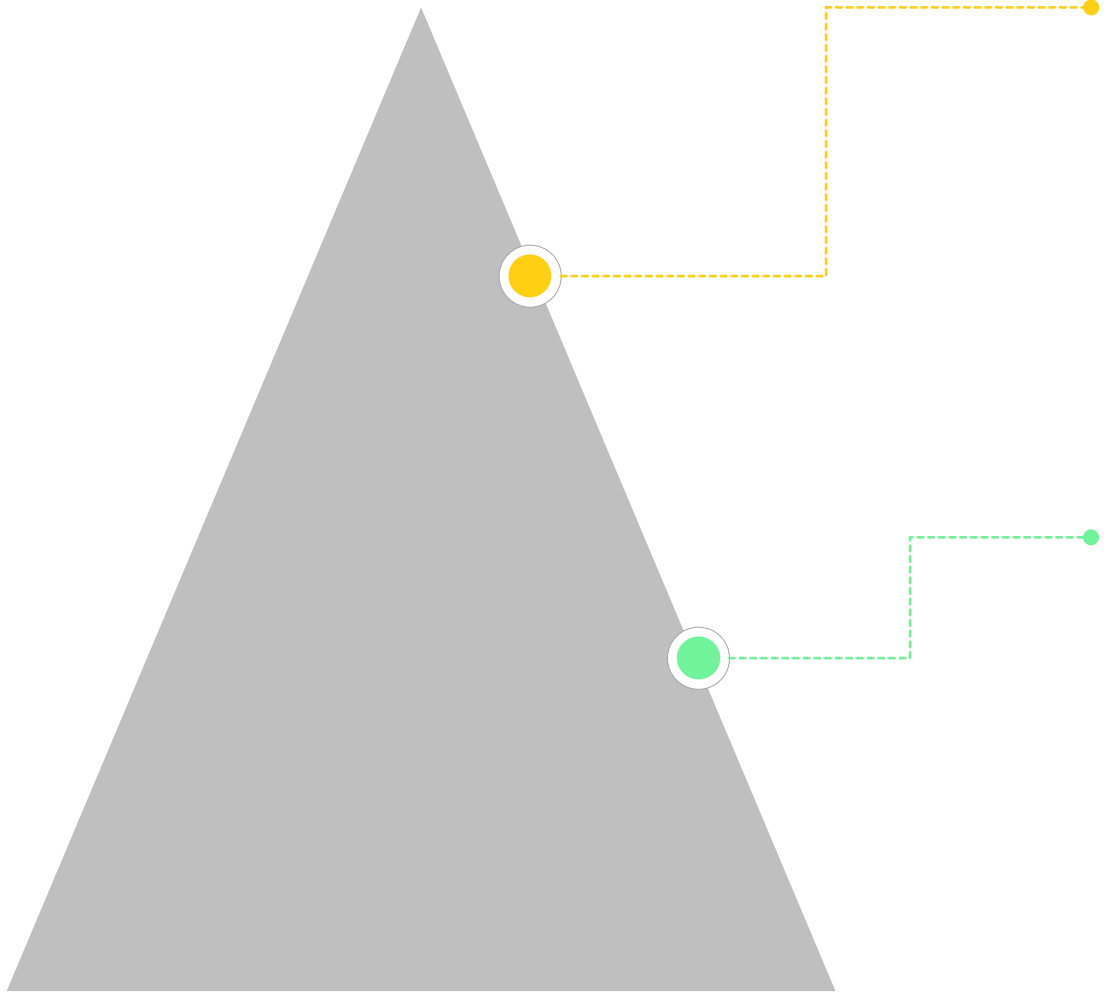
Analysis Report

Park, Pablo Chanwoo

WeissBeerger
TURNING DRINKS INTO DATA

Project Objectives



■ Primary objective

In [1]: Prove whether the hypothesis " Soju sales are increasing and cannibalizing the beer category" is acceptable.

Out [1]: *How is the relationship between demand for Soju and beer?*

■ Secondary objective

In [2]: Find opportunities to gain category share and increase beer incidents.

Out [2]: *What are the features that affect beer demand?*

The Findings

- It is difficult to state that demand for Soju is cannibalizing demand for Beer.
- Demand for Food seems to have meaningful and positive effect on demand for Beer.

DataSet

Given 16,635 rows of daily summary of transaction from 147 bars

Basic information			Volumes			Units			Revenue							Order counts				Derivatives	
Bar ID	*Date*	*Bar Segmentation*	*Beer Draught Volume (L)*	*Beer Packaged Volume (L)*	*Total Volume (L)*	*Beer Units*	*Spirits Units*	*Soju Units*	*Total Revenue*	*Beer Revenue*	*Spirits Revenue*	*Soju Revenue*	*Wine Revenue*	*Food Revenue*	*Non Alcoholic Revenue*	*# Beer Orders*	*# Beer & Food Orders*	*# Soju & Beer Orders*	*# Soju & Food Orders*	*Avg. Check Size*	*Soju Price*
Bar 1	2017-07-01	Food,Spirits	9.3	6.16	15.46	22	68	12	887,100	124,500	278,000	54,000	-	461,100	23,500	14	14	1	7	26,882	4,500
Bar 1	2017-07-02	Food,Spirits	21.3	2.5	23.8	16	63	4	745,800	167,500	256,000	18,000	-	298,300	24,000	7	7	1	2	46,613	4,500
Bar 1	2017-07-03	Food,Spirits	9.3	2.5	11.8	12	57	5	575,100	87,000	237,000	27,000	-	234,100	10,000	7	7	2	4		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...						
Bar 147	2017-12-30	Food	44.9	1.98	46.88	91	10	-	1,577,712	318,639	36,360	-	-								
Bar 147	2017																				

☐ Dataset

Size: 16,635 x 22
Period: July 2017 ~ December 2017
Source: POS from 147 bar
Types: identifiable information,
volumes sold in liter (limited to beer),
units sold (beer and spirits(soju) category only),
revenue for a day (Food + beverage categories),
order counts (limited to specific conditions),
and derivatives from above variables

☐ Able to...

