Database Foundations for Business Analytics

Project – Iowa Liquor Sales Relational Database

Group Members

1. Ajay Ryan Anand - ara180001
2. Bhagyashri Raghunath Gadkari - brg210002
3. Pakshal Shah - pxs210075
4. Saloni Choksi - sxc210012
5. Shambhavi Sharma - sxs210243

Dataset

This dataset contains the spirits purchase information of Iowa Class “E” liquor licensees (stores) by product and date of purchase from June 1 – October 25, 2021.[1]

Business Understanding

According to a 2020 study, alcohol sales exceed $200 billion per year. Additionally, liquor sales rise by about 4.3% per year in USA. With the information given, it’s safe to assume that the alcohol industry plays a huge role in the U.S. economy. This Kaggle dataset can be used to analyze total spirits sales in Iowa of individual products at the store level. In Iowa, only “Class “E” liquor licensees” (holders of liquor control license) are authorized to sell and deliver alcoholic liquor in the original, sealed, and unopened container to consumers and Class “A”, Class “B”, and Class “C” liquor licensees for consumption off the premises. Our dataset gives us access to the data related to orders placed by these licensees’ category. Based on the orders placed by Class “E” liquor licensees (stores), we can gauge the alcohol demand as they supply to the rest of the state. The goals and targets that we are trying to achieve by studying this dataset for the given time period are:

1. What category of liquor gets consumed the most?
2. What brand tops the charts?
3. Which alcohol vendor is the most preferred?
4. Which city has the highest sales and consumption?
5. Which stores drive the sales?

This can help in understanding the market structure for the alcohol industry in Iowa.

[1] https://data.iowa.gov/Sales-Distribution/2021-Iowa-Liquor-Sales/cc6f-sgik

Data Understanding - description

Original Column Name	Original Data Type	Modified Column Name	SQL Data Type used in database	Modification	Description	Missing Values
Invoice/Item Number	Plain Text	Invoice_id	BIGINT	Invoice_id modified to numeric by removing text characters from original data	Unique identifier(no duplicates) for the individual liquor products included in the store order	None
Date	Date & Time	Invoice_date	Date		Date of order	None
Store Number	Plain Text	Store_id	Integer		Unique number assigned to the store which ordered the liquor	None
Store Name	Plain Text	Store_name	Varchar(255)		Name of store which ordered the liquor	None
Address	Plain Text	Address	Varchar(255)		Address of store who ordered the liquor	None
		City_id	Integer	City_id added to function as primary key for city table	Unique number assigned to the city where store is located	None
City	Plain Text	City	Varchar(255)		City where the store who ordered the liquor is located	None
Zip Code	Plain Text	Zip_code	Integer	Original text data converted to integer	Zip code where the store who ordered the liquor is located	None
Store Location	Point			Column dropped due to missing values	Location of store which ordered the liquor	12%
Category	Plain Text	Category_id	Integer	Original text data converted to integer	Category code associated with the liquor ordered	None
Category Name	Plain Text	Category_name	Varchar(255)		Category of the liquor ordered	None
Vendor Number	Plain Text	Vendor_id	Integer	Original text data converted to integer	The vendor number of the company for the brand of liquor ordered	None
Vendor Name	Plain Text	Vendor_name	Varchar(255)		The vendor name of the company for the brand of liquor ordered	None
Item Number	Plain Text	Item_id	Integer	Original text data converted to integer	Item number for the individual liquor product ordered	None
Item Description	Plain Text	Item_description	Varchar(255)		Description of the individual liquor product ordered	None
Pack	Number	Pack	Integer		The number of bottles in a case for the liquor ordered	None
Bottle Volume (ml)	Number	Bottle_volume_ml	Integer		Volume of each liquor bottle ordered in milliliters	None
State Bottle Cost	Number	Bottle_cost_$	Float		The amount that Alcoholic Beverages Division paid for each bottle of liquor ordered	None
State Bottle Retail	Number	Bottle_Price_$	Float		The amount the store paid for each bottle of liquor ordered	None
Bottles Sold	Number	Bottles_sold	Integer		The number of bottles of liquor ordered by the store	None
Sale (Dollars)	Number	Sale_$	Double		Total cost of liquor order (number of bottles multiplied by the state bottle retail)	None
Volume Sold (Liters)	Number	Sale_volume_liters	Float		Total volume of liquor ordered in liters. (i.e. (Bottle Volume (ml) x Bottles Sold)/1,000)	None

Data Understanding - Summary statistics

Column Name	Mean	Min	Max	Range	Unique No of observations
Invoice_id	-	-	-	-	1048575
Invoice_date	-	6/1/2021	10/25/2021		
Store_id	-	-	-	-	1879
Store_name	-	-	-	-	
Address	-	-	-	-	
City_id	-	-	-	-	440
City	-	-	-	-	440
Zip_code	-	-	-	-	
Category_id	-	-	-	-	56
Category_name	-	-	-	-	
Vendor_id	-	-	-	-	196
Vendor_name	-	-	-	-	
Item_id	-	-	-	-	3624
Item_description	-	-	-	-	
Pack		1	120	119	
Bottle_volume_ml		20	3500	3480	
Bottle_cost_$		$0.66	$1949.02	$1948.36	
Bottle_Price_$		$0.99	$2923.53	$2922.54	
Bottles_sold		1	13,200	13,199	
Order_value_$	$168.90	$1.34	$250,932	$250,930.66	
Sale_volume_liters	9.73	.02	13,200	13,199.98	

The overall size of the dataset is 247 mb. The data has 1,048,575 rows, all of which are uniquely identified by the invoice id.

