Mining Iowa Liquor Sales

Team members:
Nicholas Stafford
Alexander Sueppel
Alex Thomas
Kevin Vick

University of Colorado Boulder

Data Preprocessing - Reduction & Cleaning

- Data Reduction:
  - Ignore attribute "Store Location"
    - 10% missing values (lat/long coordinates of some stores)
    - Similar information already in "city" location
      - Unnecessary to be this precise geographically
  - Ignore attribute "Invoice/Item Number"
    - Specific to one transaction - no relationship to others
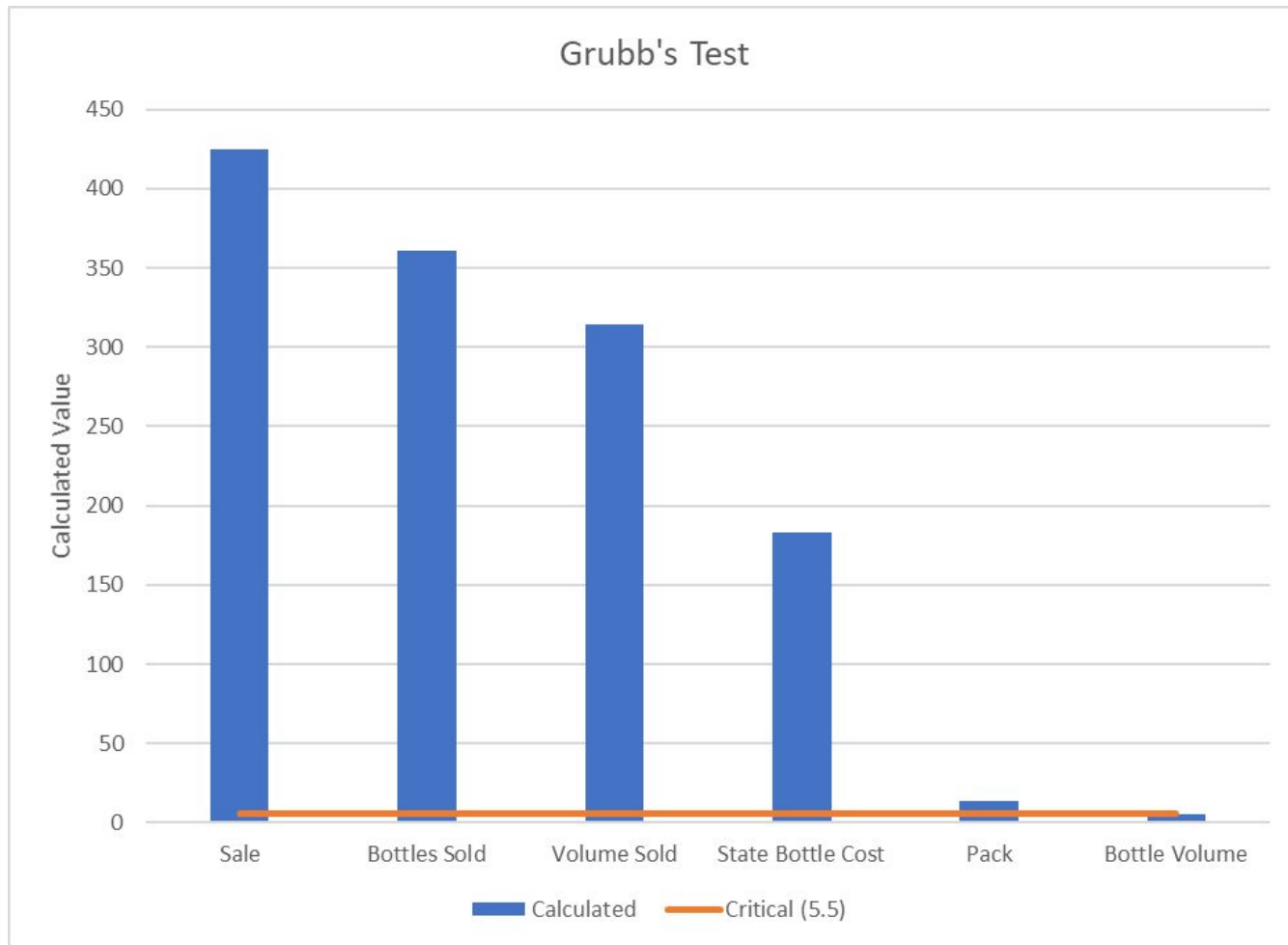- Data Cleaning:
  - Outlier analysis on numerical attributes
    - e.g. "Bottles Sold", "Volume Sold"
  - Resolving inconsistencies
    - Varying capitalization

