

# **Analysis of Iowa Liquor Market in 2012-2018**

## **-A Professional Consulting Report**

**Course & Section: DAT 560M-Sec 23**

**Group 28 members: Zilong Ding (499900)**

**Yiyuan Ding (499242)**

**Jinhao Deng (499965)**

**Yuanqi Fu (500880)**

**April Ge (502452)**

## Content

<b>Executive Summary</b> .....	3
<b>Description of Data</b> .....	3
<b>Problem statement</b> .....	4
1. What are you trying to do? .....	4
2. What are your research questions? .....	4
<b>Why is This Big Data?</b> .....	4
<b>Methods&amp; Results</b> .....	4
1. Process Overview .....	4
2. Integral Analysis .....	5
3. Detailed Analysis .....	5
<b>Conclusion</b> .....	10
<b>Appendix &amp; Codes</b> .....	10

# Big Data Final Project

Dataset Choice: <https://www.kaggle.com/sibmike/iowaliquorsales2020>

## Executive Summary

Our report's purpose is to make a consultant's report to a liquor vendor who wants to join the Iowa Liquor Market. We want to analyze the market data by using big data tactics to grasp a comprehensive view of Iowa Liquor Market to help our customer with their commercial decisions.

We raised six questions regarding total volume trend, top vendors, top selling cities, popular and high profit-making liquor, potential correlation between retail price and unit sales and interaction between seasons and total volume. Those questions can help our customers to get a better understanding on market size, competitors, goods category, pricing strategy and entering time.

Our team utilizes Hive, MapReduce, Hadoop, Tableau, Excel as major tools to process, analyze and visualize corresponding data.

The report generates integral and detailed analyses to help customer identify competitors, a series of strategies and sales performance indicators to enter Iowa liquor markets. We compared the data of other vendors who performed well in previous years.

Overall, we did market size analysis, competitor analysis, goods category analysis, pricing strategy, and launching time analysis. Based on our analysis, we have four recommendations for our customers. To be specific, firstly, our customers can choose Des Moines, Cedar Rapids, etc. Secondly, we suggest our customer choose Canadian Whiskey, which has the highest profit-making ability. Thirdly, our customers should have products of the whole price range. Lastly, our customers should sign a long contract with their suppliers.

## Description of Data

### 1. Overview

The dataset describes the liquor sales data in Iowa during the year 2012-2018. Specifically, it includes every piece of transaction related to liquor sales in Iowa. There are 24 columns in this dataset, which including columns like Purchase date, store name, liquor name, liquor category, vendor name, category name, bottle retails, sales (dollars), etc.

### 2. Details

The total size of the dataset is 4.77GB. The whole dataset is collected and downloaded from Kaggle (link: <https://www.kaggle.com/sibmike/iowaliquorsales2020>). It was uploaded and updated in November 2020 by sibmike, who is from California, United States.

## **Problem statement**

### **1. What are you trying to do?**

Our research focuses on how to provide some consulting advice for new vendors who want to enter the Iowa liquor market. To achieve this goal, we represent a new vendor who has a bunch of various liquor products. Firstly, we do the integral analysis, focusing on the market capability of the Iowa liquor market pattern for the last five years. Secondly, we do the competitor analysis, using MapReduce to get the top 5 total sales vendors. Thirdly, we set our sights on cities' market capability. After we choose the target market, we want to determine the optimal products allocations, which are the products that have the highest profit-making ability. After determining the market and product category, we anticipate that the new vendor has  $n$  units to sell. So, we need to figure out the product pricing strategy by digging out the relation between the retail price and unit. Finally, we will determine the time to put on the market, given that the volume consumed varies with season.

### **2. What are your research questions?**

- a. Hive: is there a decrease or increase on total volume for the last 5 years
- b. MapReduce: Find the top 5 vendors having the highest Sales for last 3 years
- c. Hive: find the top 5 city having the highest sales for last 3 years
- d. Hive: which categories of wine has the top 5 profit making ability for last 3 years
- e. Hive: is there any correlations between retail price and unit sales for last 3 years
- f. Hive: interactions between seasons and total volume sold for recent 3 years

## **Why is This Big Data?**

Our dataset describes the liquor sales data in Iowa during the year 2012-2018. This data is crucial for vendors and stores because they can determine their price and liquor allocation. This dataset is extensive with 4.77 GB and 19.66 million columns. Every Vendors need this big data before distributing and launching their products to analyze the sale performance, profit making ability, total volume sold for each liquor brands, cities. With those data and analysis, vendors could make decisions on how much to produce, areas that need to mainly distribute, and categories that are popular in each city. All those decisions need big data and user analysis to support.