The original dataset had 24 columns. We excluded county number and county name from the original dataset because it was not related to our business understanding. We also excluded the store location as it had missing values. Additionally, we excluded the sale volume in gallons and only kept the sale volume in liters as gallon amount can be deduced by simple multiplication which can be performed when retrieving data from the tables.

We added the city id column to identify each city, despite each city being unique in Iowa so that there are fewer errors in data entry as it is easier to commit spelling mistakes when making character entries in data tables. Because of these changes, we decided to go with 21 columns in our dataset.

The relationship between these columns is of the form:

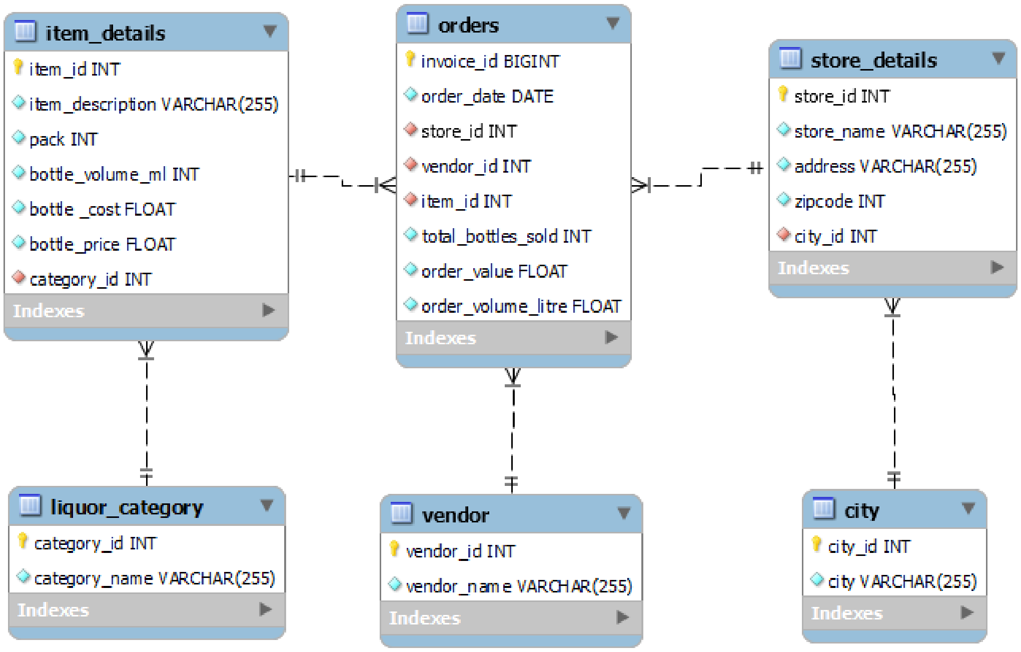
Store located in a city places orders from the vendor for a particular item/product belonging to a category on a date and all this information is captured in the invoice. Based on this relationship structure we could see that there are functional dependencies between these columns, therefore we broke the larger dataset into 6 tables to minimize the functional dependency.

Design a Database

The 6 tables which we constructed, and its ER diagram are:

Table No	Table Name
1	Orders
2	Store_details
3	City
4	Liquor_category
5	Vendor
6	Item_details

ER Diagram



After this we performed checks for identifying whether our schema is in BCNF (Boyce-Codd Normal Form)

F={		Check for BCNF		Functional dependency
	{Invoice_id}→{Store_id}	Invoice_id and store_id are in table 1, Invoice_id is key	√	Based on initial decomposition
	{Store_id}→{City_id}	store_id and city_id are in table 2, store_id is key	√	Based on initial decomposition
	{City_id}→{City_name}	city_id and city_name are in table 3, city_id is key	√	Based on initial decomposition
	{Store_id}→{City_name}	Store_id and City_name are not in the same table	√	Inferred
	{Invoice_id}→{City_id, City_name}	Invoice_id and city_id and city_name are not in the same table	√	Inferred
	{Invoice_id}→{Vendor_id}	Invoice_id and vendor_id are in table 1, Invoice_id is key	√	Inferred
	{Vendor_id}→{Vendor_name}	Vendor_id and Vendor_name are in table 5, Vendor_id is key	√	Based on initial decomposition

	{Invoice_id}→{Vendor_name}	Invoice_id and vendor_name are not in the same table	√	Inferred
	{Invoice_id}→{Item_id}	Invoice_id and item_id are in table 1, Invoice_id is key	√	Based on initial decomposition
	{Item_id}→{category_id}	Item_id and category_id are in table 6, Item_id is key	√	Based on initial decomposition
	{Invoice_id}→{category_id}	Invoice_id and category_id are not in the same table	√	Inferred
	{Category_id}→{Category_name}	Category_id and category_name are in table 4, Category_id is key	√	Based on initial decomposition
	{Invoice_id}→{Category_name}	Invoice_id and category_name are not in the same table	√	Inferred
	{Invoice_id}→{Pack, Bottles_sold, Sales_$, Sale_volume_litres}	Invoice_id, Pack, Bottles_sold, Sales_$, Sale_volume_litres are in table 1, Invoice_id is key	√	Based on initial decomposition
	{Store_id}→{Store_name, Address, Zip_code}	Store_id, Store_name, Address, Zip_code are in table 2, Store_id is key	√	Based on initial decomposition
	{Item_id}→{Item_Description, Pack, Bottle Volume (ml), bottle_Cost, bottle_price }	Item_id, Item_Description, Pack, bottle_volume_ml, Bottle_price, bottle_cost are in table 6, Item_id is key	√	Based on initial decomposition
	{Invoice_id}→{Store_name, Address, Zip_code}	Invoice_id, Store_name, Address, Zip_code are not in the same table	√	Inferred
	{Invoice_id}→{Item_Description, Pack, bottle_volume_ml, bottle_cost, bottle_price }	Invoice_id, Item_Description, Pack, bottle_volume_ml, bottle_cost, bottle_price are not in the same table	√	Inferred
}				

The ER diagram is in BCNF, and we proceed with loading the data into MySQL server.

