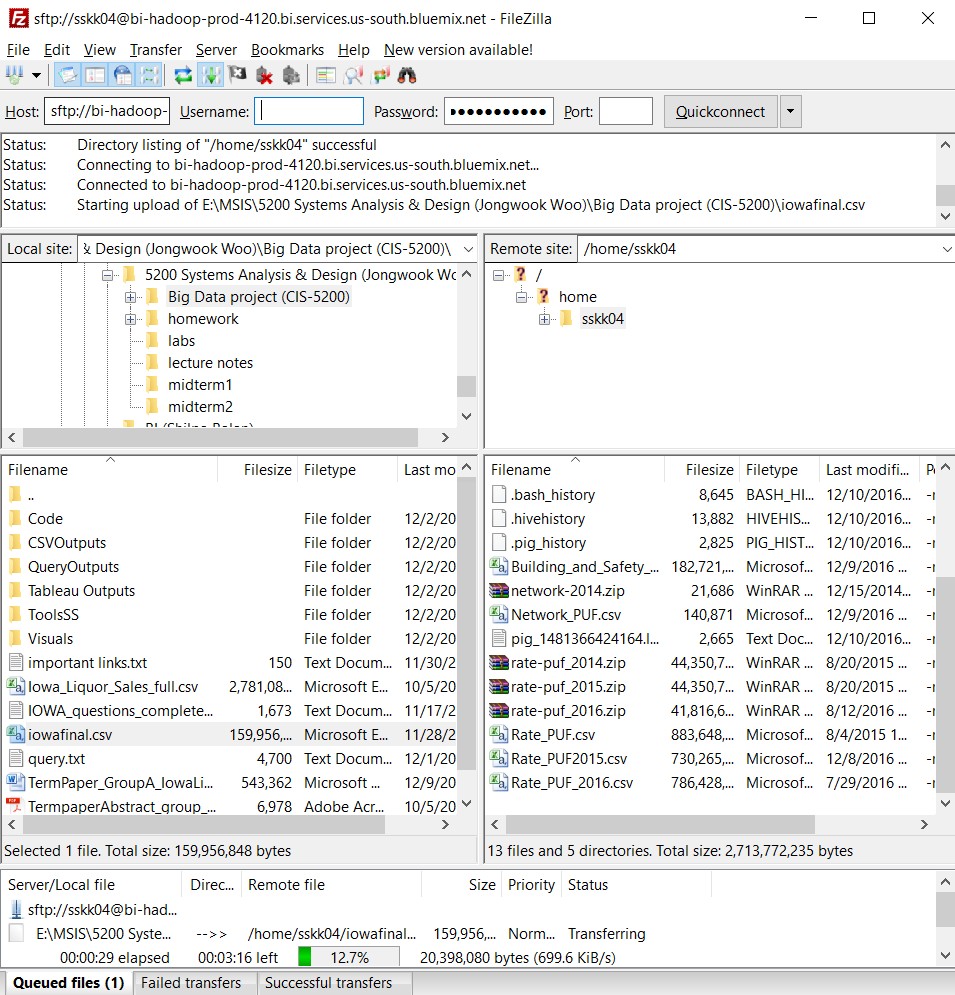
* Select all by pressing **Ctrl+A** then press **Ctrl+F** to replace all commas (,) with space or nothing. As this is a csv file, the hive gets confused with the ‘commas between the columns’ and ‘commas within the content of the columns’. So we replace all the commas as shown in the figure.



Save the file with name iowafinal.csv

**Uploading the file to Local of Hadoop using FileZilla FTP**

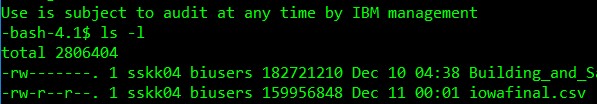
Drag and drop the file as shown in the screenshot



**Iowa liquor Sales Data loaded into BigInsights**

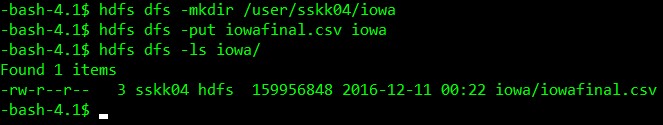
You need to remotely access your BigInsights that you executed in your Bluemix account using ssh. Enter the given command in the shell to see the uploaded file **iowafinal.csv**:

|  |
| --- |
| $ ls –l |



You have to upload this file to hdfs folder **iowa**. Run the following HDFS commands to create and list **iowa** directory in HDFS:

|  |
| --- |
| $ hdfs dfs –mkdir /user/sskk04/iowa  $ hdfs dfs –put iowafinal.csv /user/sskk04/iowa  $ hdfs dfs –ls /user/sskk04/iowa |



**Creating Hive table to Query Liquor Sales Data**

Open hive shell environment as follows:

|  |
| --- |
| $ hive |

The following Hive statement creates an external table for Iowafinal. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.

In the hive shell CLI, you need to copy and paste the following HiveQL code to create an external table **iowa1**.

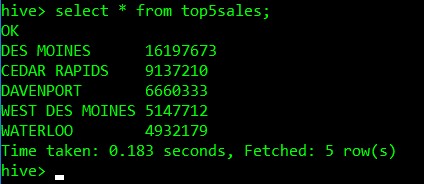
|  |
| --- |
| CREATE EXTERNAL TABLE IF NOT EXISTS iowa1(  invoice STRING,  year INT,  dat STRING,  storename STRING,  city STRING,  zip INT,  catname STRING,  vendorname STRING,  itemname STRING,  pack BIGINT,  bottlevol BIGINT,  bottlesold BIGINT,  sale BIGINT,  volume DECIMAL)  ROW FORMAT DELIMITED  FIELDS TERMINATED BY ','  STORED AS TEXTFILE LOCATION '/user/sskk04/iowa/'  tblproperties ('skip.header.line.count'='1'); |

Create different directories in the hdfs to store the output of the queries:

|  |
| --- |
| $ hdfs dfs –mkdir q1  $ hdfs dfs –mkdir q2  $ hdfs dfs –mkdir q3  $ hdfs dfs –mkdir q4  $ hdfs dfs –mkdir q5  $ hdfs dfs –mkdir q6  $ hdfs dfs –mkdir q7  $ hdfs dfs –mkdir q8  $ hdfs dfs –mkdir q9 |

