I. Motivation:

Food waste is and always has been a widespread issue that continues to be tackled. In the United States, 119 billion pounds of food waste is produced each year. According to the EPA (Environmental Protection Agency), an estimated 58% of fugitive methane emissions in landfills are from landfilled food waste. Fresh produce waste greatly contributes to this statistic as Americans waste 46% of all fruits and vegetables whether it be from poor preservation methods or concluding that they were inedible due to their physical appearances. With this in mind, we aim to see how rainfall in major fruit production states affects the producer price index (PPI) for fruits, and in turn, affects the amount of predicted waste of fruits. PPI measures the average change in time in selling prices, or the price change from the seller's perspective. Specifically, we are looking at annual sales of grapes, apples, and oranges and how they are affected by annual precipitation in the US. We hope that with this detailed analysis, we will be able to predict future amounts of waste as well as future producer price indexes.

II. Data:

We had several main datasets. The first dataset we focused on was annual precipitation. We found this dataset on Kaggle, titled "US Weather Events (2016 - 2022)," which provided a countrywide dataset of 8.6 million weather events from the years 2016 to 2022. This contained 14 columns of information—including the type of precipitation (rain, snow, fog, etc.), start time, end time, state, timezone, and average precipitation in inches (among other information). For our purposes, we only needed the type of precipitation, start time, state, timezone, and average precipitation in inches. However, since the dataset was so large, we first had to filter the dataset before exporting the data. The first step in this was to filter by the type of precipitation, which we narrowed down to only include "Rain." Next, we split up the data by year. To achieve this, we split the "Start Time" column (which included Month, Day, Year, and Time) into their own respective columns. From there, we exported the data from each year separately into Excel.

After the data was filtered into a more manageable size, we imported it into Deepnote for further cleaning. Starting with the 2016 dataset, we created a new column called "Combined," which combined the month and year for each row of data. From there, we applied the `group_by` function to the dataset, grouping by "Combined," "State," and "TimeZone" to take the mean of the precipitation. Afterward, we dropped all the additional columns in the dataset, leaving us with only "Combined," "State," "TimeZone," and "Precipitation (in)." We repeated this process for each of the years 2017 to 2022. Once all the data was compiled and cleaned, we used `concat` to compile all the years into one dataset.

Our second dataset focused on dollar sales of fresh fruit focusing on grapes, various types of apples, and various types of oranges. We found one dataset on the USDA's Economic Research Service (ERC) website, titled "Fruits and Tree Nuts Data," encompassing data for apples and oranges. Grapes were not available in this dataset, so we used another dataset from the Federal

Reserve Bank of St. Louis specifically for grapes. Once we combined these two datasets to create one large dataset with apples, oranges, and grapes, we had data containing the type of fruit, year, month, and Producer Price Index (PPI). After combining the datasets, we had an Excel spreadsheet titled "142 ppi," which we then read and exported into Deepnote, making it ready for data cleaning and modeling. We first cleaned this data by combining the year and month columns/rows into one column, titled "year_month" to make our time splitting easier—this is due to the years being in multiple rows, while the months were stored in columns. Next, we placed all the data points of PPI into one column, titled "ppi." Lastly, we removed the rows containing data for Rome Apples, as there were no values listed for this type of fruit. Once we had all of this data cleaned, we were left with three columns, titled "type of fruit," "year_month," and "ppi."

Once both the precipitation and the PPI dataset were created, we combined the tables on the "Date" value (which for precipitation was titled "Combined" and for PPI was titled "year_month") into a new table titled **rainppi**. From there, we added a "Season" column, classifying each month as one of the four seasons, so that each season contained three months.

Following this, we have our third dataset, titled "food waste." This data was primarily found on the United States Department of Agriculture (USDA) Economic Research Service (ERS), which we downloaded as an Excel spreadsheet and loaded into our Notebook 3. This dataset contained information such as the years (1970-2022), primary weight of the fruit, loss from primary weight, retail weight, loss from retail weight, consumer weight, edible weight, cooking loss, and fruit. To clean this dataset, we first limited our food waste statistics to 2019 and upwards to match the years on our rainfall and PPI dataset (**rainppi**). We then merged these two datasets into one dataset, titled "food_rainppi," additionally filtering out the year 2022 to match our years and merge the datasets accordingly. Next, we selected the columns "Year," "Fruit," "TimeZone," "Precipitation(in)," "Primary weight2," and "Retail weight." We then created a new column titled "Food_Loss" which consisted of the difference between our "Primary weight2" and "Retail weight" columns. Lastly, in our final step of cleaning this dataset, we dropped the "Primary weight2" and "Retail weight" columns, as we had their difference stored in our "Food_loss" column that we had just created.

III. Analytic Models:

Before beginning to build any of our models, we first split our first combined precipitation dataset into a training and testing set, using train_test_split, and using the 'ppi' column as our dependent "y" variable and everything else as our independent "X" variables. We then created a baseline model, finding the average PPI and baseline accuracy to be 195.0583 and 1747.29546, respectively.

We repeated this process for our second combined dataset, found in Notebook 3, titled "food_rainppi," using the "Food_Loss" column as our dependent "y" variable and the rest of our columns as our independent "X" variables. Our average food loss was found to be around 0.55915 and baseline accuracy was found to be around 0.046537. For both datasets, our test size was 0.33, indicating our test sets to be ⅓ of our total data and the training sets to be the remaining ⅔ of our data. Additionally, we plotted our dummy variables for both, using `colorbar` and background gradient techniques to do so.

1) **OLS**

Starting with OLS, otherwise known as Ordinary Least Squares, we began our regression model. By fitting OLS onto our y and X variables, we printed a summary of our OLS Regression Results, finding both our regular and adjusted R-squared values being 0.972. Given these values, we can see that using OLS indicates a good fit.

