**Jie Cui, Gokulan Thedchana, Xuemei Xie**

Iowa Liquor Sales Analysis

DATA SCIENCE – 3253 (TERM PROJECT)

## **Table of Contents**

1. **OBJECTIVE**
2. **DATA SOURCES**
3. **DATA PREPARATION**
4. **SALES ANALYSIS**
5. **LSTM ANALYSIS & FORECASTING**
6. **CONCLUSIONS**
7. **CHALLENGES**
8. **APPENDIX**

## **Objective**

Our objective in this report is to analyze liquor sales across Iowa in the US for predicting spikes and dips in sales throughout the years to encourage proactive preparation of inventory, promotions, scheduling. We also seek to predict any general trends in sales to help retailers understand which focus areas would be most impactful. We also seek to evaluate markets on the statewide, county, vendor basis based on sales and gross profit to identify potentially lucrative markets and sales volatility

## **Data Sources**

**Source Data** – Our sourced data is a large dataset containing information on the name, kind, price, quantity and location of sale for alcoholic beverages in the state of Iowa in the United States. All stores included must hold a class “E” liquor license and all sales are registered with the IOWA Department of Commerce in the Commerce Department System.

Link: <https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy>

## **Data Preparation**

* **Data Dilemma and Solution**

1. Dilemma – The Data set is too large for analysis on most devices. For ease of use we will need to reduce the size of the dataset or partition it in a way for relatively quick processing
2. Solution

* All irrelevant dimensions are removed, i.e. All descriptive data and effectively duplicate columns as well. A full list is provided below
* Loading in original data piece by piece
* **Data Preparation**

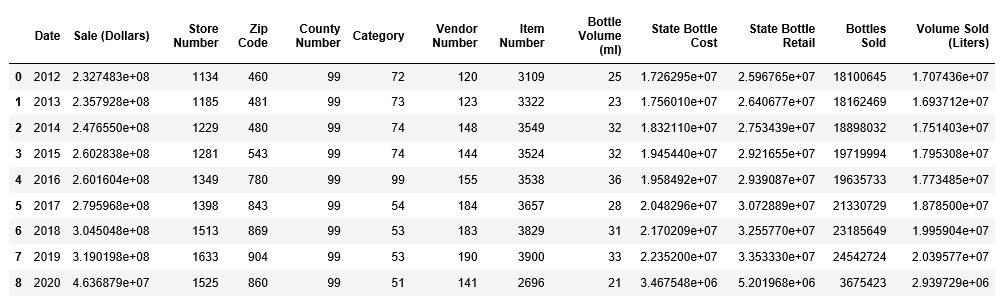
1. Removal of all descriptive and duplicate fields
2. Conversion of the ‘date’ field from a String datatype to a DateTime datatype
3. Conversion of all currency fields into floats with the understanding that all currency fields use USD
4. Partition the dataset by year
5. Separation of store location into latitude and longitude attributes

* **Removed fields:**

1. Invoice/Item Number
2. Store Name
3. Address
4. City
5. Store Location
6. Country
7. Vendor Name
8. Volume Sold (Gallons)

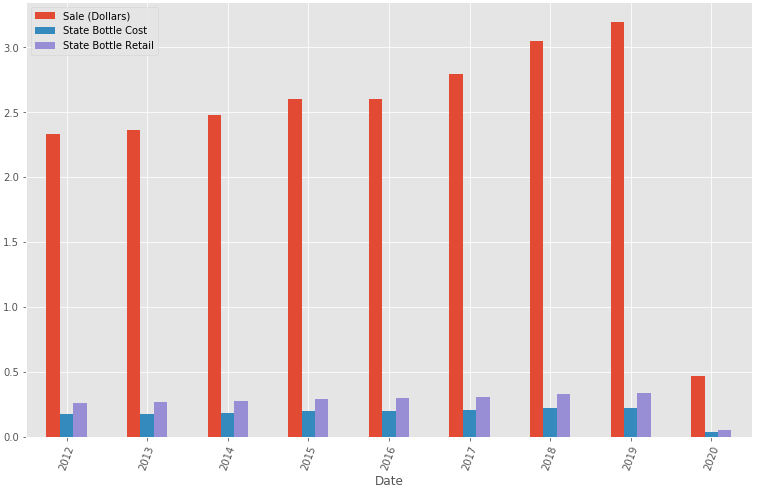
* **Generated the referenced tables –** all descriptive data that is removed is placed into these tables for use when compiling reports
* **First glance of the data**