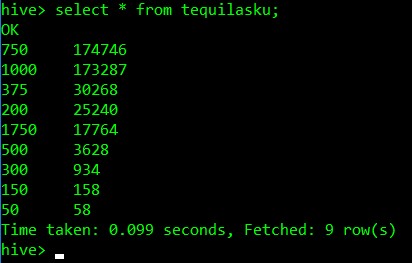
**QUERY 1**: We need to create a table top5sales to represent Top 5 Alcohol Consuming Cities in Iowa:

|  |
| --- |
| create table top5sales  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q1"  as  select city, sum(sale) as totalsale from iowa1 group by city order by totalsale desc limit 5;  select \* from top5sales; |



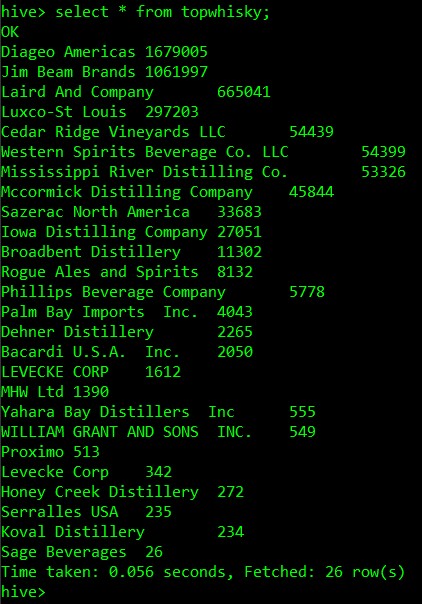
**QUERY 2**: Now, we need to create a table tequilasku to represent relation between SKU size and bottles sold for Tequila in Iowa:

|  |
| --- |
| create table tequilasku  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q2"  as  select bottlevol, sum(bottlesold) as totaltequila from iowa1 where catname = "TEQUILA" group by bottlevol order by totaltequila desc;  select \* from tequilasku; |



**QUERY 3**: Now, we need to create a table topwhiskey to represent top companies for Blended Whiskies according to Dollar Sales:

|  |
| --- |
| create table topwhisky  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q4"  as  select vendorname, sum(sale) as totalsales from iowa1 where catname = 'BLENDED WHISKIES' group by vendorname order by totalsales desc;  select \* from topwhisky; |



**QUERY 4**: Now, we need to create a table topstore to represent highest selling of ‘Blended Whiskies’:

|  |
| --- |
| create table topstore  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q5"  as  select sum(sale) as totalsales, storename from iowa1 where catname = 'BLENDED WHISKIES' group by storename order by totalsales desc limit 50;  select \* from topstore; |



**QUERY 5**: Now, we need to create a table halloween to represent cities with highest Liquor Sales on Halloween in 2013:

|  |
| --- |
| create table halloween  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q6"  as  select city, sum(sale) as totalsales from iowa1 where dat = '2013-10-31' group by city order by totalsales desc limit 5;  select \* from halloween; |



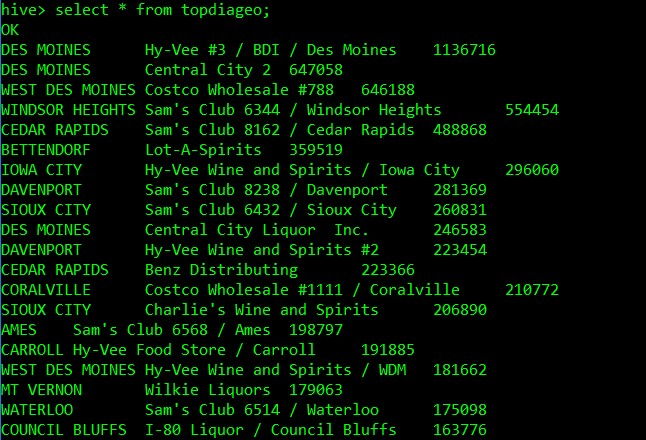
**QUERY 6**: Now, we need to create a table impvodkasale to represent cities with highest Liquor Sales on Halloween in 2013:

|  |
| --- |
| create table impvodkasale  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q7"  as  select city, sum(sale) as s1, sum(volume) from iowa1 where catname='IMPORTED VODKA' group by city order by s1 desc;  select \* from impvodkasale; |



**QUERY 7**: Now, we need to create a table topdiageo to represent top dealers for ‘DIAGEO AMERICAS’ Blended Whiskies:

|  |
| --- |
| create table topdiageo  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q8"  as  select city, storename, sum(sale) as s1 from iowa1 where vendorname = 'Diageo Americas' group by storename, city order by s1 desc limit 100;  select \* from topdiageo; |



**QUERY 8**: Now, we need to create a table everydaysale to represent everyday sales in dollars:

|  |
| --- |
| CREATE TABLE everydaysale  ROW FORMAT DELIMITED  FIELDS TERMINATED BY ","  STORED AS TEXTFILE  LOCATION "/user/sskk04/q9"  AS  select dat, sum(sale) from iowa1 group by dat order by dat;  select \* from everydaysale; |



