Project 2
Liquor Sales + Linear Regression

Cheng Ji

Contents

- The Data
- EDA
- Mine
- Refine
- Model

The Data

- 10% sample of transactions from 2015 to 2016 Q1 for all stores in Iowa that have a class E liquor license.
- Goal: predict sale based on location (County), time (season), average price, and bottles sold.

EDA & Mining

- Handle missing values: over 27 thousand obs, less than 2000 missing values. 0.6% missing values, I can live without them.

- Removing $ and convert Cost, Retail, and Sale to float type.

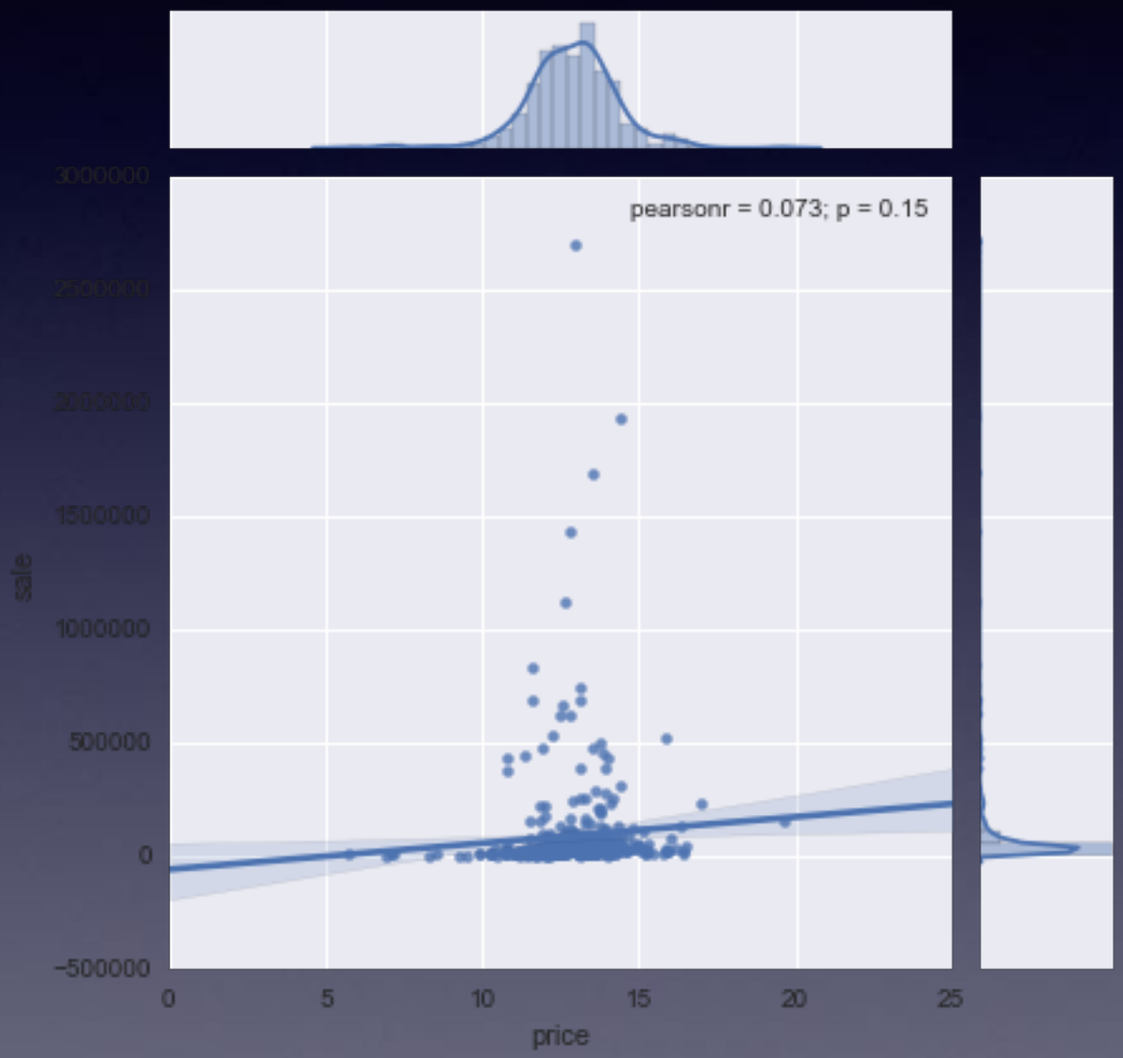
- Create a ‘season’ column taking values from 1 to 4 based on ‘Date’

