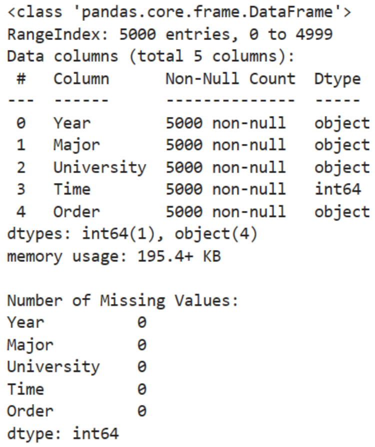
TechPoint Artificial Intelligence Assessment Report

By: Aaron Wong

1 Exploratory Analysis

In this section, we walk through the exploratory analysis I did on the data set, as well as go over and patterns and trends that I saw.

First, I imported the data set and looked at the values that are in the data set. It is noted that we are given 5 columns, Year, Major, University, Time, and Order. These columns are all object data-types except for time, which is an int data type. This will be addressed later.

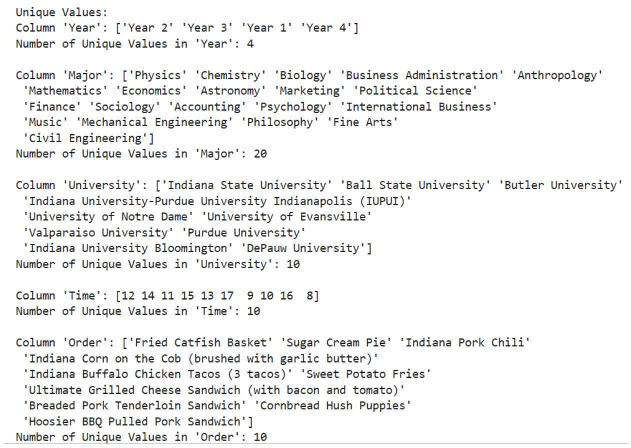


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Year        5000 non-null   object
 1   Major       5000 non-null   object
 2   University  5000 non-null   object
 3   Time        5000 non-null   int64 
 4   Order       5000 non-null   object
dtypes: int64(1), object(4)
memory usage: 195.4+ KB

Number of Missing Values:
Year          0
Major         0
University    0
Time          0
Order         0
dtype: int64
```

Figure 1: Summary of Data Types and Missing Values

Next, I wrote a script to print out the number of unique values in each column as well as the unique values themselves. This allows to give a summary of the type of data we are working with. As you can see, we are dealing with the year in college the student is in. The college the student is going to. Their major, as well as the time they spent ordering. But the biggest help, was seeing that we have a fixed menu. This means that we only have to worry about those fixed selections, and not about different combinations.



```
Unique Values:
Column 'Year': ['Year 2' 'Year 3' 'Year 1' 'Year 4']
Number of Unique Values in 'Year': 4

Column 'Major': ['Physics' 'Chemistry' 'Biology' 'Business Administration' 'Anthropology'
 'Mathematics' 'Economics' 'Astronomy' 'Marketing' 'Political Science'
 'Finance' 'Sociology' 'Accounting' 'Psychology' 'International Business'
 'Music' 'Mechanical Engineering' 'Philosophy' 'Fine Arts'
 'Civil Engineering']
Number of Unique Values in 'Major': 20

Column 'University': ['Indiana State University' 'Ball State University' 'Butler University'
 'Indiana University-Purdue University Indianapolis (IUPUI)'
 'University of Notre Dame' 'University of Evansville'
 'Valparaiso University' 'Purdue University'
 'Indiana University Bloomington' 'DePauw University']
Number of Unique Values in 'University': 10

Column 'Time': [12 14 11 15 13 17  9 10 16  8]
Number of Unique Values in 'Time': 10

Column 'Order': ['Fried Catfish Basket' 'Sugar Cream Pie' 'Indiana Pork Chili'
 'Indiana Corn on the Cob (brushed with garlic butter)'
 'Indiana Buffalo Chicken Tacos (3 tacos)' 'Sweet Potato Fries'
 'Ultimate Grilled Cheese Sandwich (with bacon and tomato)'
 'Breaded Pork Tenderloin Sandwich' 'Cornbread Hush Puppies'
 'Hoosier BBQ Pulled Pork Sandwich']
Number of Unique Values in 'Order': 10
```

Figure 2: Unique Labels and Number of Unique Labels for Each Column

1.1 Analysis of Plots

1.1.1 Scatter Plots

I then created 4 separate scatter plots. All of which plot the data represented in the column by the food that that particular person ordered.