Further, we did this for our combined dataset titled "food_rainppi," where we merged our food waste, precipitation, and ppi datasets. Here, our adjusted R-squared value was slightly lower than that of our first dataset with just rainfall, with this one being 0.991, meaning a good R-squared reading. Moving onto linear regression, we found that our accuracy for "food_rainppi" was 0.988957, indicating very strong accuracy with this model, in addition to finding our intercept, coefficients, Mean Squared Error (MSE), Root Mean Squared Error, and Mean Absolute Error. Since this adjusted R squared was above 0.8 as well, we can see this is a good fit.

2) **CART**

We then moved on to using Decision Trees with CART, or a classification and regression tree, as a model to show how our outcome's variables, ppi, can be predicted based on our independent variables. By fitting our model onto our X and y training sets and predicting our X test set, we found the Mean Squared Error of our Decision Tree Model to be around 1179.55728. Following this, we visualized our Decision Tree, plotting our tree to have around 61 nodes. The large MSE suggests that the model may have struggled to accurately predict PPI based on the given independent variables, highlighting potential challenges or complexities in the underlying data relationships.

We also used this modeling method for our combined food waste, rainfall, and ppi dataset, with a Decision Tree MSE of around 0.0022357. We then visualized this, finding that we had the most splits on the fruit type, precipitation, and state. This implies that these factors play a significant role in predicting the combined outcome. The low MSE suggests that the model performed well in capturing the patterns within the dataset, potentially indicating a more straightforward relationship between the selected independent variables and the target outcome.

3) Elastic Net

We first used an elastic net on our "**rainppi**" dataset, finding our best hyperparameters to be "alpha" with a value of 0.1, and "l1_ratio" with a value of 1. Additionally, we found our MSE to be around 635.916, Root MSE at around 25.217, and MAE at around 19.0079684, suggesting a reasonable predictive performance.

We also used an elastic net on the "**food_rainppi**" set, as well, as finding our best variables and hyperparameters in the process. Using this method, we found our best hyperparameters to be "alpha" at a value of 0.1 and "l1_ratio" at 0.3, with an MSE of around 0.046, Root MSE at around 0.21455137, and an MAE of 0.198506 in contrast to the "rainppi" dataset, indicating a notably improved predictive accuracy for this specific dataset.

4) Random Forest

Another model we chose for our modeling was Random Forests—an algorithm used for classification and regression by combining multiple decision trees that reduce overfitting by averaging these decision trees. Additionally, we wanted a model that is lower in sensitivity to outliers and noise in our data for greater accuracy.

For our first dataset, found in Notebook 1, we used Random Forests to fit and predict our Mean Squared Error (MSE) on our training and test sets. Our results outputted an MSE of around 725.09075, which was a significant improvement on our baseline model, which had an MSE of 1747.295461. Since Random Forests had a smaller MSE than our baseline, we can conclude that this modeling technique was successful.

For the next dataset, found in Notebook 3, we replicated this process and found our MSE to be around 0.00072529. This is, again, higher than our baseline model, whose MSE value was found to be at 0.04653716642.

5) K-means clustering

For our first merged dataset, we plotted our clustering algorithm to visualize our silhouette scores better and find the best value of K to use. By looking at our resulting graph, we found that the best K to use was K = 4, and used this information to further cluster our data. We began doing this by first printing out a graph where we fit our k-means clustering algorithm on a reduced version of our dataframe, and plotting these data points. We then printed out the average precipitation for each cluster, doing so from cluster 0 to 3, the total precipitation in inches, the sum for each state, time zone, type of fruit, and season.

We used k-means clustering similarly for our combined food waste, rainfall, and ppi dataset modeling, where we found the ideal K to be K = 4, and were able to plot clusters 0 to 3.

The model that had the most success/lowest MSE for the first merged dataset, which was titled "rainppi," was Logistic Regression, with a value of 636.1746891. This indicates that this model was the best fit.

IV. Impact:

Using our methods and findings, major potential impacts regarding our problem include a decrease in produce waste, higher cost efficiency for producers, and a reduction in consumer costs for the said produce. A year with lower average rainfall results in a higher producer price index since producers have to make up for the lack of rain with their own resources. This would consequently increase the consumer price index, limiting the sale of the produce to the minority who can afford spiked prices. Ultimately, this will result in all the unpurchased produce going to waste and contribute to the current food waste statistic. It also leaves producers losing money and consumers unable to purchase these basic groceries. With our work, producers can have an idea of rainfall expected for the year and whether or not there will be an increase or decrease in PPI and CPI. They can plan accordingly for the lack or abundance of rain. The consumers will then have a better chance of purchasing higher quality produce for cheaper prices since producers can make arrangements for the upcoming seasons.

In efforts to expand our impact on our problem of tackling food waste, an area to explore would be finding additional factors that impact PPI. We are aware of how unpredictable weather is, so another option would be finding additional factors that significantly affect the quality and price of produce in addition to the rainfall data.

Citations

Moosavi, S. (2023, May 23). US Weather Events (2016 - 2022). Kaggle. https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

Producer price index by Commodity: Farm Products: Table grapes. FRED. (n.d.). https://fred.stlouisfed.org/series/WPU011102281

United States Environmental Protection Agency. (n.d.). Facts and figures about materials, waste and recycling. United States Environmental Protection Agency. https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling

USDA. (n.d.). Producer price indexes. USDA ERS - Data Products. https://data.ers.usda.gov/reports.aspx?ID=17848

USDA. (n.d.). Food availability (per capita) Data System. USDA ERS - Food Availability (Per Capita) Data System. https://www.ers.usda.gov/data-products/food-availability-per-capita-data-system/food-availability-per-capita-data-system/#Loss-Adjusted%20Food%20Availability