ALCOHOL CONSUMPTION ON CORRELATING FEATURES

Jonathan Li
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
jhl3xn@virginia.edu

Nicholas Talton
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
nrt3xs@virginia.edu

Justin Zhang
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22904
jmz8rm@virginia.edu

December 9, 2022

Contents

List of Figures

1 Abstract

Alcohol consumption has a significant impact on a person's life and can have far-reaching consequences. Our goal was to create a machine learning model that relates socioeconomic factors in a particular region to the prevalence of alcohol consumption in that area, with the aim of helping to inform policy decisions on which aspects of a community should be targeted for the prevention of alcohol abuse and addiction. We used two datasets: Virginia socioeconomic data from the U.S. Census Bureau and VABC revenue by store in Virginia. After preprocessing, we tried various models we learned about in class for regression, such as linear, random forest, and neural networks. We found the most important factors to alcohol consumption to be being divorced or being widowed. Overall, we can treat the findings of our model as areas of interest, but our model is not a definitive authority on what factors affect alcohol consumption.

2 Introduction

2.1 Overview

Alcohol consumption has a significant impact on a person's life and can have far-reaching consequences. Physical health, happiness, quality of relationships, and economic status are all variables that are directly or indirectly connected to alcohol use. On average, the higher the alcohol consumption in any given region, the more attention is needed to address the individual and collective effects of alcoholism. To this end, our goal was to create a model that relates socioeconomic factors in a particular region to the prevalence of alcohol consumption in that area, with the aim of helping to inform policy decisions on which aspects of a community should be targeted for the prevention of alcohol abuse and addiction. Such policies may include increased access to treatment facilities, increased public awareness campaigns, and increased access to social and medical services, among others. By focusing on the most influential factors, we can work towards a future of reduced alcohol consumption and improved quality of life for all.

We plan to quantitatively test correlations between socioeconomic factors and the alcohol consumption that influences the overall well-being of a community. From the patterns and correlations we find, we hope this data can be used to influence future policy around Virginia and make it more effective.

2.2 Hypothesis

By using machine learning we hope to gain a new interpretation of correlations with communities of people that directly relate to alcohol consumption. Using machine learning allows us to find new correlations that we perhaps did not realize were correlated with one another as if we were to use just manual analysis of this dataset. We do know that alcoholism is higher in groups such as the less educated, financially unstable, and socially/romantically unstable, so we expect the top factors to include at least some of these.

2.3 Related Work

An overview of the genetic susceptibility to alcoholism[3]

Classification of EEG signals to detect alcoholism using machine learning techniques[5]

Using Machine Learning to Identify and Investigate Moderators of Alcohol Use Intervention Effects in Meta-Analyses[4]

From our preliminary research, it seems that most research on alcoholism using machine learning has been done on an individual level, where certain features of an individual are used to predict a patient's susceptibility to alcoholism. However, there has not been much research using machine learning to examine larger demographic trends in alcoholism to inform policy.

3 Method

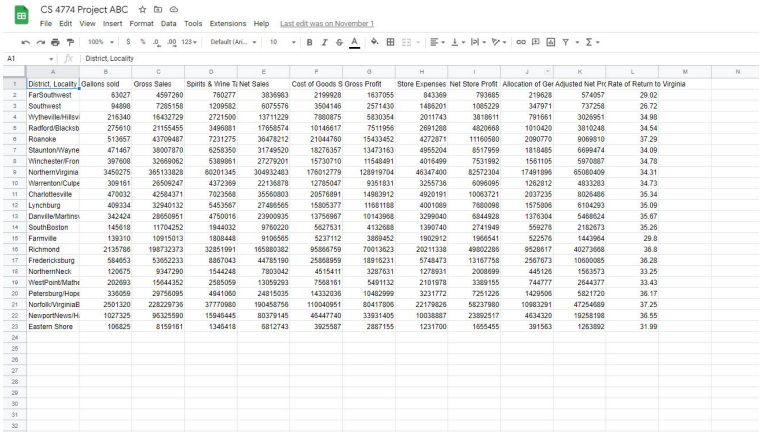
3.1 Data

We used two datasets: Virginia socioeconomic data from the U.S. Census Bureau and VABC revenue by store in Virginia. We combined these datasets with relevant features that could be correlated with alcohol consumption and abuse. The links to our datasets are: VABC sales[6], Income[1], Age and sex[2].