Data Cleaning and Database Testing

The dataset is mostly clean with only minor inconsistencies like the columns with characters sometimes entered in uppercase or lowercase. Fortunately for us, these did not pose any problems in loading the data which was done in the later stages of the project. The date variable had to be brought into the format acceptable to MySQL, so that it loaded properly. Also, the comma separated numerical values format had to be changed to non-comma separated because MySQL was truncating the data at comma leading to incorrect data being loaded into tables.

Constraint Check:

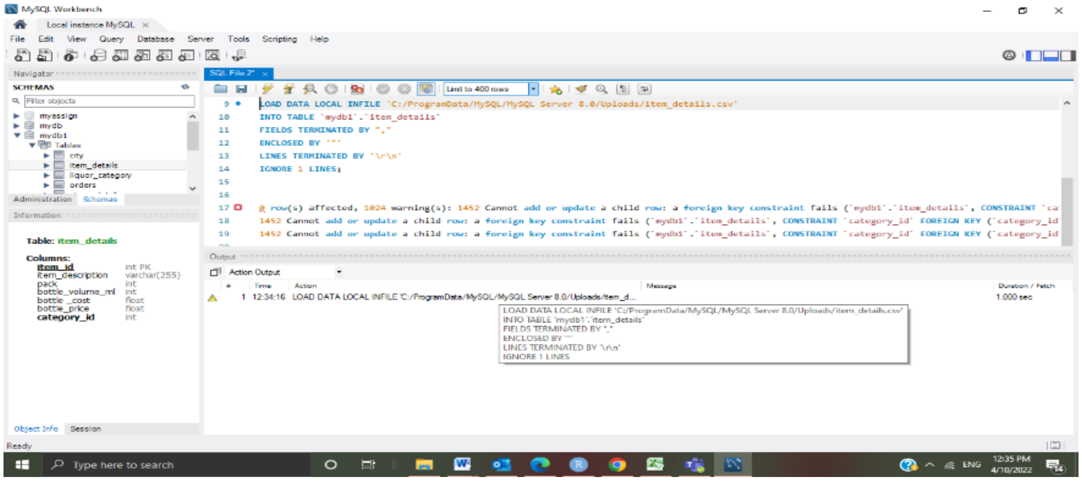


Table: item_details

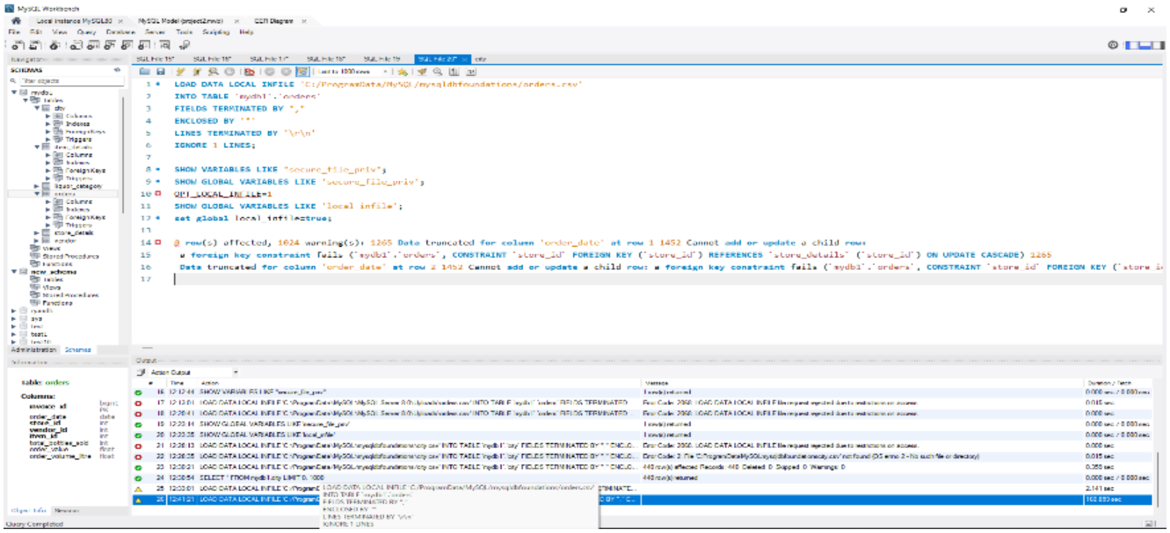
Columns:

item_id	int PK
item_description	varchar(255)
pack	int
bottle_volume_ml	int
bottle _cost	float
bottle_price	float
category_id	int

```
LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/item_details.csv'
INTO TABLE `mydb1`.`item_details`
FIELDS TERMINATED BY ","
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES;


0 row(s) affected, 1024 warning(s): 1452 Cannot add or update a child row: a foreign key constraint fails (`mydb1`.`item_details`, CONSTRAINT `ca
1452 Cannot add or update a child row: a foreign key constraint fails (`mydb1`.`item_details`, CONSTRAINT `category_id` FOREIGN KEY (`category_id
1452 Cannot add or update a child row: a foreign key constraint fails (`mydb1`.`item_details`, CONSTRAINT `category_id` FOREIGN KEY (`category_id
```

#	Time	Action	Message	Duration / Fetch
1	12:34:16	LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/item_d...		1.000 sec