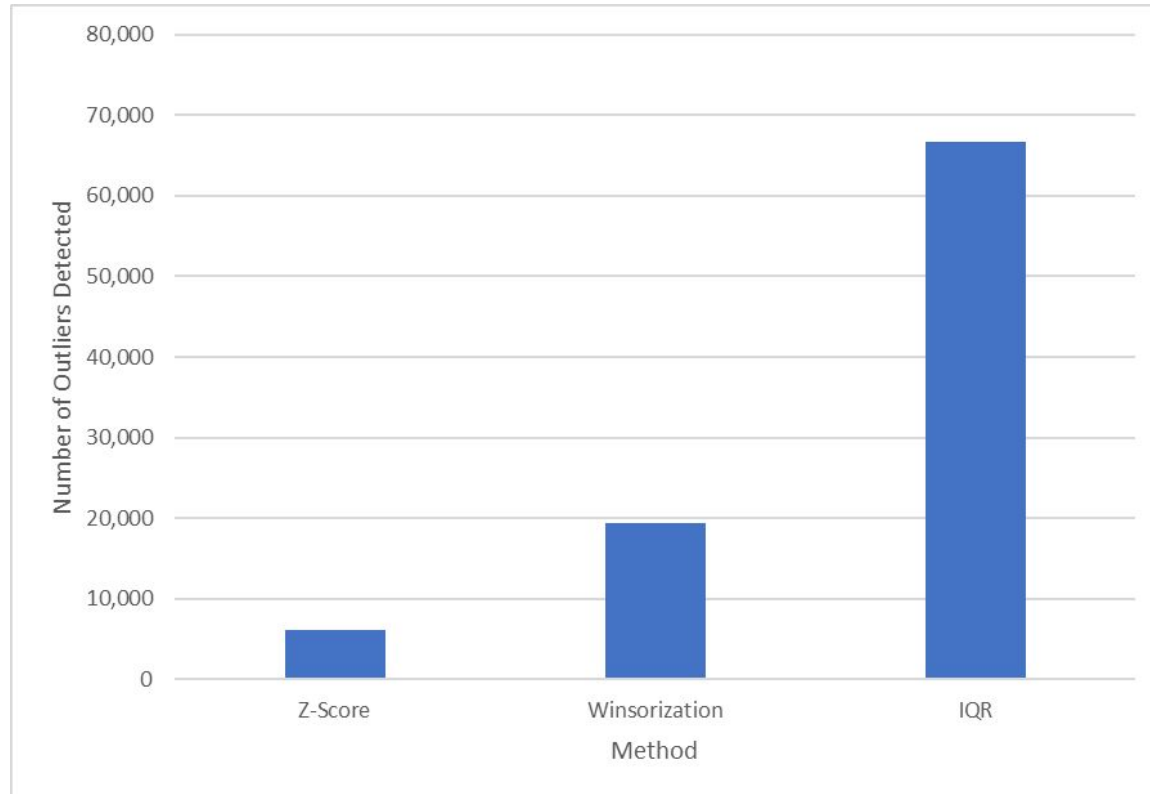| Store Lo... |
| --- |
| POINT (-91.4... |
| POINT (-93.8... |
| POINT (-91.7... |
| POINT (-93.6... |
| POINT (-93.3... |
| POINT (-96.3... |
| |