1. Yearly information, including sales (dollar, bottle, volume), clear counts (stores, zip code, counties, categories, vendors, items)

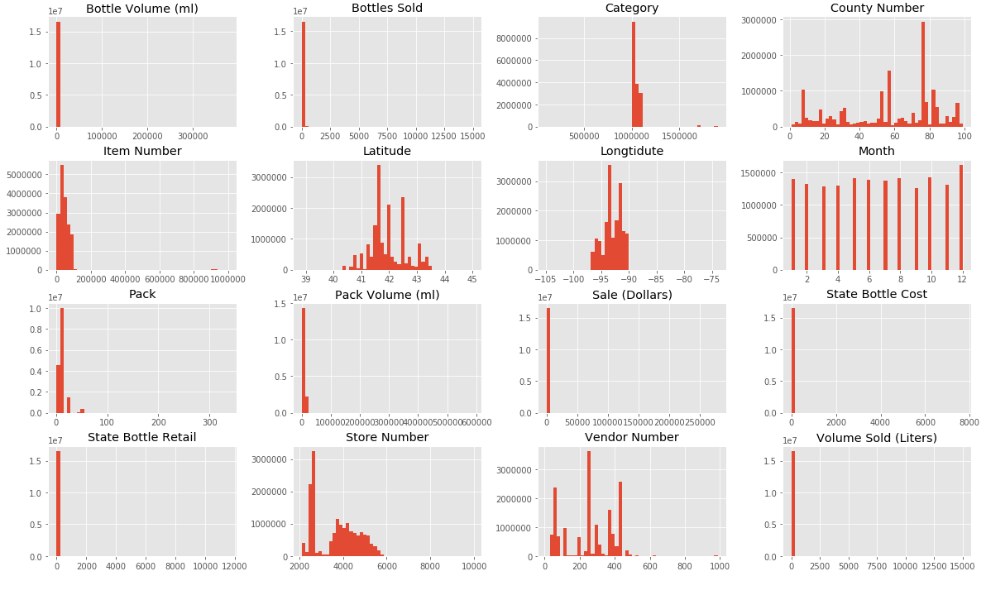


* Stores gradually increased yearly. In the last 2 years (2018, 2019), it jumped over 100 stores each year
* Total 99 counties contributed the data
* Total Zip code areas had a huge jump from 2015 to 2016, and them increased smoothly each year after
* Liquor categories had a significant difference in 2016, jumped a lot compared with prior years and decreased a lot compared with following years. could see what type of Liquors is popular
* Bottle Volume had the max 36 types, interested to see what size is the most popular
* Cost & Retail price is increasing yearly, as well as the yearly sales (graph below)

1. Yearly Sale (Dollars) vs. State Bottle Cost vs. Sate Bottle Retail



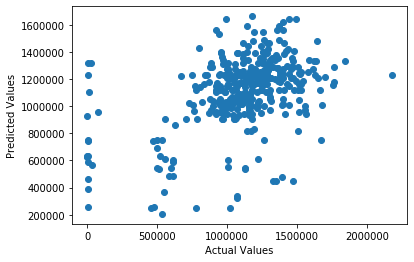
1. Histogram view
   * there is 1 liquor category has the highest frequency
   * county, store, latitude, longitude have related each other, and there is 1 store number, related county we believe, with the significant highest frequency
   * Vendor has 3 with the relatively high frequency



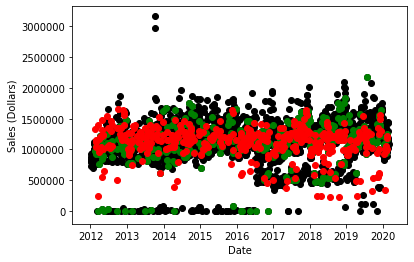
For reference, please see **IowaLiquorData\_preparation\_final.ipynb**.