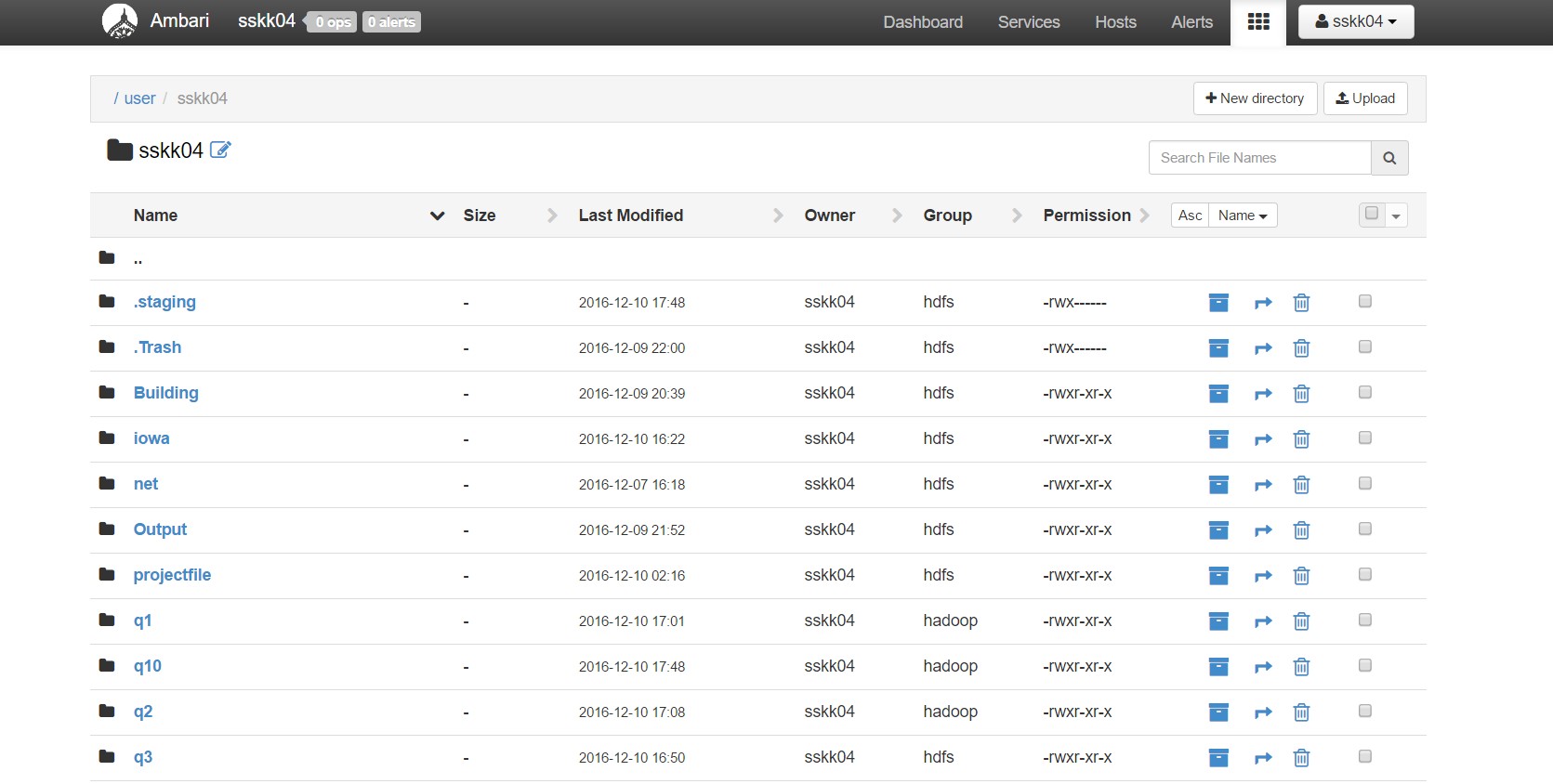
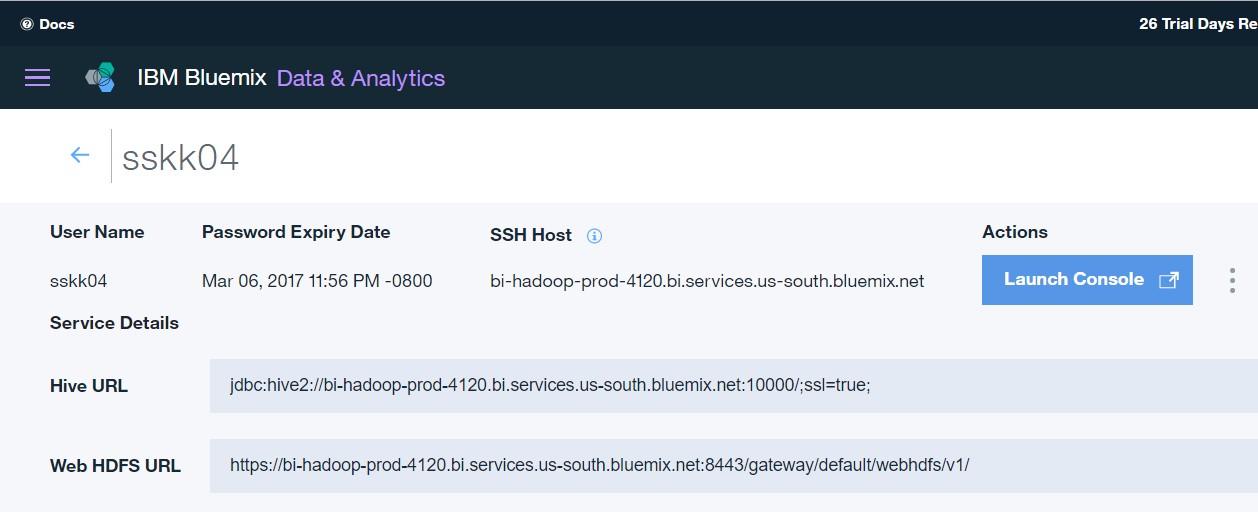
**QUERY 9**: Now, we need to create a table everydayvol to represent everyday sales in volume:

|  |
| --- |
| create table everydayvol  row format delimited  fields terminated by ","  stored as textfile  location "/user/sskk04/q10"  as  select dat, sum(volume) from iowa1 group by dat order by dat;  select \* from everydayvol; |



**Downloading the query outputs from Ambari**

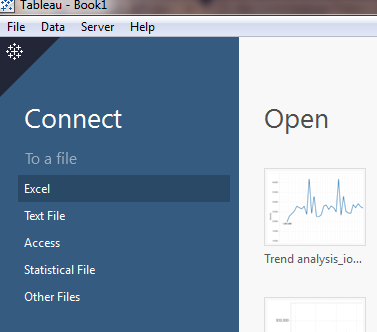
Now go to the BigInsights cluster page to open Ambari web page.