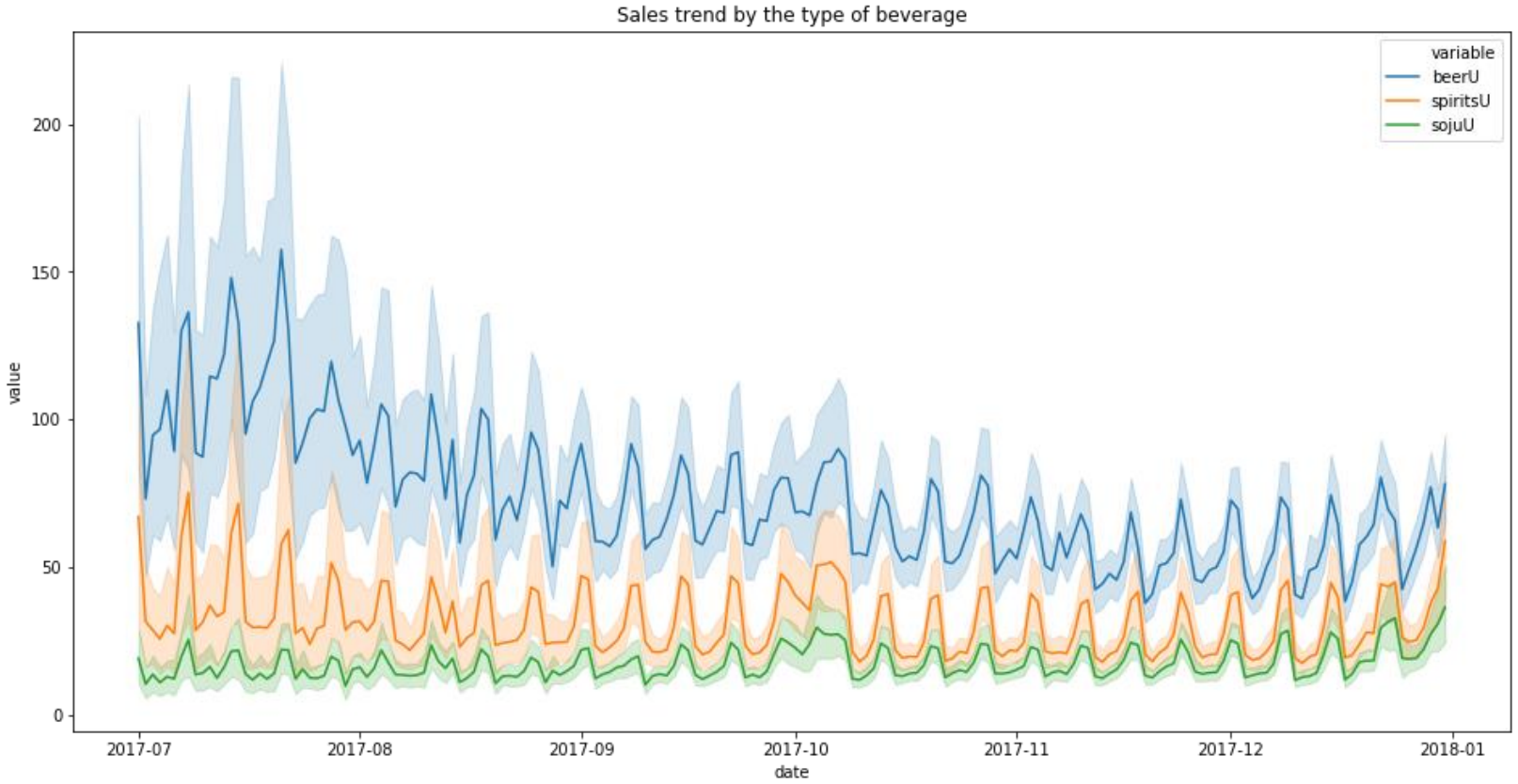
- ✓ Understand basic information about daily transactions
- ✓ Track sales volume of beverage by category
- ✓ Approximate price given revenue and units

☐ Unable to...

- ✗ Track record of products sold
- ✗ Identify composition of items in a single order
- ✗ Identify how items are categorized into *(What else but soju?)*

Explanatory Data Analysis

Beer demand seems to be in downward trend, but it cannot be so sure of.



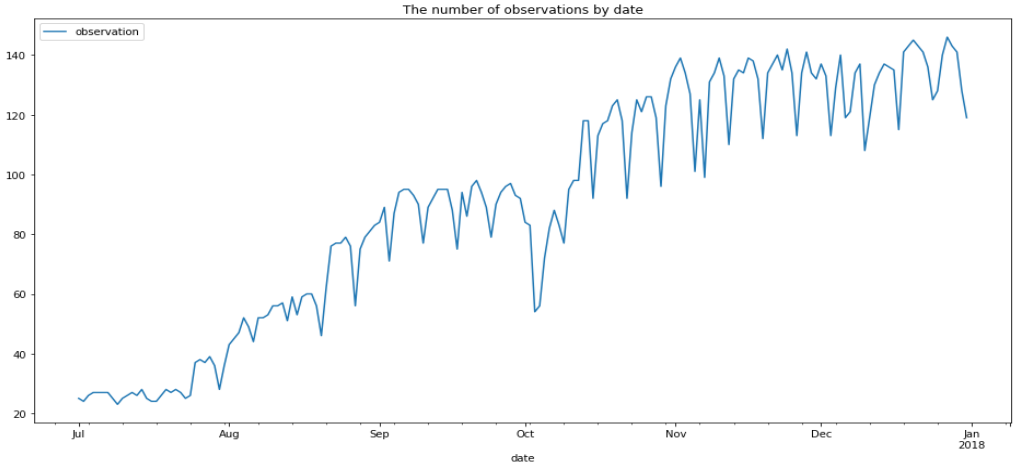
Beer Units	July	August	September	October	November	December
mean	109.9497	80.82576	70.82791	65.07439	54.49539	58.70654
std	**127.7843**	95.14461	75.71621	71.73081	58.72303	**62.45338**

- The line plot certainly shows downward trend for beer units, while other categories does not seem to be fluctuating to much.
- Despite the downward trend, I could also note that the variance is relatively large at the initial stage of observations.