Polk
POLK

Outlier Analysis - Grubb's Test

Outlier Analysis - Methods

Outlier Analysis - Results
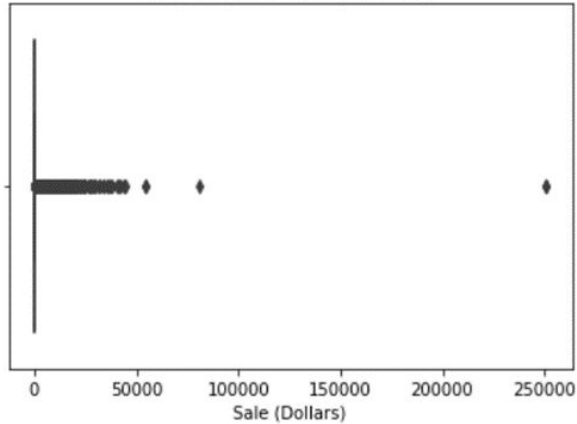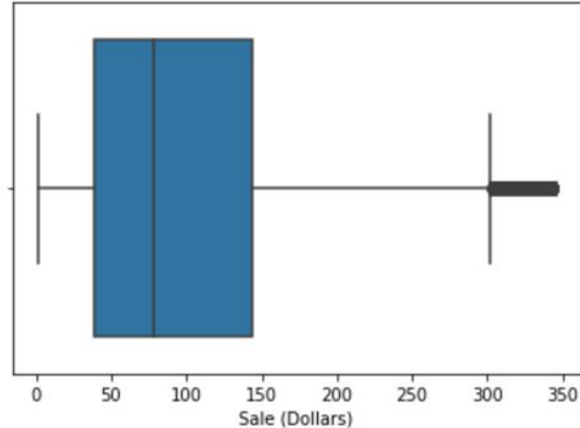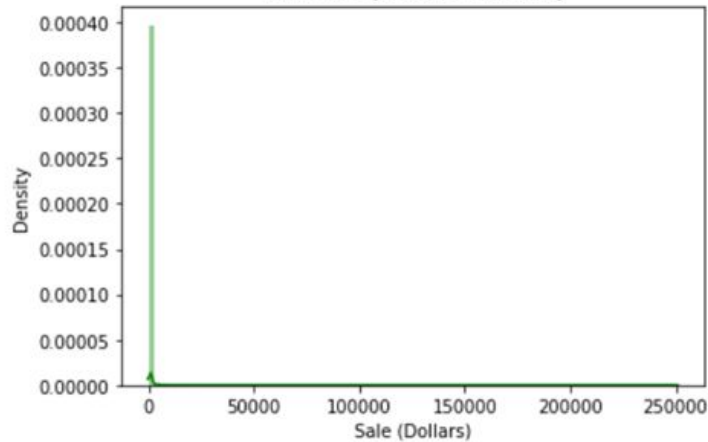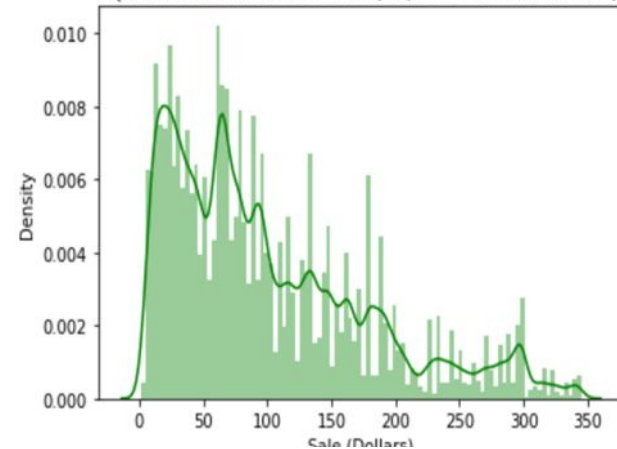
Trends: Counties With Highest Sales