## **Sales Analysis**

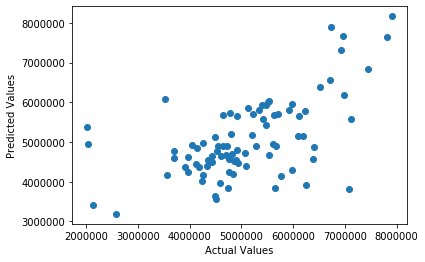
* The condensed data table was used to determine if sales can be predicted using a Random Forest Regression method with a training set consisting of 80% of the data. The dataframe for the daily (aggregate) statewide sales data was obtained as follows:
  1. Daily Aggregate Sales (Statewide): n estimators = 1000
* Prior to training the model, the following additional features were introduced:
  1. Week (of the year)
  2. Day (of the week)
  3. Month
* The Day and Month features were categorical features onto which one hot encoding was applied to convert the data into binary vectors
* Features were scaled using the preprocessing.scale() function
* Results
  1. Daily Aggregate Sales (Statewide)
     + RMSE = 314811.03
     + Actual Values vs Predicted Values:



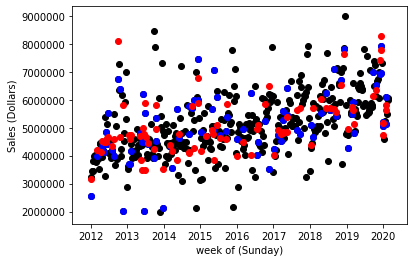
* + - Time series plot showing the original aggregate sales (black), test set (green), and predicted values (red)



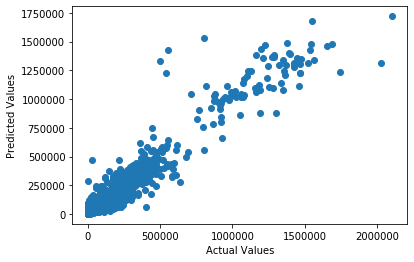
* Due to the high variation between day to day sales and considering weekly cycles, random forest regression was applied to weekly aggregate sales data, with each week starting on a Sunday.
* For scenarios broken down by county and/or vendor, one hot encoding was applied to the county and vendor IDs to convert the categorical data into binary vectors.
* Weekly aggregate data was not broken down by item due to computational limits (as the dataframes become too large).
* The number of estimators for each scenario were set as follows:
  1. Weekly Aggregate Sales (Statewide): n\_estimators = 100
  2. Weekly Aggregate Sales by County: n\_estimators = 100
  3. Weekly Aggregate Sales by County and Vendor: n\_estimators = 10
  4. Weekly Aggregate Sales by Vendor: n\_estimators = 100
* The results of the of applying the random forest regression technique are as follows:
  1. Weekly Aggregate Sales (Statewide): n\_estimators = 100
     + RMSE = 985491.59
     + Actual Values vs Predicted Values Plot:



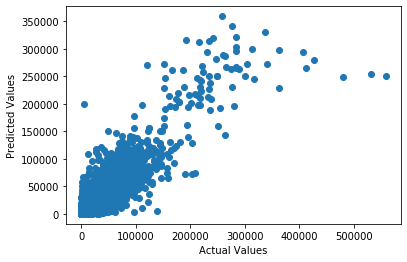
* + - Time series plot showing the original aggregate sales (black), test set (blue), and predicted values (red):

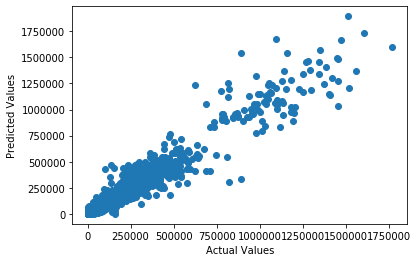


* 1. Weekly Aggregate Sales by County
     + RMSE = 34977.24
     + Actual Values vs Predicted Values:



