Alcohol Sales in the US: Drink less, Forecast more

Names: Davide Giovanardi, Ishaan Khanna

1. Description/Executive summary:

Alcohol consumption in the United States is the second highest across the world. As per a leading industry database IWSR, the total alcohol market in 2017, including beer, wine and spirits, was 300 million hectoliters in volume terms and $157 billion in value terms. However, recent industry trends indicate a gradual shift in consumption patterns. Within the alcohol segment, beer has been losing its share to other categories such as spirits and wines. Moreover, the rising popularity of craft beer has been a constant threat to some of the major players in the alcohol beverage industry. Outside the alcohol segment, there have also been concerns that legalization of cannabis for recreational purposes may lead consumers to move away from consuming alcohol.

Previous work:
James Fogarty and Derby Voon (2018), in their paper titled *"Alcohol Consumption in the United States: Past, Present, and Future Trends"* study the long-run changes in alcohol consumption patterns in the US at the state-level. Using ARIMA methodology, the paper develops forecasts of per capita consumption of beer, wine and spirits. It also examines the impact of alcohol policy settings and tax rate levels to describe the changes in historical consumption patterns. It was found that there was no systematic relationship between alcohol taxes or other alcohol policy settings and forecast future consumption changes.

2. Data Set Introduction

We are interested in predicting monthly alcohol sales in the US. Towards this, we will estimate multiple models such as different combinations of ARMA, Seasonal Dummies, VAR with possible predictors such as lagged values of monthly sales, unemployment rate, CPI of food and beverage, CPI of tobacco and smoking products, and disposable income, to explore the impact of these variables on the monthly alcohol sales in the US. Data is obtained from FRED website. We use monthly data for each of these variables starting from 1/1/1992 to 12/31/2018. We have also split the datasets into two samples: training sample, starting from 1/1/1992 to 12/31/2017 and the test sample, starting from 1/1/2018 to 12/31/2018.

Descriptive statistics:

Alcohol Beverage Sales Descriptive Statistics	
Mean	7877.70679
Standard Error	161.9066182
Median	7438.5
Mode	4974
Standard Deviation	2914.319127
Sample Variance	8493255.973
Kurtosis	-0.806233869
Skewness	0.392405956
Range	12448
Minimum	3031
Maximum	15479
Sum	2552377
Count	324
	-3.6646E-249

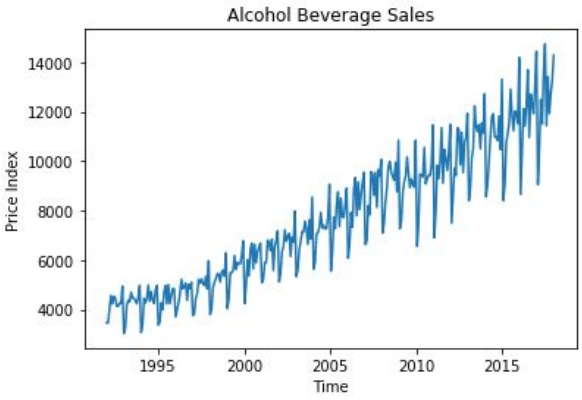
Descriptive Statistics - Unemployment Rate	
Mean	4.805246914
Standard Error	0.064140676
Median	4.9
Mode	3.8
Standard Deviation	1.154532165
Sample Variance	1.332944521
Kurtosis	-0.720602944
Skewness	0.140431562
Range	5.4
Minimum	2.5
Maximum	7.9
Sum	1556.9
Count	324
	-3.6646E-249

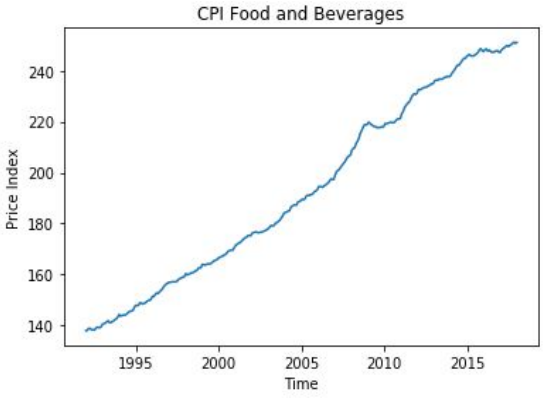
Descriptive Statistics: Food&Beverage CPI	
Mean	84.40709877
Standard Error	1.943132669
Median	86.4
Mode	40.4
Standard Deviation	34.97638804
Sample Variance	1223.34772
Kurtosis	-1.335387801
Skewness	0.082017962
Range	108.7
Minimum	34.6
Maximum	143.3
Sum	27347.9
Count	324
	-3.6646E-249

