STOR 565 Group Project

By Barrett Buhler, Ethan Wood, Daniel Meskill, Love Myoung, Hamza Khan

Project Overview

In this project, we wanted to discover trends within grocery store density that could help us predict health outcomes as well as outline areas that could benefit from access to healthy foods. Specifically, the noticeable disparities in health outcomes that occur in food deserts, which we define as areas with low access to healthy foods. To do so, we employed machine learning techniques such as linear models, extreme gradient boosting, random forests, and tree-based methods to find a model with a root mean squared error of 0.001782068 on the testing dataset.

We explored data from the National Neighborhood Data Archive (NaNDA): Grocery Stores by County 2003-2017, created by the Social Science department at the University of Michigan. It includes counts of grocery, specialty, and warehouse stores, as well as number of residents and land ratios of each census tract which is a kilometer shaped block. Speciality stores can be defined as stores such as meat, seafood, and produce markets as well as bakeries or spice stores. Warehouse stores can be defined as "buy-in-bulk" stores such as Costco or Sam's Club. To this dataset, we joined data from the Institute for Health Metrics and Evaluation which has information regarding mortality rates by causes of death in America by county using the county FIPS codes. In order to join the kilometer block level FIPS codes to county mortality rate we concentrated our analysis to the county level to correct this. We found that 53.35% of variability when we graphed our predicted values of overall mortality rate based on grocery store data and we compared it to the actual values. This shows that there are significant health disparities in areas without access to grocery stores. With that said, these differences exist, but several other factors should be considered.

AI Usage

I attest that this project made use of AI in the following ways:

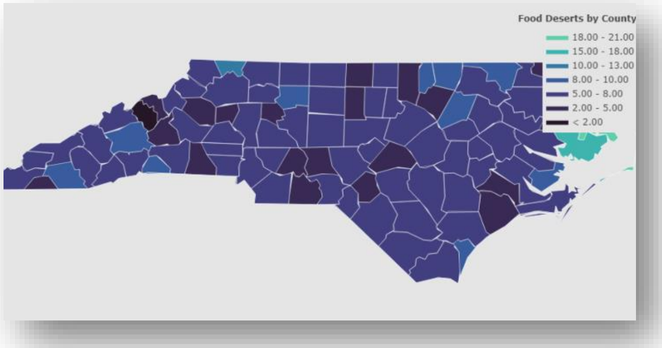
Usage	Tool Used (e.g. ChatGPT-4)	How you edited the output, if at all	Conversation Link (if available)
Research	ChatGPT 3.5	We utilized ChatGPT solely to identify syntax errors in our code, largely in the data merging and plotting	N/A

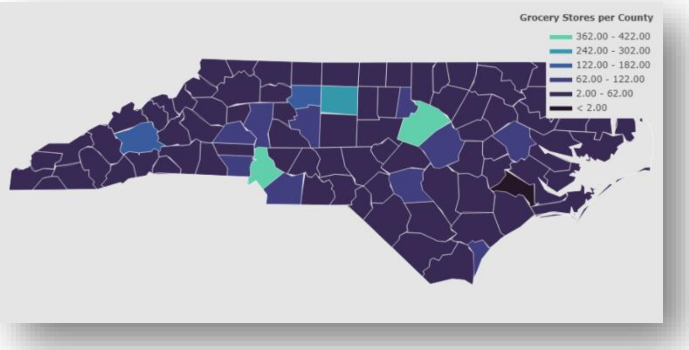
Exploratory Data Analysis

Firstly, what is a food desert? A food desert is defined by the United States Department of Agriculture as an area with low access to healthy foods, often within low-income communities. This is measured by population density across various regions of the US, with a rural food desert being defined as areas with a grocery store every 33 miles and an urban food desert as areas with a grocery store every one mile. To do some exploratory data analysis, we mapped all the counties in North Carolina by number of grocery stores as well as another map showing areas defined above as food deserts. We did this to see if we could draw any similarities between areas of high grocery store concentration and areas marked as food deserts.

```
    fig = ff.create_choropleth(
        fips=fips,
        values=values,
        scope=[state],
        show_state_data=True,
        colorscale=colorscale,
        binning_endpoints=endpts,
        round_legend_values=True,
        plot_bgcolor='rgb(229,229,229)',
        paper_bgcolor='rgb(229,229,229)',
        legend_title='Grocery Stores per County',
        county_outline={'color': 'rgb(255,255,255)', 'width': 0.5},
        exponent_format=True,
    )
make_state_plot('North Carolina',2017,df2,'37')
```