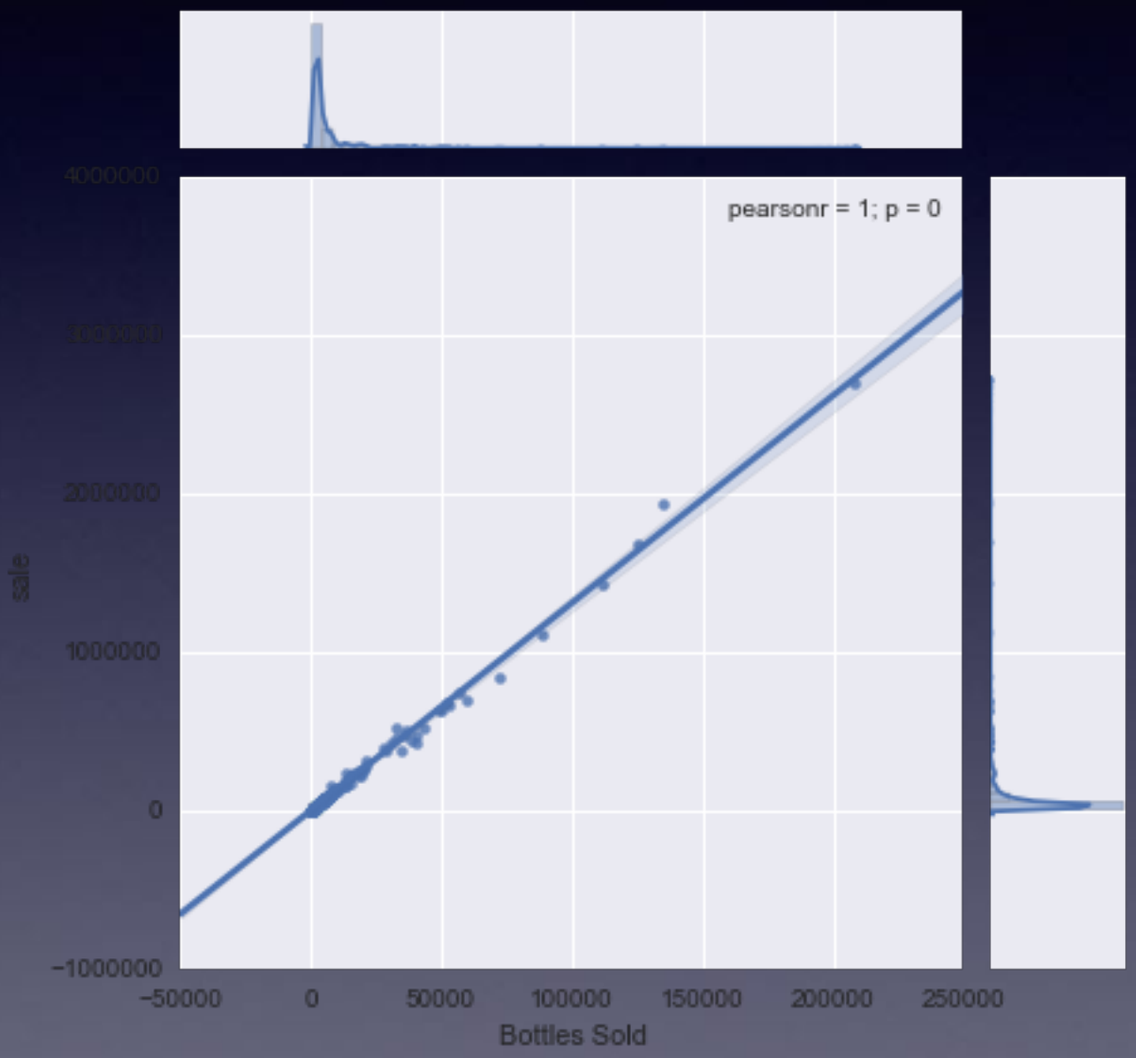
EDA & Mining

- Dropping irrelevant columns, and sum sale, bottles sold, cost, and average price based on location and season.

	County Number	season	sale	Volume Sold (Liters)	Bottles Sold	price
0	1.0	1	22019.02	1664.16	1945	11.320833
1	1.0	2	10820.66	842.73	894	12.103647
2	1.0	3	11974.03	961.33	921	13.001118
3	1.0	4	10730.49	891.62	890	12.056730
4	2.0	1	4450.17	342.37	375	11.867120

EDA & Mining





Drop outliers over 3std

Refine the data

- Perfect correlation between sale and bottles sold.
- Just like you are trying to predict income tax, and you have income as X. Problem Solved.

	County Number	season	sale	Volume Sold (Liters)	Bottles Sold	price
County Number	1.000000	0.007773	0.057974	0.052634	0.047641	0.070149
season	0.007773	1.000000	-0.050248	-0.063149	-0.051682	0.092015
sale	0.057974	-0.050248	1.000000	0.997190	0.995112	0.100977
Volume Sold (Liters)	0.052634	-0.063149	0.997190	1.000000	0.995004	0.094661
Bottles Sold	0.047641	-0.051682	0.995112	0.995004	1.000000	0.057417
price	0.070149	0.092015	0.100977	0.094661	0.057417	1.000000

Refine the data

- Convert categorical data (County Number, season)

```
categorical = preprocessing.OneHotEncoder(categorical_features = [0,1])
X = modeldata[['County Number', 'season', 'price', 'Bottles Sold']]
y = modeldata['sale']
X = categorical.fit_transform(X)
```

- Multicollinearity

Models

- X = county, season, price, bottles sold
- Train test split
- Use Lasso to drop the redundant dummy
- R^2 = 0.9979, rmse=6193 (mean=63848)

Models

- X without bottles sold
- R^2 = 0.9364, rmse=34220
- R^2 may not be a good measure for regularized model.