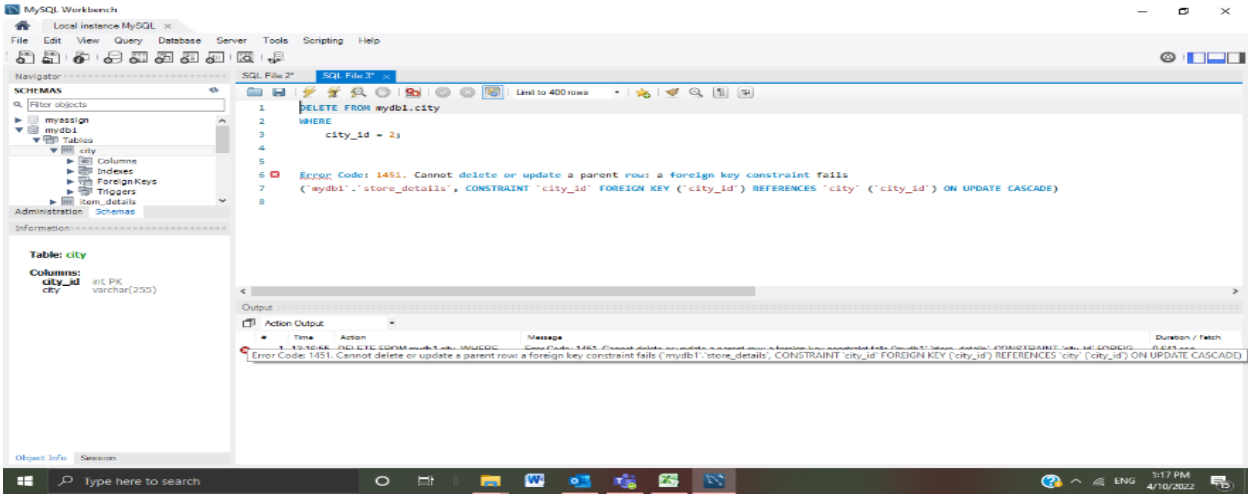
```
LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/item_details.csv'
INTO TABLE `mydb1`.`item_details`
FIELDS TERMINATED BY ","
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES
```



```
LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/mysqldbfoundations/orders.csv'
INTO TABLE `mydb1`.`orders`
FIELDS TERMINATED BY ","
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES;

SHOW VARIABLES LIKE "secure_file_priv";
SHOW GLOBAL VARIABLES LIKE 'secure_file_priv';
OPT_LOCAL_INFILE=1
SHOW GLOBAL VARIABLES LIKE 'local infile';
set global local_infile=true;

0 row(s) affected, 1024 warning(s): 1265 Data truncated for column 'order_date' at row 1 1452 Cannot add or update a child row:
 a foreign key constraint fails (`mydb1`.`orders`, CONSTRAINT `store_id` FOREIGN KEY (`store_id`) REFERENCES `store_details` (`store_id`) ON UPDATE CASCADE) 1265
 Data truncated for column 'order_date' at row 2 1452 Cannot add or update a child row: a foreign key constraint fails (`mydb1`.`orders`, CONSTRAINT `store_id` FOREIGN KEY (`store_i
```

```
#   Time      Action                                                                                                   Message                                                                                      Duration / Fetch
16  12:12:44  SHOW VARIABLES LIKE "secure_file_priv"                                                                   1 row(s) returned                                                                            0.000 sec / 0.000 sec
17  12:13:01  LOAD DATA LOCAL INFILE 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\orders.csv' INTO TABLE `mydb1`.`orders` FIELDS TERMINATED ...  Error Code: 2068. LOAD DATA LOCAL INFILE file request rejected due to restrictions on access.   0.015 sec
18  12:20:41  LOAD DATA LOCAL INFILE 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\orders.csv' INTO TABLE `mydb1`.`orders` FIELDS TERMINATED ...  Error Code: 2068. LOAD DATA LOCAL INFILE file request rejected due to restrictions on access.   0.000 sec
19  12:23:14  SHOW GLOBAL VARIABLES LIKE 'secure_file_priv'                                                            1 row(s) returned                                                                            0.000 sec / 0.000 sec
20  12:23:35  SHOW GLOBAL VARIABLES LIKE 'local_infile'                                                                1 row(s) returned                                                                            0.000 sec / 0.000 sec
21  12:26:13  LOAD DATA LOCAL INFILE 'C:\ProgramData\MySQL\mysqldbfoundations\city.csv' INTO TABLE `mydb1`.`city` FIELDS TERMINATED BY ',' ENCLO...  Error Code: 2068. LOAD DATA LOCAL INFILE file request rejected due to restrictions on access.   0.000 sec
22  12:28:35  LOAD DATA LOCAL INFILE 'C:\ProgramData\MySQL\mysqldbfoundations\city.csv' INTO TABLE `mydb1`.`city` FIELDS TERMINATED BY ',' ENCLO...  Error Code: 2. File 'C:\ProgramDataMySQLmysqldbfoundationscity.csv' not found (OS errno 2 - No such file or directory)   0.015 sec
23  12:30:21  LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/mysqldbfoundations/city.csv' INTO TABLE `mydb1`.`city` FIELDS TERMINATED BY ',' ENCLO...  440 row(s) affected Records: 440  Deleted: 0  Skipped: 0  Warnings: 0   0.359 sec
24  12:30:54  SELECT * FROM mydb1.city LIMIT 0, 1000                                                                   440 row(s) returned                                                                          0.000 sec / 0.000 sec
25  12:33:01  LOAD DATA LOCAL INFILE 'C:/ProgramD...                                                          ...TERMINATE...                                                                                              2.141 sec
26  12:41:21  LOAD DATA LOCAL INFILE 'C:/ProgramD...                                                          ...D BY ',' C...                                                                                             160.859 sec

LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/mysqldbfoundations/orders.csv'
INTO TABLE `mydb1`.`orders`
FIELDS TERMINATED BY ","
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES
```

Data in tables with foreign key failed to load till the table with the foreign key as primary key was populated. After populating the data in the tables in the correct order, we checked for delete and update constraints.

