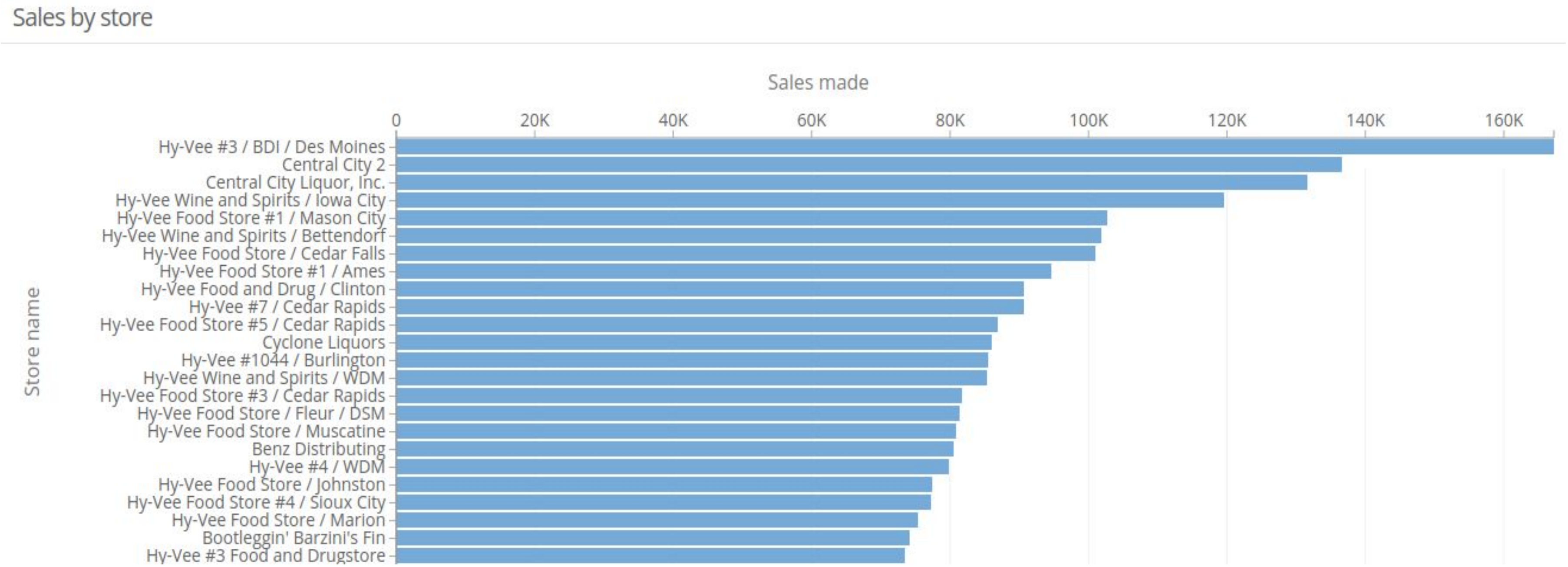
Capstone Project
Iowa Liquor Retail Sales

Applied Data Science by IBM/Coursera

Introduction - Business problem

- Predict future liquor retail store sales (stockout prediction) from data on
  - Past sales
  - Weather
- Help store owners make smarter decision on restocking
- Using public data
  - Available from the Alcoholic Beverage Division (Commerce) of Iowa
  - Also a part of Google's Public Datasets Program

Introduction - Business problem



Stores sorted by total sales reported in 2012-2020

Data

- Each row in the data set is a wholesome liquor sale made by a store in Iowa since 2012
    - Total data set includes 19.4M rows with 24 columns each
- In this project, we assume the role of store owner
    - The store: Hy-Vee #3 / BDI / Des Moines
    - Has made more than 160K sales from 2012 to 2020
- Liquor data set is joined with weather data from NOAA's Global Historical Climatology Network (GHCN)

Data wrangling & preprocessing

1. Quite a few steps needed to be taken to prepare the data for modeling
    1.1. Filter the liquor data set only for the store’s sales
    1.2. Find the weather station closest to the store that is still operating
        1.2.1. Station code: USW00014933 @ 5.3km
    1.3. Query the weather data to include only weather elements of interest from that station
        1.3.1. Precipitation, minimum and maximum temperature and snowfall
    1.4. Cross join the liquor sales data and the weather data by timestamp
    1.5. Compute lag values for the sales and the weather
        1.5.1. To aid our predictive performance, we use the sales from 1, 2, 3 and 12 months ago relative to the period for which we are predicting
        1.5.2. Similarly, we take note of the weather conditions from 1 and 12 months ago