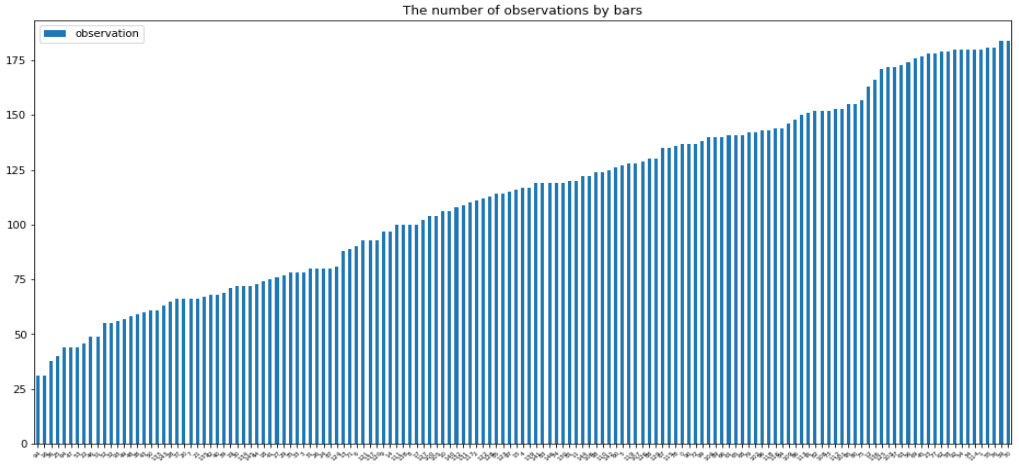
  "Additional study on date variable required"

Explanatory Data Analysis

The data is not complete by date, and by bars, as there seem to be irregular patterns in missing data.



Date	counts
2017-07-09	23
2017-07-02	24
2017-07-17	24
2017-07-16	24
2017-07-01	25