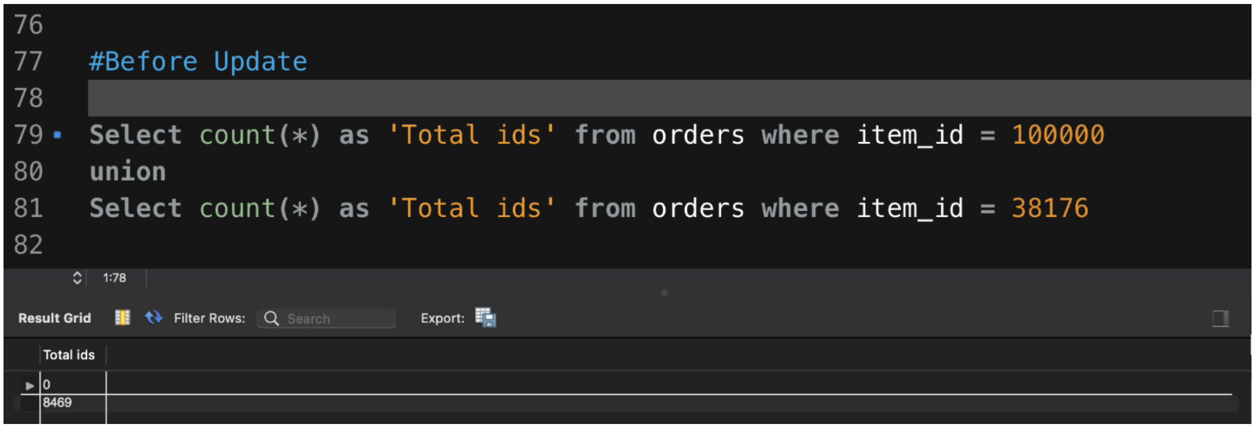


```
DELETE FROM mydb1.city
WHERE
    city_id = 2;


Error Code: 1451. Cannot delete or update a parent row: a foreign key constraint fails
(`mydb1`.`store_details`, CONSTRAINT `city_id` FOREIGN KEY (`city_id`) REFERENCES `city` (`city_id`) ON UPDATE CASCADE)
```

```
#  Time      Action                           Message                                                                                     Duration / Fetch
1  13:16:56  DELETE FROM mydb1.city WHERE     Error Code: 1451. Cannot delete or update a parent row: a foreign key constraint fails (`mydb1`.`store_details`, CONSTRAINT `city_id` FOREIGN KEY (`city_id`) REFERENCES `city` (`city_id`) ON UPDATE CASCADE)   0.641 sec
```

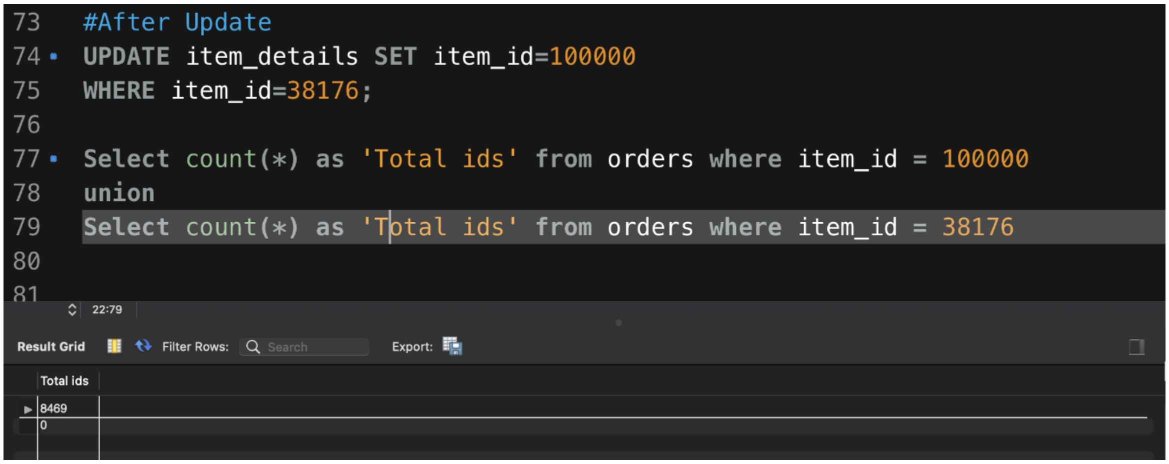
Table: city

Columns:
city_id int PK
city varchar(255)



```
#Before Update

Select count(*) as 'Total ids' from orders where item_id = 100000
union
Select count(*) as 'Total ids' from orders where item_id = 38176
```