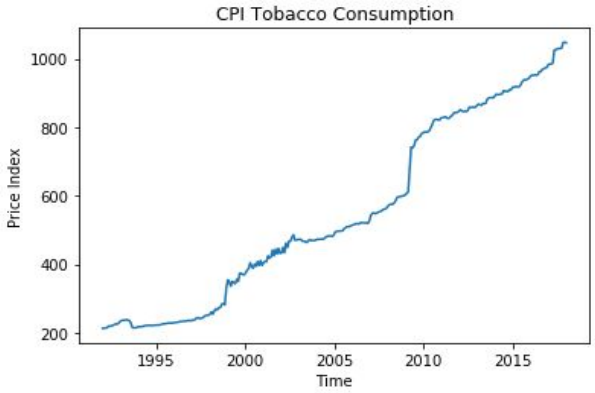
Descriptive Statistice - Disposable Personal Income	
Mean	1218.348148
Standard Error	46.26585503
Median	891.2
Mode	3019.2
Standard Deviation	832.7853906
Sample Variance	693531.5068
Kurtosis	-0.456801235
Skewness	0.885596208
Range	2849.6
Minimum	351.5
Maximum	3201.1
Sum	394744.8
Count	324
	-3.6646E-249

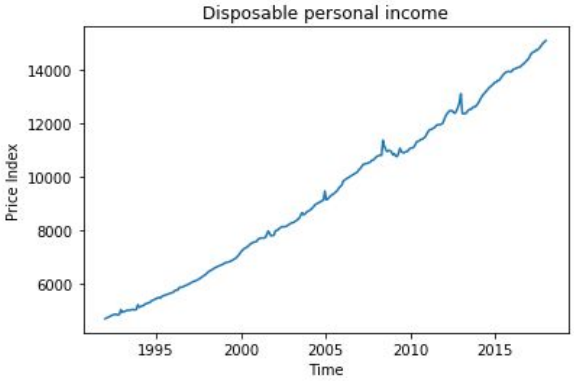
Descriptive Statistics - CPI Tobacco Consumption	
#Observations	312
Min	212.6
Max	251.13
Mean	550.6957372
Variance	71290.64391
Skewness	0.299619621
Kurtosis	-1.310200725

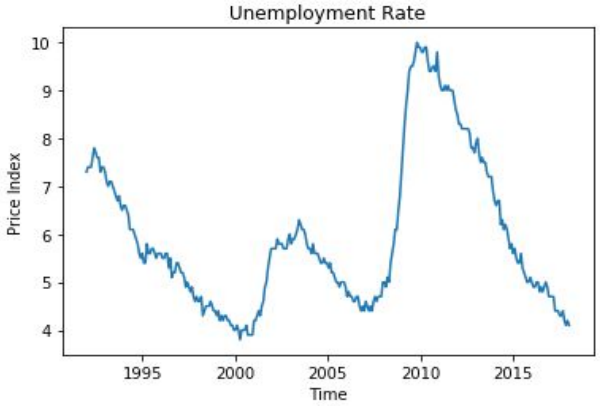
Below are the plots of all our variables including target and predictor variables:











ADF Test results for the training sample:

To check whether the data is stationary, we perform an ADF test. The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary. We check the result using p-value at a threshold level of 5%. The results are presented below:

Variables	p-value
CPI Food and Beverages	0.70
CPI Tobacco Consumption	0.52
Disposable Personal Income	0.21
Unemployment Rate	0.97

Since all the p-values of all the variables are above the 0.05, we fail to reject the null hypothesis and conclude that the variables are non-stationary.

Log-Differencing:

The datasets, Alcohol Sales, CPI Food & Beverage, CPI Tobacco Consumption, and Disposable Personal Income are log differenced first before we start conducting our analyses. Unemployment Rate is not log differenced since it is already representative of a growth pattern.

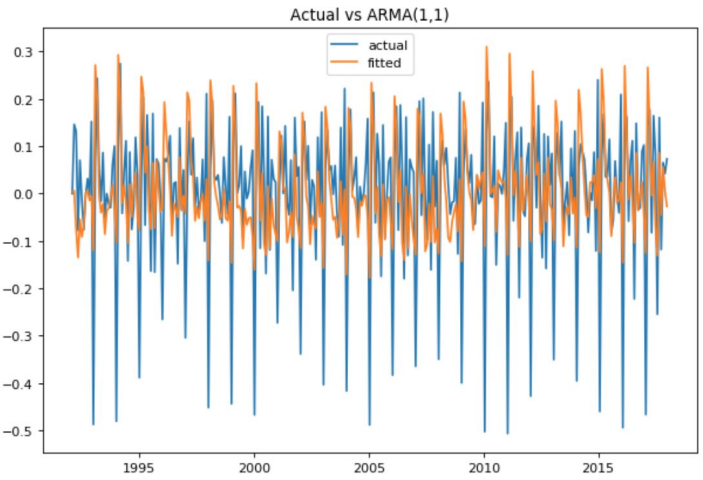
3. Estimation

The first part of this section consists in building a benchmark model. In this case, the benchmark will be an ARMA(1,1) model. We'll conduct our estimation in Python. As we'll see, there are multiple useful results from this first estimation that we will try to improve in the next models.
First, we can look at the summary statistics:

Model:	ARMA	BIC:	-409.2112
Dependent Variable:	y	Log-Likelihood:	216.09
Date:	2019-03-09 11:04	Scale:	1.0000
No. Observations:	311	Method:	css-mle
Df Model:	3	Sample:	0
Df Residuals:	308		1
Converged:	1.0000	S.D. of innovations:	0.120
No. Iterations:	29.0000	HQIC:	-418.191
AIC:	-424.1704		

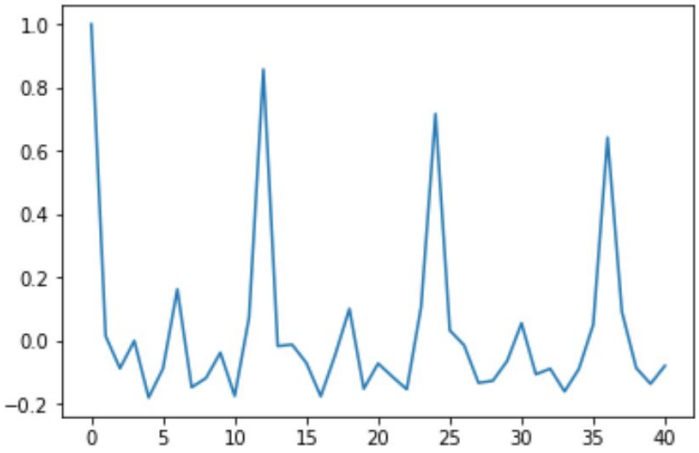
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	0.0039	0.0003	15.1860	0.0000	0.0034	0.0044
ar.L1.y	0.1462	0.0582	2.5130	0.0125	0.0322	0.2602
ma.L1.y	-0.9722	0.0140	-69.3207	0.0000	-0.9997	-0.9447

The BIC and AIC scores are important indicators of the goodness of the model; they are known as Information Criteria and allow to evaluate the model based on two metrics: estimation error and fit; specifically, they allow to assess the trade-off between reducing estimation error by introducing new predictors and overfitting the model causing poor out-of sample performance; indeed, these metrics feature a penalization term which increases as new predictors are added to the model. In the ARMA(1,1) the BIC and AIC are -409.21 and -424.17 respectively. We will use this metric to compare the goodness of our next models.
We can visualize this result by plotting the ARMA(1,1) on the actual observations:



Secondly, we can already see how significant our lag predictors are: this seems compelling at first glance. To confirm these results though, we shall study the residual correlation, which should be 0, meaning that they are independently distributed; if this is not the case, the model could be misspecified and we would get an inconsistent estimation of the betas.
To study the behavior of the residuals, we will plot the autocorrelation function:

Residuals Autocorrelation



We can already see that the residuals are not uncorrelated. This hints that the model has to be improved. Next, we'll compute the Box-Pierce test to get an extra metric to evaluate the residuals autocorrelation.
The Box-Pierce test is a type of statistical test that tests the null of whether any of a group of autocorrelations of a time series are different from zero. In our case we'll test the autocorrelations of 12 lags, in line with the fact that we have monthly observations.
We can now confirm our hypothesis that the residuals are not uncorrelated. Indeed, the p-value is almost equal to zero (1.23e-51). This signals that there is almost a zero chance that we observed the residuals in the case that the null hypothesis is true. Again, in the Box-Pierce test the null hypothesis affirms that the data are independently distributed. Therefore, we can conclude that the residuals are not uncorrelated.

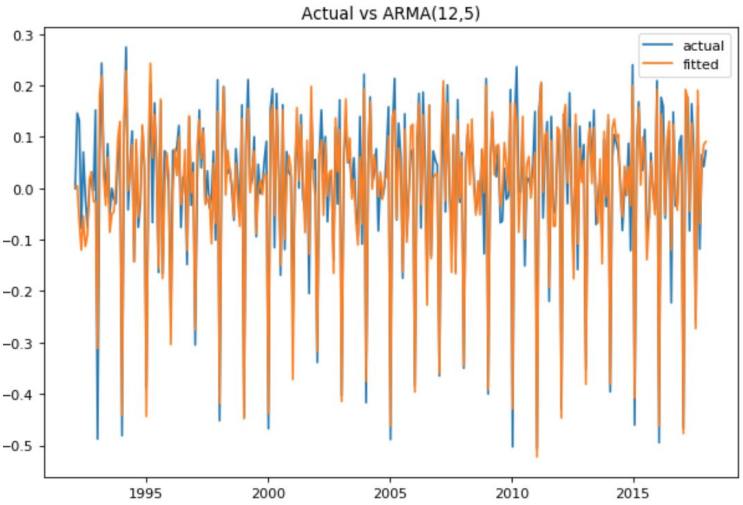
Armed with this knowledge, we proceed to the next step of the estimation.