## **Methods& Results**

### **1. Process Overview**

- a. Process data after downloading: Firstly, we uploaded the data directly from local. However, it appeared to be mismatched. We realized that there are ',' in the vendor's name and address. So, we uploaded it to the server and reset the new delimiter which is pipe.
- a. Upload to server.
- b. Upload to Hive.
- c. Analysis Data by using MapReduce.
- d. Analysis Data by using SQL in Hive.

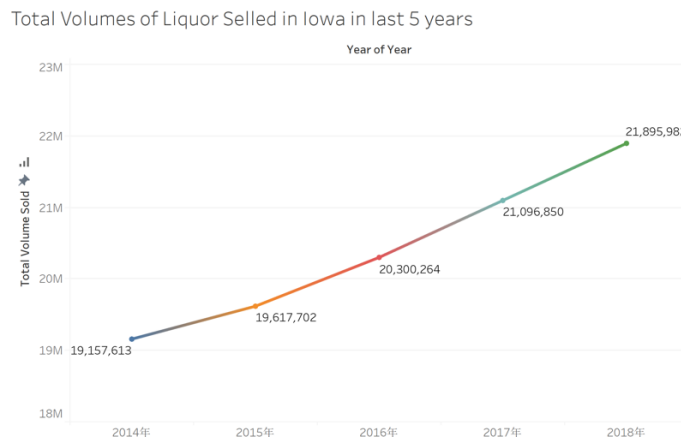
e. Visualization by using Tableau.

## 2. Integral Analysis

a. **Hive:** is there a decrease or increase on total volume for the last 5 years

By running queries on Hive, we process data and group-by and order-by on year to see five consecutive years' total volume sales. We also use Tableau to visualize the growth trend. We can discover when the year goes by and the total alcohol sales increase.

We can find that for the last 5 years, there is a steady increase in the Iowa liquor market, which means that the market is still far from saturation. It's a good time for our customer to enter the Iowa liquor market.



## 3. Detailed Analysis

a. **MapReduce:** Find the top 5 vendors having the highest Sales for last 3 years

Using MapReduce in python environments, we process data and transform all data to lowercase to prevent different cases influencing our results. Then, we use the sort function to derive our top 5 vendors. We also use Tableau to draw the pie chart and text cloud of the top 5 vendors.

After finding the top five competitors, our customer can do more research on their potential target, pricing methods, marketing strategy, and competitive advantage to better understand advantage and strategies of each competitor before entering the market.

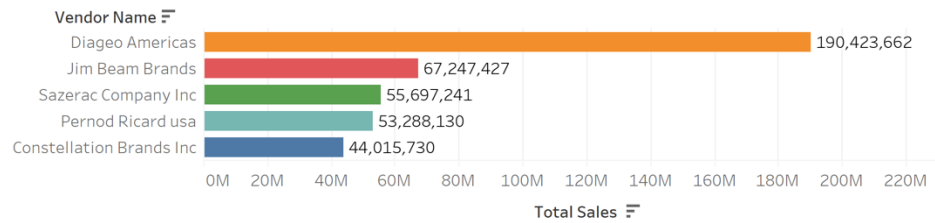
Based on the pie chart, other vendors occupy 57% of the whole market. It represents that the Iowa liquor market is not a monopoly market.

```
diageo americas,190423661.54
jim beam brands,67247427.3801
sazerac company inc,55697240.68
pernod ricard usa,53288129.58
constellation brands inc,44015730.16
```

(Screenshot of MapReduce output—top 5 vendors)