Total ids
0
8469

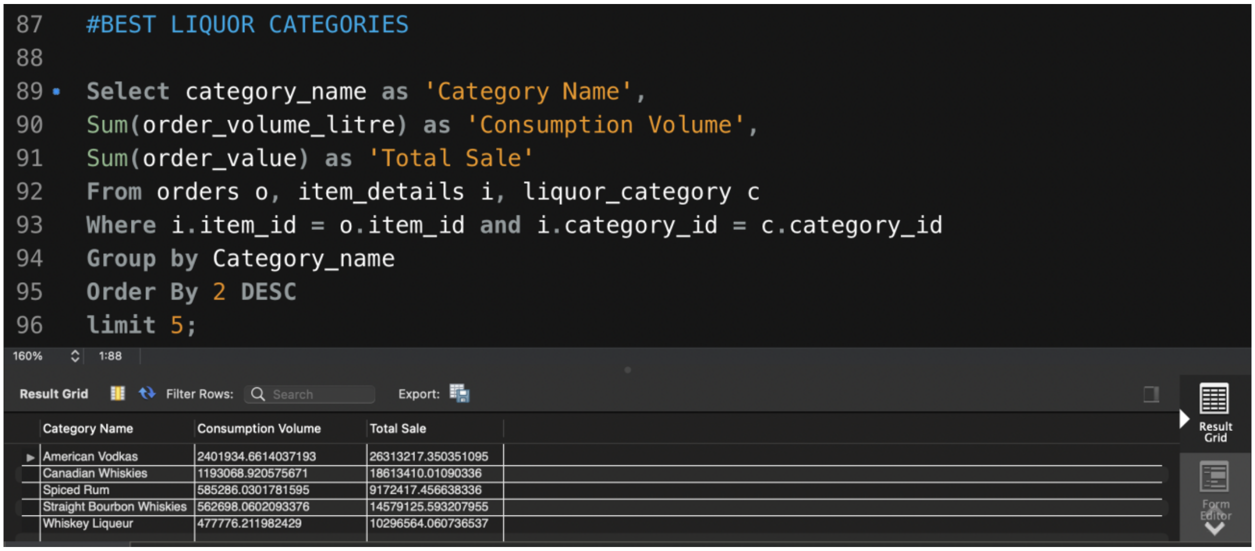


```sql
#After Update
UPDATE item_details SET item_id=100000
WHERE item_id=38176;

Select count(*) as 'Total ids' from orders where item_id = 100000
union
Select count(*) as 'Total ids' from orders where item_id = 38176

```

Total ids
8469
0

After checking for all the constraints we ran queries to answer the business understanding questions:

1. What type of liquor gets consumed the most?
   Vodka and whiskey are the most consumed categories of alcohol.

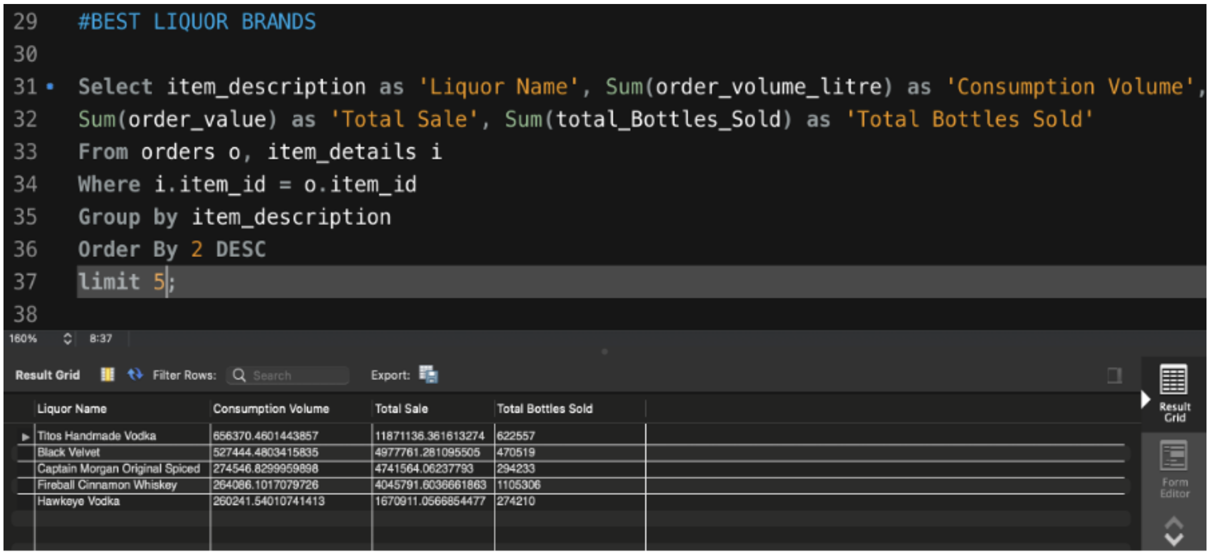


```sql
#BEST LIQUOR CATEGORIES

Select category_name as 'Category Name',
Sum(order_volume_litre) as 'Consumption Volume',
Sum(order_value) as 'Total Sale'
From orders o, item_details i, liquor_category c
Where i.item_id = o.item_id and i.category_id = c.category_id
Group by Category_name
Order By 2 DESC
limit 5;
```

Category Name	Consumption Volume	Total Sale
American Vodkas	2401934.6614037193	26313217.350351095
Canadian Whiskies	1193068.920575671	18613410.01090336
Spiced Rum	585286.0301781595	9172417.456638336
Straight Bourbon Whiskies	562698.0602093376	14579125.593207955
Whiskey Liqueur	477776.211982429	10296564.060736537

2. What brand tops the charts?

Titos Handmade Vodka is the most consumed product (based on consumption volume) in our dataset, followed by Black Velvet Canadian Whiskey. This ties in with the categories of liquor that get consumed the most.