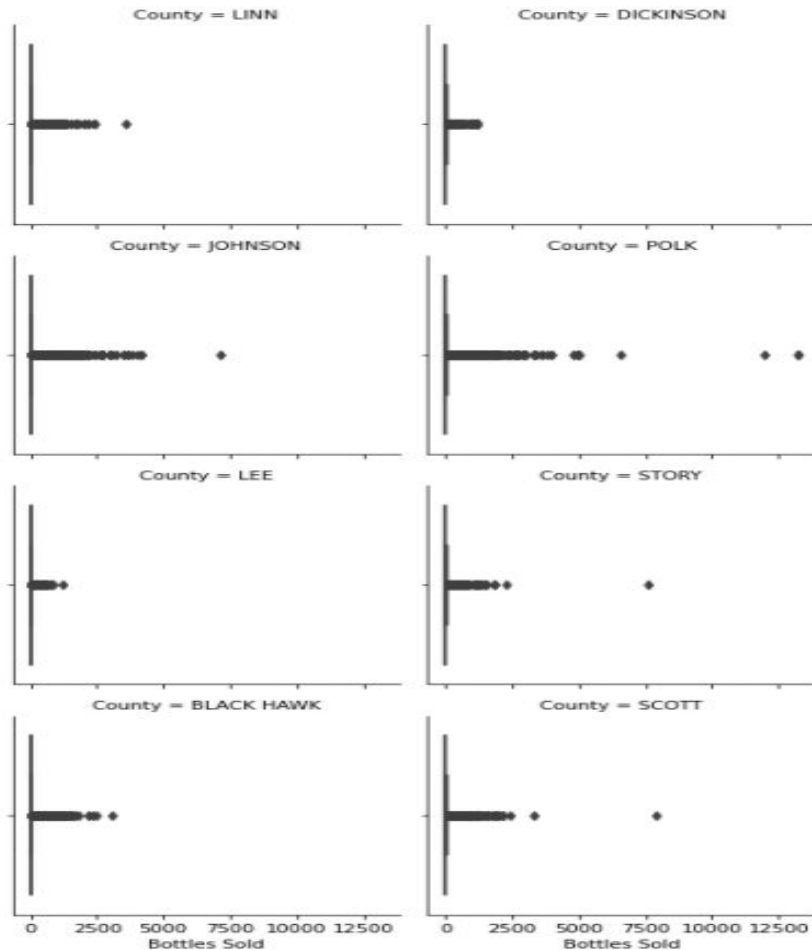
Counties:
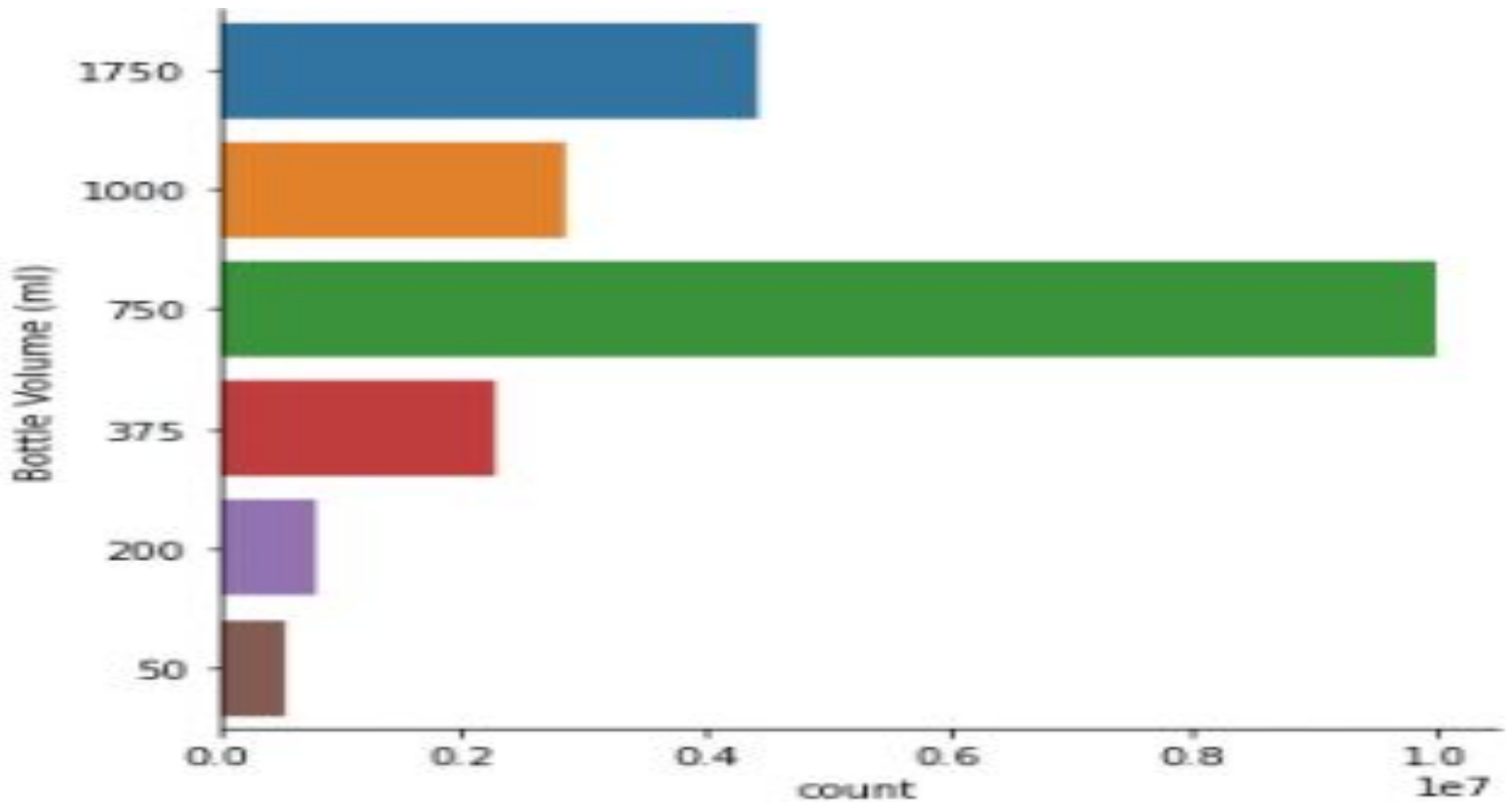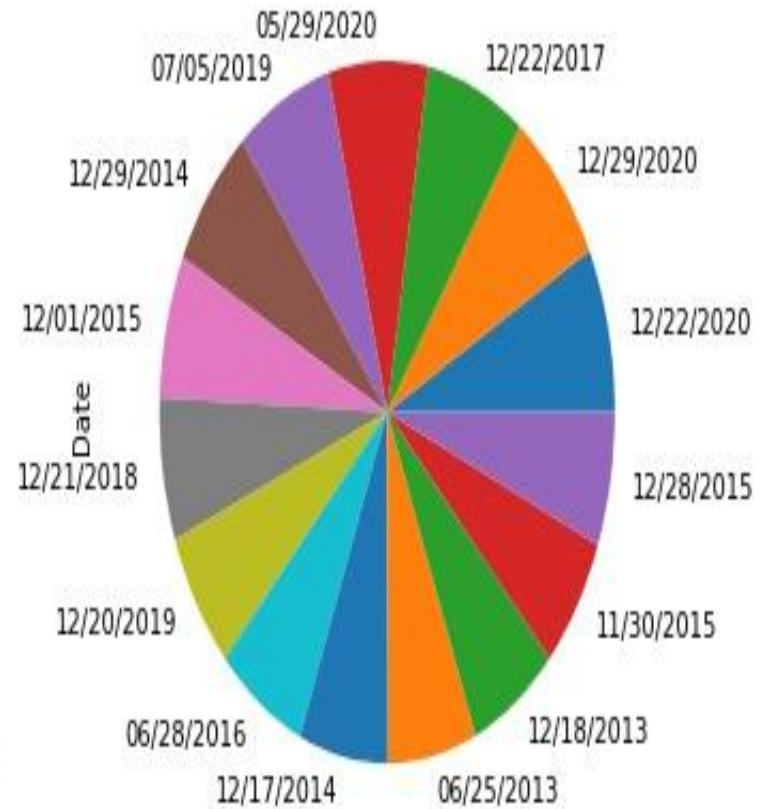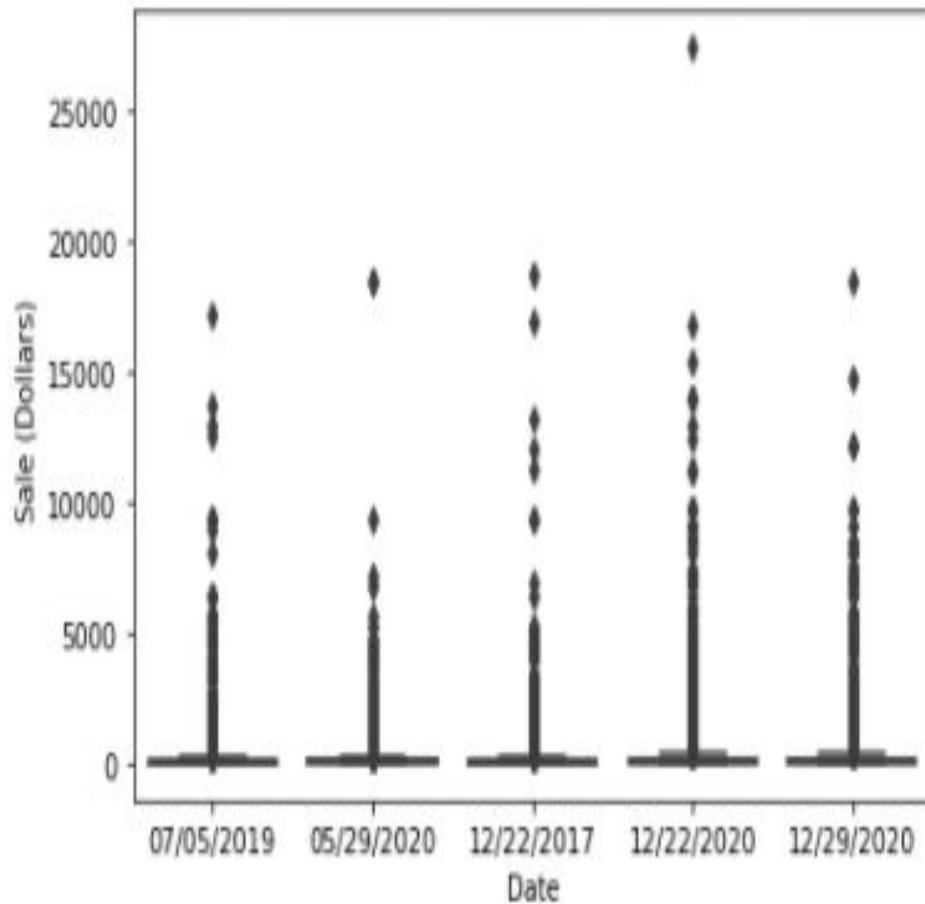 -POLK
  -JOHNSON
  -LINN
  -SCOTT
  -DICKINSON
  -BLACK HAWK
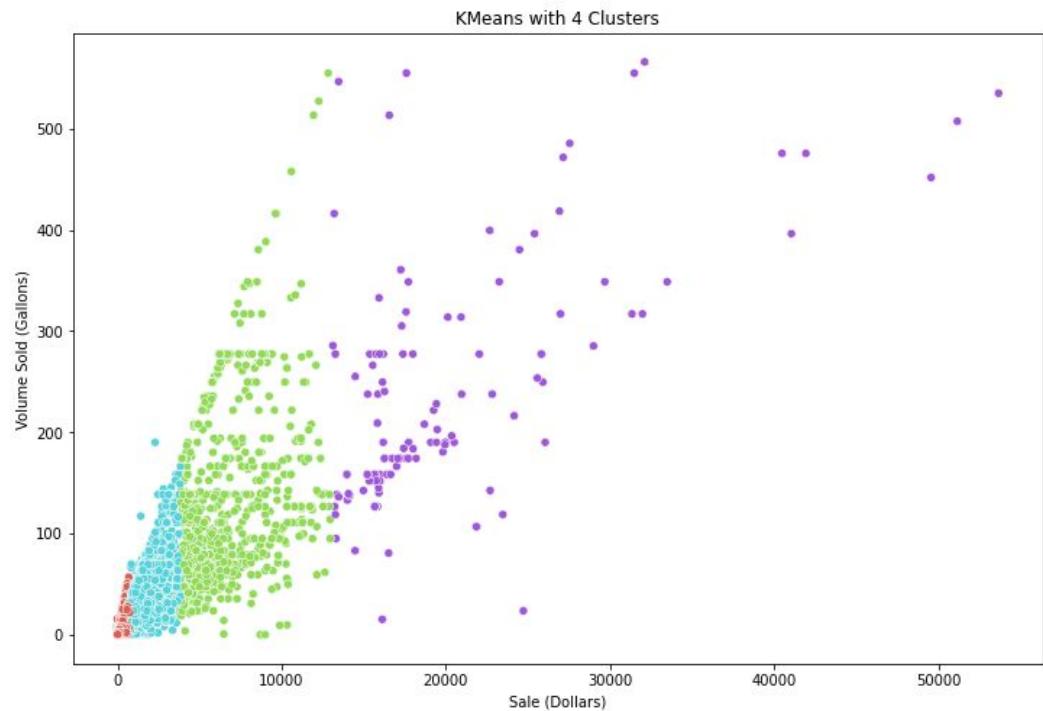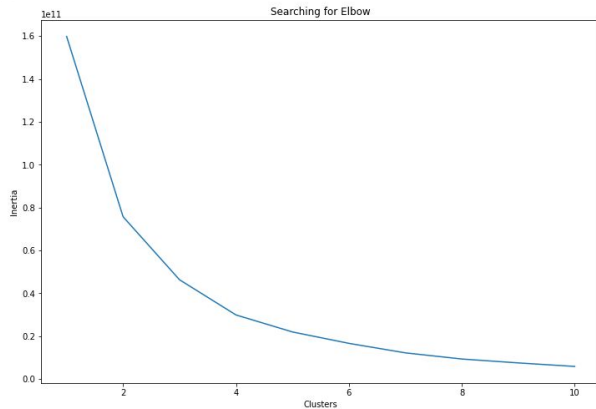 -STORY
  -LEE

Trends: Bottle Volume Sold

Trends: Top Sale Dates

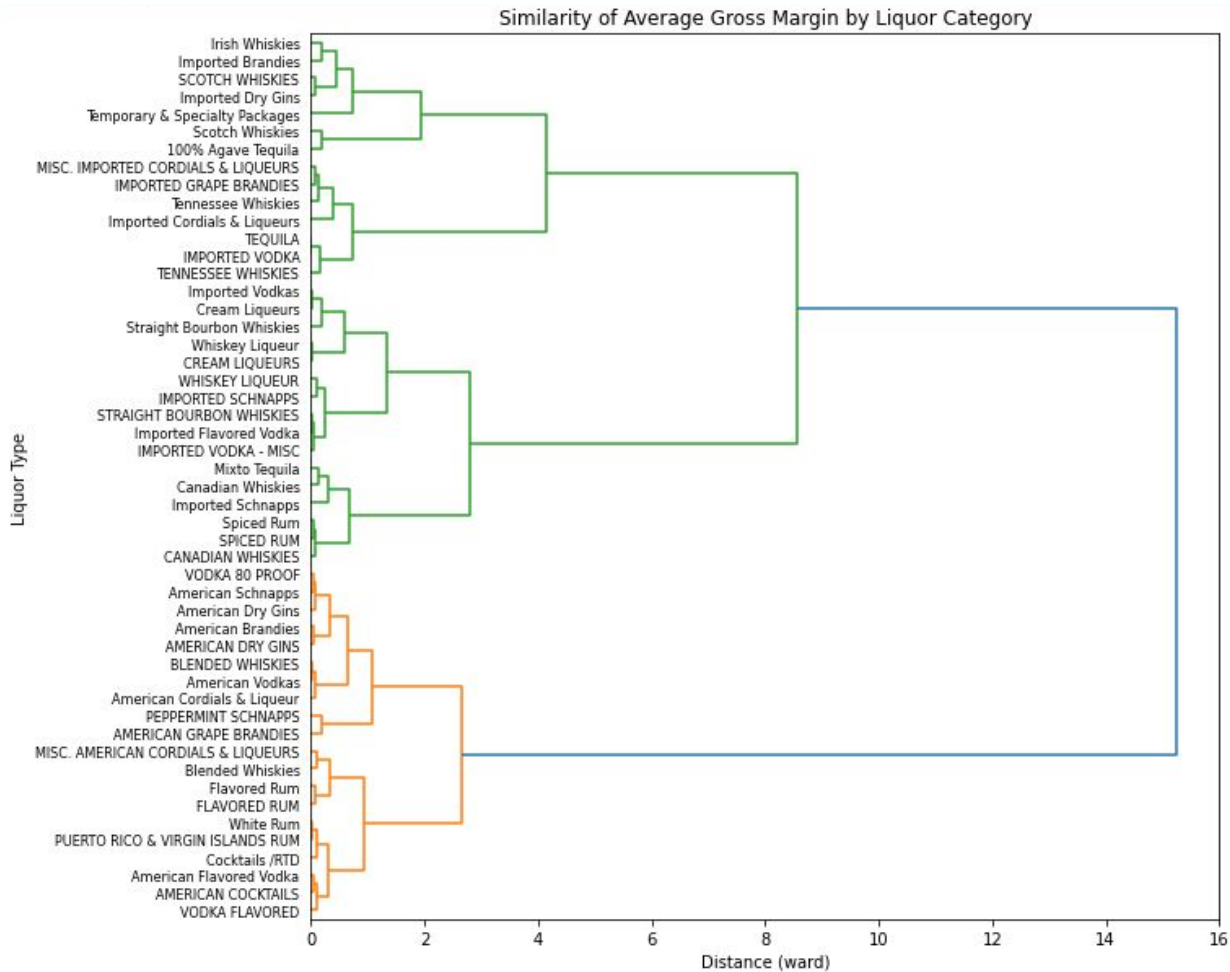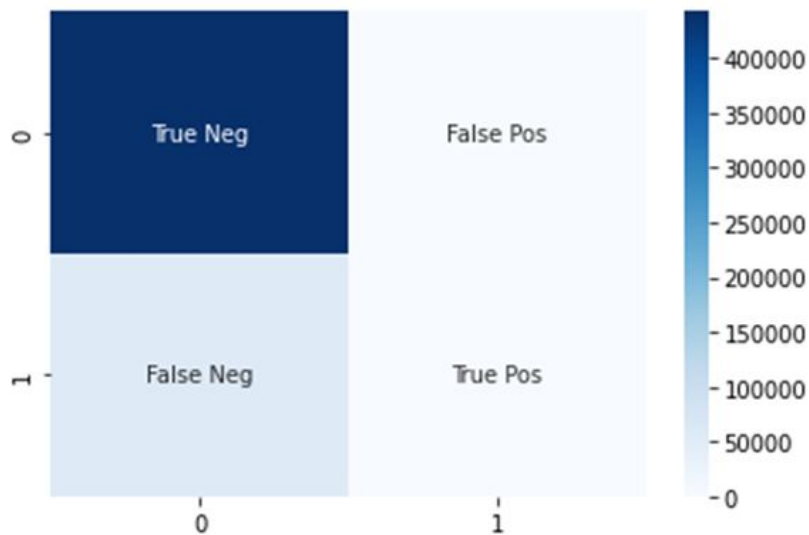# Clustering: Price and Volume

Clustering: Gross Margin

Similarity of Average Gross Margin by Liquor Category

Naive Bayes Classifier

Confusion Matrix	
442594	0
56626	0

	Precision	Recall	F1-score	Support
FALSE	0.89	1.00	0.94	442594
TRUE	0.00	0.00	0.00	56626

	Precision	Recall	F1-score	Support
Macro Avg	0.44	0.50	0.47	499220
Weighted Avg	0.79	0.89	0.83	499220

K-Nearest Neighbor Classifier
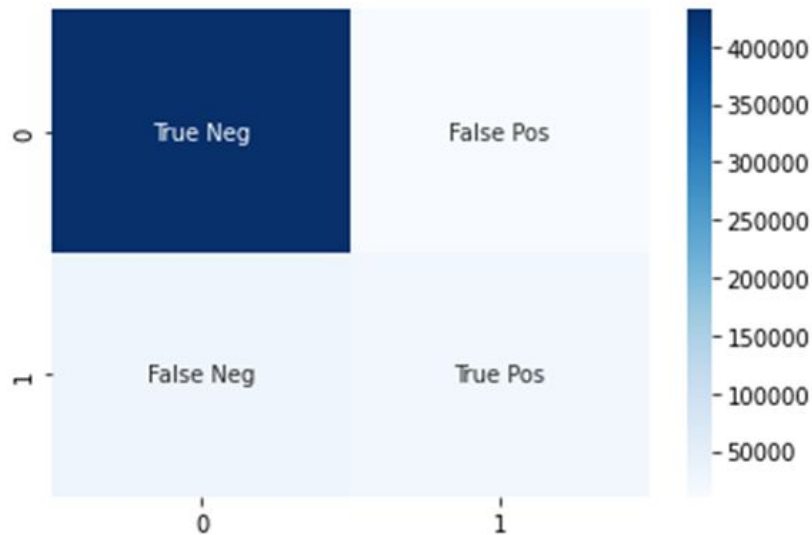
Confusion Matrix	
431812	10974
31839	24595

	Precision	Recall	F1-score	Support
FALSE	0.93	0.98	0.95	442786
TRUE	0.69	0.44	0.53	56434

	Precision	Recall	F1-score	Support
Macro Avg	0.81	0.71	0.74	499220
Weighted Avg	0.90	0.91	0.91	499220