The demographic data we have is from the American Community Survey (ACS), a monthly survey (ran by the Census Bureau) of around 1% of homes in the US. The survey asks questions more detailed than the Census, so it would help us find more features that might be useful. As the Census website says itself, the ACS "provides local and national leaders with the information they need for understanding local issues and conditions," which reflects the purpose of this project.

The biggest issue we had is that our data does not have a 1-to-1 fit between the two data sets. Census data can be split into regions and counties, but the data from Virginia ABC stores are not split in the same manner. Because of this, we knew we needed to come up with a better way to pair each data with each other. Without the format of the rows of data matching as close as possible, the results may have not been accurate. We decided to split the data from the ACS and VABC stores into regions, as that both datasets could be sorted into regions. We cleaned the data and were able to get it into a Pandas format. We combined all of the datasets we have (ACS data and ABC data) into one CSV file with Pandas for convenience. The Virginia ABC revenue data had no missing values to impute, and because it's the thing we are trying to predict, we don't need to scale the data. From ACS data, we imported all four datasets we plan on using (tables DP02-DP05 on data.census.gov separated by counties in Virginia) into Pandas and combined all of the datasets into one table.

We found some missing values that needed to be imputed, so we imputed with the median where necessary. We did a one-hot encoding on all categorical data. Finally, we scaled all of our data (except our output labels) to have a mean of 0 and a variance of 1. Our labels, the values being predicted, were gallons of alcohol sold from ABC stores. We do realize that ABC stores are not the only sellers of alcohol but for this experiment, but we will assume the majority of alcohol sales are from ABC stores. Features we believed were extraneous and could either be captured by other variables or were irrelevant for our purpose of policy-making, such as use of computers, language spoken at home, place of birth, etc, were dropped.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	District, Locality	Gallons sold	Gross Sales	Spirits & Wine Ta	Net Sales	Cost of Goods S	Gross Profit	Store Expenses	Net Store Profit	Allocation of Ger	Adjusted Net Pr	Rate of Return to Virginia		
2	FarSouthwest	63027	4597260	760277	3836983	2199928	1637055	843369	793685	219628	574057	29.02		
3	Southwest	94898	7285158	1209582	6075576	3504146	2571430	1486201	1085229	347971	737258	26.72		
4	Wytheville/Hillsvi	216340	16432729	2721500	13711229	7880875	5830354	2011743	3818611	791661	3026951	34.98		
5	Radford/Blacksb	275610	21155455	3496881	17658574	10146617	7511956	2691288	4820668	1010420	3810248	34.54		
6	Roanoke	513657	43709487	7231275	36478212	21044760	15433452	4272871	11160580	2090770	9069810	37.29		
7	Staunton/Wayne	471467	38007870	6258350	31749520	18276357	13473163	4955204	8517959	1818485	6699474	34.09		
8	Winchester/Fron	397608	32669062	5389861	27279201	15730710	11548491	4016499	7531992	1561105	5970887	34.78		
9	NorthernVirginia	3450275	365133828	60201345	304932483	176012779	128919704	46347400	82572304	17491896	65080409	34.31		
10	Warrenton/Culpe	309161	26509247	4372369	22136878	12785047	9351831	3255736	6096095	1262812	4833283	34.73		
11	Charlottesville	470032	42584371	7023568	35560803	20576891	14983912	4920191	10063721	2037235	8026486	35.34		
12	Lynchburg	409334	32940132	5453567	27486565	15805377	11681188	4001089	7680098	1575806	6104293	35.09		
13	Danville/Martinsv	342424	28650951	4750016	23900935	13756967	10143968	3299040	6844928	1376304	5468624	35.67		
14	SouthBoston	145618	11704252	1944032	9760220	5627531	4132688	1390740	2741949	559276	2182673	35.26		
15	Farmville	139310	10915013	1808448	9106565	5237112	3869452	1902912	1966541	522576	1443964	29.8		
16	Richmond	2135786	198732373	32851991	165880382	95866759	70013623	20211338	49802286	9528617	40273668	36.8		
17	Fredericksburg	584653	53652233	8867043	44785190	25868959	18916231	5748473	13167758	2567673	10600085	36.28		
18	NorthernNeck	120675	9347290	1544248	7803042	4515411	3287631	1278931	2008699	445126	1563573	33.25		
19	WestPoint/Mathe	202693	15644352	2585059	13059293	7568161	5491132	2101978	3389155	744777	2644377	33.43		
20	Petersburg/Hope	336959	29756095	4941060	24815035	14332036	10482999	3231772	7251226	1429506	5821720	36.17		
21	Norfolk/VirginiaB	2501320	228229736	37770980	190458756	110040951	80417806	22179826	58237980	10983291	47254689	37.25		
22	NewportNews/H	1027325	96325590	15946445	80379145	46447740	33931405	10038887	23892517	4634320	19258198	36.55		
23	Eastern Shore	106825	8159161	1346418	6812743	3925587	2887155	1231700	1655455	391563	1263892	31.99		
24														
25														
26														
27														
28														
29														
30														
31														
32														