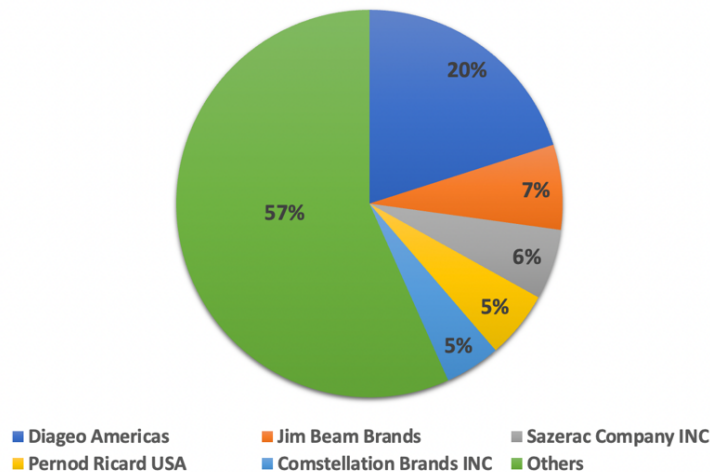
Total Sales from 2016-2018: 948194452.7282178 (Generated by Hive)

The Top 5 vendors having the highest Sales for last 3 years



Constellation Brands Inc  
 Sazerac Company Inc  
 Diageo Americas  
 Pernod Ricard usa Jim Beam Brands  
 (Text cloud of top 5 Vendors)

Share of Sales for the last 3 years

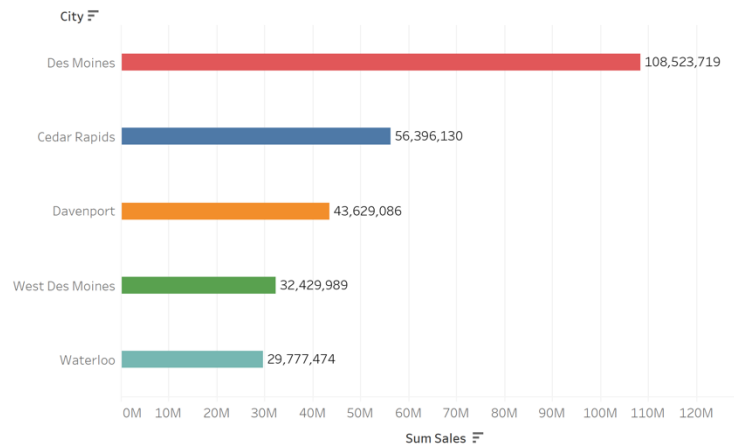


b. **Hive:** find the top 5 city having the highest sales for last 3 years

After determining our major competitor vendors, we decided to analyze which cities are the largest liquor markets for sale.

We used Hive to complete the sorting and computation job, and ultimately get the following five cities. So that, to better avoid risks, our customer can choose Des Moines, Cedar Rapids, Davenport, West Des Moines and Waterloo as primary markets where they should mainly distribute their goods and services and build up their supply chains and storage facilities.

Top 5 city having highest sales in 3 years



(Bar chart showing top 5 cities with highest Sales during 2016-2018)

c. **Hive:** which categories of wine has the top 5 profit making ability for last 3 years

After identifying the cities where vendors can make higher profit, we want to find the categories of wines that have better profit-making ability.

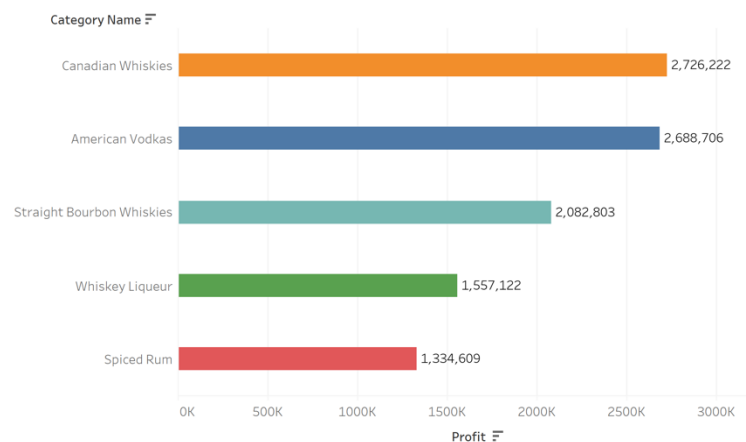
We used Hive to do the sorting and computation job, we used sum of 'state bottle retail' minus 'state bottle cost' to represent each category's profit-making ability. We selected the top 5 profit-making abilities using Hive. Based on this result, we can recommend the following highly profitable types of wines to our customer. Meanwhile, Since the top 5 category of liquor only sum up to 30% of total profit of the market, we also suggest our customer to consider keeping categories of goods diverse to lower potential risks (such as sudden shortage of certain ingredients).