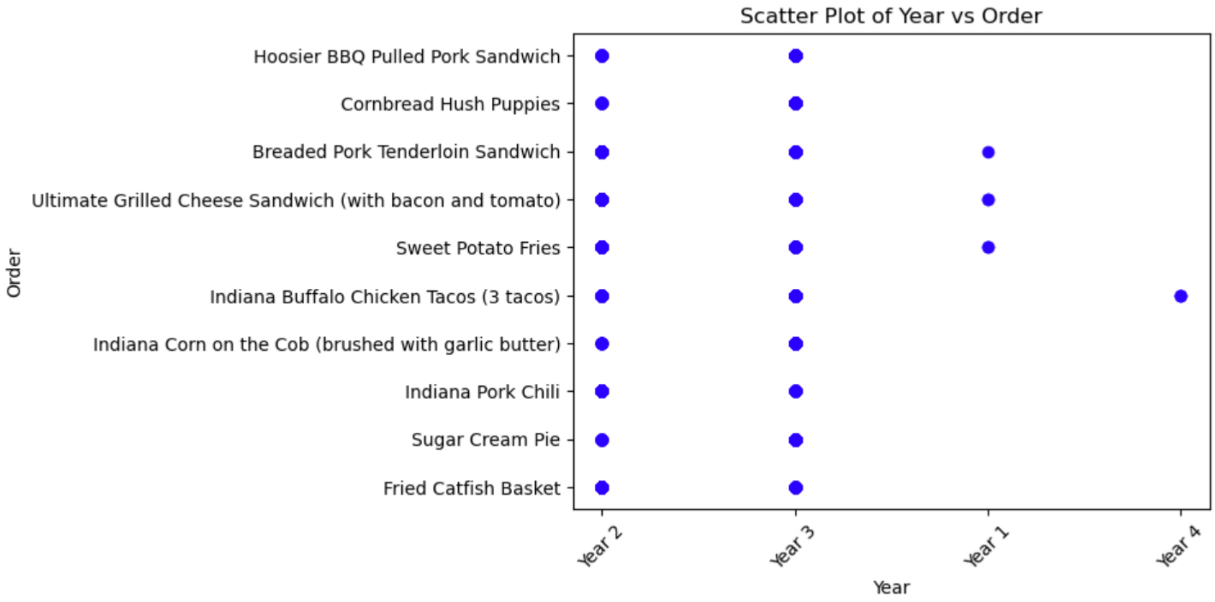


Figure 3: Scatter Plot of Year vs Order

In Figure 3, we can see that all forth year students ordered the Indiana Buffalo Chicken, and that all 1st years ordered one of 3 foods. This can be useful in predicting which year would order what. We can consider that the student being in Year 1 and Year 4 should favor those foods more heavily when creating our model.

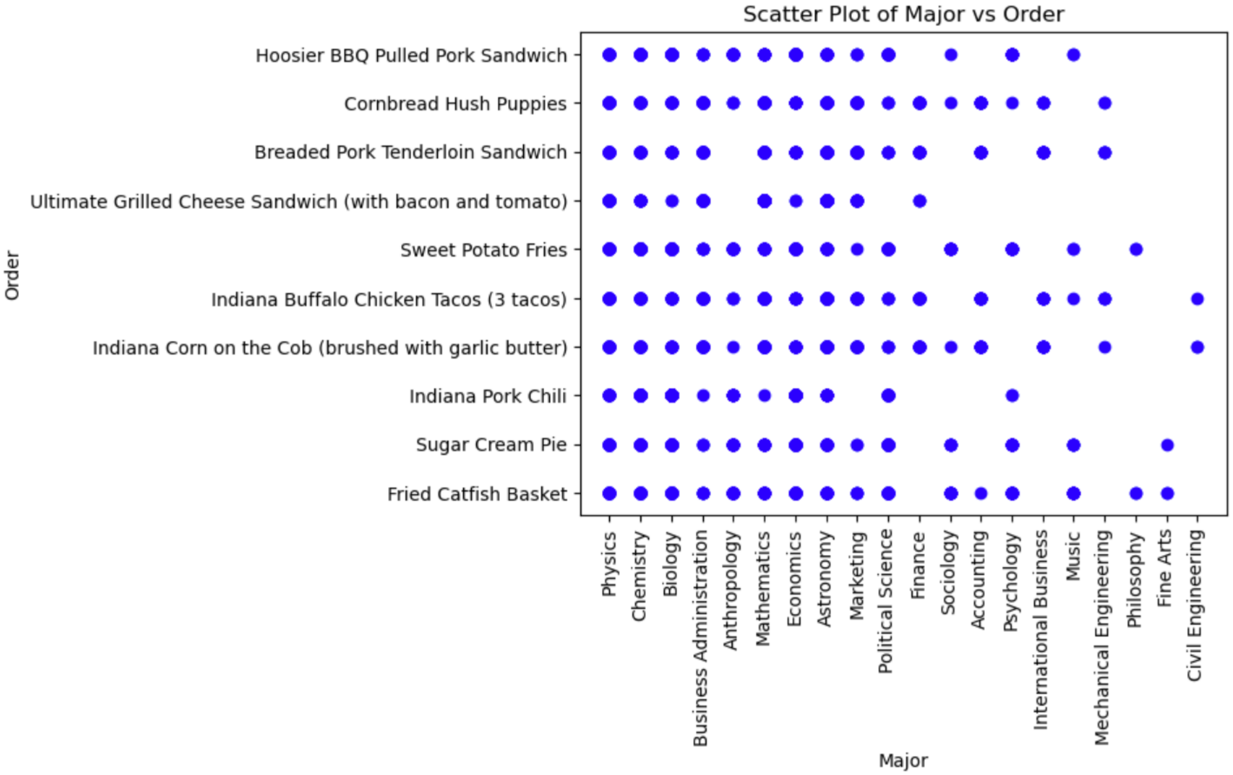


Figure 4: Scatter Plot of Major vs Order

In Figure 4, we can see that everyone of every major basically ordered at least 1 of each item. But, we can see that some of the majors near the end ordered mostly just 2 items. We can consider that the people who have those majors will be more likely to order those items

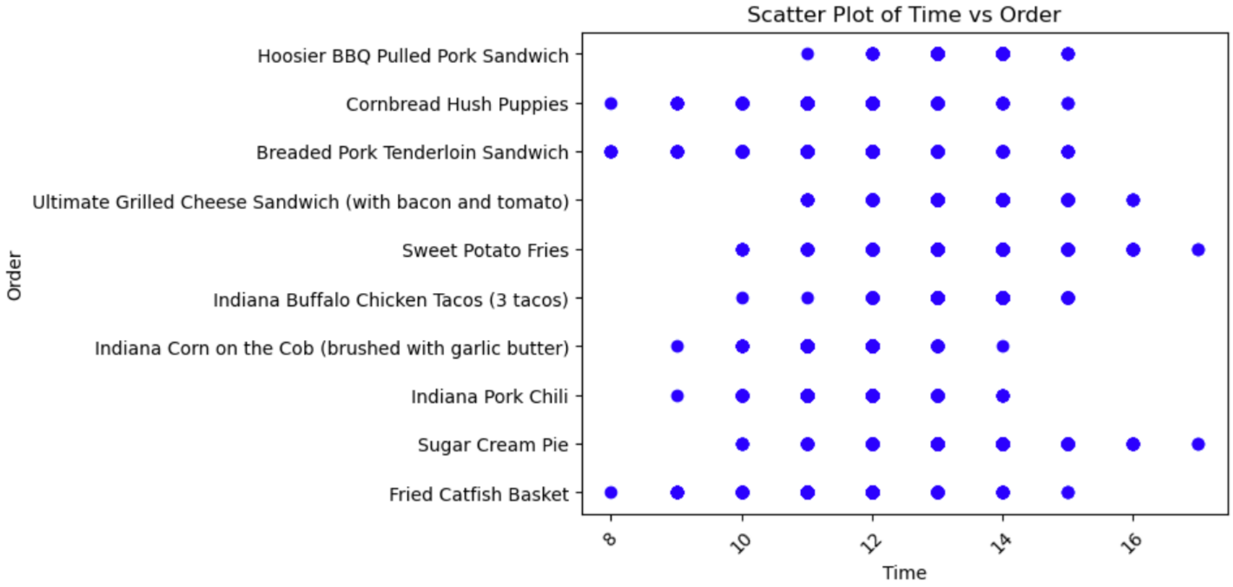


Figure 5: Scatter Plot of Time vs Order

In Figure 5, we see that there are some items that are ordered more frequently when the time to order is longer and when the time to order is shorter. We can consider this when making our predictions.