Figure 1: VABC Data

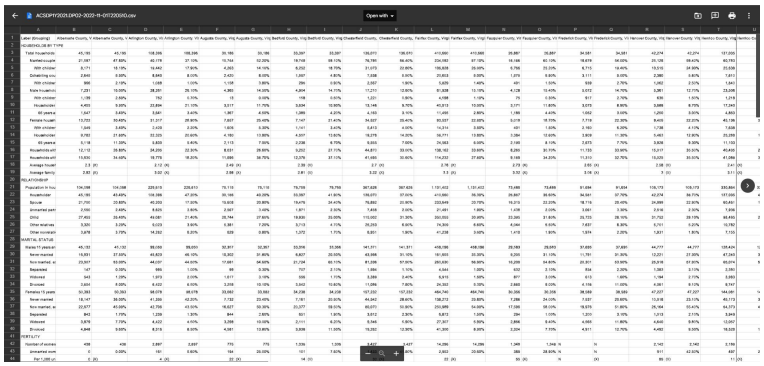
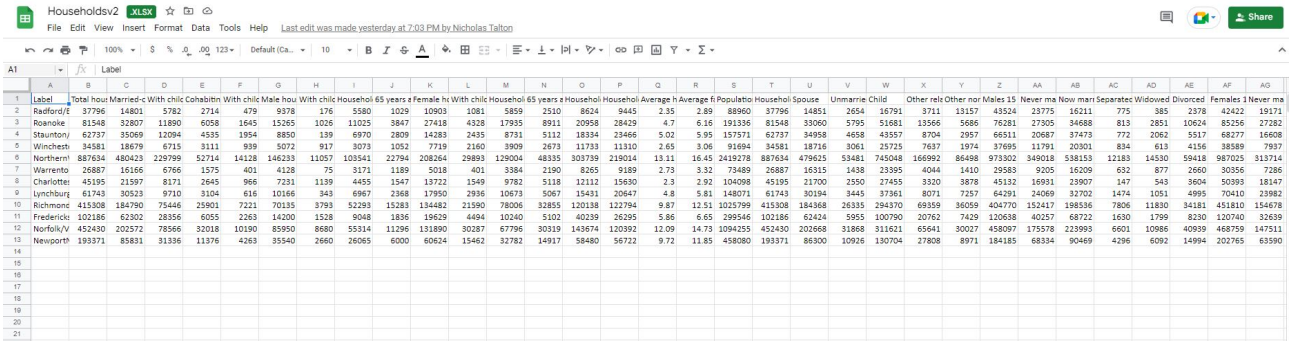