## Top 5 Profit Making Category

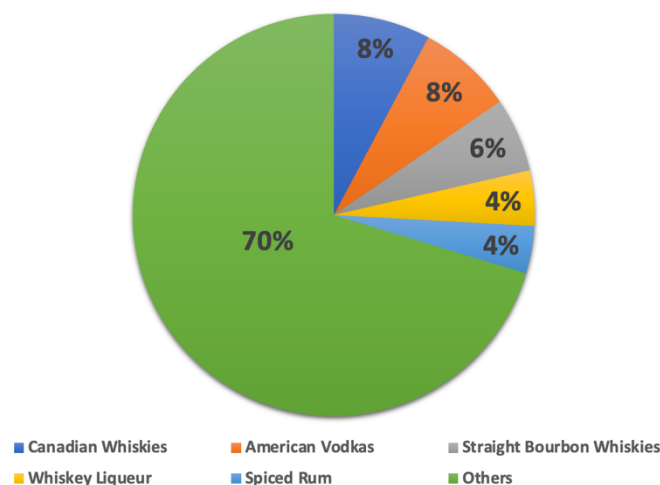
Category Name	
Canadian Whiskies	2,726,222
American Vodkas	2,688,706
Straight Bourbon Whiskies	2,082,803
Whiskey Liqueur	1,557,122
Spiced Rum	1,334,609

(Highlight table of top 5 profit-making categories by total profit)

Top 5 Profit Making Category



Profit Making Category



d. **Hive:** is there any correlations between retail price and unit sales for last 3 years

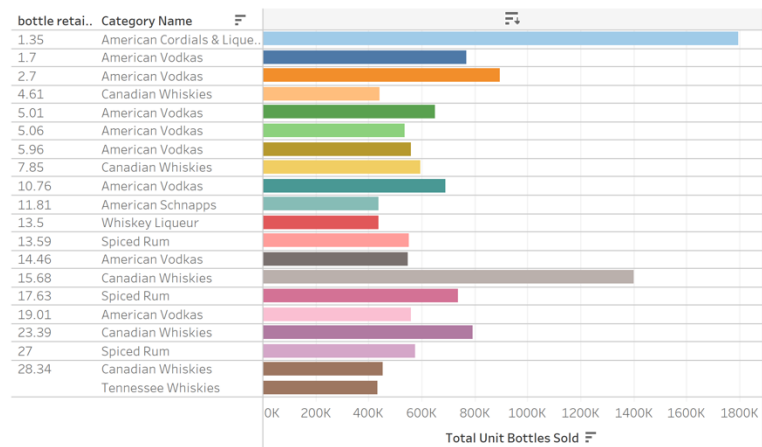
By running SQL in Hive, we can find the retail price and total bottle sold for each category and retail price. Different categories have various kinds of liquors in it, which is the reason that there is repetitive category name with different retail price with it. Doing this analysis could help vendors to set their retail price and find the proper price to set based on their cost and sales performance.

We selected the first 20 retail price and category name with the top 20 bottle sold, and we don't find a clear correlation between retail price and bottle sold.

As we can see the product with the cheapest price (\$1.35) has the highest sale, but the second highest sale have price \$15.68. Thus, our customer can't rely on correlation between retail price and unit sales to determine its retail price. Instead, they should determine their retail price based on the product cost.



Association between Retail Price and Bottle Sold

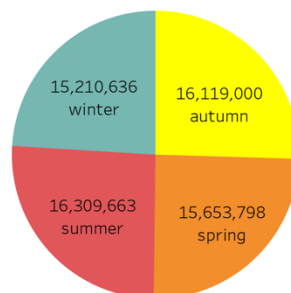
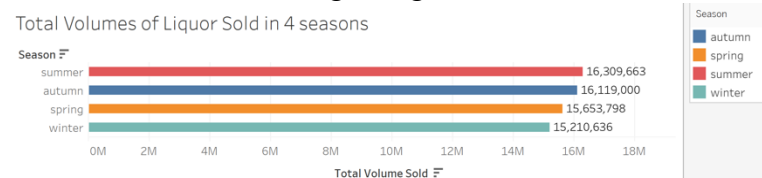


e. **Hive:** interactions between seasons and total volume sold for recent 3 years

Finally, after we finished the competitor analysis, market analysis, and determining the product category and pricing, we need to figure out when the new vendor should enter the market. In other words, which season the total market of Iowa liquor expands.