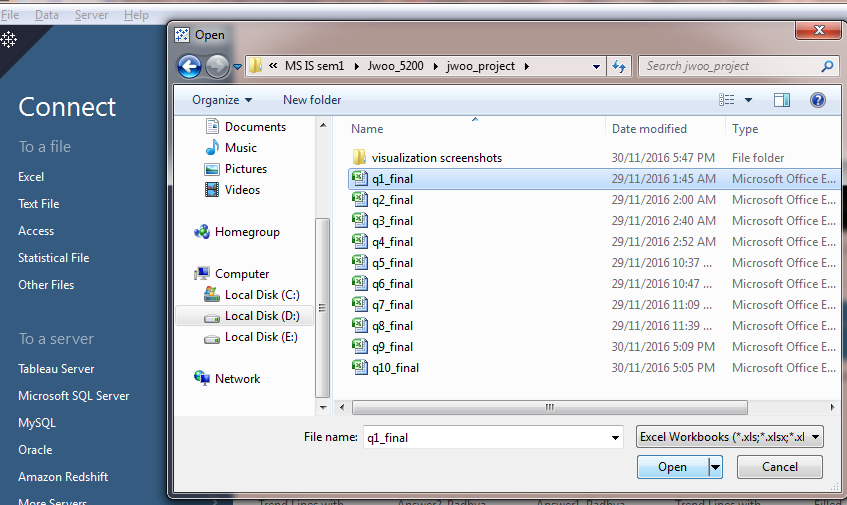
And, you need to go and open “File Browser” to find out files 000000\_0 in directories q1-q10 and click on the file to download to your local computer in order to open in Tableau.

**Loading Data into Tableau**

Open your Tableau to connect with the output files stored in your computer as Excel workbook. You need to select Connect to a File and in that an Excel file.

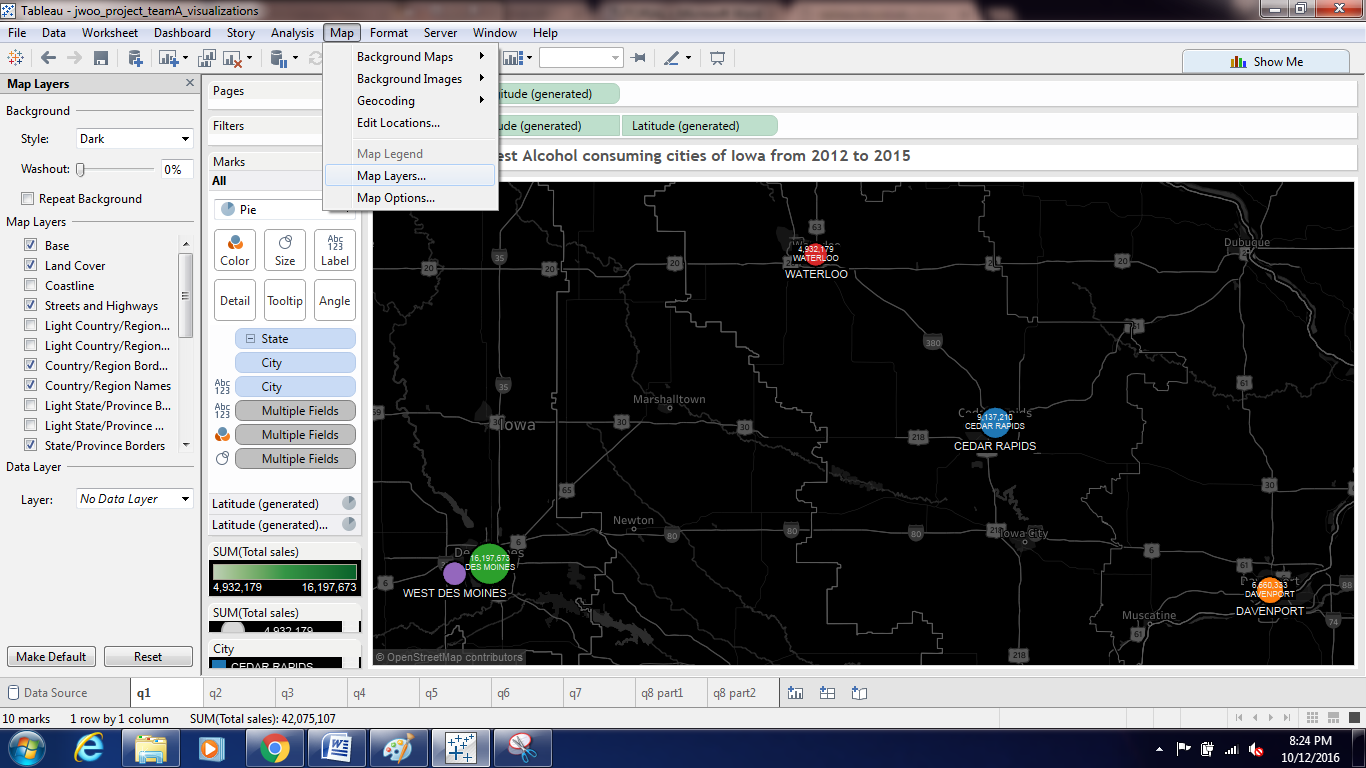


Then, open the Excel output file of the first query in Tableau as follows:



For Query 1 Visualization:

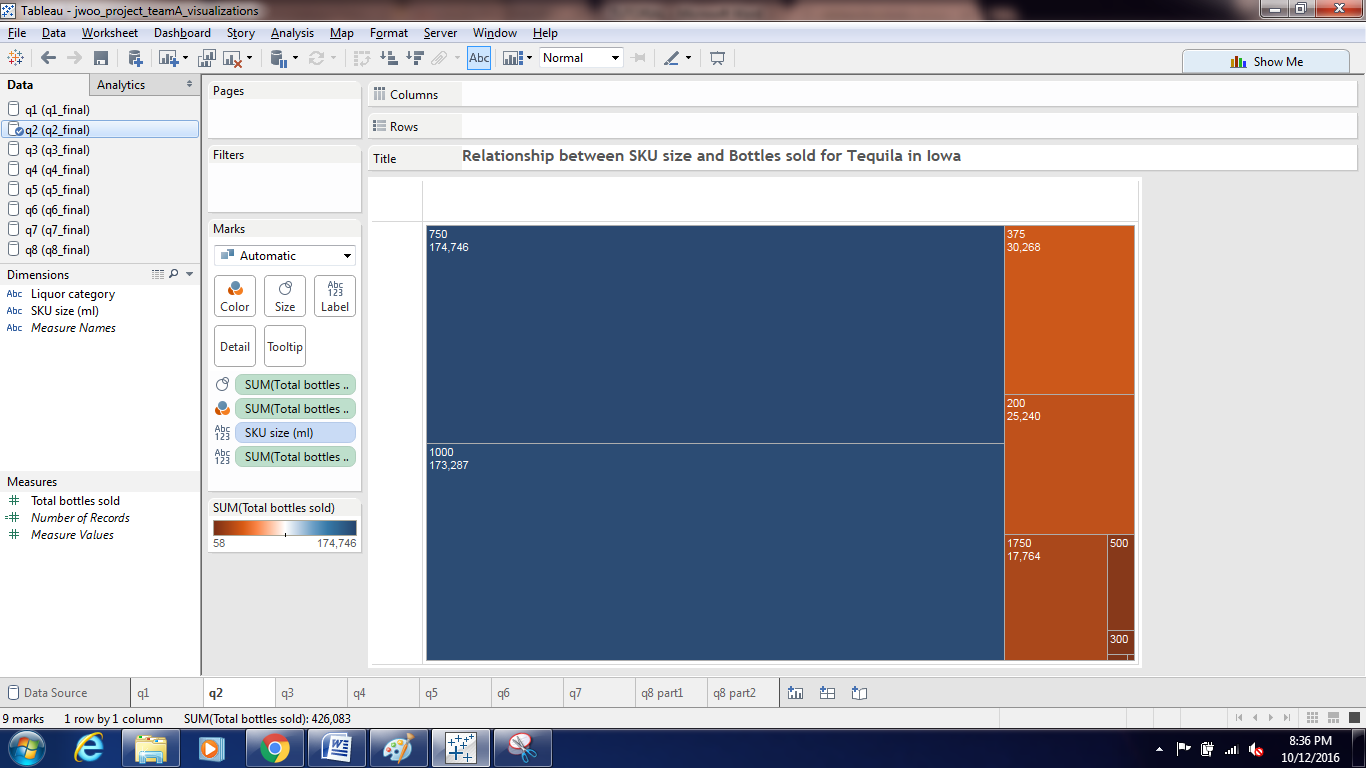
Now you need to select sheet 1 to open a worksheet. Then, drag the Longitude to columns and Latitude generated from Measures tab to Rows. Drag State, city and Sum (Total Sales) to Labels on the Marks card. Also, drag city to Colors and Total sales to Size on the marks card. Select symbol maps from the Show Me bar in the right top corner of the screen. Further, Select Map layers by clicking the Map tab and tick check the boxes as shown in the below screenshot.



This visual shows the top five cities of Iowa according to total Sales in Dollars.

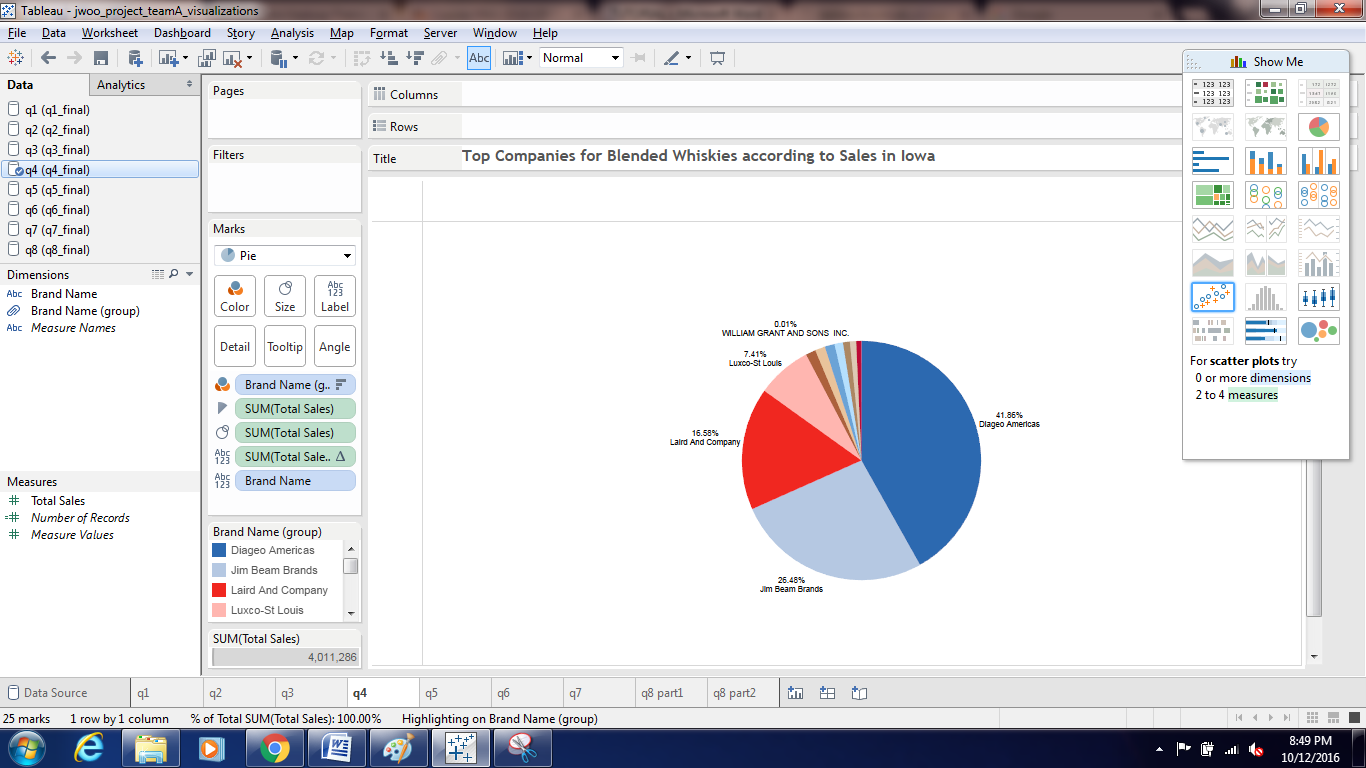
Visualization for Query 2:

Open the Excel output file for the second query in Tableau. Drag SKU size to columns and Total bottles sold to Rows. And then select Tree Maps from Show Me bar. Put Total bottles sold to Size and Color, and SKU size and Total bottles sold to Labels on the Marks card.