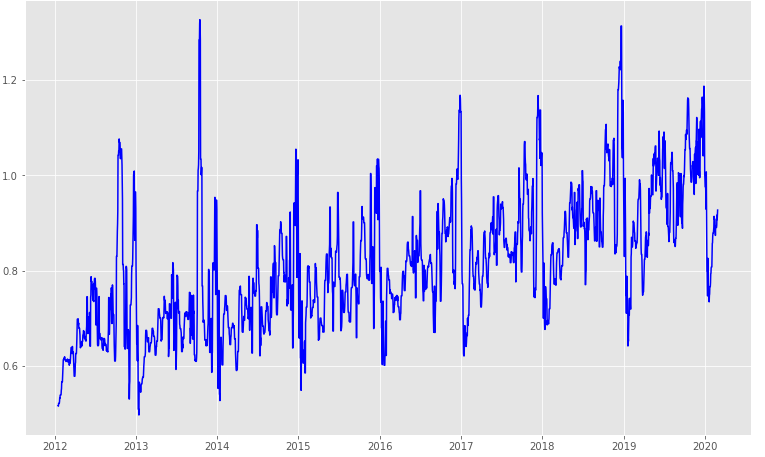
* 1. Weekly Aggregate Sales by County and Vendor
     + RMSE = 3344.59
     + Actual Values vs Predicted Values:



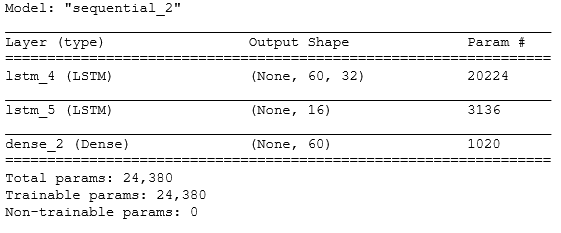
* 1. Weekly Aggregate Sales by Vendor
     + RMSE = 36989.16
     + Actual Values vs Predicted Values:
     + 

## **LSTM Analysis & Forecasting**

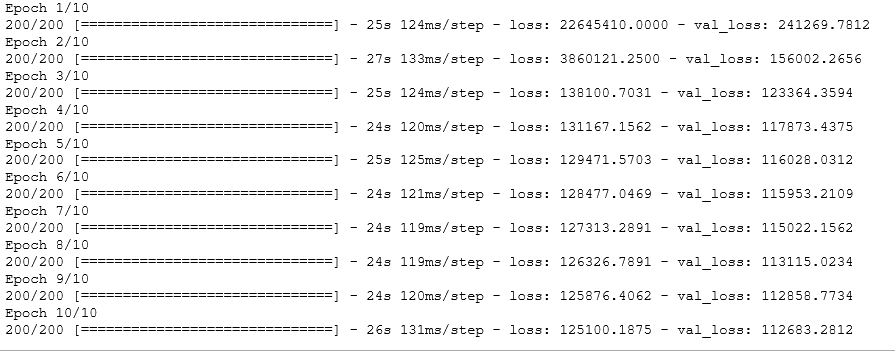
* **LSTM** (Long Short-Term memory) model will detect both long-term and short-term dependencies in the data, while other models more consider short term and more likely to lose some long-term memory or information. Base on the inspection of the data, we might consider a little bit more of long-term dependencies.
* **Converting counties and categories name attributes into binary vectors by using onehot encoder transformation**
* Using resample function to generate timeseries data
* Applying 14- days rolling total sale data to fit the data model, after checking daily, weekly data check. The rolling data will be helpful to smooth the curve.



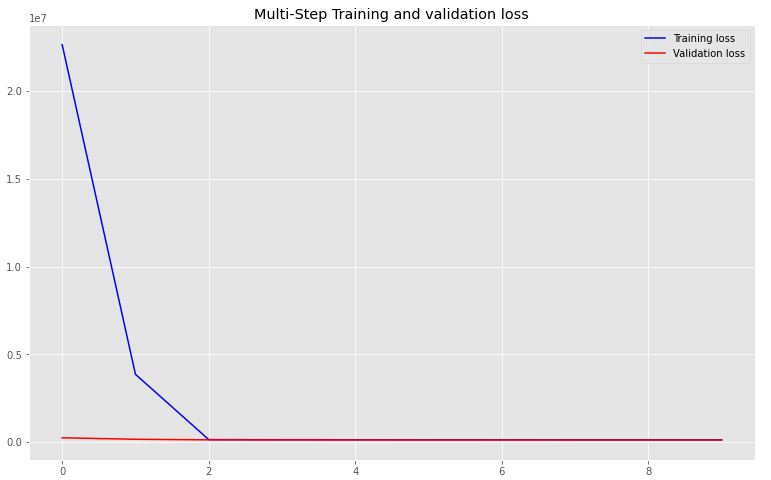
* Standardizing data before applying them into the model
* Using ‘Relu’ as the activation function
* Compiling the model with RMSprop optimizer along with ‘Mean Absolute Error’ loss function
* The result after setup LSTM model