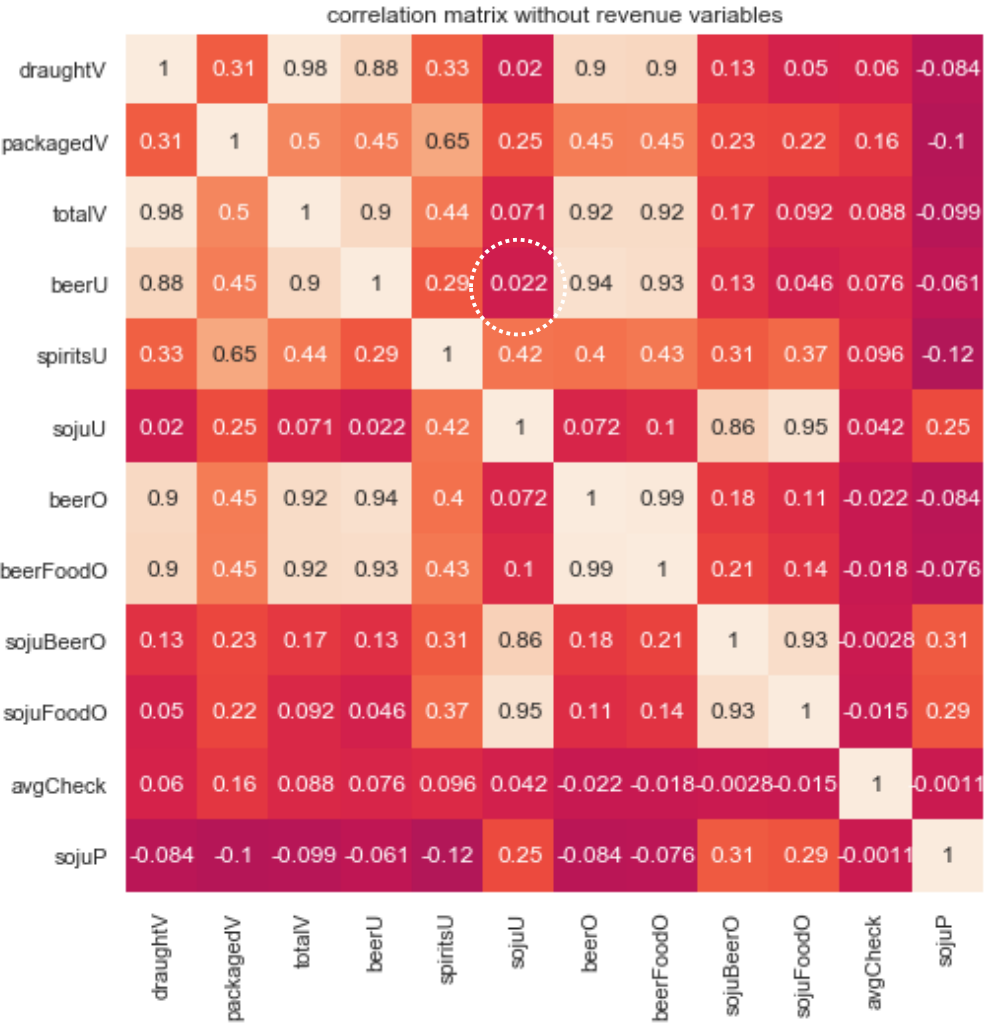


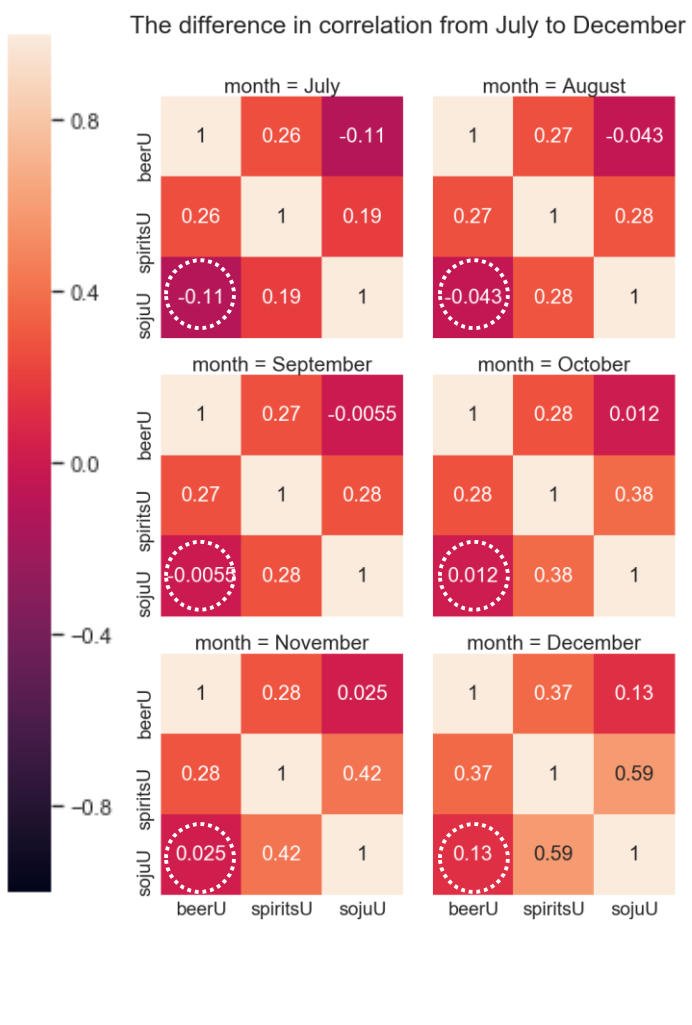
Bar ID	count
Bar 51	31
Bar 113	31
Bar 131	38
Bar 121	40
Bar 24	44

- ☐ Looking at the number of observations by date, and by Bar ID, it has come to my attention that the data is not complete
- ☐ Out of 184 days, not a day had full record from 147 bars. The 9th July had only 23 observations, which would have caused such a large variance.
- ☐ Out of 147 bars, only two had full record of 184 days. Unfortunately, two of the bars did not seems to have sold Soju at all, to compare relationship with beer.

Explanatory Data Analysis

Correlations between soju and beer turns positive over time, while some variables are highly correlated.

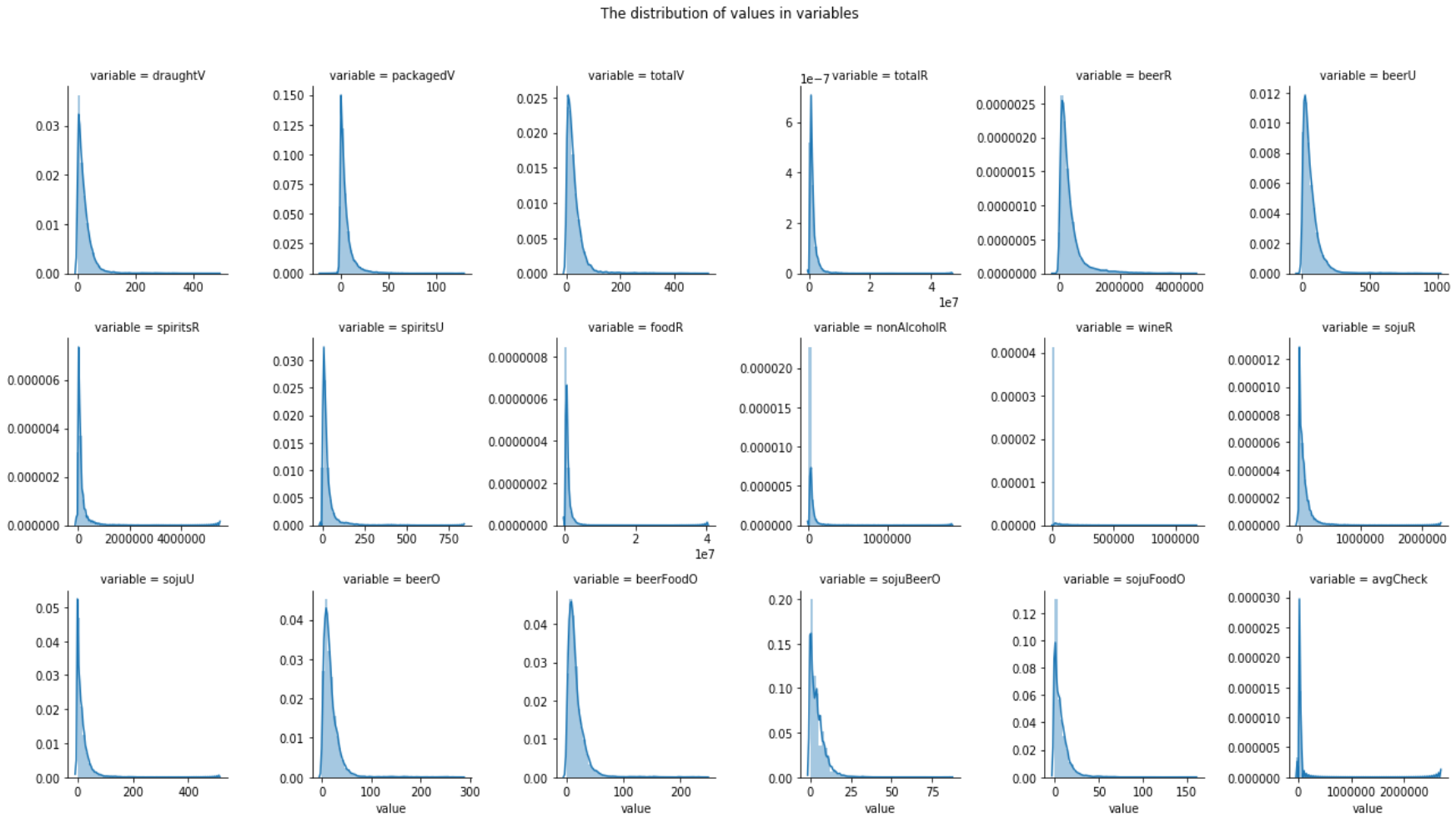




- ☐ Running correlation analysis on all variables, there seems to be strong correlations among volume, revenue, unit and orders. It is obvious in deed, and using all variables would distort the effect on each other.
- ☐ Comparing correlation coefficient of “Soju Units” and “Beer Units” by month, I could notice that they are becoming more positively correlated, showing weak evidence of complimentary goods.
  [-0.11, -0.04, -0.01, 0.01, 0.03, 0.13]

Explanatory Data Analysis

The variables are all in right-skewed distribution, which needs transformation into normal distribution upon linear regression.