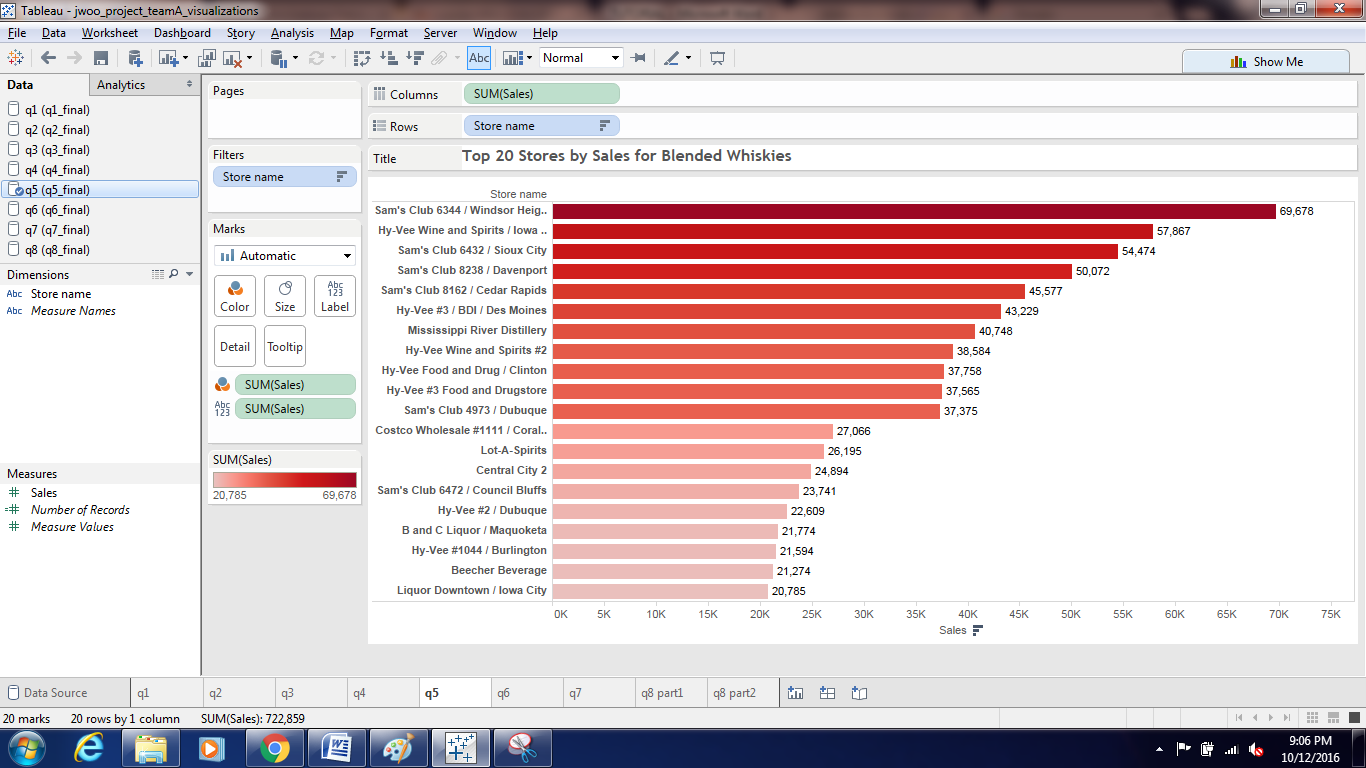
Visualization for Query 3:

Open the Excel output file for the Query 3 in Tableau. Drag Brand name to columns and total sales to Rows. Select Pie chart from the Show me bar. Next, drag Brand Names and (Group brands with lower value of Sales) to colors and Labels, and Total sales to size and Labels on the Marks Card.



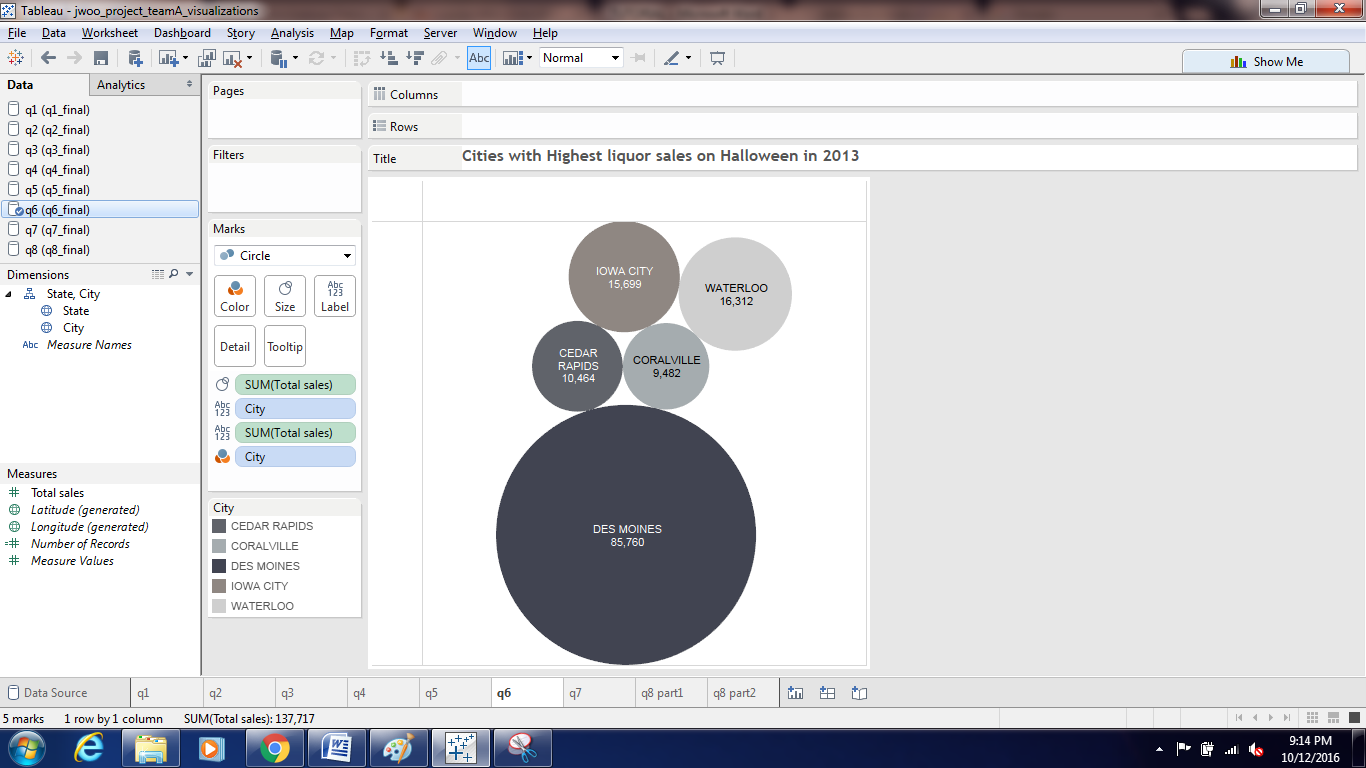
Visualization for Query 4:

Drag Sales to Columns and Store name to Rows and Filter by Top 25 stores by Sales value. Select Horizontal Bar chart from Show me bar. Drag Sales to color and Labels on the Marks card.



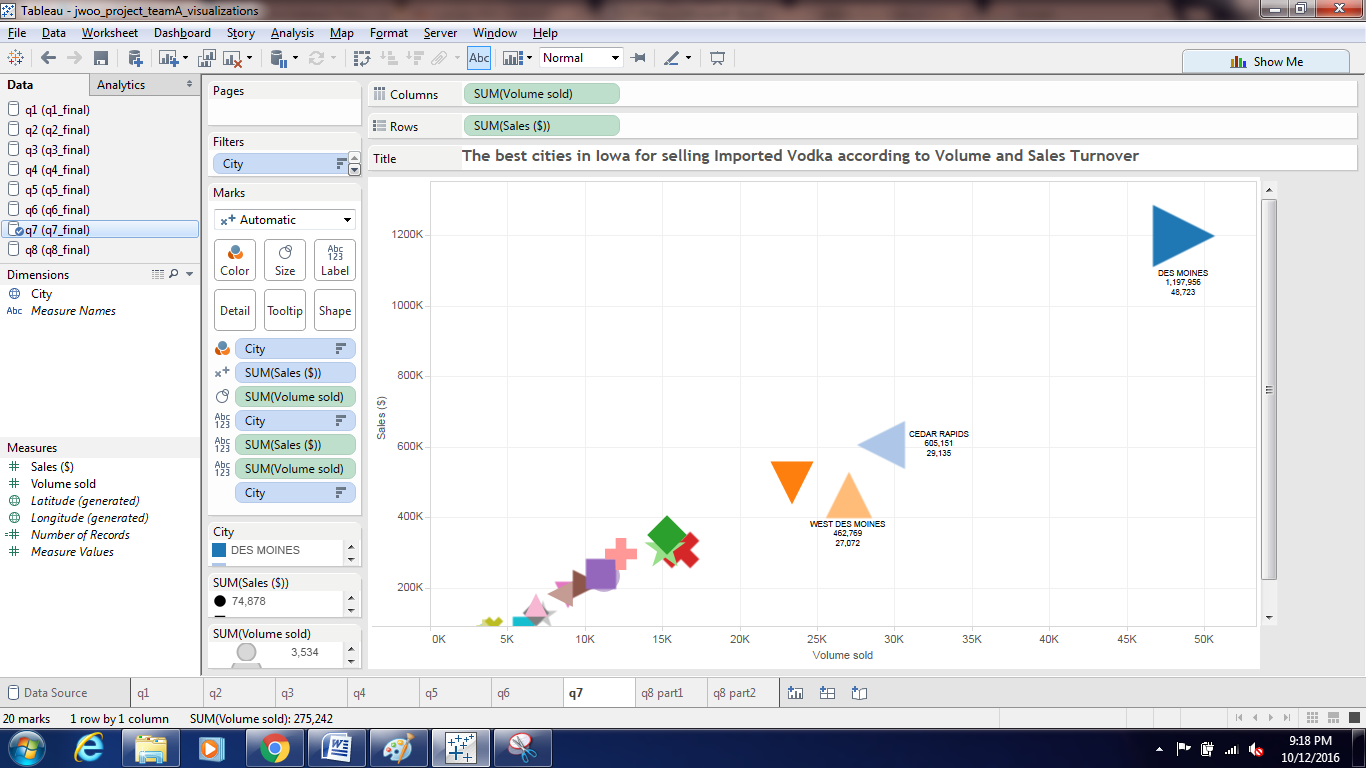
Visualization for Query 5:

Drag City to Columns and Total sales to Rows. Select packed Bubbles chart from the Show me bar. Drag city to color and labels, and Total sales to Labels on the Marks tab.