Because the residuals are correlated, our aim is now to find the correct number of lags of the ARMA(p,q) model. In other words, we'll search for the best combination of p and q.
To achieve this objective, we run a nested for loop in which we evaluate an ARMA(p,q) for each combination and we store the AIC, BIC results of each pair of AR-MA lags. Specifically, we'll test all the possible combinations of up to 12 and 2 AR and MA lags. Finally, we'll pick the model with the lowest sum of AIC and BIC. This model turned out to be the ARMA(12, 2) with a AIC and BIC scores of -1018.03, -958.19, a major improvement from the ARMA(1,1).
Because the best combination of lags is on the edge, we further test the model with more lags, namely we test the ARMA(12, q) where 2 < q < 13. We then select the best model based on the lowest BIC-AIC and get the ARMA(12, 5) as our bet result.
Here is the summary statistics for the ARMA(12, 5):

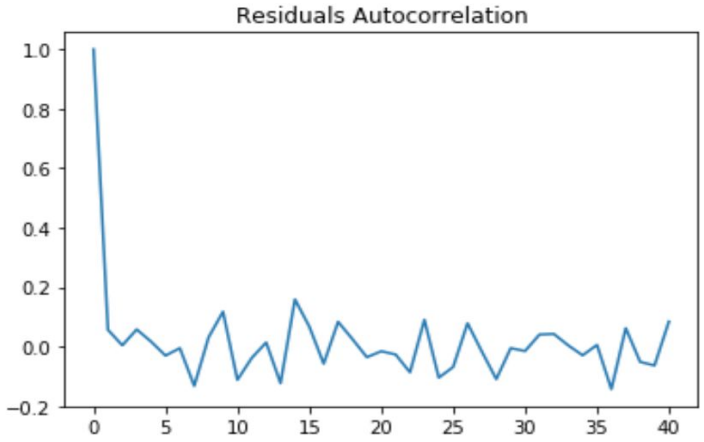
Model:	ARMA	BIC:	-996.2985
Dependent Variable:	y	Log-Likelihood:	552.68
Date:	2019-03-09 12:43	Scale:	1.0000
No. Observations:	311	Method:	css-mle
Df Model:	18	Sample:	0
Df Residuals:	293		1
Converged:	0.0000	S.D. of innovations:	0.039
No. Iterations:	500.0000	HQIC:	-1038.952
AIC:	-1067.3545		

	Coef.	Std.Err.	t	P>|t|
const	0.0034	0.0005	6.7812	0.0000
ar.L1.y	-0.2885	0.0726	-3.9754	0.0001
ar.L2.y	-0.2346	0.0762	-3.0786	0.0023
ar.L3.y	-0.3147	0.0688	-4.5767	0.0000
ar.L4.y	-0.2721	0.0706	-3.8528	0.0001
ar.L5.y	-0.2859	0.0735	-3.8917	0.0001
ar.L6.y	-0.2586	0.0714	-3.6241	0.0003
ar.L7.y	-0.3248	0.0703	-4.6193	0.0000
ar.L8.y	-0.2488	0.0722	-3.4471	0.0006
ar.L9.y	-0.2711	0.0711	-3.8113	0.0002
ar.L10.y	-0.3314	0.0697	-4.7540	0.0000
ar.L11.y	-0.2032	0.0738	-2.7554	0.0062
ar.L12.y	0.6553	0.0688	9.5203	0.0000
ma.L1.y	-0.7567	0.0814	-9.2949	0.0000
ma.L2.y	0.0140	0.0652	0.2141	0.8306
ma.L3.y	0.5419	0.0735	7.3682	0.0000
ma.L4.y	-0.4942	0.0673	-7.3446	0.0000
ma.L5.y	0.4607	0.0755	6.1020	0.0000

We can see how the predictors are almost all significant, which allows the BIC and AIC to substantially decrease from the ARMA(1,1) model, meaning that the estimation improvement far outpaces the fact that we introduce more parameters.
Plotting the ARMA(12,5) on the actual values we can notice how the model is better specified but possibly more prone to overfitting; we'll study this tradeoff in the forecasting section of this project:



We can now check the correlations of the residuals and perform the Box-Pierce test to see if we managed to improve the independence of the residuals and thus correctly specify the model.