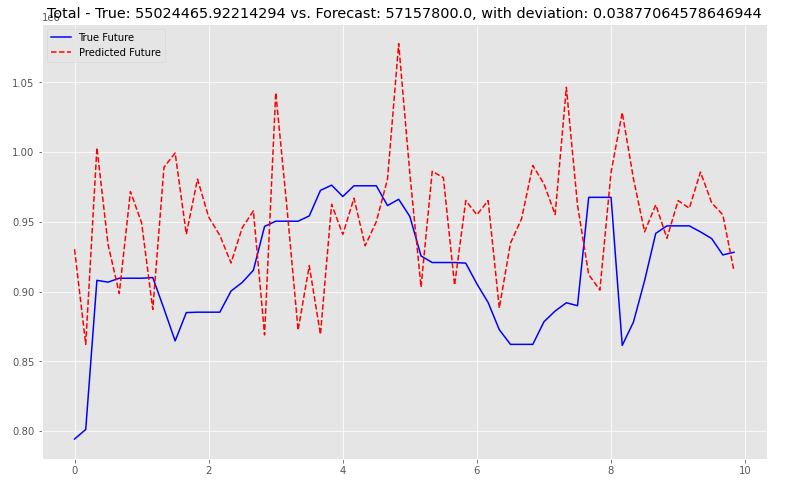
* Fitting the model with the standardized training data

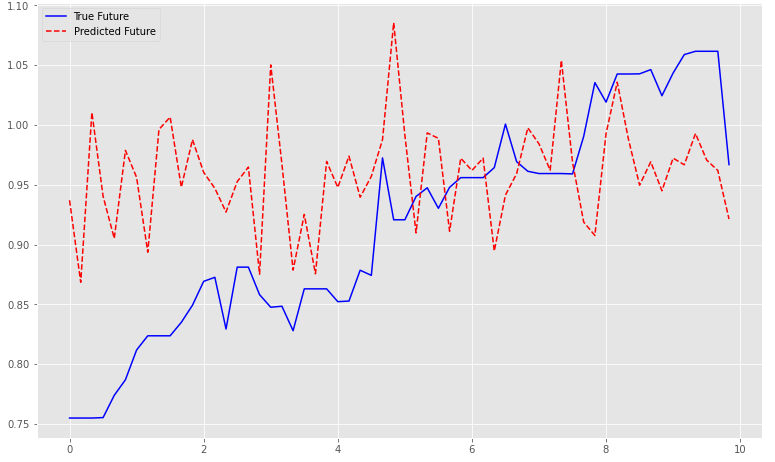


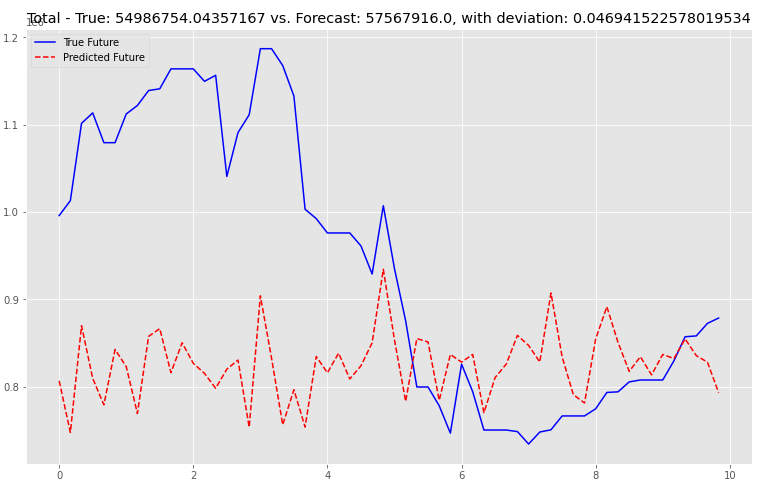
* From the training history result below shows, the gap of Training loss and Validation loss is getting smaller and finally almost converge, which means the performance on validation dataset is reasonable.

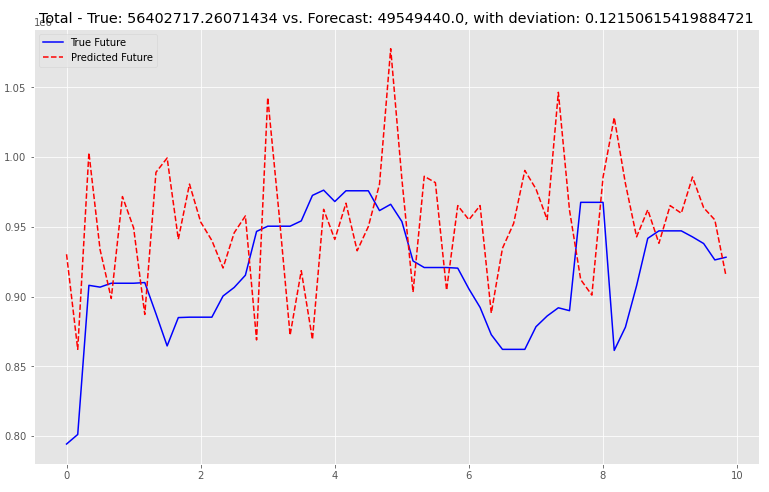


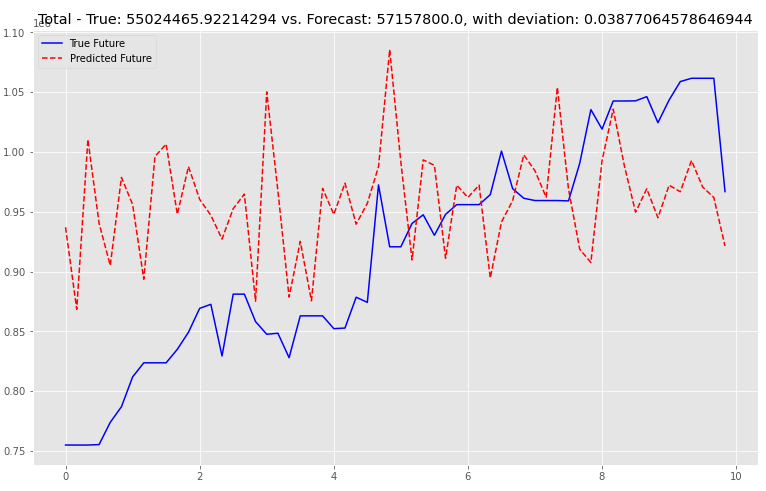
* Applying test data with trained model, for sampled 10 forecasting, the result as below:
* accuracy mean: 0.9339574386523521 (i.e. inaccuracy mean: 0.06604256134764781)
* Deviation: 0.03646670941507
* Some results











## **Conclusion**

* For statewide daily and weekly sales, the predicted values overlap the general region covered by actual test values reasonably well. However, the Random Forest Regression model does not perform as well as it does when county and/or vendor is considered. The Actual Values vs Predicted Values plots appear to show a correlation between the predicted data and the actual data.
* LSTM model works well regarding time series forecasting 60 days sales. However, it tends to be worse if forecasting longer future steps.

## **Challenges with Data/Risks/Other Considerations**

* More focus on sales dollars, could applying model into cost and retail price to make profit analysis
* Only onehot encoder applied to convert category attributes, which results in many input features. It slowed down the performance. Like textbook mentioned, could find alternative meaningful number, like county could be represented by population, etc.

## 

## **Appendix**

* **Python files:**
* IowaLiquorData\_preparation\_final.ipynb
* IowaLiquorData\_Model\_daily\_final.ipynb
* IowaLiquorData\_Model\_weekly\_final.ipynb
* IowaLiquorData\_Model\_LSTM\_final.ipynb
* **PPTx file:**
* IowaLiquorSaleAnalysis.pptx