Figure 2: Census Data



Label	Total hous	Married-c	With child	Cohabitin	With child	Male hou	With child	Household	65 years a	Female ho	With child	Household	65 years a	Household	Household	Average h	Average fa	Populatio	Household	Spouse	Unmarrie	Child	Other rela	Other nor	Males 15	Never ma	Now marr	Separatec	Widowed	Divorced	Females 1	Never ma
Radford/E	37796	14801	5782	2714	479	9378	176	5580	1029	10903	1081	5859	2510	8624	9445	2.35	2.89	88960	37796	14851	2654	16791	3711	13157	43524	23775	16211	775	385	2378	42422	19171
Roanoke	81548	32807	11890	6058	1645	15265	1026	11025	3847	27418	4328	17935	8911	20958	28429	4.7	6.16	191336	81548	33060	5795	51681	13566	5686	76281	27305	34688	813	2851	10624	85256	27282
Staunton/	62737	35069	12094	4535	1954	8850	139	6970	2809	14283	2435	8731	5112	18334	23466	5.02	5.95	157571	62737	34958	4658	43557	8704	2957	66511	20687	37473	772	2062	5517	68277	16608
Winchest	34581	18679	6715	3111	939	5072	917	3073	1052	7719	2160	3909	2673	11733	11310	2.65	3.06	91694	34581	18716	3061	25725	7637	1974	37695	11791	20301	834	613	4156	38589	7937
Northern	887634	480423	229799	52714	14128	146233	11057	103541	22794	208264	29893	129004	48335	303739	219014	13.11	16.45	2419278	887634	479625	53481	745048	166992	86498	973302	349018	538153	12183	14530	59418	987025	313714
Warrento	26887	16166	6766	1575	401	4128	75	3171	1189	5018	401	3384	2190	8265	9189	2.73	3.32	73489	26887	16315	1438	23395	4044	1410	29583	9205	16209	632	877	2660	30356	7286
Charlottes	45195	21597	8171	2645	966	7231	1139	4455	1547	13722	1549	9782	5118	12112	15630	2.3	2.92	104098	45195	21700	2550	27455	3320	3878	45132	16931	23907	147	543	3604	50393	18147
Lynchburg	61743	30523	9710	3104	616	10166	343	6967	2368	17950	2936	10673	5067	15431	20647	4.8	5.81	148071	61743	30194	3445	37361	8071	7257	64291	24069	32702	1474	1051	4995	70410	23982
Richmonc	415308	184790	75446	25901	7221	70135	3793	52293	15283	134482	21590	78006	32855	120138	122794	9.87	12.51	1025799	415308	184368	26335	294370	69359	36059	404770	152417	198536	7806	11830	34181	451810	154678
Frederick	102186	62302	28356	6055	2263	14200	1528	9048	1836	19629	4494	10240	5102	40239	26295	5.86	6.65	299546	102186	62424	5955	100790	20762	7429	120638	40257	68722	1630	1799	8230	120740	32639
Norfolk/V	452430	202572	78566	32018	10190	85950	8680	55314	11296	131890	30287	67796	30319	143674	120392	12.09	14.73	1094255	452430	202668	31868	311621	65641	30027	458097	175578	223995	6601	10986	40939	468759	147511
Newport	193371	85831	31336	11376	4263	35540	2660	26065	6000	60624	15462	32782	14917	58480	56722	9.72	11.85	458080	193371	86300	10926	130704	27808	8971	184185	68334	90469	4296	6092	14994	202765	63590

Figure 3: Joined Table (VABC and Census)

3.2 Models run

Since we are trying to predict alcohol sales/consumption with features, this is a regression problem and we needed to train regression models. We planned on using linear regression, random forest, and neural networks for this problem. Because of the limited number of regions in Virginia (<10), we need to prevent overfitting. To do this, we should reduce the complexity of the models while looking for the same performance, and conduct cross-fold validation or a train-test split, depending on how long our models take to train and run.

4 Experiments

In order to gauge the baseline performance of our model, the simpler models of linear regression and random forest regression were trained using 12-fold cross-validation on the Joined Table data. We chose these simpler models first as opposed to other more advanced models because they are relatively simpler models and do not require many hyper-parameters to tune. This gave us a general understanding of how these simple models fit the data.

```
[33]  1 from sklearn.linear_model import LinearRegression
      2 from sklearn.model_selection import cross_val_score
      3
      4 regr = LinearRegression()
      5 regr.fit(data_x_sklearn, data_y_sklearn)
      6 lin_scores = cross_val_score(regr, data_x_sklearn, data_y_sklearn,
      7                               scoring="neg_mean_squared_error", cv=12)
      8 lin_rmse_scores = np.sqrt(-lin_scores)
      9 display_scores(lin_rmse_scores)
```

```
Scores: [0.10355929 0.03123252 0.02089243 0.07005791 0.48675726 0.04317711
 0.0518543  0.01625493 0.03236278 0.10101525 0.37247457 0.10207441]
Mean: 0.11930939612050494
Standard deviation: 0.1438708105097129
```

```
1 from sklearn.ensemble import RandomForestRegressor
2
3 regr = RandomForestRegressor()
4 regr.fit(data_x_sklearn, data_y_sklearn)
5 rf_scores = cross_val_score(regr, data_x_sklearn, data_y_sklearn,
6                               scoring="neg_mean_squared_error", cv=12)
7 rf_rmse_scores = np.sqrt(-rf_scores)
8 display_scores(rf_rmse_scores)
9
```

```
Scores: [0.10102256 0.01019727 0.03633315 0.04607457 1.19152649 0.05325424
 0.13119278 0.12141895 0.15456234 0.03214606 0.55380402 0.31134998]
Mean: 0.22857353511934161
Standard deviation: 0.32530604238381244
```

Figure 4: Sklearn models: Linear Regression and RandomForestRegressor

After these preliminary experiments, moving forward, we used more advanced models that can capture more patterns from our features. We took our dataset, marked a section of our dataset as labels, and split the test data into training and testing.

We ran a more complex model, a regression artificial neural network (ANN). We ran this experiment in hopes to find a better correlation to work with for analysis. Our design for the architecture of the neural network consisted of taking in all features initially and then consolidating the features into 120 neurons, then 60, then into a single one-dimensional output. We arrived at this architecture by choosing an estimate for the optimal number of nodes to be 2/3 the input size. We hoped that this result would be able to predict a better model than our preliminary experiments and better show any features that stand out from other ones.

```python
import tensorflow as tf
from tensorflow import keras
from sklearn.metrics import confusion_matrix, accuracy_score

model = keras.Sequential([
    keras.layers.Dense(units=448),
    keras.layers.Dense(120, activation='relu'),
    keras.layers.Dense(60,activation='relu'),
    keras.layers.Dense(1,activation='linear'),
])

model.compile(loss='mean_squared_error',metrics=['mean_squared_error'],optimizer=keras.optimizers.Adam(learning_rate=0.0001),)
history = model.fit(X_train, y_train, epochs=150, validation_data=(X_val,y_val))
```

Figure 5: Architecture of the neural network we used

5 Results

After we ran the model for 150 epochs we found the model was returning a very small mean squared error. The data we had was scaled so it makes sense for the mean squared error to be very small as a result (MSE = 0.0168). As it turns out, the ANN model seemed to perform rather well compared to the previous models we have tried. Because of this, we believe that ANNs would be good at predicting alcohol consumption from demographic data.

Because neural networks introduce nonlinearity into the data, feature importance is less defined. However, anyone can use a goal demographic and use that as an input into the model and see how that would affect the predicted alcohol consumption (ex. feed the ANN data with demographics with better education and see if predicted alcohol consumption goes up or down).

```
Epoch 146/150
1/1 [==============================] - 0s 38ms/step - loss: 3.0206e-08 - mean_squared_error: 3.0206e-08 - val_loss: 0.0168 - val_mean_squared_error: 0.0168
Epoch 147/150
1/1 [==============================] - 0s 46ms/step - loss: 1.9815e-08 - mean_squared_error: 1.9815e-08 - val_loss: 0.0168 - val_mean_squared_error: 0.0168
Epoch 148/150
1/1 [==============================] - 0s 40ms/step - loss: 3.7671e-08 - mean_squared_error: 3.7671e-08 - val_loss: 0.0168 - val_mean_squared_error: 0.0168
Epoch 149/150
1/1 [==============================] - 0s 43ms/step - loss: 2.6357e-08 - mean_squared_error: 2.6357e-08 - val_loss: 0.0168 - val_mean_squared_error: 0.0168
Epoch 150/150
1/1 [==============================] - 0s 36ms/step - loss: 3.5426e-08 - mean_squared_error: 3.5426e-08 - val_loss: 0.0168 - val_mean_squared_error: 0.0168
```

Figure 6: MSE of the model after 150 epochs

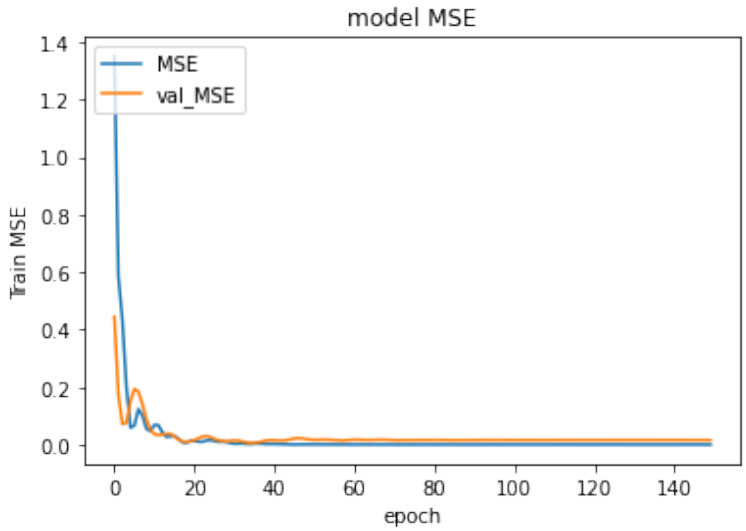


Figure 7: Performance of the neural network over the epochs

6 Discussion & Conclusion

6.1 Discussion

We found the most important factors to alcohol consumption to be being divorced or being widowed. This was expected and matched our hypothesis, as these have one thing in common: they are hardships. Hardships lead to a higher chance of alcoholism or just alcohol consumption in general.

This result shows that even with our assumptions, our model was still able to pick up on these trends, which means that our assumptions are not that problematic. However, part of the problem with our model was the low number of data points. The default way the VABC store data was organized was into huge regions, of which there were less than 10. Because of this, we had very limited data. However, overfitting didn't seem to be that much of a problem, because our validation error still went down as our training error went down, so we still feel that we can analyze our results confidently.

6.2 Conclusion

Overall, we can treat the findings of our model as areas of interest, but our model is not a definitive authority on what factors affect alcohol consumption. In the future, we could try to make our model more robust with more data points by reorganizing the VABC store data or generalizing our model with ABC stores around the country to inform our analysis on Virginia alcohol consumption.

6.3 Limitations

This research problem is a complex problem to solve, as predicting alcohol consumption involves a variety of factors such as cultural norms, economic factors, and demographic variables. This is possible to do, however, it would require a lot of data and perhaps multiple machine learning models to tackle. Additionally, to make the model be more precise in the future we could include more variables from other sources, such as local laws related to alcohol consumption and the availability of alcohol in stores and restaurants.

7 Team Members and Contributions

7.1 Jonathan Li

Data Preprocessing, program script writing, video recording and editing, and final report (abstract, preliminary experimentation, figures).

7.2 Nicholas Talton

Data preprocessing, video recording, final report (methods and next steps, figures, LateX outline).

7.3 Justin Zhang

Model training, acquisition of VABC and Census data, importing data to jupyter, final report (conclusions, abstract), Video recording.

8 Citations and References

1.) Bureau, U. S. C. (n.d.). S1901INCOME IN THE PAST 12 MONTHS (IN 2021 INFLATION-ADJUSTED DOLLARS). Explore census data. Retrieved December 7, 2022, from https://data.census.gov/table?q=virginia

2.) Bureau, U. S. C. (2020). S0101AGE AND SEX. Explore census data. Retrieved December 7, 2022, from https://data.census.gov/table?q=virginia

3.) Buscemi L, Turchi C. An overview of the genetic susceptibility to alcoholism. Medicine, Science and the Law. 2011;51(1_suppl):2-6. doi:10.1258/msl.2010.010054

4.) Parr, N. J., Loan, C. M., & Tanner-Smith, E. E. (n.d.). Using Machine Learning to Identify and Investigate Moderators of Alcohol Use Intervention Effects in Meta-Analyses. Academic.oup.com. Retrieved December 7, 2022, from https://academic.oup.com/alcalc/article/57/1/26/6272648?login=true

5.) Rodrigues, J. das C. et al. (2019, April 22). Classification of EEG signals to detect alcoholism using machine learning techniques. Pattern Recognition Letters. Retrieved December 7, 2022, from https://www.sciencedirect.com/science/article/pii/S0167865519301266?casa_token=RC8w7FsFSdgAAAAA %3Amd-JnjXkGogIJP9ud3fJxoovbfRAV2EEpNwDiMykr5bcjCIlgyDKW1umsMELBcRvgKJOnFHHvIag

6.) Virginia Alcoholic Beverage Control Authority. (2021). Virginia ABC. Retrieved December 7, 2022, from https://www.abc.virginia.gov/