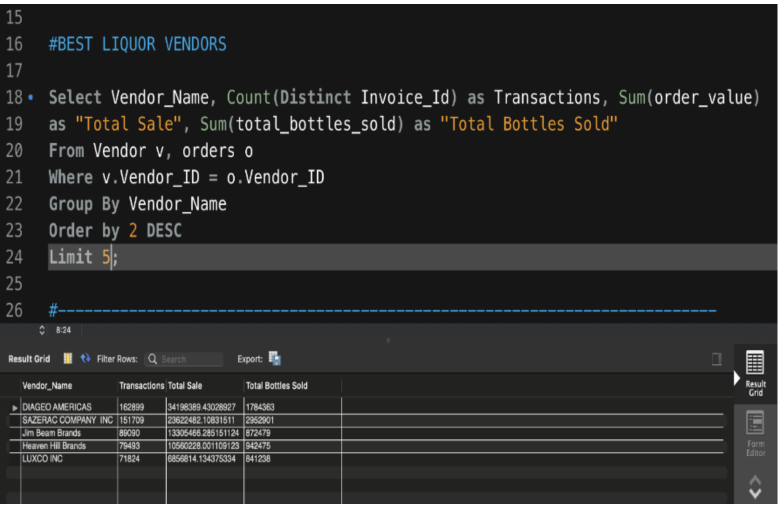


```sql
#BEST LIQUOR BRANDS

Select item_description as 'Liquor Name', Sum(order_volume_litre) as 'Consumption Volume',
Sum(order_value) as 'Total Sale', Sum(total_Bottles_Sold) as 'Total Bottles Sold'
From orders o, item_details i
Where i.item_id = o.item_id
Group by item_description
Order By 2 DESC
limit 5;
```

Liquor Name	Consumption Volume	Total Sale	Total Bottles Sold
Titos Handmade Vodka	656370.4601443857	11871136.361613274	622557
Black Velvet	527444.4803415835	4977761.281095505	470519
Captain Morgan Original Spiced	274546.8299959898	4741564.06237793	294233
Fireball Cinnamon Whiskey	264086.1017079726	4045791.6036661863	1105306
Hawkeye Vodka	260241.54010741413	1670911.0566854477	274210

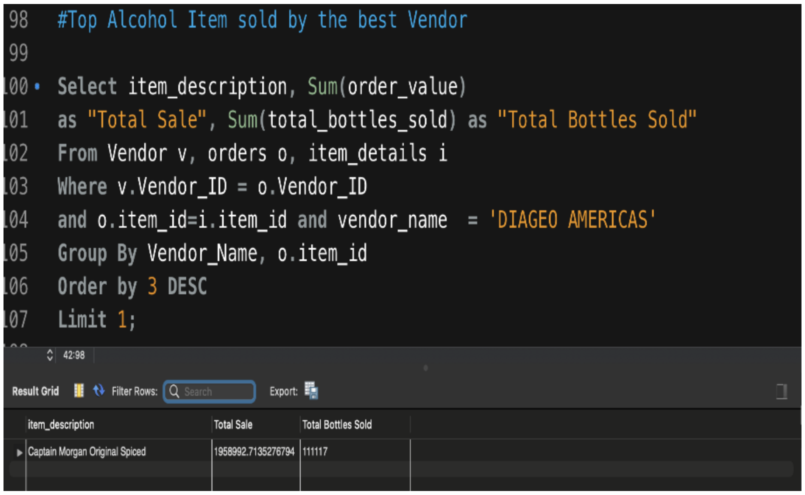
3. Which alcohol vendor is the most preferred?

The vendor Diageo Americas has the highest number of individual transactions/invoices closely followed by Sazerac Company Inc indicating their popularity. For Diageo Americas the biggest source of revenue was " Captain Morgan Spiced Rum".



```sql
#BEST LIQUOR VENDORS

Select Vendor_Name, Count(Distinct Invoice_Id) as Transactions, Sum(order_value)
as "Total Sale", Sum(total_bottles_sold) as "Total Bottles Sold"
From Vendor v, orders o
Where v.Vendor_ID = o.Vendor_ID
Group By Vendor_Name
Order by 2 DESC
Limit 5;

#---------------------------------------------------------------------------
```

Vendor_Name	Transactions	Total Sale	Total Bottles Sold
DIAGEO AMERICAS	162899	34198389.43028927	1784363
SAZERAC COMPANY INC	151709	23622482.10831511	2952901
Jim Beam Brands	89090	13305466.285151124	872479
Heaven Hill Brands	79493	10560228.001109123	942475
LUXCO INC	71824	6856814.134375334	841238

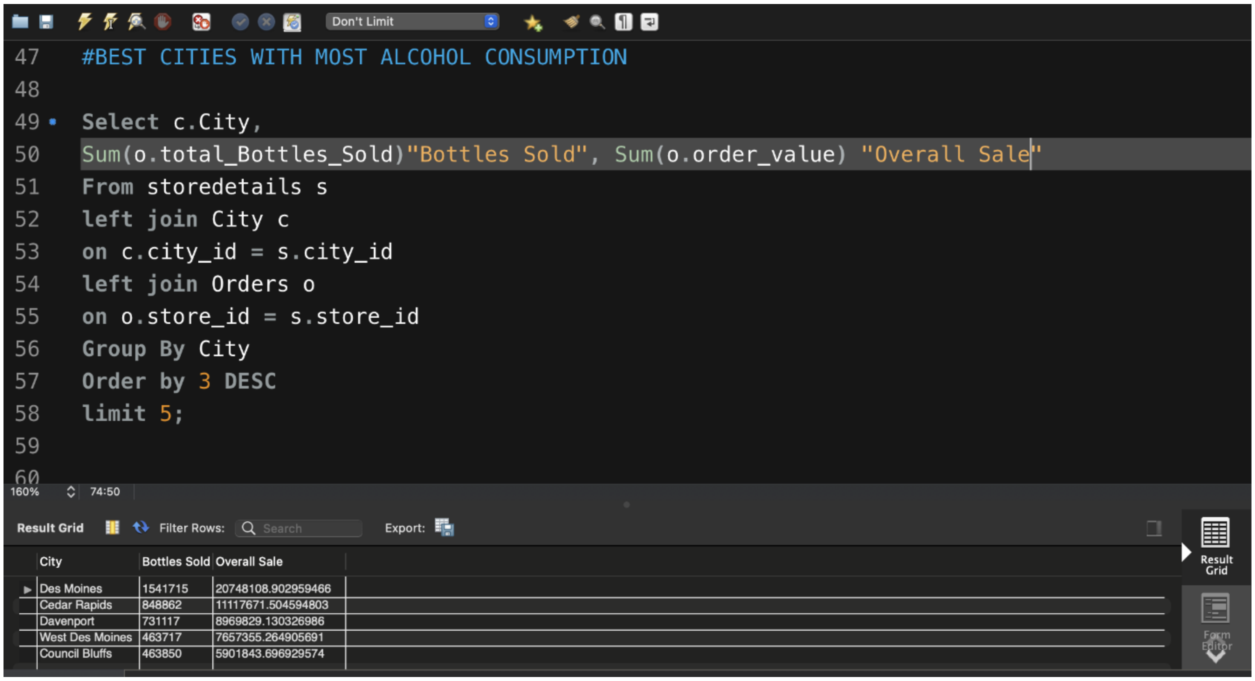


```sql
#Top Alcohol Item sold by the best Vendor

Select item_description, Sum(order_value)
as "Total Sale", Sum(total_bottles_sold) as "Total Bottles Sold"
From Vendor v, orders o, item_details i
Where v.Vendor_ID = o.Vendor_ID
and o.item_id=i.item_id and vendor_name  = 'DIAGEO AMERICAS'
Group By Vendor_Name, o.item_id
Order by 3 DESC
Limit 1;
```

