DATA PROFILING:

Graphical user interface, application

Description automatically generated

We listed down column names and data types that were coming from staging tables so it was easier while loading into dim and facts

**IOWA Sales Data**   
   
**String/Character Fields**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Remarks** |
| Item Number | 0.0% | 11,730 | 633 | x904631 | 1 | 248,845 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| County Number | 2.9% | 100 | 90 | 90 | 4,310 | 4,459,618 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Zip Code | 0.3% | 510 | 52501 | 52501 | 1 | 599,545 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Address | 0.3% | 2,822 | 123 A ST | 1510 SOUTH ANKENY BOULEVARD PRAIRIE TRAIL SUITE 106 108 | 1 | 279,983 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store Name | 0.0% | 2,940 | SAME | FORT MADISON TOBACCO AND LIQUOR OUTLET PLUS / FORT MADISON | 1 | 205,995 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| County | 0.6% | 105 | IDA | POTTAWATTAMIE | 2 | 4,568,617 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store Number | 0.0% | 2,807 | 5307 | 010025 | 1 | 205,995 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| City | 0.3% | 477 | ELY | COLUMBUS JUNCTION | 1 | 2,135,584 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store Location | 9.9% | 3,204 | POINT (-91.02618 41.59) | POINT (-91.67707300000001 41.968016000000006) | 1 | 2,450,072 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Date | 0.0% | 2,730 | 12/22/2016 | 12/22/2016 | 1 | 18,002 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Vendor Number | 0.0% | 439 | 55 | 420 | 1 | 4,127,854 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Category Name | 0.1% | 114 | MEZCAL | AMERICAN DISTILLED SPIRITS SPECIALTY | 1 | 2,376,643 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Vendor Name | 0.0% | 549 | MHW LTD | PRESTIGE WINE & SPIRITS GROUP / UNITED STATES DISTILLED PRODUCTS CO | 1 | 4,127,854 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Item Description | 0.0% | 10,745 | A+ | THE GLENLIVET TASTING KIT W/12YR 200ML, 15YR 200ML & 18YR 200ML | 1 | 631,415 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Invoice/Item Number | 0.0% | 24,847,481 | 378900006 | RINV-04236500053 | 1 | 1 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |

**IOWA Product**   
   
**String/Character Fields**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Remarks** |
| List Date | 0.0% | 720 | 10/01/2017 | 10/01/2017 | 1 | 205 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Vendor | 0.0% | 280 | 421 | 421 | 1 | 728 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Category Name | 0.0% | 49 | MEZCAL | IMPORTED DISTILLED SPIRITS SPECIALTY | 3 | 644 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Vendor Name | 0.0% | 280 | PROXIMO | AMERICAN HERITAGE DISTILLERS, LLC / CENTURY FARMS DISTILLERY | 1 | 728 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Item Description | 0.0% | 5,243 | P31 | HIGH WEST WHISKEY HUCKBERRY GIFT PACK W/WHISKEY PEAK GLASSES | 1 | 8 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Report Date | 0.0% | 1 | 10/01/2022 | 10/01/2022 | 6,350 | 6,350 |  |

**String/Character Fields**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Remarks** |
| Address | 0.0% | 2,198 | 708 MAIN | 1510 SOUTH ANKENY BOULEVARD PRAIRIE TRAIL SUITE 106 108 | 1 | 3 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store Address | 5.4% | 1,950 | POINT (-91.02618 41.59) | POINT (-93.755906 41.623602) | 1 | 124 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Name | 0.0% | 2,237 | RNDC | FORT MADISON TOBACCO AND LIQUOR OUTLET PLUS / FORT MADISON | 1 | 3 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| City | 0.0% | 458 | ELY | COLUMBUS JUNCTION | 1 | 106 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store | 0.0% | 2,277 | 5529 | 010035 | 1 | 1 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. |
| Store Status | 0.0% | 2 | A | A | 323 | 1,954 |  |
| State | 0.0% | 1 | IA | IA | 2,277 | 2,277 |  |
| Report Date | 0.0% | 1 | 10/04/2022 | 10/04/2022 | 2,277 | 2,277 |  |

Graphical user interface, text, application

Description automatically generated

**Some Insights after doing data profiling part:**  
   
1) We found there was a discrepancy in the county column between sales table and county population table. In one of the tables the word county was present whereas it was absent in the other table, so we had to make it in a standard format to perform joins.   
   
2) We also found that there were case sensitive issues. The data coming from the sales stage table had store names in capital case whereas store stage table had mixed case format. So, while loading the fact table we had to put up a data cleaning tool to make data consistent

3) By analyzing data with this query:

select storename, date, sum(bottle sold), sum(sales dollars) from sales group by storename, date

order by store name asc, date desc   
   
We found a pattern that for a every store name, there were many invoices generated on a particular date. Now using the summation tool we first populated header table. And then we populated line item fact table.

4) We also populated missing data with standard data like:   
   
-99: null (integer data types)   
‘Unknown’: null (varchar data types)   
(0,0): location:null 

5) The location column in all the staging tables had a value similar to POINT(longitude, latitude). For data viz we required (lat,long), so we had to split the data into 3 columns (location, latitude, longitude)   
   
6) Reloaded dimension: For County, Category, Vendor, City data in the staging tables had missing values (that is few cities were present in sales staging tables which were missing in cities staging tables. So, to populate dimensions with accurate and all the cities, we performed union between the left and inner join)

Diagram, schematic

Description automatically generated

 Row counts:

Graphical user interface, application, email

Description automatically generated

IOWA Part 2

Dim\_County & Dim\_Category\_Reloaded:

Diagram

Description automatically generated

FCT\_Total\_Population\_by\_City & FCT\_Total\_Population\_by\_County:

Text

Description automatically generated

Dim\_Stores & Dim\_Vendor\_Reloaded:

Diagram

Description automatically generated

Dim\_City\_Reloaded:

A picture containing diagram

Description automatically generated

Dim\_City:

Timeline

Description automatically generated

Dim\_Date:

Timeline

Description automatically generated with medium confidence

FCT\_iowa\_liquor\_sales\_header:

Diagram

Description automatically generated

FCT\_iowa\_liquor\_sales\_line\_item:

Timeline

Description automatically generated with medium confidence