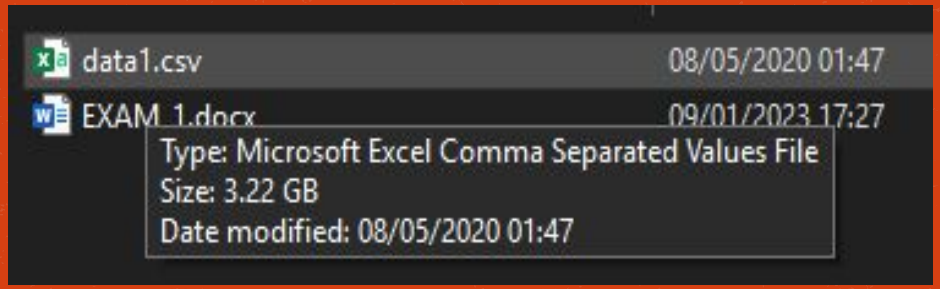
Iowa Liquor Sales

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()
```

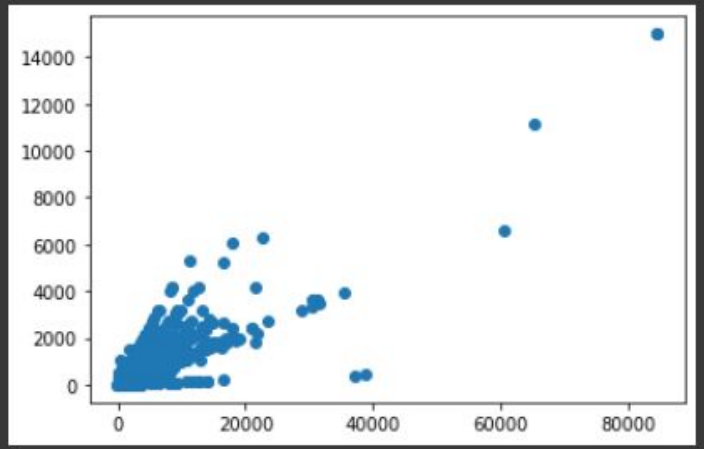
```
3491
```

```
# 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.

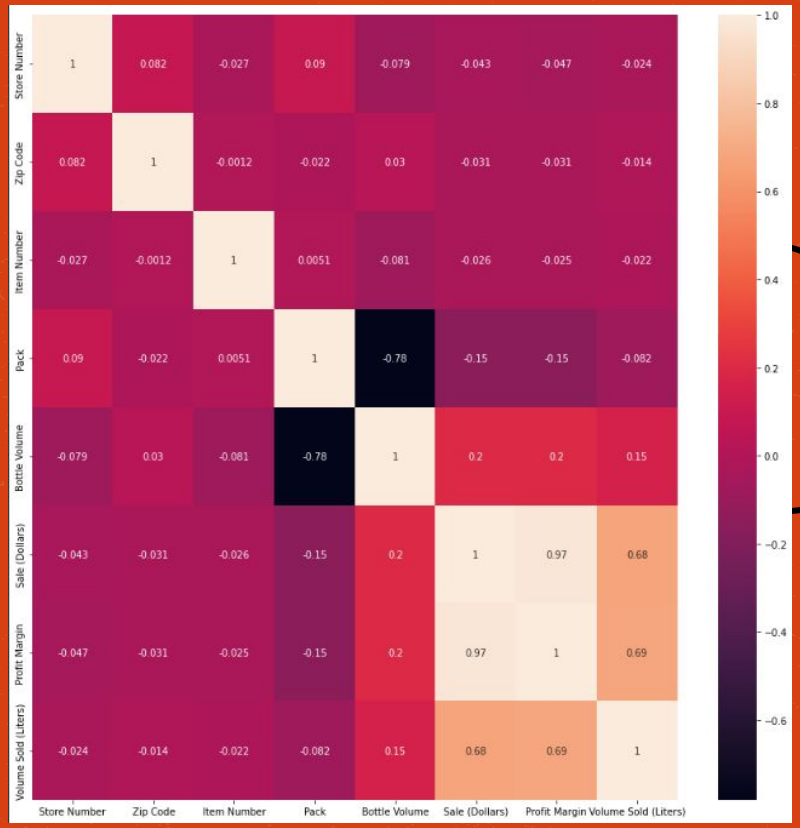
```
[ ] pearson = PearsonCorrelationSelector(min_corr=0.01, max_corr=0.9)
```

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

```
selected_columns2
```

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

Correlations



igmoid

Training time

	ML model	MSE	MAE	RMSE
0	LinearRegression()	409.773837	8.258306	20.242871
1	KNeighborsRegressor()	739.784596	6.480763	27.198982
2	RandomForestRegressor()	10.780307	0.071379	3.283338
3	DecisionTreeRegressor()	19.550262	0.066496	4.421568

Lime

Thanks

Questions?
Hope not XD