item_description	Total Sale	Total Bottles Sold
Captain Morgan Original Spiced	1958992.7135276794	111117

4. Which city has the highest sales and consumption?

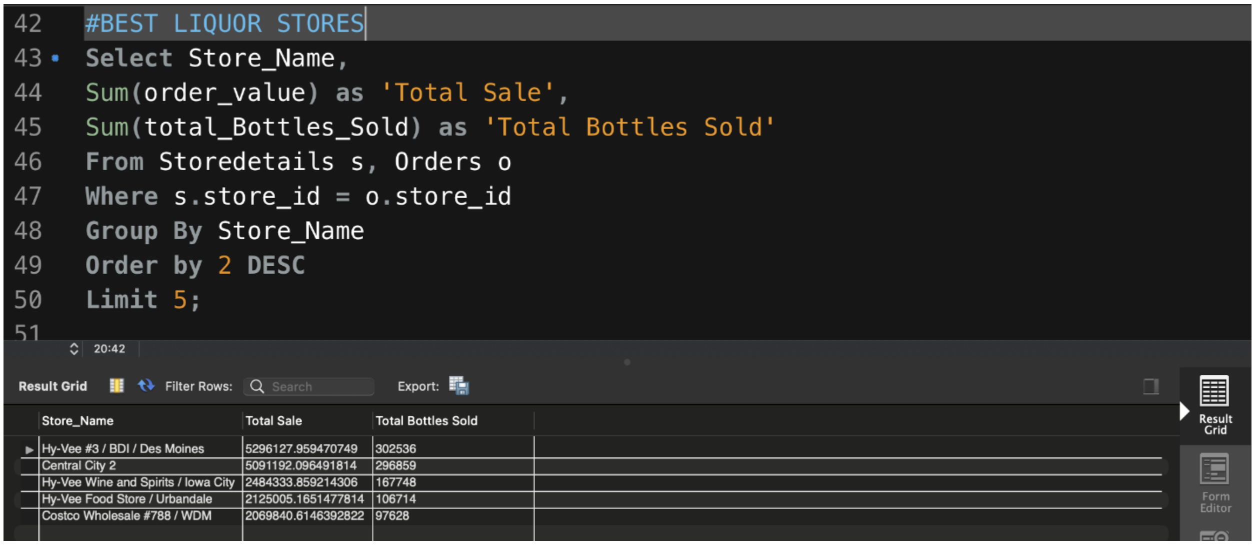
Des Moines, which is the largest city in Iowa, consumes the most alcohol.



```sql
#BEST CITIES WITH MOST ALCOHOL CONSUMPTION

Select c.City,
Sum(o.total_Bottles_Sold)"Bottles Sold", Sum(o.order_value) "Overall Sale"
From storedetails s
left join City c
on c.city_id = s.city_id
left join Orders o
on o.store_id = s.store_id
Group By City
Order by 3 DESC
limit 5;
```

City	Bottles Sold	Overall Sale
Des Moines	1541715	20748108.902959466
Cedar Rapids	848862	11117671.504594803
Davenport	731117	8969829.130326986
West Des Moines	463717	7657355.264905691
Council Bluffs	463850	5901843.696929574

5. Which stores drive the sales?
Hy-Vee #3, located in Des Moines, leads liquor sales with a total of 302,536 bottles sold. Overall, the Hy-Vee franchise drives the most liquor sales in Iowa.



```sql
#BEST LIQUOR STORES
Select Store_Name,
Sum(order_value) as 'Total Sale',
Sum(total_Bottles_Sold) as 'Total Bottles Sold'
From Storedetails s, Orders o
Where s.store_id = o.store_id
Group By Store_Name
Order by 2 DESC
Limit 5;
```

Store_Name	Total Sale	Total Bottles Sold
Hy-Vee #3 / BDI / Des Moines	5296127.959470749	302536
Central City 2	5091192.096491814	296859
Hy-Vee Wine and Spirits / Iowa City	2484333.859214306	167748
Hy-Vee Food Store / Urbandale	2125005.1651477814	106714
Costco Wholesale #788 / WDM	2069840.6146392822	97628

From our initial observations, we noticed that Iowa is driven by certain products/brands, vendors, and stores which are concentrated in the most populous cities. Further data analysis is necessary to arrive at concrete conclusions, but that is beyond the scope of this project.