By using SQL in Hive, we define the month December, January and February are winter; March, April and May are spring; June, July and August are summer; September, October and November are Autumn. By using the sum function in Hive, we find that there is no obvious difference in total volume sold for different seasons. We also use Tableau to visualize to get this conclusion directly.

So, we suppose that there is no preference for our customer to enter the market, but summer has a slightly higher volume consumed, which may be a good time. Meanwhile, we suppose that there is a constant demand through the year, so our customer can make inventory when there is a price reduction. In addition, our customer can sign long and stable contracts with its suppliers.



## Conclusion

The logic of our consulting analysis is market size analysis, competitor analysis, goods category analysis, pricing strategy, and launching time.

From our market size analysis, we can find out that as the total Sales of liquor in Iowa state over last five years increased 3.4% in average, the Iowa liquor market is still far from saturation and it's not a monopoly market. It's a good time for our customer to enter the Iowa liquor market. As for the competitors, we find the top 5 vendors of highest sales, so our customer can get a better understanding of this market environment and their competitive advantages. After determining the competitors, we set our sights on cities' market capability, choosing top 5 cities having highest sales. As for goods category analysis, we determine the optimal products allocations, which are the products that have the highest profit-making ability. Regarding pricing strategy, we don't find a clear correlation between retail price and unit sales. Finally, for the launching time analysis, we calculate volume consumed in different seasons and find there is a constant demand through the year.

Based on our analysis, we will give 4 specific recommendations to our customer:

1. Liquor market in Iowa is still far from saturation, so our customer could choose to enter the market. Our customer can choose Des Moines, Cedar Rapids, Davenport, West Des Moines and Waterloo as primary locations markets where they should mainly distribute their goods and services and build up their supply chains and storage facilities.
2. People in Iowa don't have preference on liquor, which gives our customer more flexibility to choose their products. We will suggest our customer to choose to make Canadian whiskey because it has the highest profit-making ability. However, our customer can also consider more on production cost and supply chain while choosing its products.
3. As we found that liquors across low price to high price all have pretty much sales. We would suggest our customer hat to at least publish three kinds of liquor products across price range of low, medium, and high then adjust the pricing strategy according to their real performance after entered the market.
4. There is no obvious difference on consumption of each season. Thus, our customer can choose to sign a long-term contract with local suppliers.

## Appendix & Codes

#Import data to cloud terminal

```
scp -i S_keypair.pem Iowa_Liquor_Sales.csv d.zilong@18.206.158.228:~
```

#Rename file

```
mv Iowa_Liquor_Sales.csv alc.csv
```

```
#Create file remove header
```

```
tail -n+2 alc.csv > alc_tmp.csv
```

```
#Mapreduce
```

```
#Create mapper function
```

```
nano alcmp.py
```

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip()
    year = line.split(",")[-4:]
    vender = line.split(",")[13]
    sales = line.split(",")[21]

    print '%s,%s,%s' %(vender, sales, year)
```

```
#Create reducer function
```

```
nano alcrd.py
```

```
#!/usr/bin/env python
import sys
ven_sale = {}
for line in sys.stdin:
    vender, sales, year = line.strip().split(',')
    vender = vender.lower()

    try:
        sales = float(sales)
    except ValueError:
        continue

    try:
        year = int(year)
    except ValueError:
        continue
```

```

try:
    if year == 2018 or year == 2017 or year == 2016:
        ven_sale[vender] = ven_sale[vender] + sales
except:
    if year == 2018 or year == 2017 or year == 2016:
        ven_sale[vender] = sales

for v in ven_sale.keys():
    print '%s,%s' %(v, ven_sale[v])

```

### #Create bash function to run map and reduce function

```

nano alcmr.sh
#!/bin/bash
hadoop jar /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-551.jar \
    -Dmapred.reduce.tasks=1 \
    -input /user/d.zilong/alc_tmp.csv \
    -output /user/d.zilong/final_project \
    -file alcmp.py \
    -file alcrd.py \
    -mapper "python alcmp.py" \
    -reducer "python alcrd.py"

```

### #Find top 20 sales vendors