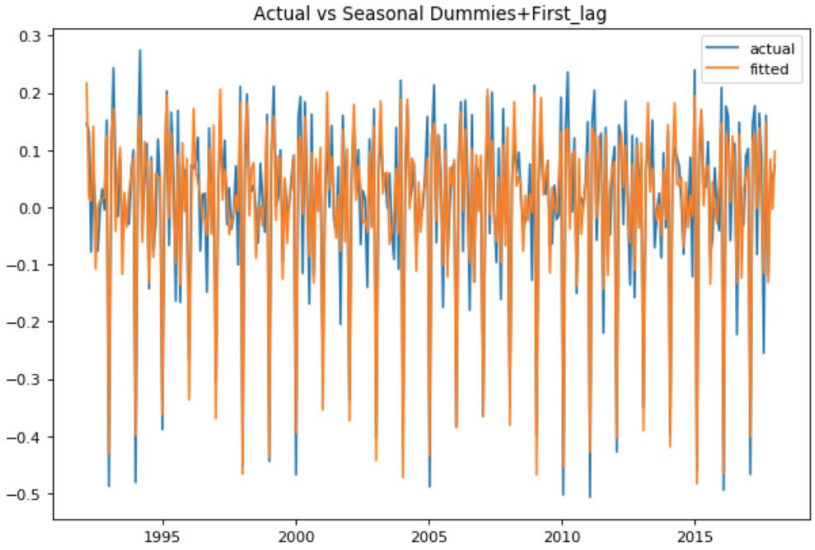
We can already see how the data is much more independent. To formally test thus conclusion, we compute Box-Pierce with 12 lags and we get a 15.64% p-value; in conclusion, we successfully managed to correctly specify the model.

In the next section, we'll estimate a different model, introducing seasonal dummy variables to account for seasonality in the alcohol consumption. Indeed, we can from the first plot that the trend might be prone to high seasonality; therefore, we'll test this assumption by estimating a model that includes 12 dummy variables (one for each month, we'll run the regression without intercept to avoid perfect multicollinearity) and the alcohol consumption first lag.

	Coef.	Std.Err.	t
a	-0.1661	0.0220	-7.5348
b	0.2166	0.0114	19.0594
c	0.1041	0.0133	7.8047
d	0.0897	0.0108	8.3322
e	0.0945	0.0115	8.2035
f	-0.0655	0.0109	-5.9859
g	-0.0056	0.0116	-0.4823
h	-0.0347	0.0110	-3.1552
i	0.0134	0.0112	1.1967
l	0.0361	0.0110	3.2678
m	0.1232	0.0108	11.4334
n	-0.3395	0.0124	-27.4814
alc_lag	-0.5996	0.0463	-12.9411

where a,...,n represent month 1,...,12. Most of the months are significant predictors and capture seasonality; specifically, February and December have a very large t-statistic, signaling that the mean effect in those months is substantially persistent.

To visualize how this model fits the data, we plotted the results against the actual values:



FInally, looking at the information Criteria we get a BIC and AIC values of -906.72 and -858.14, a net improvement from the ARMA(1,1) model which is our benchmark, but a worse performance compared to the ARMA(12,5) previously estimated. Again, we'll evaluate and combine this model in the next section of the project.

Lastly, we will estimate a Vector Autoregressive model with one lag. This model includes the external predictors described at the beginning, namely Food and Beverages CPI, Tobacco CPI, personal Income, Unemployment rate.
Here is the summary statistics of the estimation:

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Sat, 09, Mar, 2019
Time:                     14:31:59
--------------------------------------------------------------------
No. of Equations:         5.00000    BIC:                   -37.2010
Nobs:                     309.000    HQIC:                  -37.4185
Log likelihood:           3641.29    FPE:                4.85759e-17
AIC:                     -37.5634    Det(Omega_mle):     4.41226e-17
--------------------------------------------------------------------
Results for equation Alcohol
================================================================================
                     coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const                   0.056701         0.034178            1.659           0.097
L1.Alcohol             -0.422805         0.057524           -7.350           0.000
L1.CPIfood            -10.623385         3.703761           -2.868           0.004
L1.CPItobacco          -0.596844         0.449867           -1.327           0.185
L1.income              -2.449809         1.094559           -2.238           0.025
L1.unemployment        -0.002954         0.005176           -0.571           0.568
================================================================================
```

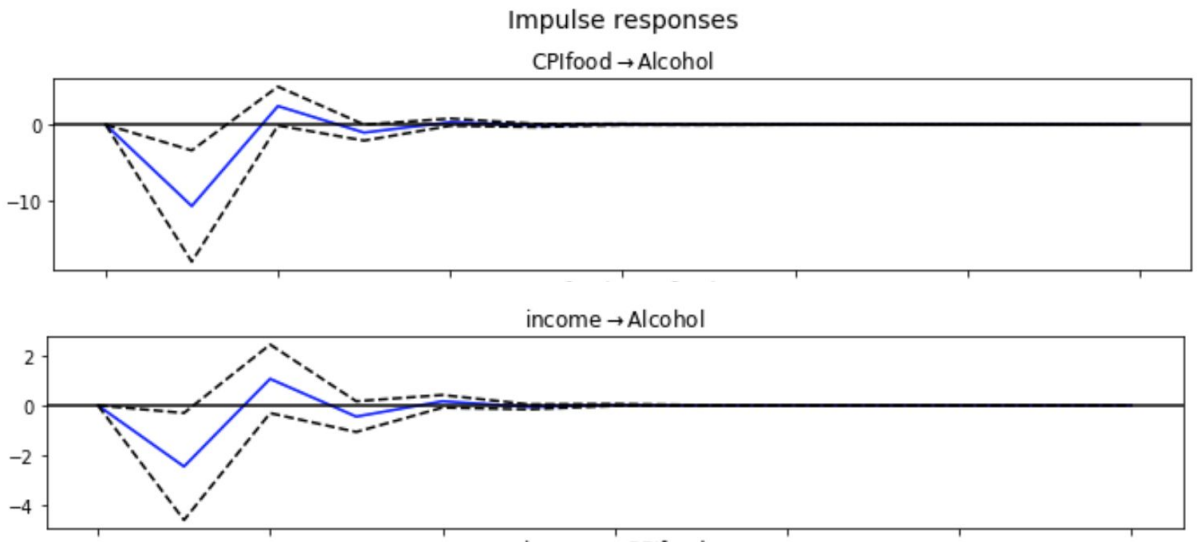
We can see how Lag Food, Income, and Alcohol are the most significant
According to BIC and AIC this model performs worse than the ARMA(1,1); we will assess this performance in the out of sample forecast. We also estimated a VAR(2) with 2 lags but the performance of this latter model are worse than VAR(1) in terms of BIC, AIC so we'll stick with our previous model. We believe that this model could still be useful in a forecasting combination.

3b. Forecasting

VAR(1)

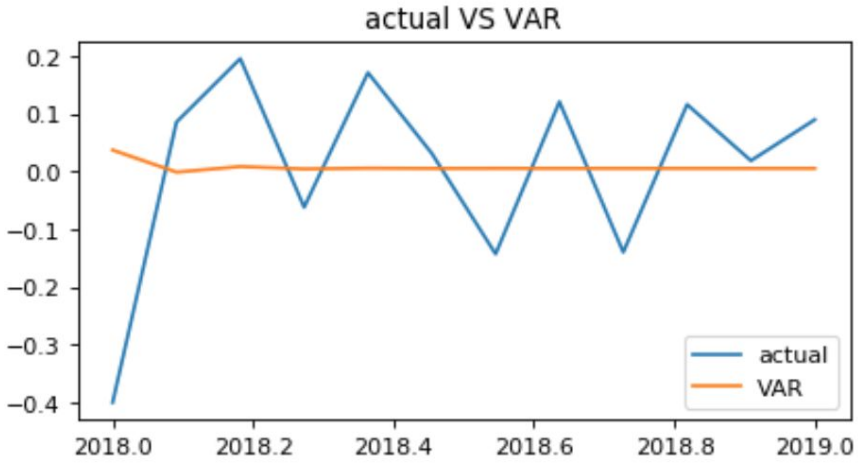
In the following section we will forecast the Alcohol sales one year ahead, for the period starting January 2018 and ending December 31st 2018, using our estimated models, namely VAR(1), ARMA(1,1), ARMA(12,5), Seasonal Dummies and lag Alcohol. We expect eachmodel to give different results. We will compute the root mean squared error as we proceed with the forecasting and finally we'll combine the models and evaluate the predictive accuracy.

Before doing the actual forecast, we are interested in studying the impulse response of the two most significant variables within the VAR(1), namely CPIfood and Income on Alcohol. We do this by estimating computing the impulse response function for 12 periods ahead. This function will use the estimated VAR(1) to forecast the values 12 months from now, then shock one variable by one unit, and study the effects on the other variables.
The two impluses are shown below:



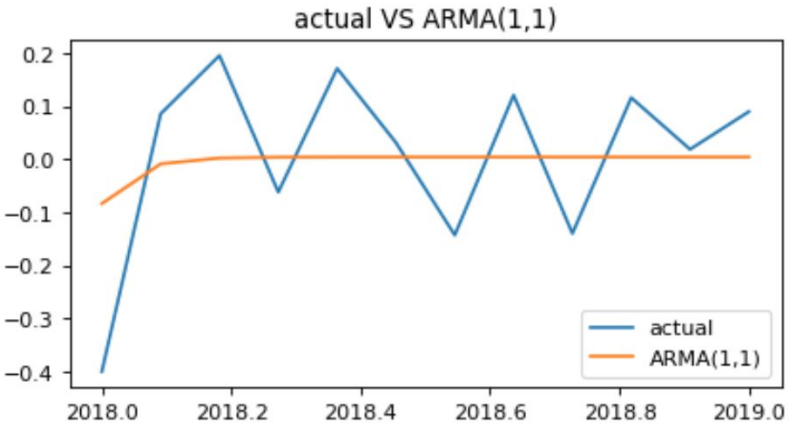
It's considerably interesting to notice that the two variables have an amazingly similar quick effect on Alcohol that has different magnitude: approximately -10 for Food and Beverage and -2 for Income. While for the former there is a clear economic explanation, namely that when the prices of Food and Beverage rises Sales of Alcohol shrink, for the latter impulse, an explanation is less obvious; we could suppose that as income goes up, people feel happier and therefore buy less alcohol.