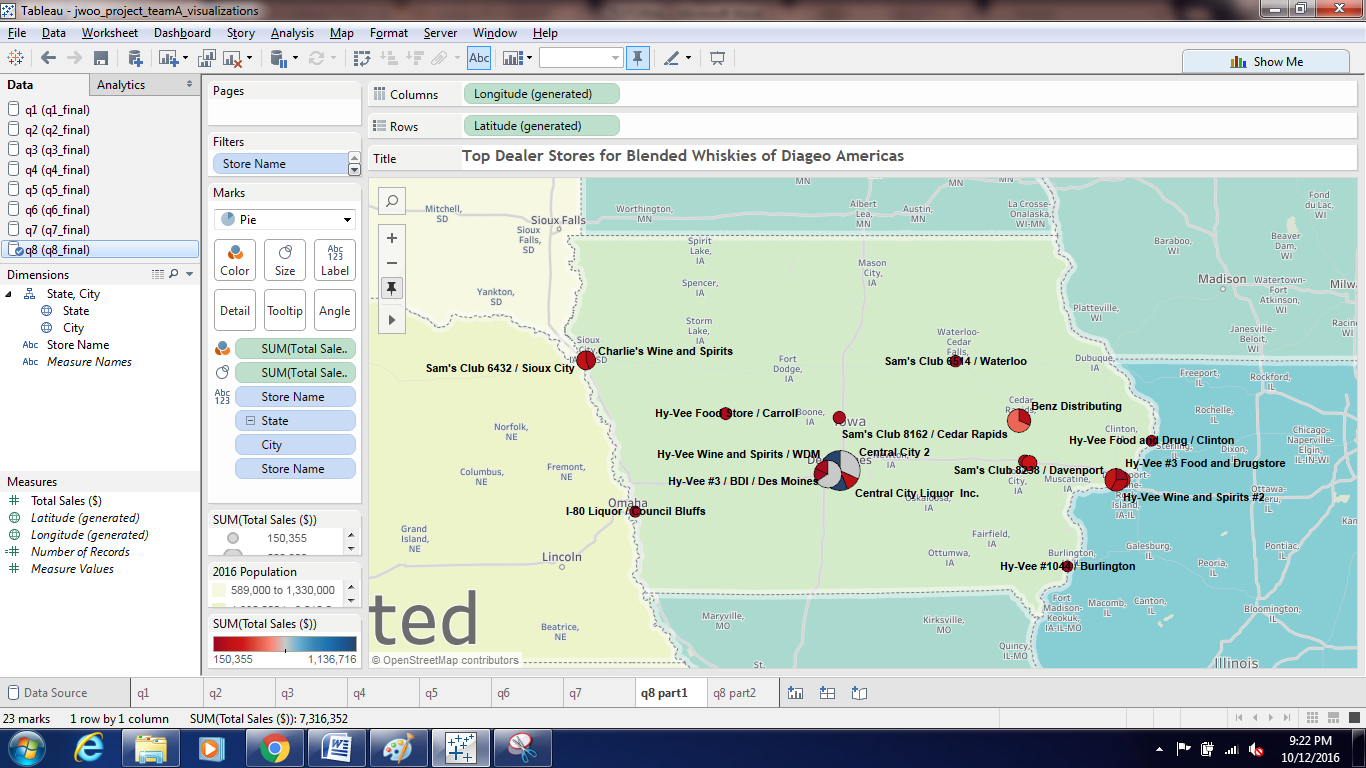
Visualization for Query 6:

Open the Excel output file of Query 6 in Tableau and Drag Volume sold to Columns and Sale($) to Rows. Select a Scatter plot from Show me bar. Put city to Labels and Color on the Marks card. Put Sale($) to Shape on the Marks card. Filter by City having maximum Sale($) for Imported Vodka.



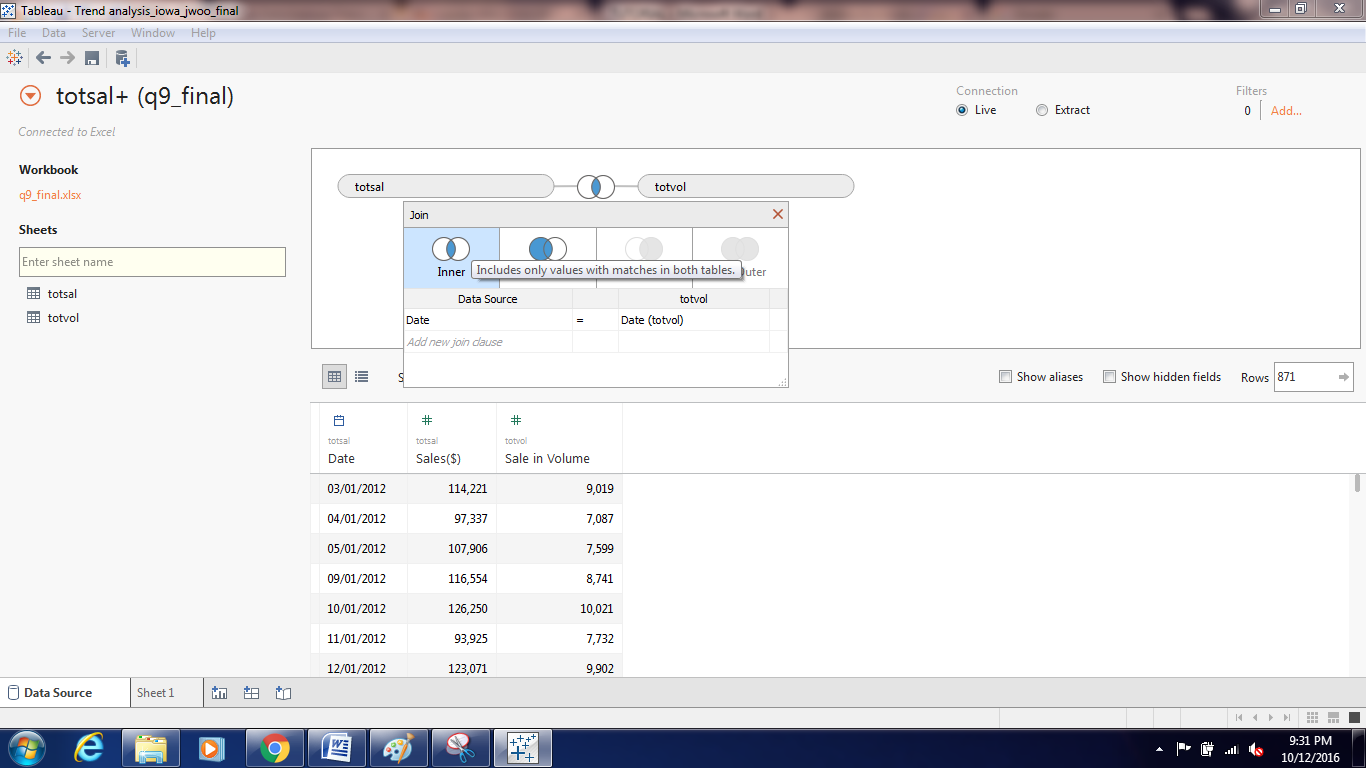
Visualization for Query 7:

Open the output file of Query 7 and then drag Longitude to Columns and Latitude to Rows and select a symbol map from the Show me bar. Filter by store name by Total sales for top 20 stores. Put Total sales to color and size and select a pie chart for the same from the marks card.Place store names to the Labels on the Marks card.



Visualization fro Query 8 and 9:

Open the Excel output file in Tableau and drag Totsal and totvol workbooks and join both datasets by having an inner Join as shown below.



Then drag Date to Columns and select Months from the dropdown. Drag Sale ($) and Sale in Volume to Rows; and select Lines chart from the Show me bar. Synchronize the Axis for Sale in Volume. Click the Analytics tab as shown in the below screen shot and drag Forecast in the workspace.

