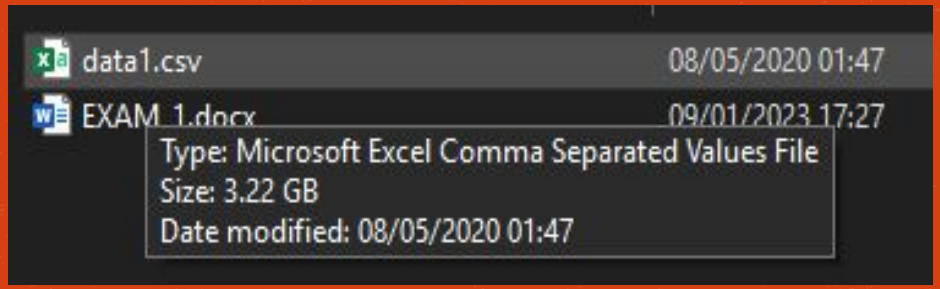
Iowa Liquor Sales

Țărnă Cristina

First thought




```
[ ]  # let's see the size of the dataset
     df.shape

     (12591077, 24)
```

Cleaning process

let's see if actualy is a lot, or meaning less for the model.

```
[ ]  df['County'].isnull().sum()*100/df.shape[0]
```

```
0.6288421554407141
```

As we can see there is not much data missing, under a percent.

There are records where the Sale, Cost and Retail are 0. So I need to get rid of this errors.

```
[ ]  # Total numbers of ambiguous data.
     df['Sale (Dollars)'].isin([0]).sum()
```

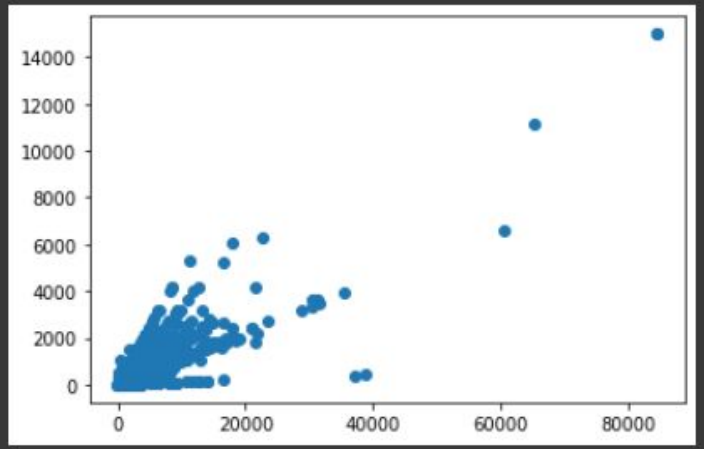
```
3491
```

```
# Getting rid of $ in each value of column and change the type to floats
df['State Bottle Cost'] = df['State Bottle Cost'].str.replace('$', '', regex=True).astype('float')
df['State Bottle Retail'] = df['State Bottle Retail'].str.replace('$', '', regex=True).astype('float')
df['Sale (Dollars)'] = df['Sale (Dollars)'].str.replace('$', '', regex=True).astype('float')
df.info()
```

Outliers

Let's the outliers

```
[ ]  plt.scatter(df['Profit Margin'], df['Volume Sold (Liters)'])
     plt.show()
```



Kydavra comes to save us

I will try an algorithm from kydavra library.