1.1.2 Histograms

In this subsection, I will show show the frequencies of each feature in histogram format:

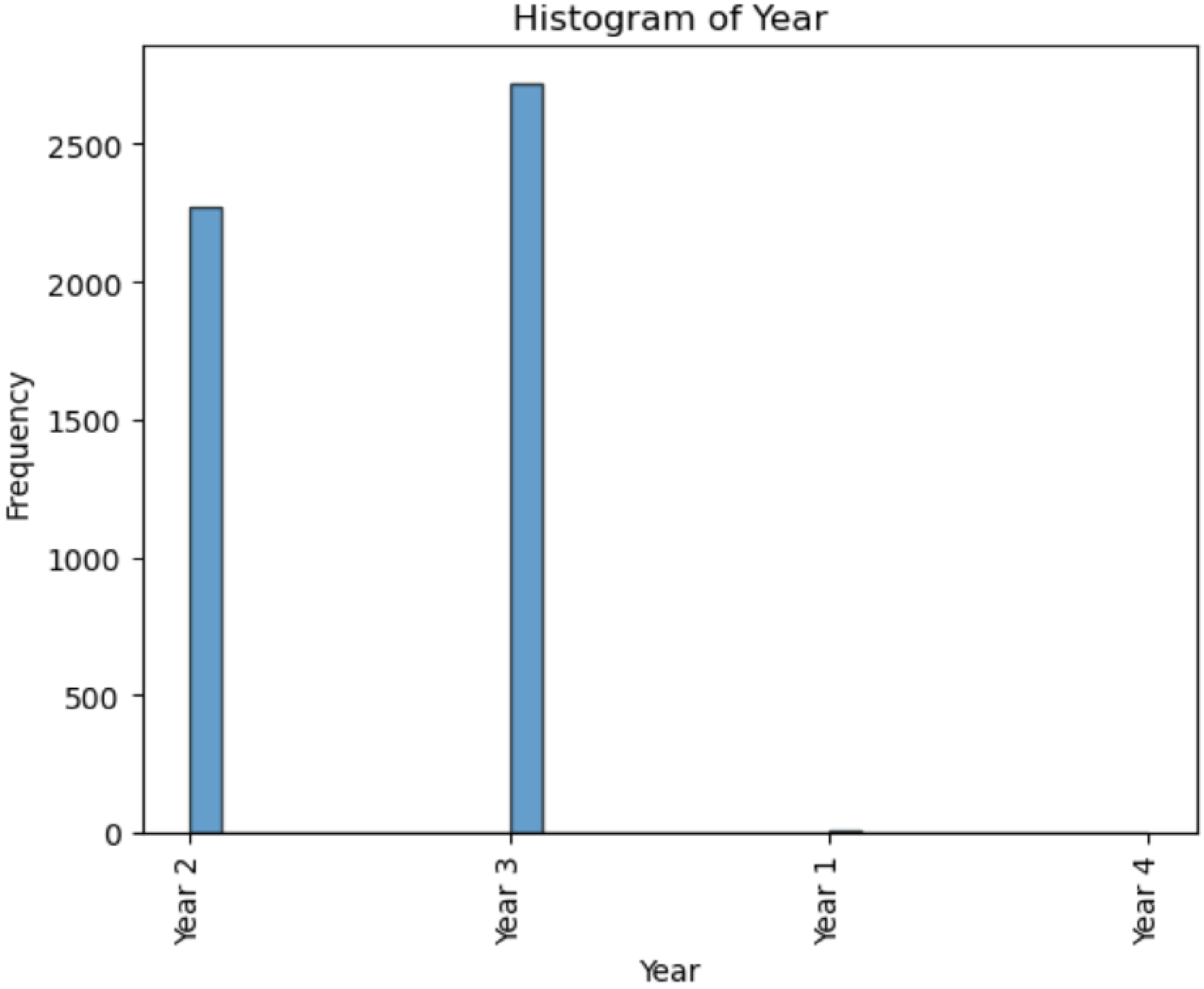


Figure 6: Histogram of Year

Now that we have the distribution of all the frequencies of each feature, we can analyze how many people ordered what. If you refer to Figure 6, notice that the vast majority of students ordering at the food truck are Year 2 and Year 3 students. For major, we have half of the majors ordering way more than the other half of majors. For University we can see that only have of the university really have students going to the food trucks. We can also see what looks like some sort of a normal distribution for the time it takes to order food. Which is expected since we have a large sample size. Finally, the type of food that is ordered is widely spread out, so we don't have much information there.

As a result, this can help us identify possible outliers and flukes in our algorithm. For example, just because we had one fourth year order tacos, does not mean that all fourth years will order tacos.

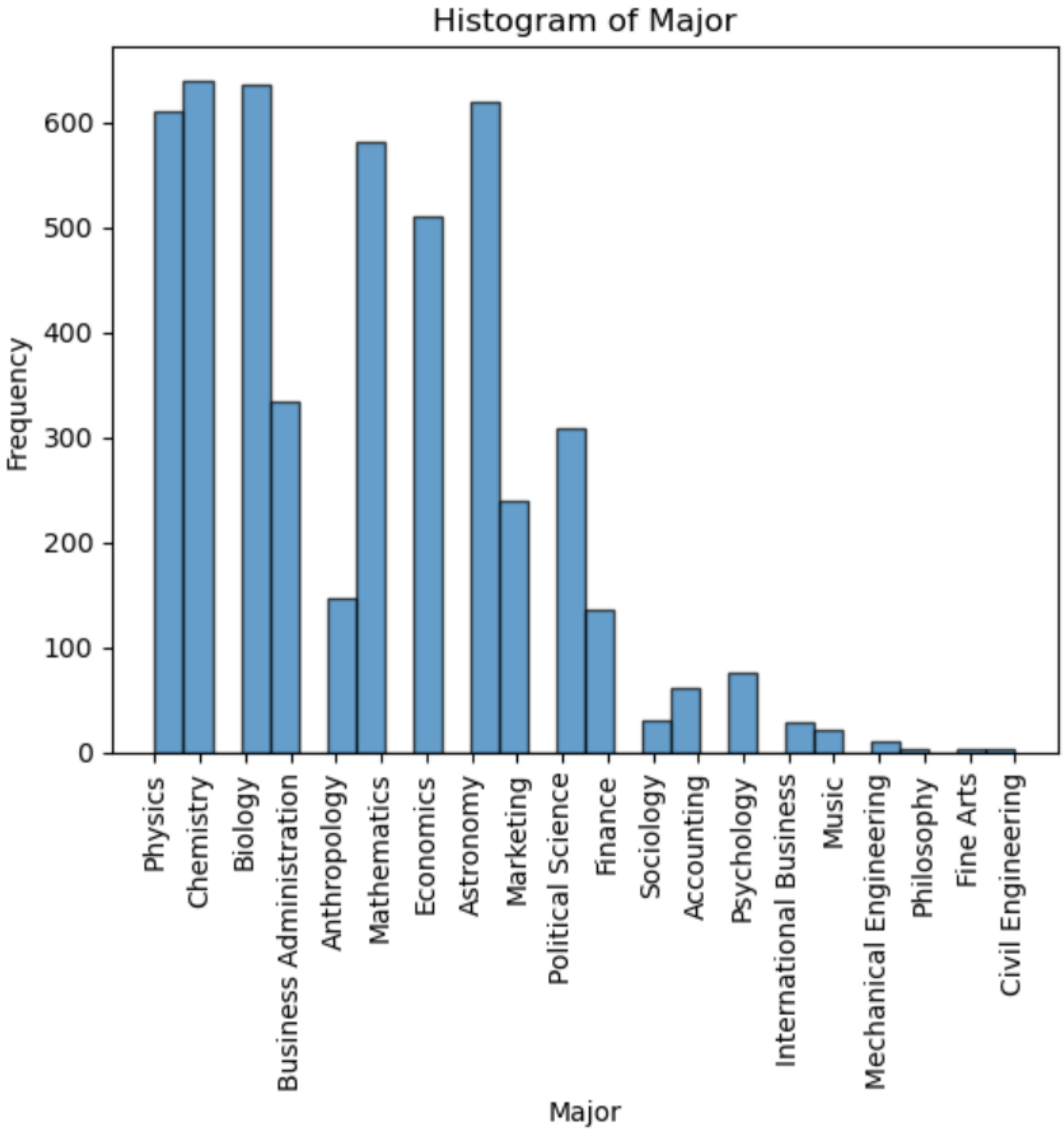


Figure 7: Histogram of Major

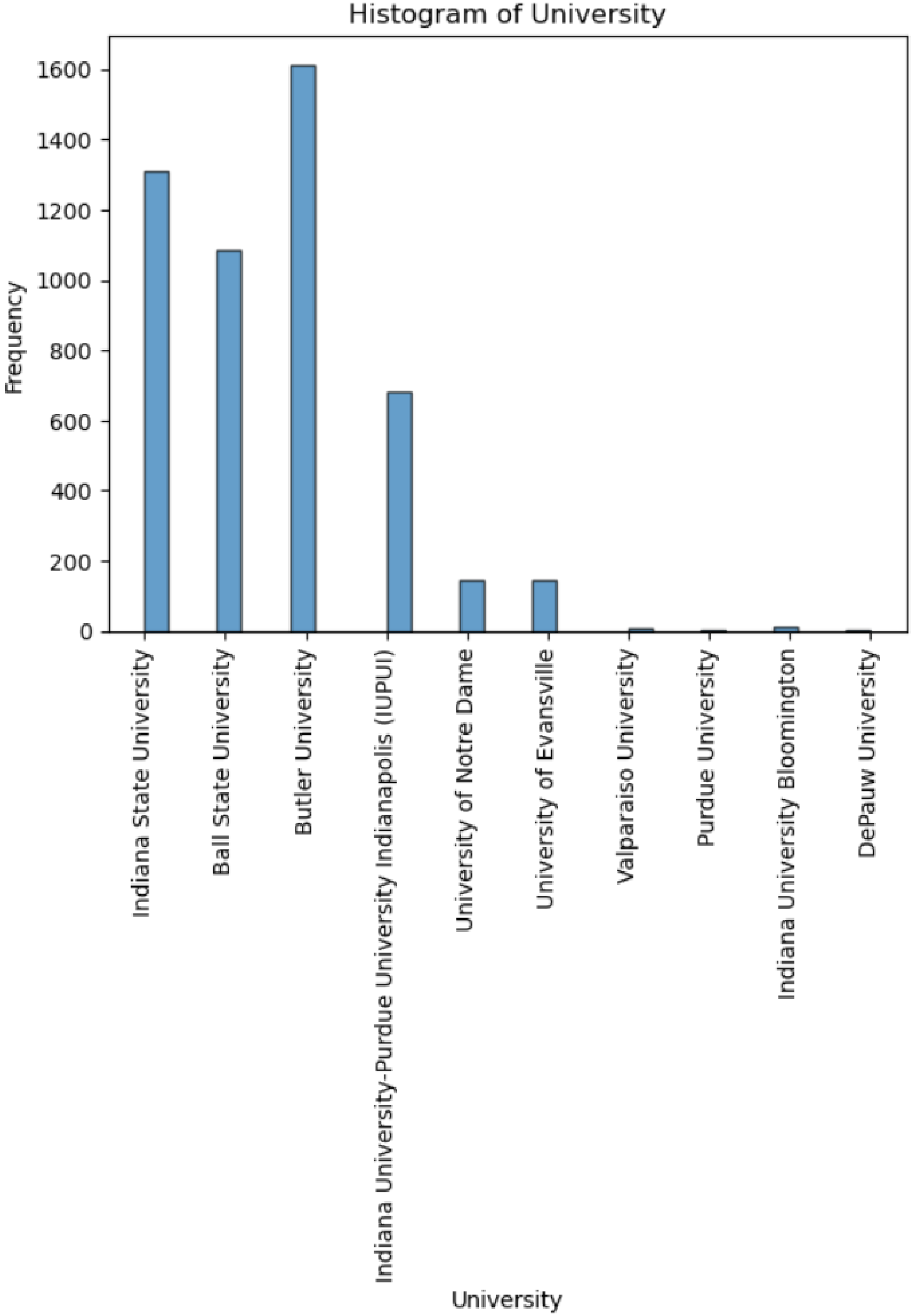


Figure 8: Histogram of University

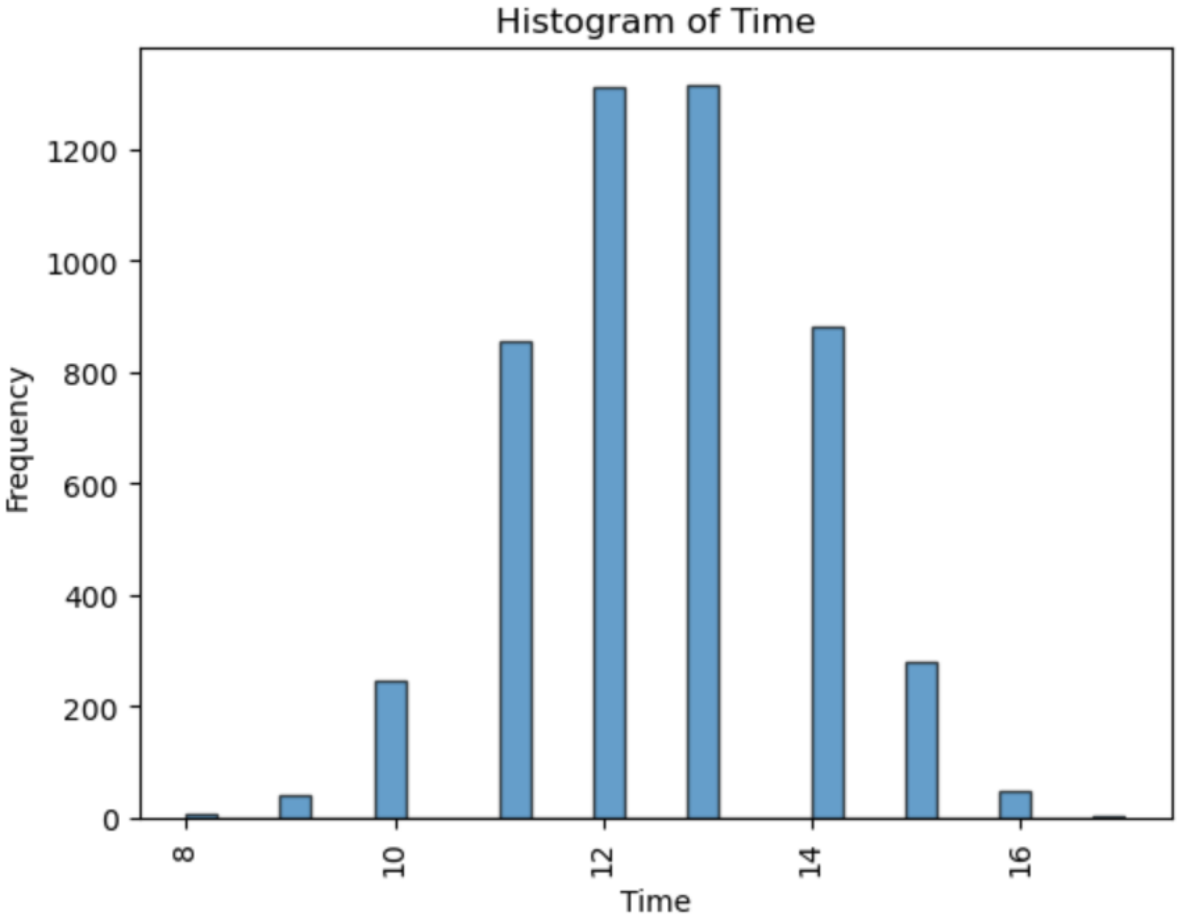


Figure 9: Histogram of Time

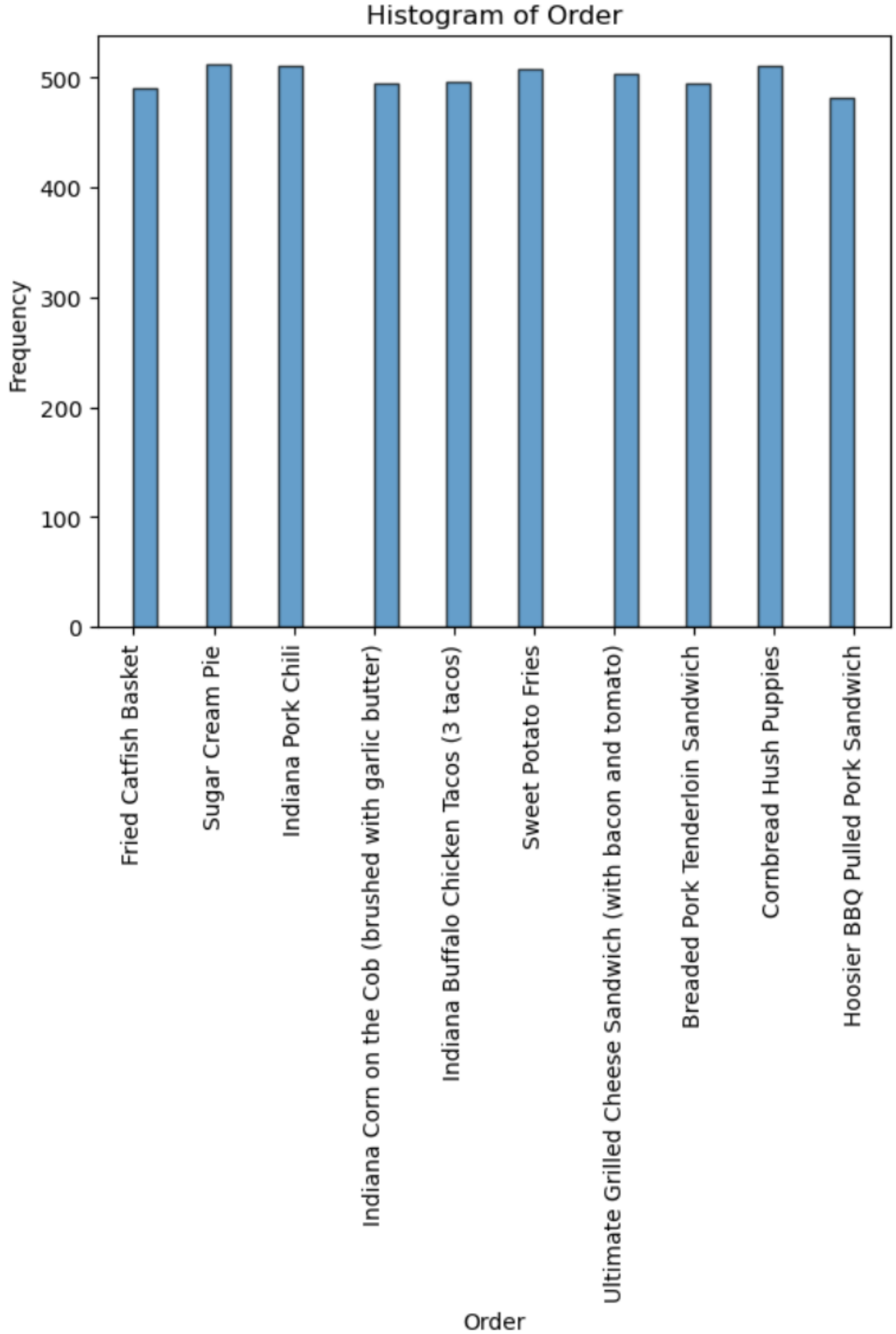


Figure 10: Histogram of Order

1.2 Suggestions for Other Use of Data

In addition to attempting to predict the purchases of students based off of their features, there are other things we can find with this data set as well. For example, we notice from the histograms that the clientele of the food trucks varies greatly. This could mean that the food truck can try and cater to the missing groups of people who are not eating there. This can also let the company know what clientele to focus on, what majors to focus on. The data could also be used to predict the year the student is, their major, or the university they attend based off of what food they ordered. That could offer for some more clients as well.

2 Implications

2.1 Ethical Implications

2.1.1 User Privacy

The first ethical implication we have to consider would be user privacy. We are collecting their information and therefore, might accidental leak their information when they would rather keep it private. This means that we have to be sure to anonymize their data so it can not be traced back to them. We also have to inform the users that we are using their data but the data will not be traced back to them. We also have to make sure that we are not making users uncomfortable by giving us data that they don't want to give. Therefore informed consent is paramount for our data collection

2.1.2 Bias

Given that our data is from college students, or models and predictions might not fair well when used on a population that is not predominately college students from around the same area. We have to make these considerations when attempting to scale up our operation to different demographics.

2.2 Business Implications

2.2.1 Guessing the Correct Food

The goal of our model is to accurately guess the food that the person orders so that we do not lose more money by giving away a free 10 percent discount. Right now, the current practice is to just randomly choose, but we are wrong most of the time. If the model is even right 25 percent of the time, 50, even more, we can end up saving way more money rather than just give away discounts over and over again. We might also garner more clients since our new model can be perceived as sort of a magic trick.

2.2.2 Clientele

The data that we collected also allows us to see what kind of people are ordering for us. As stated before, a disproportionate number of customers end up being just Year 2 and Year 3 people. We can then cater towards them more, or make strides to somehow include the other years as well. This goes the same with the majors and university too. This allows us to have potential for growth into new sectors of customers that we previously did not have before

2.2.3 Efficiency

Using machine learning which allows us to predict who will order what can help us streamline the process. If we know what people would be coming on a certain day, we could make more of the same foods that they like so we do not have to worry about being under stocked or over stocked. This can give us an advantage to other food trucks who rely on eyeballing and guesswork when determining what foods to sell.

2.3 Technical Implications

2.3.1 Data Storage and Preservation

When dealing with large amounts of data, we have to make sure that we have an infrastructure that can support scaling up, as well as data failures. If we were to scale into an even larger company, we have to be able to keep up with the new influx of data. This means buying more storage, backup storage, etc.

2.3.2 Algorithm Selection

Although in my report I only compared 2 different algorithms, there are many many more that can be used, and we should not stay fixed on just one algorithm when there could be one that performs better. We also have to consider the way that we cleaned and converted the data, there maybe better ways to do that as well.

3 Model

3.1 Determining a Model

The process I used in determining which model to use was by simply trial and error. I used two algorithms that I was familiar with, namely Naive-Bayes Classifiers and Decision Trees. For Naive-Bayes I used the GaussianNB and CategoricalNB located inside sklearn. One was better for continuous variables while the other is better for categorical variables. I then chose the better performing model out of the two and went with that one, although there are reasons for both models as to why one might be better than the other. I will go into how I optimized these models in later sections.

3.2 Loading Dataset

In this data, remember that we are dealing with categorical data, not numerical. So one thing that these models need is that the data that you are passing in numerical. So there are two ways that I considered to encode the data. The first method was simply just labeling the first object I saw 1, then the second 2, etc. The second is one hot encoding. I attempted Gaussian Naive Bayes using the first labeling method but ended up with a score of 0.468. Which is better than randomly guessing but I believed we can do better. So I tried again with one-hot and got better results. The reason why I believe this is the case, is because when performing the first labeling method, values that have no sense of rank in a column are given a meaning of rank. But, one-hot encoding avoids this method.

3.3 Designing My Model

3.3.1 Naive Bayes Classifier

When designing my model let's start off with Naive Bayes. I used a Gaussian Naive Bayes model and also Categorical Naive Bayes model. The categorical model performed better because Gaussian models are better for continuous variables, while Categorical Naive Bayes is better for non numerical items. Which in this case, we are dealing with words and therefore Categorical Naive Bayes is better. I then cycled through different seeds of splitting the test set to find highest score, which I found to be a seed of 82 with an accuracy of 0.6827. However, this is not really that reliable since changing the seed does not necessarily mean that your model has improved because only the test and training sets are changed.

The final best accuracy I achieved using Categorical Naive Bayes was 0.682. Which is better than a 50/50 chance and a lot better than just randomly guessing between 10 different food options, which is a 0.1 chance.

3.3.2 Decision Trees

My next model that I created was a decision tree. I again encoded my data using one-hot and then simply ran the sklearn decision tree algorithm. This ended up giving me a model with an accuracy of 1. However, this might not be the best since we might be overfitting the data. If there was new data that we haven't seen yet, we might get it wrong. In order to overcome this, I attempted to use cross validation to try and prune the tree. I used different hyperparameters to attempt to prune the tree to see if the accuracy changes, and as as result, we ended up with the same tree that we started with. These hyperparameters were used to determine how much of the tree we should prune, in which we then train the model again and if we get a better result we keep the pruned tree. Meaning that cross validation was unsuccessful in reducing the overfitting. This however, does not mean that the model is bad, we would not know unless we have new data to test it against.

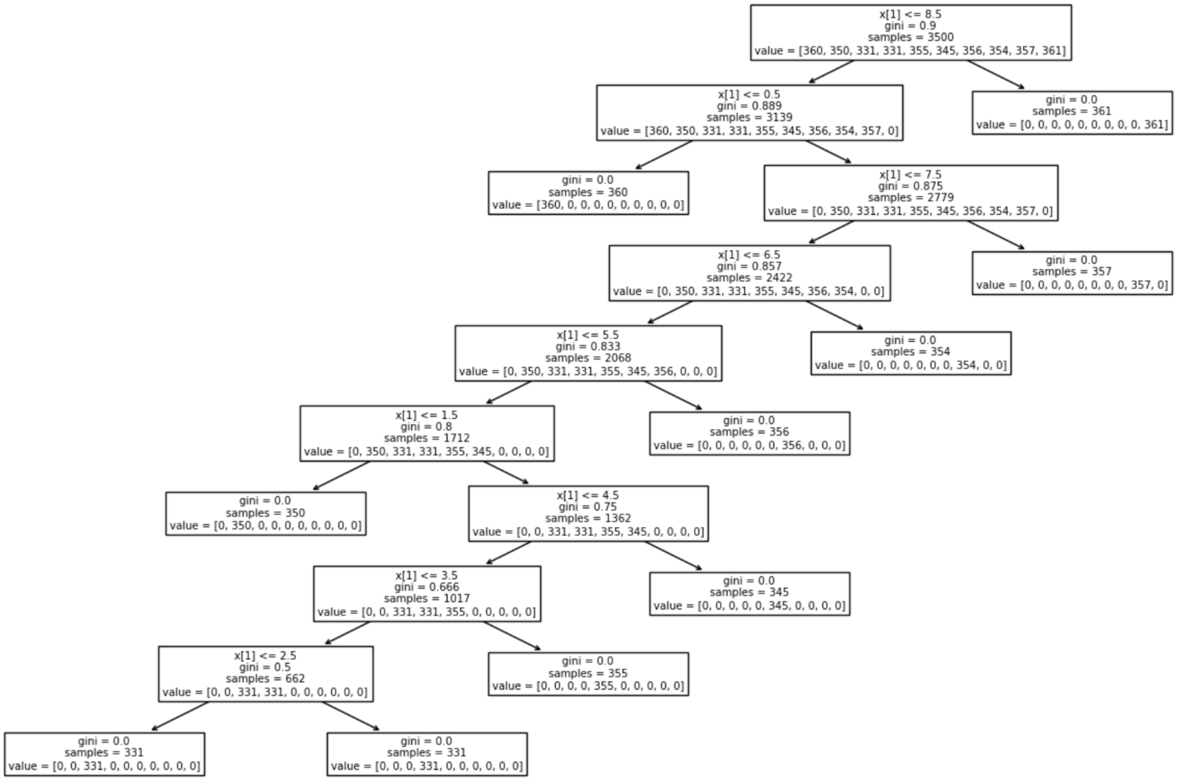

x[1] <= 8.5
gini = 0.9
samples = 3500
value = [360, 350, 331, 331, 355, 345, 356, 354, 357, 361]
x[1] <= 0.5
gini = 0.889
samples = 3139
value = [360, 350, 331, 331, 355, 345, 356, 354, 357, 0]
gini = 0.0
samples = 361
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 361]
gini = 0.0
samples = 360
value = [360, 0, 0, 0, 0, 0, 0, 0, 0, 0]
x[1] <= 7.5
gini = 0.875
samples = 2779
value = [0, 350, 331, 331, 355, 345, 356, 354, 357, 0]
x[1] <= 6.5
gini = 0.857
samples = 2422
value = [0, 350, 331, 331, 355, 345, 356, 354, 0, 0]
gini = 0.0
samples = 357
value = [0, 0, 0, 0, 0, 0, 0, 0, 357, 0]
x[1] <= 5.5
gini = 0.833
samples = 2068
value = [0, 350, 331, 331, 355, 345, 356, 0, 0, 0]
gini = 0.0
samples = 354
value = [0, 0, 0, 0, 0, 0, 0, 354, 0, 0]
x[1] <= 1.5
gini = 0.8
samples = 1712
value = [0, 350, 331, 331, 355, 345, 0, 0, 0, 0]
gini = 0.0
samples = 356
value = [0, 0, 0, 0, 0, 0, 356, 0, 0, 0]
gini = 0.0
samples = 350
value = [0, 350, 0, 0, 0, 0, 0, 0, 0, 0]
x[1] <= 4.5
gini = 0.75
samples = 1362
value = [0, 0, 331, 331, 355, 345, 0, 0, 0, 0]
x[1] <= 3.5
gini = 0.666
samples = 1017
value = [0, 0, 331, 331, 355, 0, 0, 0, 0, 0]
gini = 0.0
samples = 345
value = [0, 0, 0, 0, 0, 345, 0, 0, 0, 0]
x[1] <= 2.5
gini = 0.5
samples = 662
value = [0, 0, 331, 331, 0, 0, 0, 0, 0, 0]
gini = 0.0
samples = 355
value = [0, 0, 0, 0, 355, 0, 0, 0, 0, 0]
gini = 0.0
samples = 331
value = [0, 0, 331, 0, 0, 0, 0, 0, 0, 0]
gini = 0.0
samples = 331
value = [0, 0, 0, 331, 0, 0, 0, 0, 0, 0]

Figure 11: Decision Tree

4 Final Remarks

In conclusion, the model that I created was able to more accurately predict what people were going to order based off their data. Before, the workers had to guess on each person, but now, the model has made it so that it is correct 62.8% of the time, which is way better than a 1 in 10 chance. And with the decision tree, granted that the tree did not over fit the data, was accurate to determine what the person was going to eat 100% of the time.

The considerations I would make would be first, is my model accurate enough to be used. And I believe the answer is yes. Now the next question is, would this model be profitable. In the problem statement, it is stated that they are getting more customers but the workers are not happy with guessing orders and we are getting a lot wrong. Nothing is really stated about money, so since my model is more accurate than them guessing, I believe that this is another reason why this model should be implemented. Therefore the return on investment has to be determined to be enough in order to make this model worth it

Another consideration I would make is would I be able to have the infrastructure needed to make accurate predictions. For example, a big enough storage unit, a way to continuously get new data from new customers in order to grow our data set and make our model more accurate.

Another thing we have to worry about would be scalability. Since our model was only trained with this set of data, what happens if we have new demographics, or if we open a food truck in another area different than the areas that we are in now. We might then at that point have to create a new model, or tweak the existing one in order to work well for the new data.