- Bars, restaurants, and most of the commercial outlets share right-skewed distribution of variables related to its revenue. (most of sales are made below average level, and sales boost up when externalities like holiday, weekend play in)
- The best way to deal with right-skewed data would be log transformation in order to prevent exaggerating the effect of certain variable.

Preprocessing

Zero values referring to “we do not sell” do not tell information about the relationship, while zero values referring to “we can not sell” do tell something about its relationship with other items

	totalR	beerR	beerU	sojuR	sojuU
Bar 131	10947900	2384500	548	0	0
Bar 145	16508000	2696000	670	0	0
Bar 2	981541700	173039900	26144	0	0
Bar 93	47318500	24366500	3652	0	0
Bar 23	93985355	20195544	3729	0	0

“15 Bars without Soju”

	date	totalR	beerR	beerU	sojuR	sojuU
Bar 1	2017-07-14	1129000	194500	29	31500	5
Bar 1	2017-07-15	938000	146500	29	22500	4
Bar 1	2017-07-16	225900	62500	11	10000	2
Bar 1	2017-07-17	182300	25000	3	0	0
Bar 1	2017-07-18	664800	153000	34	18000	2

“Bars trying to sell Soju”

- ☐ There is no need for data that does not explain the relationship between beer and Soju. Such observations must be removed as they will likely to tell different story.
- ☐ The kind of data to be omitted can be defined as “the data from bars with zero sales of soju throughout the entire observation period.”

Preprocessing

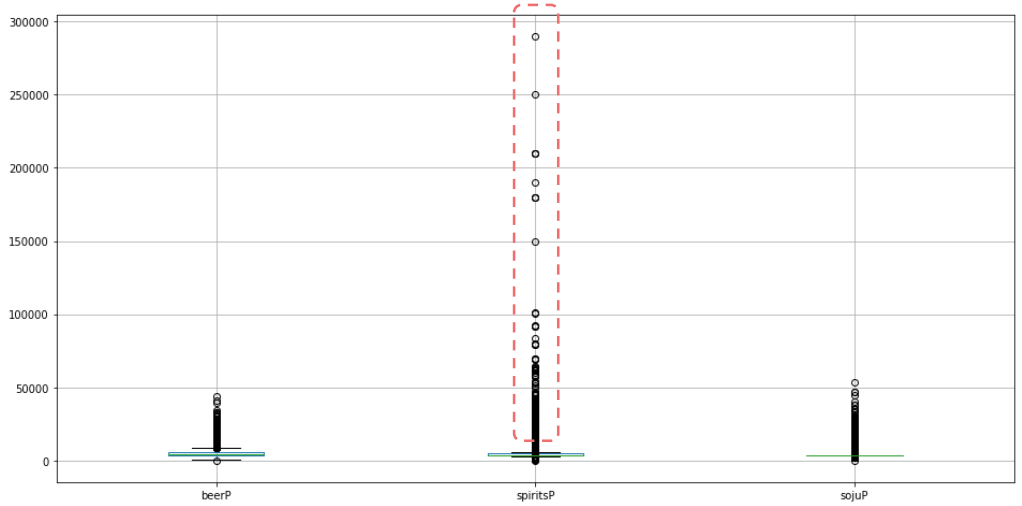
There are values impossible to explain. The values may indicate right record, but cannot be explained from the dataset and mislead without new features to explain. (i.e. beer selling at over $40/unit when the average is under $5)

Negative values

	date	beerR	beerU	spiritsR	spiritsU	foodR
Bar 3	2017-12-18	**-151200**	**-23**	210000	53	565500
Bar 47	2017-09-06	111000	15	**-69000**	**-11**	717700
Bar 50	2017-11-08	51820	8	0	**-12**	1991866
Bar 67	2017-10-31	0	0	0	0	**-65456**
Bar 91	2017-09-25	327700	46	**-38500**	**-5**	389100

It can have negative sales in real life. (i.e. refund, waste) However, without variables explaining negative sales, it is an impossible value to consider for the analysis. (7 records)

Outliers



There are values way outside normal boundary. In order to detect anomalies, 1) I generated ratio-base derivatives (price), which follows normal distribution, 2) observations with price exceeding 99.7% confidence interval (3 std) are considered as outliers. (593 records)

Preprocessing

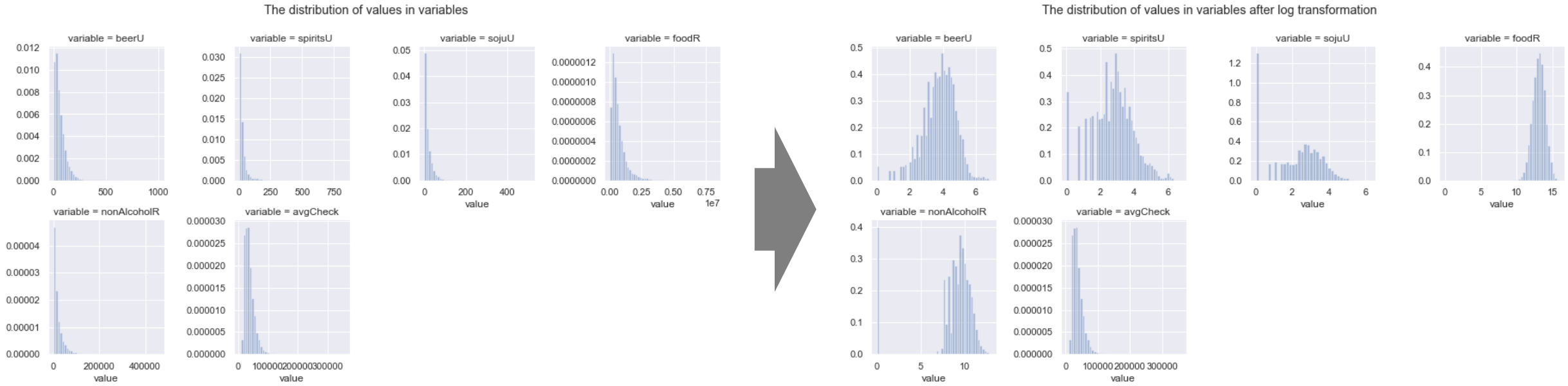
As there seems to be multicollinearity among variables, a careful selection of variables is needed to study the true effect of variables to beer demand.