```
cat 2020_result.csv | sort -rnk2 -t"," | head -5
```

### #Hive: find the top 5 city having the highest sales for last 3 years

```



Select city, sum (`sale (dollars)`)
From group28_liquor
Where `date` like '%2018' or `date` like '%2017' or `date` like '%2016'
Group by city
Order by sum (`sale (dollars)`) desc
Limit 5;

```

	city	_c1
1	Des Moines	108523718.98001269
2	Cedar Rapids	56396129.830000624
3	Davenport	43629085.64000338
4	West Des Moines	32429989.390001565
5	Waterloo	29777473.820003565

#Hive: which categories of wine has the top 5 profit making ability for last 3 years




```
Select `category name`, (sum (`state bottle retail`) - sum (`state bottle cost`)) as profit
From group28_liquor
Where `date` LIKE '%2016' or `date` LIKE '%2017' or `date` LIKE '%2018'
Group by `category name`
Order by profit DESC
Limit 5;
```



	category name	profit
1	Canadian Whiskies	2726221.6000026707
2	American Vodkas	2688706.2799820052
3	Straight Bourbon Whiskies	2082803.130008209
4	Whiskey Liqueur	1557121.7699975185
5	Spiced Rum	1334608.7999954373

#Hive: is there a decrease or increase on total volume for the last 3 years

```
Select case when `date` like '%2018' then 2018
      when `date` like '%2017' then 2017
      when `date` like '%2016' then 2016 End AS year, sum (`volume sold (liters)`)
From group28_liquor
Where `date` like '%2018' or `date` like '%2017' or `date` like '%2016'
Group by case
      when `date` like '%2018' then 2018
      when `date` like '%2017' then 2017
      when `date` like '%2016' then 2016
end;
```



	year	volume
1	2014	19157612.509997543
2	2015	19617702.469996393
3	2016	20300264.009996142
4	2017	21096850.109992508
5	2018	21895982.16999236

#Hive: is there any correlations between retail price and unit sales for last 3 years

```
Select `state bottle retail`, sum (`bottles sold`)
from group28_liquor
Where `date` like '%2018' or `date` like '%2017' or `date` like '%2016'
group by `state bottle retail`
```

order by `state bottle retail`  
limit 1000;

	category name	state bottle retail	_c2
1	American Cordials & Liqueur	1.35	1797060
2	Canadian Whiskies	15.68	1401492
3	American Vodkas	2.70	894902
4	Canadian Whiskies	23.39	793853
5	American Vodkas	1.70	769250
6	Spiced Rum	17.63	740137
7	American Vodkas	10.76	691668
8	American Vodkas	5.01	650534
9	Canadian Whiskies	7.85	596421
10	Spiced Rum	27.00	578179
11	American Vodkas	5.96	561967
12	American Vodkas	19.01	561548
13	Spiced Rum	13.59	551965
14	American Vodkas	14.46	549754
15	American Vodkas	5.06	538427
16	Canadian Whiskies	28.34	455018
17	Canadian Whiskies	4.61	443757
18	Whiskey Liqueur	13.50	439198
19	American Schnapps	11.81	436926
20	Tennessee Whiskies	28.34	434525

#Hive: interactions between seasons and total volume sold for recent 3 years

```

Select sum (volume sold (liters)), case
when date like '12%' or date like '01%' or date like '02%' then 'winter'
when date like '03%' or date like '04%' or date like '05%' then 'spring'
when date like '06%' or date like '07%' or date like '08%' then 'summer'
when date like '09%' or date like '10%' or date like '11%' then 'autumn'
End as 'season'
From group28_liquor
Where date like '%2018' or date like '%2017' or date like '%2016'
Group by case
when `date` like '12%' or `date` like '01%' or `date` like '02%' then 'winter'
when `date` like '03%' or `date` like '04%' or `date` like '05%' then 'spring'
when `date` like '06%' or `date` like '07%' or `date` like '08%' then 'summer'
when `date` like '09%' or `date` like '10%' or `date` like '11%' then 'autumn'
else 'other season'

```

end;

Query History			Saved Queries			Results (4)		