Next, we forecast the Alcohol consumption one year ahead with the estimated VAR(1) model.
We then compute the Root Mean Squared Error (RMSE) and obtain a value of 0.5841.
To have a better picture of how our forecast looks like we plot the predicted values against the actual:



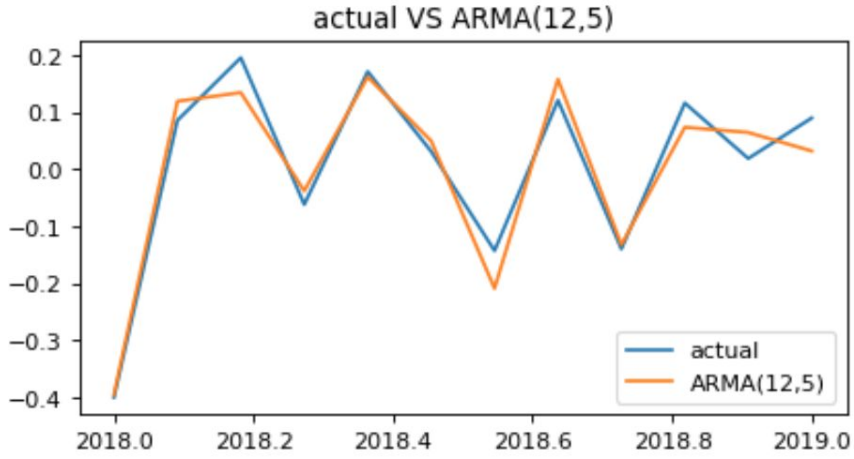
The forecast provided by the VAR is not informative as it generalizes the trend excessively. We will try to integrate this model later when we'll combine the forecasts.

ARMA(1, 1)
Next, we'll evaluate the benchmark, namely the ARMA(1,1).
After forecasting for 12 periods ahead we plot the results:

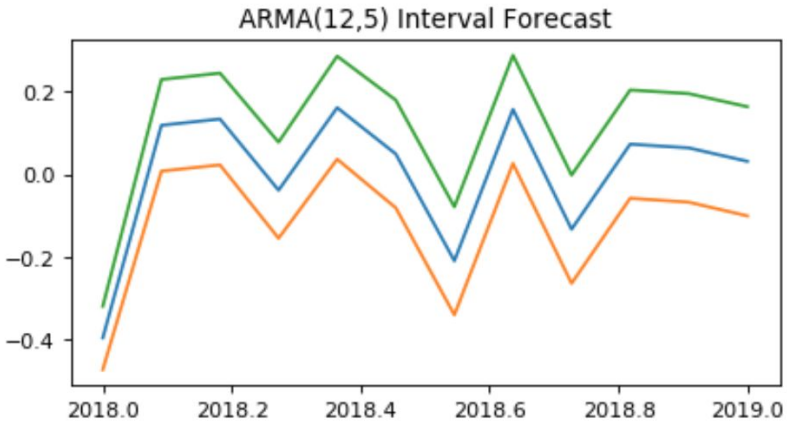


We can immediately see how this forecast fails to capture the trend; indeed, it doesn;t have enough predicting power and after few periods it converges to the mean value equal to zero.
The RMSE for ARMA(1,1) results to be 0.5048, a slight improvement from the VAR specification

ARMA(12, 5)
This was our best in sample model according to the Information Criteria (BIC, AIC); we know want to assess the goodness and accuracy in the out of sample forecast.



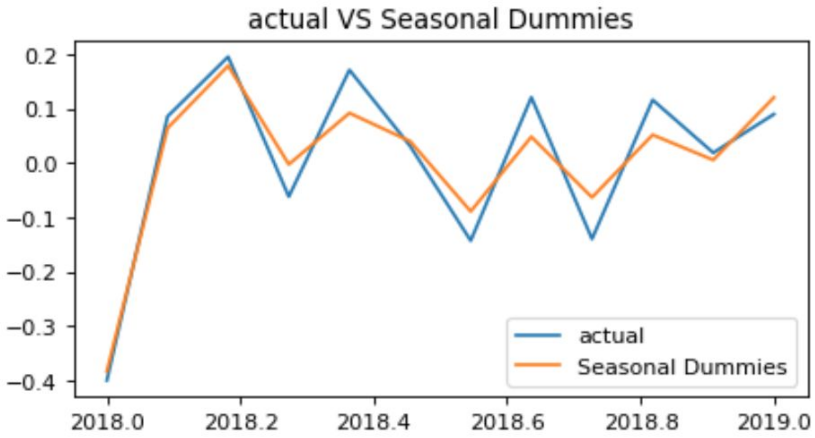
We can immediately see that the forecasted values are very well tied to the actual realization of the Alcohol sales growth.
To strengthen this result we compute the RMSE which results to be 0.1372, a net improvement from the previous two models.
We are now interested in plotting the intervals of this forecast, so that we have a better measure of the confidence interval of the prediction:



This model is the best so far. We will further test it and try to improve it with forecasting combinations later in the project.

Seasonal Dummies

We finally test the Seasonal Dummies plus the first lag of Alcohol consumption. This forecast demanded slightly more complex computations. We first created a matrix of dummies representing the 12 months in the forecast horizon; next, we produced new data points by iterating through a for loop and each time we used the lag forecasted value of alcohol sales.
The plot below shows the result of the forecast:



The RMSE resulted to be 0.1735, great result if compared to ARMA(1,1) or VAR(1), but slightly worse than the ARMA(12,5).
This forecast presents some dissimilarities if compared to the ARMA(12,5) and this signals that a combination between the two could lower the RMSE. We assess this in the next section.

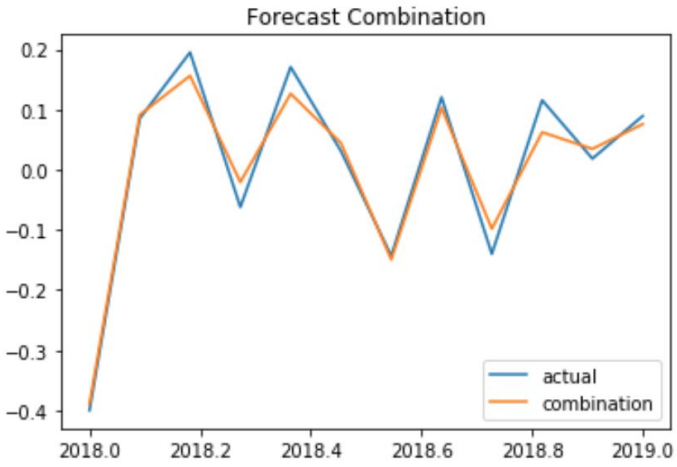
Forecast evaluation
In this section we'll implement a method that allows us to evaluate the goodness of each forecast. Specifically, we'll use the test of equal loss by Diebold Mariano.
Firstly, we are interested in evaluating the two best in-sample models, namely the ARMA(12,5) and Seasonal Dummies + Lagged Alcohol. To do so, we square the forecasting errors of each model and we take the difference between the outcome. Then, we regress this difference on a constant.