The top map is the map of food deserts and on the bottom, a map of counties by grocery stores. Immediately, we can see that the areas around the northern outer banks are considered food deserts by the food desert map and show a very low number of grocery stores on the grocery store map. This tracks with intuition as the areas with lower grocery store counts seem more likely to be food deserts. Similarly, that region of North Carolina is sparsely populated to begin with. As we look towards the more central counties as well, we see lower measures on the food desert index and higher number of grocery stores. This gives us enough confidence in our dataset to move forward.

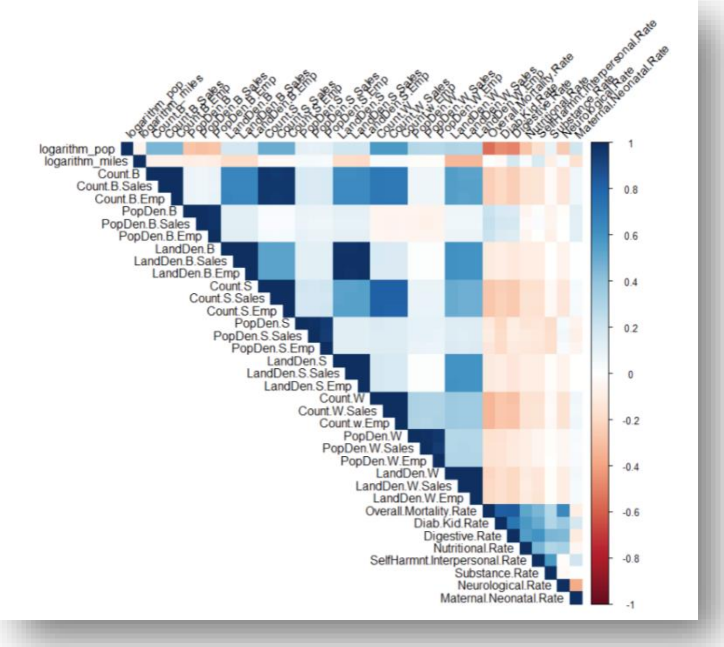




When looking at specific variables in our dataset, we felt it was appropriate to rescale some of the data. For instance, when looking at the population sizes for various counties, we find extreme skewing which may artificially affect any models made with that data. This is also true for county sizes by square mileage. We performed log-transformations to combat this and provide more normalized data.

Finally, before we attempted to make predictive models, we ran a correlation heat-map to visually check for relationships between variables, depicting the strength and direction of correlations through color intensity. We found that many of the variables had interesting behaviors when interacting with each other. In particular, the various mortality rates were showing ties with our store data. Below is the correlation heat-map.

```r
cor_matrix <- cor(datagrocery)
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45)
```

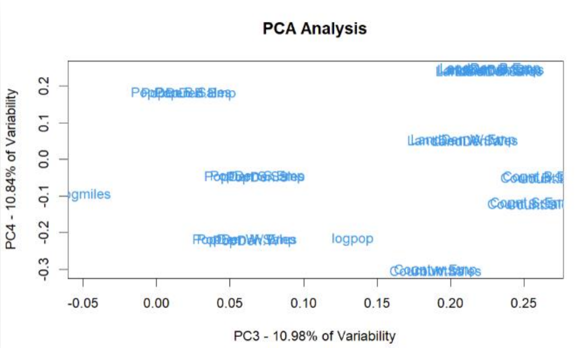


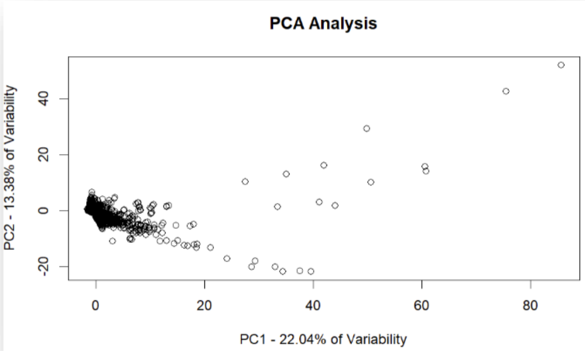
It is of note that we considered the effects of population size on our models. Intuitively, counties with higher population counts may inherently behave differently from lower population centers. To work around this, we ran our following models with total population with a log transformation as well as with data scaled for population density. We found no meaningful differences.