Predicting sales

- We will build regression models on our data set to predict sales for 2020
  - Methods: linear regression, Ridge regression and Lasso regression
  - Use R-squared value as evaluation measure
- For training, we will use the data up to December 2019
  - Because our lag values require data up to a year in the past, the sales for 2012 will be dropped because they do not have relevant data for it
- We would expect the results to start high for the first month of 2020 and then to continue decreasing
  - This is because the sales are highly correlated with their immediate close values

Results - R^2

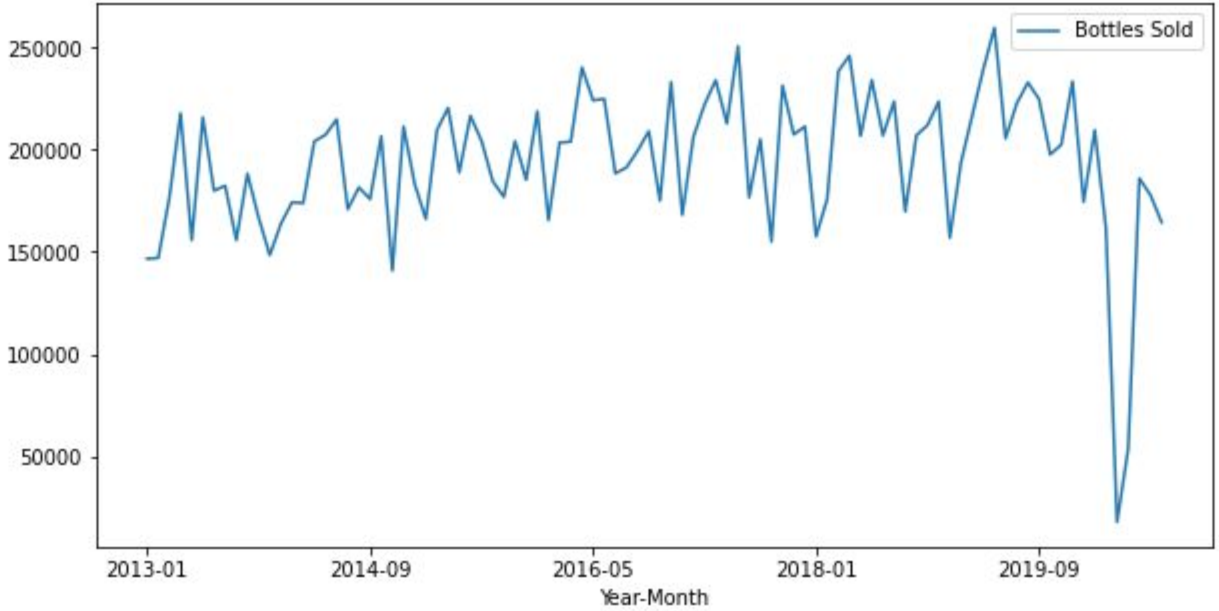
Predicting for month	Linear regression	Ridge regression	Lasso regression
2020-01	0.8731	0.8737	0.8731
2020-02	0.8461	0.8462	0.8461
2020-03	0.7181	0.7180	0.7181
2020-04	-0.6544	-0.6518	-0.6544
2020-05	-1.3093	-1.3200	-1.3093
2020-06	0.6268	0.6269	0.6268
2020-07	0.7392	0.7392	0.7392
2020-08	0.7848	0.7848	0.7848

R-squared is the proportion of the variance in the sales that is predictable by our data set.
We want this number to be as close to 1 as possible, and higher than 0 in any case.
The negative values for April and May indicate that these sales can not be predicted at all.

Discussion

- All three methods generate models of similar performance - we would prefer Linear regression for its simplicity
- Results are expected
    - Despite the very sharp and sudden drop for April and May 2020 which is easily explainable by the COVID-19 pandemic
- This kind of work can be used by many retail stores worldwide
    - I hope to be able to present this to some local merchants and see it applied

Sales plot



This plot shows the total sales by month from January 2013 up to September 2020 **for all stores** in the data set. On the far right we can clearly see the steep decrease in sales for April and May 2020 which is due to the COVID-19 pandemic. The liquor retail stores in Iowa have for sure suffered great financial losses during this period, with total liquor sales going down for more than 90%!