Looking at the t-stat, we get a value equal to approximately 1, which doesn't allow for strong significance, but is not negligible too. Because we took the Seasonal Dummies as our first member of the difference, a positive coefficient signals that the ARMA(12, 5) produces better forecasts, as it is confirmed by its lower RMSE.
Secondly, we compare the ARMA(12,5) to the VAR. Again we subtract the errors of the ARMA(12,5) from the VAR errors and we get a positive coefficient and tst-stat value of 1.77 which is stronger than before and confirms ARMA(12,5) to produce better forecasts.
Because the intercept is not significant in the Diebold Mariano test, we are now extremely interested in exploring the possible forecast combinations between the different models.

Forecast combinations
We are now interested to see if we can improve the RMSE by combining different forecasts.
The first combination is equal weighted includes the Seasonal Dummies and the VAR,with a RMSE of 0.35122. This implies that the VAR generalizes too much and fails to produce an acceptable forecast even if combined with the seasonal Dummies model.

Next, we combine the two best models, namely ARMA(12, 5) and Seasonal Dummies + Alcohol First Lag. This combination is equal weighted as well. Delightfully, we obtain a RMSE of 0.1045 which is the lowest we got so far! The two forecasts produce a better prediction than either taken alone. We test this for difference weights such as 0.7 (ARMA) - 0.3 (Dummies) and obtain very similar RMSE.

Finally, combine 3 models, namely ARMA(12,5), Seasonal Dummies, VAR(1) with equal weights and obtain a RMSE of 0.2416.
Here is the plot of our best combination: ARMA(12, 5) - Seasonal Dummies



We can see how the forecast is amazingly more precise than both models taken alone.

Conclusion

This project leaves us with many useful insights. First and foremost, we are delighted to have achieved a very reliable 1 year out of sample forecast with 2 of our models, namely ARMA(12,5) and Seasonal Dummies+Lagged Alcohol. The ARMA model provided an amazing benchmark and when 12 and 5 AR and MA lags are included, it provided a precise model to forecast the Alcohol sales growth. The seasonal Dummies featuring the lagged alcohol also captured the main trend that characterizes the target variable; specifically, we found that February and December have the most significant betas, signaling that the mean effect in this 2 months is highly persistent and thus help forecast the next period. On another note, the VAR model including other predictors such as Food & Beverages, Tobacco, Income and Unemployment, resulted weaker than ARMA ans seasonal dummies, failing to capture the main trend of Alcohol growth. Finally, the combination of the ARMA(12, 5) and Seasonal Dummies+Lagged Alcohol models produced an even better forecast, which is a fantastic result considering that they were both already great predictors; the Diebold Mariano test helped identify which models were most suitable to the combination through an easy and efficient way of assessing forecast accuracy.