Clustering

One method that could be used to find food deserts would be classification. However, given that there is no variable that identifies a county as a food desert or not, this would be impossible.

Given that our goal is to find clusters of food deserts, we will only be clustering the grocery store dataset, without any of the health variables. To begin exploring our data via clustering, we first need to visualize it. Using Principal Component Analysis (PCA), we can reduce the variability in our data down to two dimensions so it can be viewed in a plot.





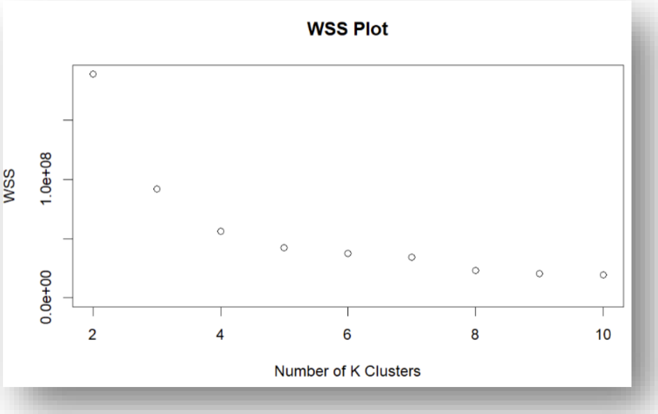
The first PC explains 22.04221% of our data, and the second PC explains 13.38344% of our data. We see that higher counts of special and basic stores, and higher land densities of these stores will make a county lie further to the right. As we can see when plotting the observations, many points lie around 0, indicating low counts of stores and land densities in these counties, while a few counties lie far away. Note the many variables that appear to lie on top of one another. This is due to the total counts, counts of stores with more than one employee, and counts of stores with more than 0 sales all having very similar variables.

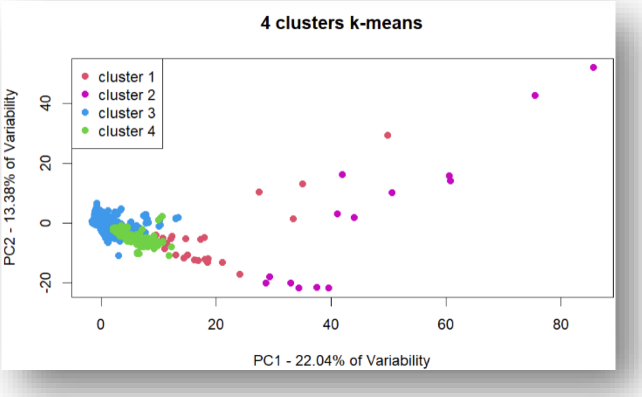
Reference the heatmap in the EDA to examine the strong correlation between these variables. Because these variables are so correlated and almost identical, we see that reducing dimensions keeps these variables together. We decided to scale our data before performing PCA due to most variables being in different units (# stores, # stores/1000 people, # stores/mile, log miles, etc). There were some variables in the same units, such as counts of basic stores and counts of basic stores with more than one employee, but because of their high correlation with one another we did not lose much information.

We see clusters of size 23, 14, 4306, and 118. We see the large cluster in blue lies around 0, with low counts of stores, then the cluster in green with slightly higher counts of stores. The two other small clusters in red and purple seem to indicate outliers with very large counts of stores. This is telling us we have a distinct cluster of most counties that have a low number of stores and land densities, and two small clusters that act as outliers for our dataset since they have distinctly different counts of stores and land densities. The last cluster is an intermediate cluster that contains points that are significantly different from

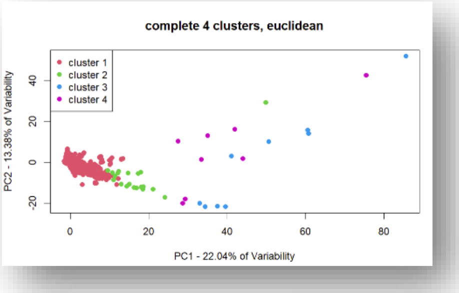
Now that we can properly view our data points, we can view our clusters. Starting with k-means using Euclidean distance, we reference the within sum squares plot to see what a proper number of clusters might be. The "elbow" or change in the slopes at k=4 clusters indicates that the distance between points in each cluster does not reduce as much when we add more clusters, meaning that 4 is likely the most significant. We use Euclidean distance for this preliminary cluster since it would measure the total distance between points. This will likely be low with counties that are food deserts.