Demand for Beer	=	***Demand for other product***	+	***Characteristics of Bar***	+	***Characteristics of Dates***
Beer Units		**Soju Units**		Bar 1		Month
		Spirits Units		…		Day
		Food Revenue		…		(in dummies)
		Non Alcoholic Revenue		Bar 147		
				(in dummy)		

- ☐ There are three types of variables/features explaining demand for beer given dataset; units, volume, revenue explains Demands, bar ID explains physical environment which customers are exposed to, and dates which explains the moment of drinking experience.
- ☐ Demand related variables are highly correlated and need to carefully choose set in order to avoid duplication of effect. (multicollinearity)
- ☐ For this model, demand for beer is to be explained in combination of 1) demand for other products, 2) characteristics of bar, and 3) characteristics of date. What we are interested in particular from the purpose of study is the effect of Soju to Beer.

Preprocessing

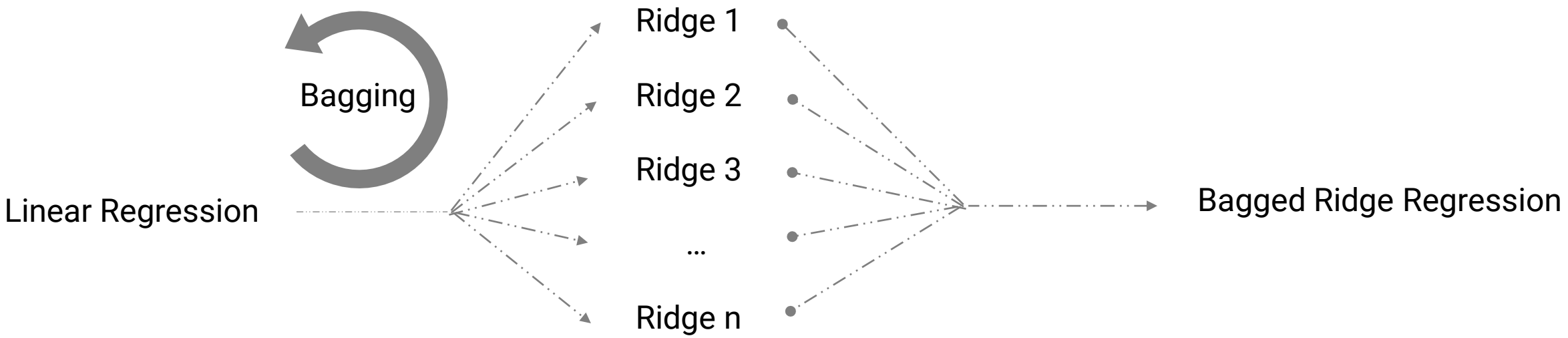
For the purpose of regression models, I want to avoid right-skewed dataset. Log transform to fit the dataset for the appropriate model.



- ☐ Regression assumes its variables to be normally distributed. It is less relevant when sample size is large, but CAN (consistent and asymptotically normal) variables are always good for fit.
- ☐ Log transformation is a good way to normalize skewed observations.

Analysis

Bagged Ridge Regression model has been implemented to test the effect of independent variables to demand for beer



- ☐ While other machine learning algorithms offer fine accuracy and precision, linear regression explains effect of independent variables to dependent variable.
- ☐ Bagging (multiple sampling for averaging) and Ridge (penalty to weight) generalize linear model to return robust output minimizing variation.

Analysis

There is no evidence to state that Soju demand is taking shares and cannibalizing beer demand, under bagged ridge regression model.

$$\textit{Beer Units} = \mathbf{-0.003\ Soju\ Units} + 0.059\ \textit{Spirits Units} + 0.744\ \textit{Food Revenue} - 0.001\ \textit{Non-Alcoholic Revenue} + \alpha$$