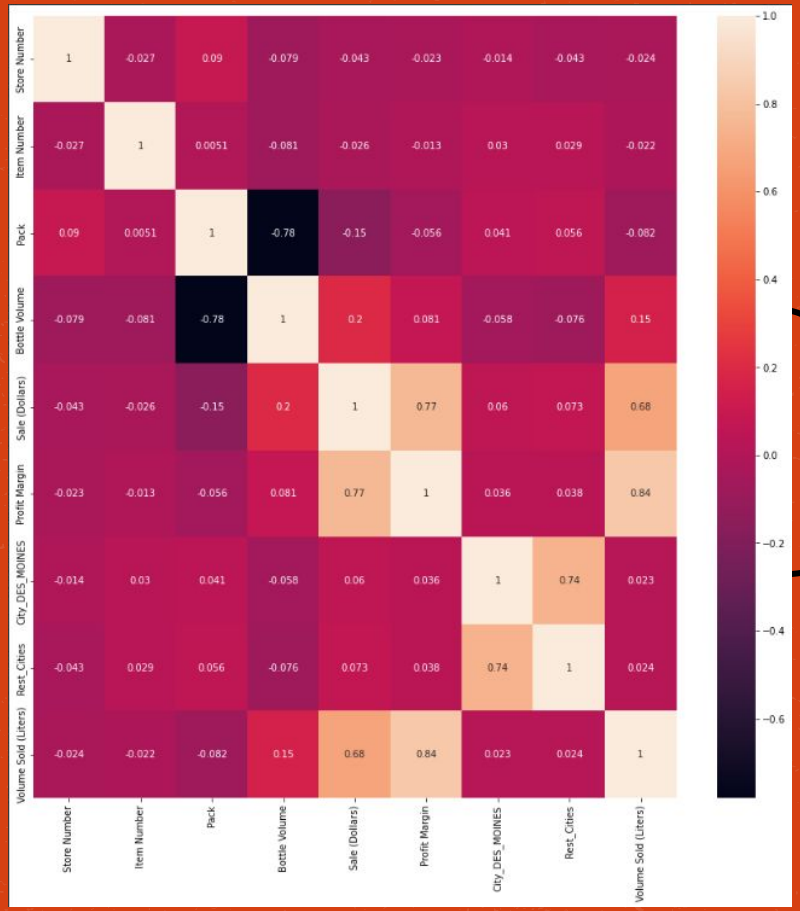
```
[14] pearson = PearsonCorrelationSelector(min_corr=0.02, max_corr=0.9)
```

```
[15] selected_columns2 = pearson.select(df, 'Volume Sold (Liters)')
```

```
[16] selected_columns2
```

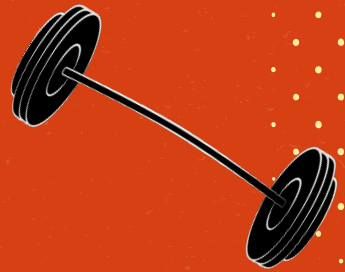
```
['Store Number',
 'Item Number',
 'Pack',
 'Bottle Volume',
 'Sale (Dollars)',
 'Profit Margin',
 'City_DES_MOINES',
 'Rest_Cities']
```

Correlations



igmoid

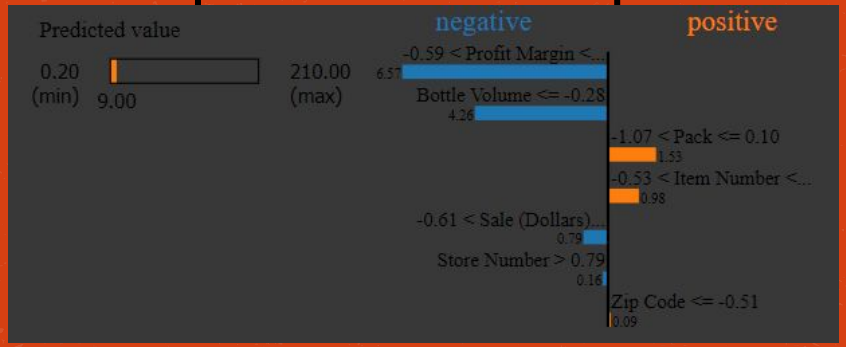
Training time



	ML model	Score	MSE	MAE	RMSE
0	LinearRegression()	0.720062	228.583409	3.849922	15.118975
1	KNeighborsRegressor()	0.911344	72.391744	0.366536	8.508334
2	RandomForestRegressor()	0.986845	10.741954	0.063238	3.277492
3	DecisionTreeRegressor()	0.979420	16.804915	0.057573	4.099380
4	RandomForestRegressor() without selecting columns	0.990049	7.793316	0.061354	2.791651

Lime

Feature	Value
Profit Margin	-0.63
Bottle Volume	-1.32
Pack	1.75
Sale (Dollars)	-0.66
Item Number	-1.32
Zip Code	-0.49
Store Number	0.98



Feature	Value
Profit Margin	-0.19
Bottle Volume	-0.28
Pack	0.10
Item Number	-0.13
Sale (Dollars)	-0.21
Store Number	0.89
Zip Code	-0.67

Thanks

Questions?
Hope not XD