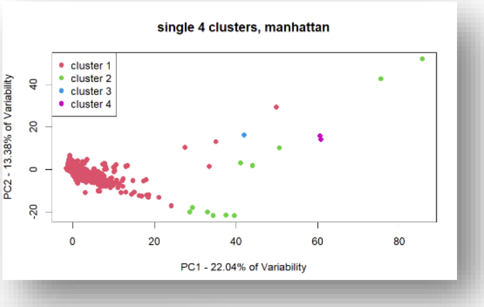
Most counties, but not as different as the counties in the other two clusters. However are these clusters legitimate? Or are we seeing them by chance?

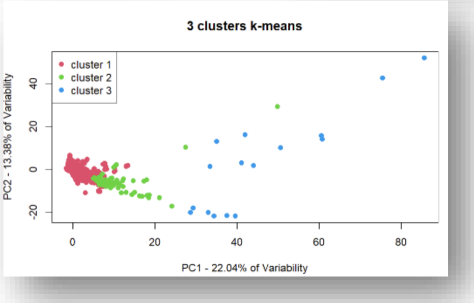




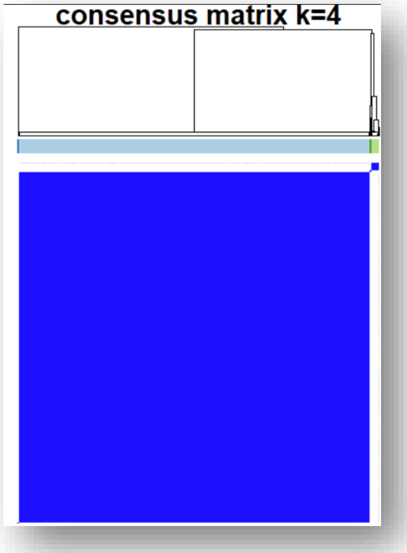
To examine the validity of this clustering, we will reference three different things. First, clustering using different methods and distances, another using consensus clustering, and lastly a real-world interpretation of the clustering. When viewing different distances and methods for clustering, we will look to see if the same clusters appear. If they do appear across multiple distances and methods, then it is likely that these clusters are not appearing randomly. Viewing k=3 and k=5 means clustering using Euclidean distances, we view similar clusters. K means clusters do not subset the clusters, so the fact we do not see vastly different clusters with different sizes *k* indicates legitimate clusters. With hierarchical clustering using complete and single linkage, and Euclidean and Manhattan distances for each, we also view similar clusters. However, the sizes of these clusters are much smaller in size. This makes sense because hierarchical clustering, specifically single linkage, is greatly affected by outliers. Hierarchical clustering also indicates a split between most counties with low counts of stores, and the few outliers with high counts.







Second, we will use consensus clustering. Consensus clustering is a way to test how often points appear in the same cluster by rerunning the clustering method many times with subsets of the data. The shade of blue implies the proportion of times a point lies with another point, darker meaning more often. Following the paper published by Matthew Wilkerson, we subset 80% of the data using k-means clustering and Euclidean distance and find very little overlap. We can interpret observations that overlap by seeing if there is lighter blue outside the dark blue regions. Because there is primarily blue and white outside these regions, we conclude that there is very little overlap between clusters, and that our clusters might be legitimate. However, the size of the cluster is something to note, and it is something that consensus clustering is able to point out well. The size of the main blue square indicates the cluster of the low grocery store counties, and further drives home the interpretation that this clustering separates the normal counties from the few outlier counties that have large counts of stores.



Due to these three methods supporting the legitimacy of our clusters, we feel comfortable saying that these are true clusters. Testing with more data could support this. However, these do not cluster the food deserts away from the regular counties, just the regular counties away from the food-rich counties. So, we will not continue using these clusters.