NOTE. Variables in log

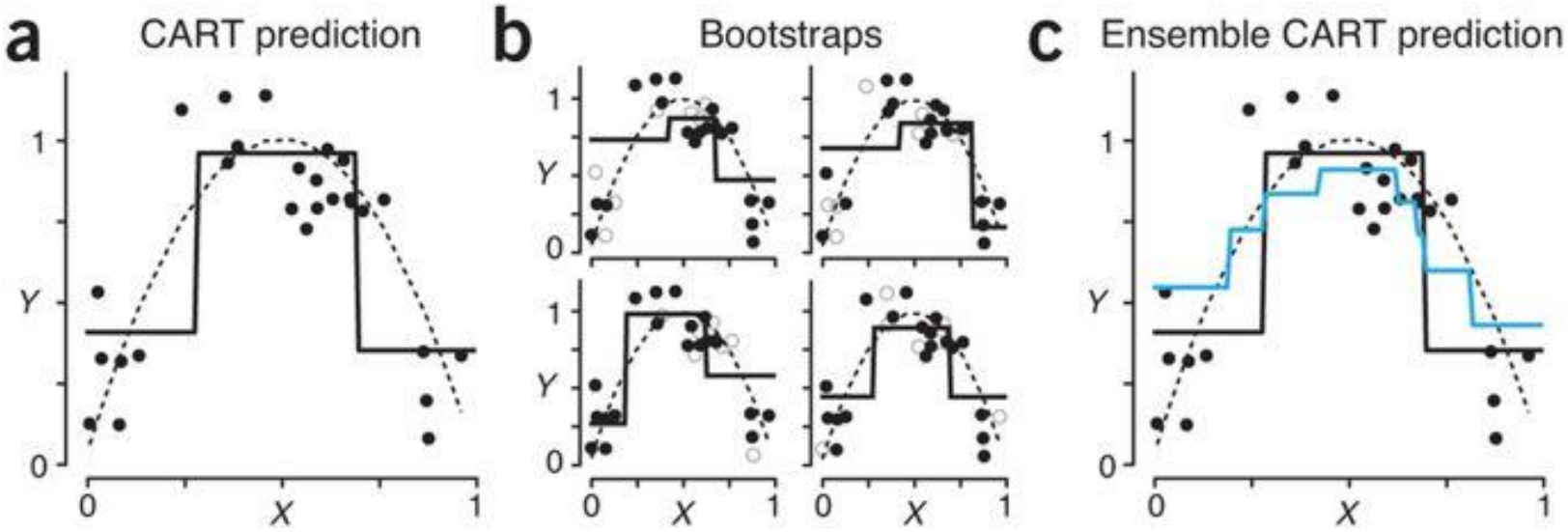
R^2 :	0.8395
MSE:	0.1632

"1% increase in Soju Units result in 0.003% decrease in Beer Units"

- ☐ The coefficient for Soju Units indicates negative relationship, but at an insignificant level. Taking account of relatively small unit volume of Soju, 0.003% seems even more disappointing.
- ☐ The model explains 83% of variances, and MSE over full dataset is 0.16, which I believe the model is trustworthy.
- ☐ Noticeable, when Food Revenue increases 1%, Beer Units increases by 0.74% alongside.

Analysis

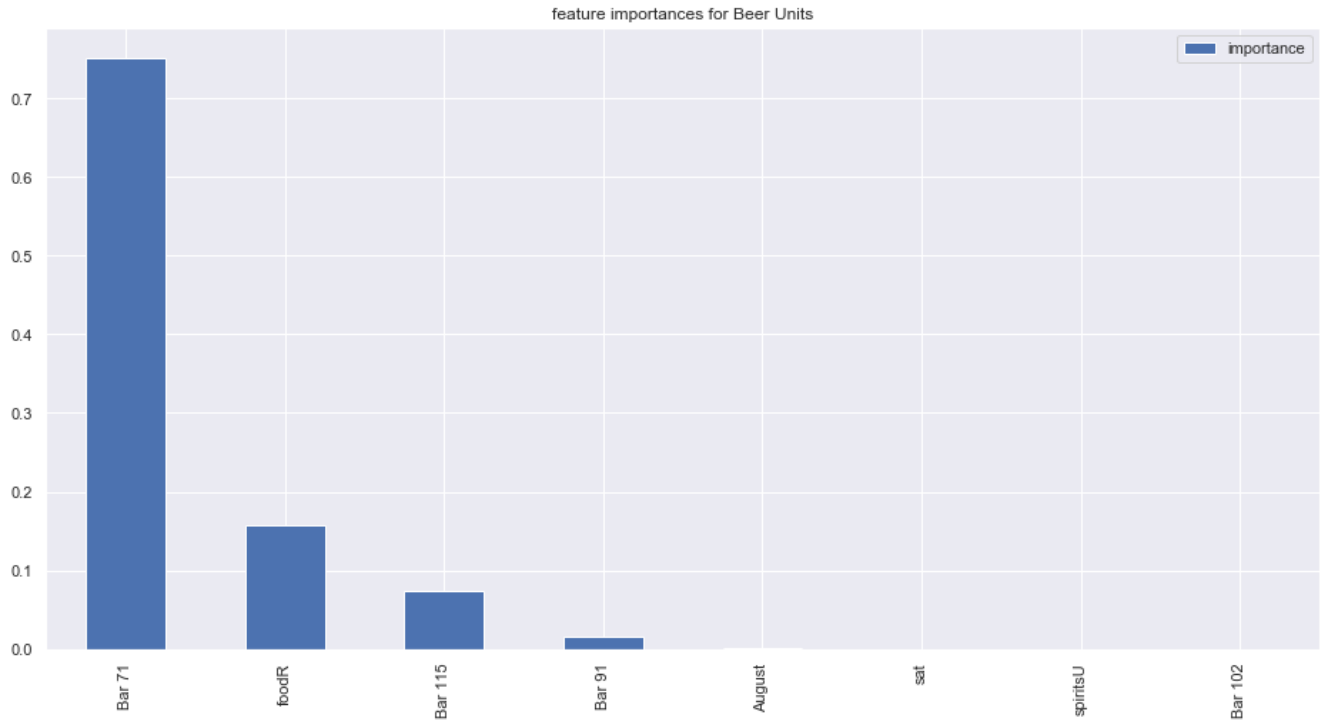
CART model fits better in modern dataset compared to regression models where decision boundaries are not linear.



- [ ] CART model is very precise and accurate, but does not provide information on how the variables are related.
- [ ] However, it provides information on which variables are determinant to forecasting dependent variable, beer units.

Analysis

Soju Units are not considered as a significant factor to determining Beer Units.

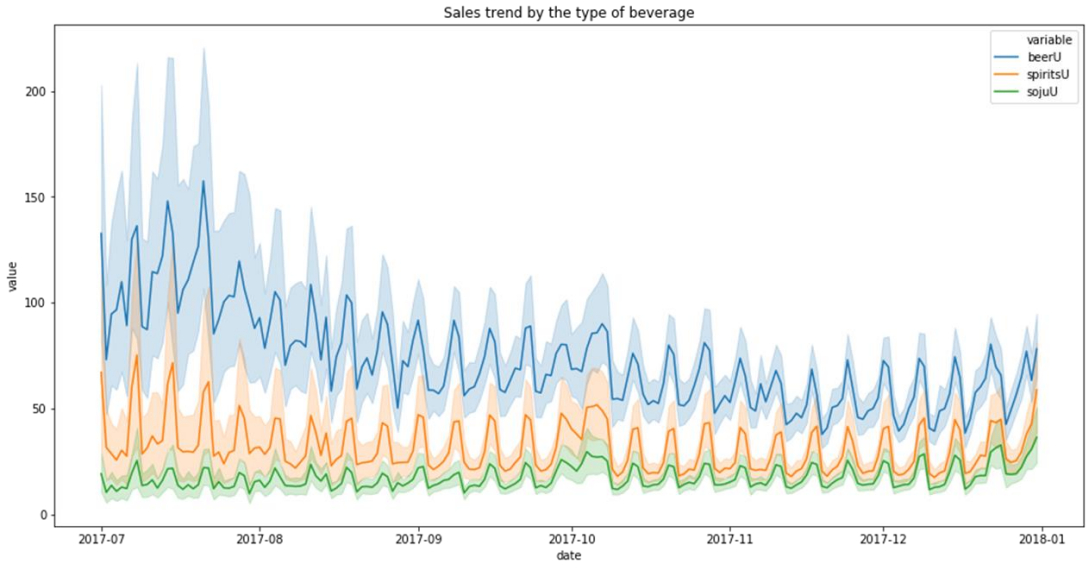


	importance
Bar 71	0.751946
foodR	0.157867
Bar 115	0.074007
Bar 91	0.015377
August	0.000522
sat	0.000282
spiritsU	0.000000
Bar 102	0.000000

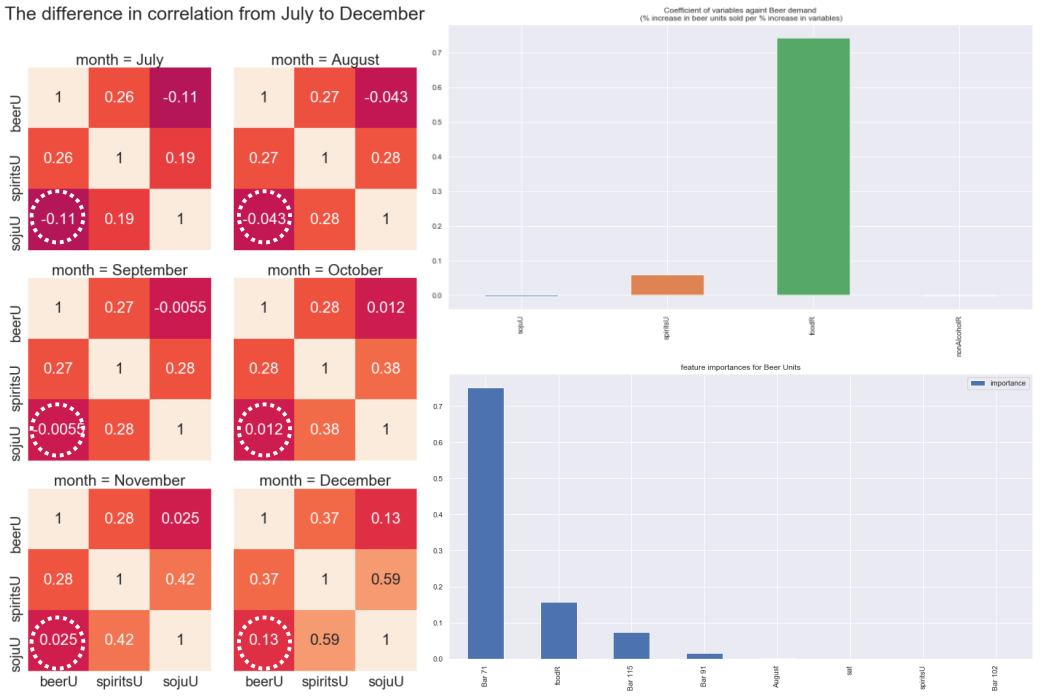
- ☐ The feature importance scores how useful or valuable each feature was in the construction of the best fitting CART model. It is not a primary metric, but something to be reference to.
- ☐ Soju Units is not listed at all, while Food Revenue is considered determinant for determining Beer Units.

Conclusion

There is no evidence to argue that Soju is cannibalizing the beer category, from the data given.



Beer Units	July	August	September	October	November	December
mean	109.9497	80.82576	70.82791	65.07439	54.49539	58.70654
std	127.7843	95.14461	75.71621	71.73081	58.72303	62.45338



1) The variance is fluctuating due to small size of sample
2) 6-month period of observation may not able to explain seasonality

“Beer sales may not be decreasing in the real world out of sample!”

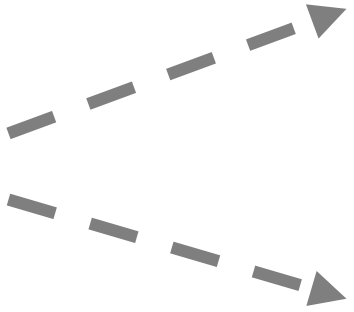
“correlation coefficient”, “regression coefficient”, “feature importance”, all three metrics all indicates that;

“demand for Soju has very little to do with the demand for beer”

Means of application

Despite unexpected outcome, there were some noticeable findings: beer & food.

“**Beer and Food** works so well together in this Zone”



Bar owners, go for decent food pairing!

Activities that trigger extra visit

Distributors, go promote dining out!

- ☐ Beer and foods have showed a strong/positive relationship. This is evident from Order count data, where 92.4% of customer who ordered beer also ordered food.
- ☐ To a bar or restaurant owners, the best way to increase beer incident and good profit would be to **develop decent cuisine and food pairing for beer.**
- ☐ To beverage distributors, I may suggest **promoting dinning out and gatherings** to increase beer consumption.

Limitation

There are a few limitations to this analysis; and also there are couple of points to be improved.

Limitation 1
“Data has been summarized on a daily basis, and it is unable to track record of products sold”

Limitation 2
“6-month period in Food and Beverage industry is not enough to take account of seasonality”

Limitation 3
“Promotion is important factor that triggers demand, which are unavailable”

Supplements 1
“Weak action plan and suggestions to increase beer incidents”

Supplements 2
“Segment variables were not taken account for while it may supplement small datasets by bars”

Supplements 3
“Order count variables were not taken account for as they were fragmented but of useful information”