Lastly, if we consider what this cluster means when we interpret this data in the real world to see if this matches our intuition. We see very many points clustered around 0 with low store counts and low land densities, with ~96.5% in that one large cluster and the few counties with large counts of stores. If we consider North Carolina for example, we see that 98 out of the 100 counties have low counts of stores, except Wake and Mecklenburg counties, which host the two largest cities in North Carolina. Our clustering found by k means seems to match this ratio of that we see in NC, and thus we see a real-world interpretation of the clustering.

Results

Linear Models

RMSE Type	Model Type with 5 Fold Cross Validation				
	Least Squares	PCR	PLS	Ridge	LASSO
Train	0.002082669	0.002162490	0.002112556	0.002111003	0.002098883
Test	0.002079253	0.002161954	0.002106576	0.002106220	0.002088258

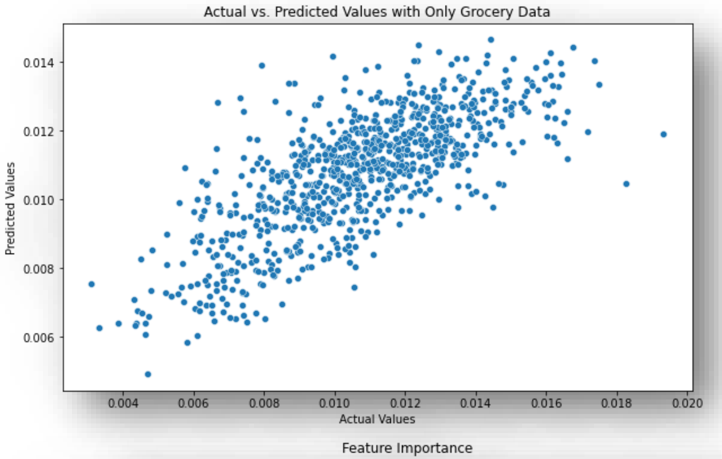
Table of Various Model RMSEs for Predicting Overall Mortality Rate

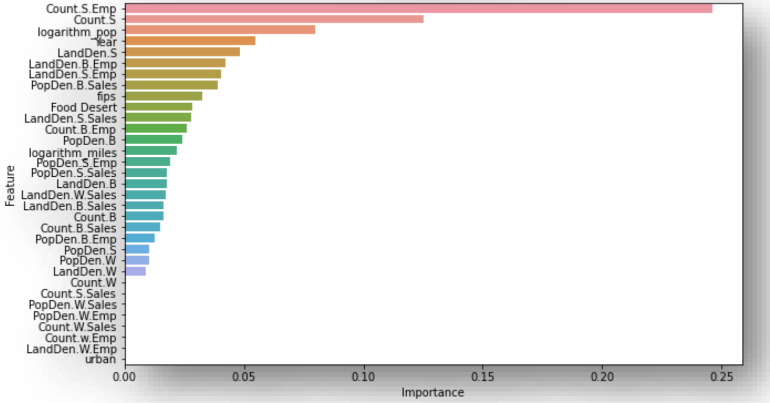
We next examine how well we can predict various health metrics with grocery store variables via common linear models. Initially, we utilized a multivariate least squares regression model as a baseline to compare against our other models of Principal Components Regression (PCR), Partial Least Squares (PLS), Ridge, and Least Absolute Shrinkage and Selection Operator (LASSO). We introduced the other models because we thought they could outperform the baseline model since many of our predictor variables were strongly correlated. We thought simplifying our model would generalize better than a model with many variables overlapping in predictive effect.

Above is a table of the Root Mean Square Errors (RMSE) for each model invoked when predicting the overall mortality rate. We applied a 5-fold cross validation across all of the models for consistency. Surprisingly, the least squares produced the best test RMSE. However, none of the models performed poorly at predicting overall mortality rate. We also note that the models generalized well signified by the little difference between train and test RMSE. Models were also created for predicting each of the other health metrics and we saw similar performance.

XGBoost Model

In our analysis, we decided to use XGBoost due to its predictive power and popularity in many machine-learning competitions. XGboost is a supervised learning algorithm that uses gradient-boosted decision trees to optimize its objective function. Out of the models constructed, XGBoost outperformed all of them significantly and we were able to obtain a root mean squared error of 0.00178 and an adjusted R squared of 51.25 percent.



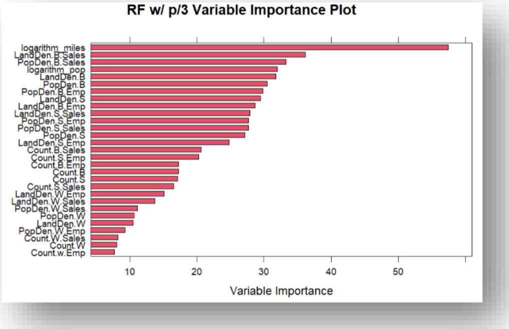


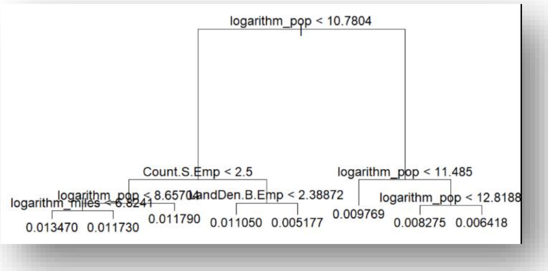
In examining the feature importance plot, we see that the most important variables in predicting the overall mortality rate are specialty stores. These types of stores are most likely found in high-income areas which could lead us to believe that income would also be a good predictor in predicting overall mortality rate.

Other Tree-Based Methods

We ran three additional tree-based methods in our analysis. We ran two random forests, one in which the number of features run was $^{p}/_{3}$ and another in which the number of features run was $\sqrt{\rho}$ with p being the number of variables we had in total. Additionally, we ran a pruned tree. The two random forest models were selected to check for any characteristics in our data that may have been missed by our XGBoost model and the pruned tree was simply selected for interpretability.

When run, our two random forest models performed very well, with testing Root Mean Square Errors of 0.00189 and 0.00193 respectively. However, they were outperformed by our XGBoost models. Our pruned tree performed the worst with a test root mean squared error of 0.00209, but this was in line with our expectations as we simply wanted an easily interpretable graph to see which variables may be most relevant. Below are the best performing Random Forest model as well as the pruned tree.





According to our best Random Forest model, the most important variables were our log-transformed square mile variable followed by the sales values by population density which is an interesting distinction from our XGBoost model. In our pruned tree, we see that the most important variables are the log-transformed population variable followed by employment numbers.

Limitations

There are three main limitations that we ran into. The first one was that there was a lack of income data. Since healthcare and grocery stores both operate with the use of monetary transactions, having information for this would help us be able to draw better conclusions. Secondly, the lack of depth in supermarket data was another limitation. The data doesn't include information regarding the nutritional content of the food, such as the availability of produce and processed foods. Lastly, we had some difficulty in trying to cluster food deserts. We can address this through adding more information about establishments other than just grocery stores.

There are three main areas where work can be done with this data such as for making policies, commercial growth and for future research. It can help in the realm of policy making due to it being simpler to track metrics like commercial establishments, land use and population instead of tracking individual health data. Businesses can benefit from this type of work as it can aid them in making decisions regarding where they should expand their business within the country. Further research can be done utilizing data from the NaNDA dataset which gives more information regarding stores like convenience stores, dollar stores and more which is useful in making additional predictive models.

All Code can be found on our GitHub

GitHub for the Project: https://github.com/barrettb/STOR565

Citations

Finlay, Jessica, Li, Mao, Esposito, Michael, Gomez-Lopez, Iris, Khan, Anam, Clarke, Philippa, and Chenoweth, Megan. National Neighborhood Data Archive (NaNDA): Grocery Stores by Census Tract, United States, 2003-2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-10-01. https://doi.org/10.3886/E123001V1

Institute for Health Metrics and Evaluation (IHME). United States Mortality Rates by Causes of Death and Life Expectancy by County, Race, and Ethnicity 2000-2019. Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME), 2023

Dutko, Paula, et al. “Characteristics and Influential Factors of Food Deserts - USDA ERS.” Www.Ers.Usda.Gov, Aug. 2012, www.ers.usda.gov/webdocs/publications/45014/30940_err140.pdf.

Wilkerson, Matthew D. “ConsensusClusterPlus (Tutorial).” Https://Bioconductor.Org/, 24 Oct. 2023, bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf.