Final Analysis
ITP 449 Final

Food Team

Rachel Litz
Sabrina Tong
Sophia Chen
Nigel Egrari
Philipp Hulsch
Ailsa Taylor
Gianna DiGiovanni
Natalie Miller
Drew Osherow

Survey Design and Data Collection

The food group set out to determine the optimal number of choices for a menu to have. We created a survey -- our test subject was told they would receive a meal and to pick the one they wanted the most, as well as how confident they were of their choice and how easy it was to make the choice.

After the data was collected, the team filtered out those who had allergies, and the number of choices was instantly restricted based on his or her allergy.

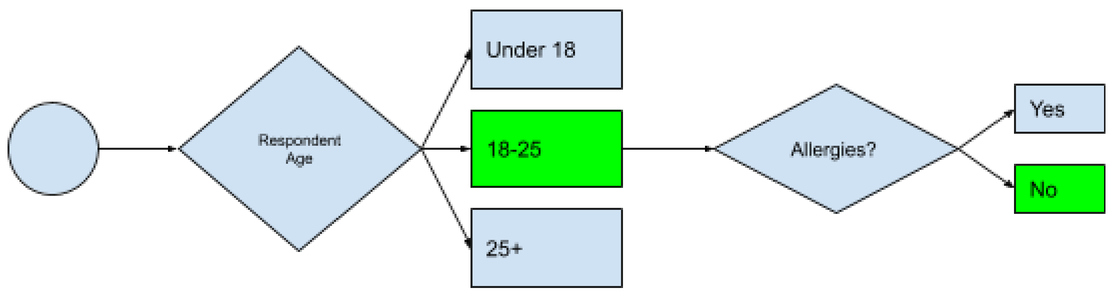


Figure 1: Process flow diagram of the data we included in our analyses

We segmented the surveys into three parts: a menu with three choices, one with seven choices (the original three and four more added on), and one with twelve choices. A respondent was sent either the 3-, 7-, or 12-choice food survey, which also included a question on their age and allergies. We wanted to control for the age group and allergies to simplify the task.

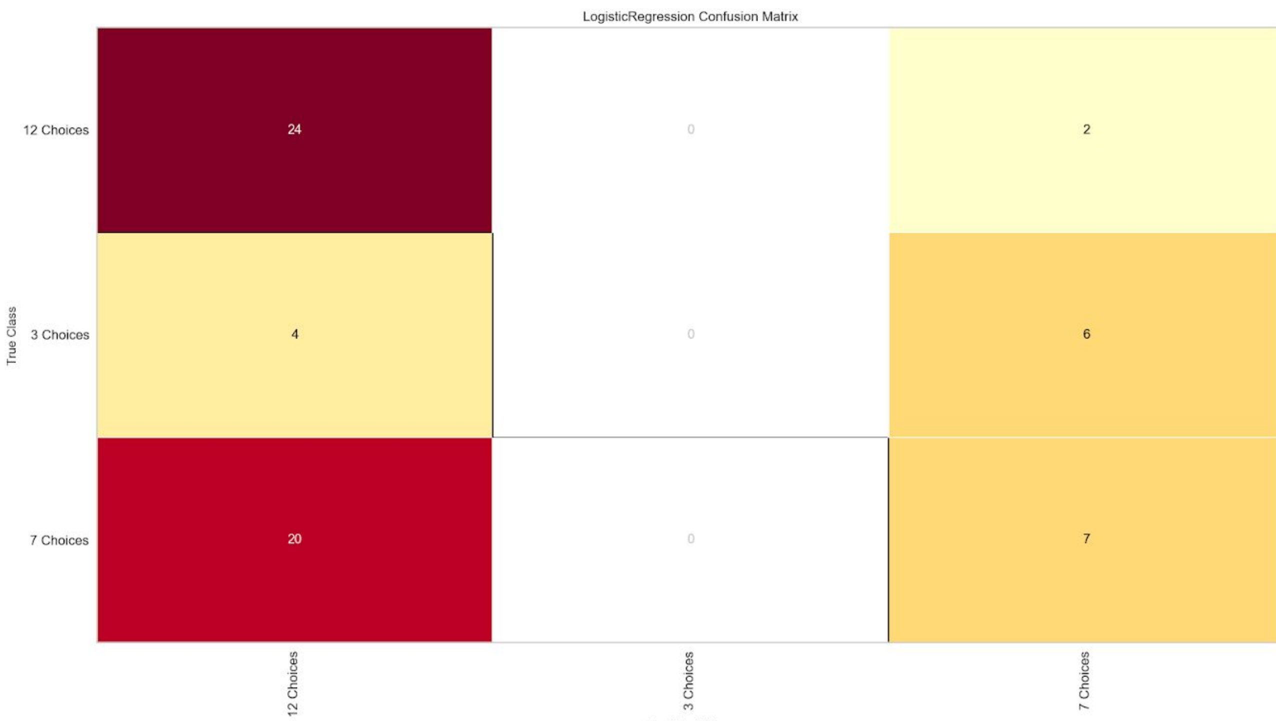
Figure 2: Screenshots of the 3-choice food survey sent out to respondents.

The menus consisted of entrees from Moreton Fig with the prices redacted. The components of each dish were also included in the menu for the respondents to see (Figure 2, middle panel). The final question asked the survey respondent how confident they were in their choice and how easy they felt making their choice was. We then analyzed the data to try to solve the paradox of choice in menus.

Logistic Regression
Sabrina Tong, Rachel Litz, Sophia Chen

We attempted to create a logistic regression to determine if the confidence was a predictor of the number of choices a survey respondent received. Initially, we tried to do a linear regression of the converse to see if the number of choices predicted confidence; however, upon running that, we realized a number of errors and decided to run a logistic model instead.

We found that the confidence a respondent reported most commonly predicted that they had 12 choices to have. However, this model was only 50% accurate, and we think it's because the 3- and 7-choice surveys were encapsulated within the 12-choice survey. Our choice sets were too close together. Furthermore, the 12-choice survey resulted in the highest variance, which would also explain why the model predicted most of the data would fall into that category.



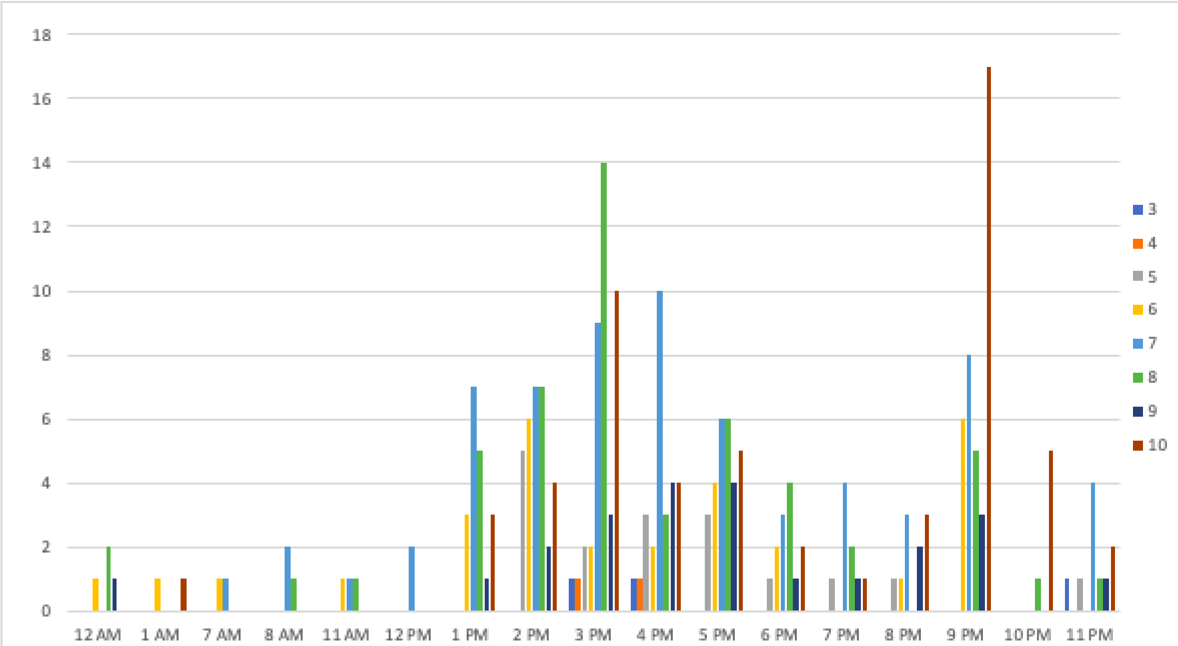
Time Series Analysis (Gianna DiGiovanni)

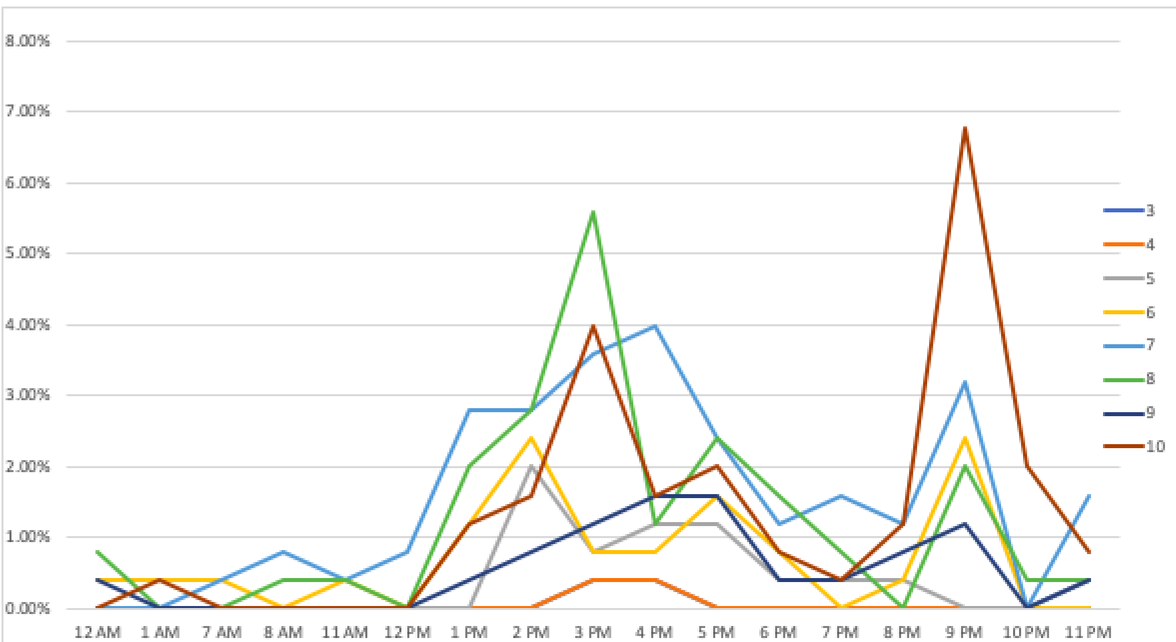
When testing the correlation between Time of Day (hour) and Confidence Level it can be seen that people gained confidence in their choice as the day progressed and the closer it was to meal time (3 PM and 9 PM).
this was conducted in excel due the lack of cooperation with PyCharm

Process -

- Brokedown the datetime stamp to date and time
- Rounded the time to the nearest hour

- Compared the confidence levels of choices made during different hours throughout the day





Count of Survey_no	Column Labels							
Row Labels	3	4	5	6	7	8	9	10	Grand Total
12 AM				1		2	1		4
1 AM				1				1	2
7 AM				1	1				2
8 AM					2	1			3
11 AM				1	1	1			3
12 PM					2				2
1 PM				3	7	5	1	3	19
2 PM			5	6	7	7	2	4	31
3 PM	1	1	2	2	9	14	3	10	42
4 PM	1	1	3	2	10	3	4	4	28
5 PM			3	4	6	6	4	5	28
6 PM			1	2	3	4	1	2	13
7 PM			1		4	2	1	1	9
8 PM			1	1	3		2	3	10
9 PM				6	8	5	3	17	39
10 PM						1		5	6
11 PM	1		1		4	1	1	2	10
Grand Total	**3**	**2**	**17**	**30**	**67**	**52**	**23**	**57**	**251**

count of choices

Count of Survey_no	Column Labels								
Row Labels	3	4	5	6	7	8	9	10	Grand Total
12 AM	0.00%	0.00%	0.00%	0.40%	0.00%	0.80%	0.40%	0.00%	1.59%
1 AM	0.00%	0.00%	0.00%	0.40%	0.00%	0.00%	0.00%	0.40%	0.80%
7 AM	0.00%	0.00%	0.00%	0.40%	0.40%	0.00%	0.00%	0.00%	0.80%
8 AM	0.00%	0.00%	0.00%	0.00%	0.80%	0.40%	0.00%	0.00%	1.20%
11 AM	0.00%	0.00%	0.00%	0.40%	0.40%	0.40%	0.00%	0.00%	1.20%
12 PM	0.00%	0.00%	0.00%	0.00%	0.80%	0.00%	0.00%	0.00%	0.80%
1 PM	0.00%	0.00%	0.00%	1.20%	2.79%	1.99%	0.40%	1.20%	7.57%
2 PM	0.00%	0.00%	1.99%	2.39%	2.79%	2.79%	0.80%	1.59%	12.35%
3 PM	0.40%	0.40%	0.80%	0.80%	3.59%	5.58%	1.20%	3.98%	16.73%
4 PM	0.40%	0.40%	1.20%	0.80%	3.98%	1.20%	1.59%	1.59%	11.16%
5 PM	0.00%	0.00%	1.20%	1.59%	2.39%	2.39%	1.59%	1.99%	11.16%
6 PM	0.00%	0.00%	0.40%	0.80%	1.20%	1.59%	0.40%	0.80%	5.18%
7 PM	0.00%	0.00%	0.40%	0.00%	1.59%	0.80%	0.40%	0.40%	3.59%
8 PM	0.00%	0.00%	0.40%	0.40%	1.20%	0.00%	0.80%	1.20%	3.98%
9 PM	0.00%	0.00%	0.00%	2.39%	3.19%	1.99%	1.20%	6.77%	15.54%
10 PM	0.00%	0.00%	0.00%	0.00%	0.00%	0.40%	0.00%	1.99%	2.39%
11 PM	0.40%	0.00%	0.40%	0.00%	1.59%	0.40%	0.40%	0.80%	3.98%
Grand Total	**1.20%**	**0.80%**	**6.77%**	**11.95%**	**26.69%**	**20.72%**	**9.16%**	**22.71%**	**100.00%**

percent of grand total

Data Visualization:

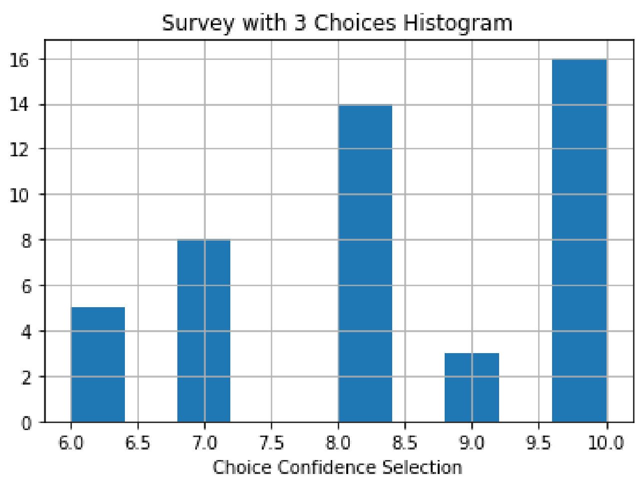
Nigel Egrari, Drew Osherow, Philipp Hultsch, Ailsa Taylor

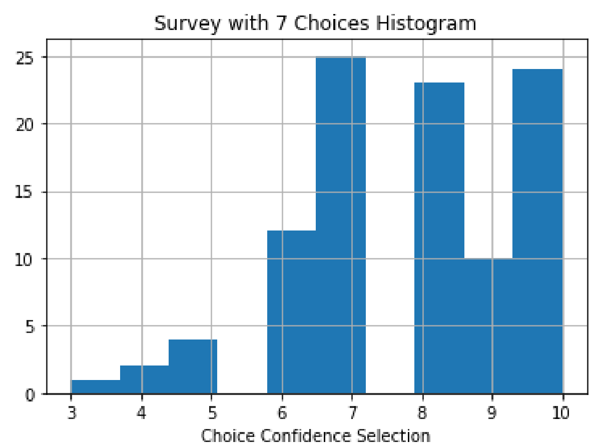
The main goal of our visualizations was to compare choice confidence with the number of choices available. We saw that there were very slight increases in mean choice confidence as the number of choices decreased. This goes along with the paradox of choice theory. The scatter correlation between easiness of choice and confidence levels do not show much of a connection between the two as the answers for one (on a 1-10 scale) can vary tremendously from the other.

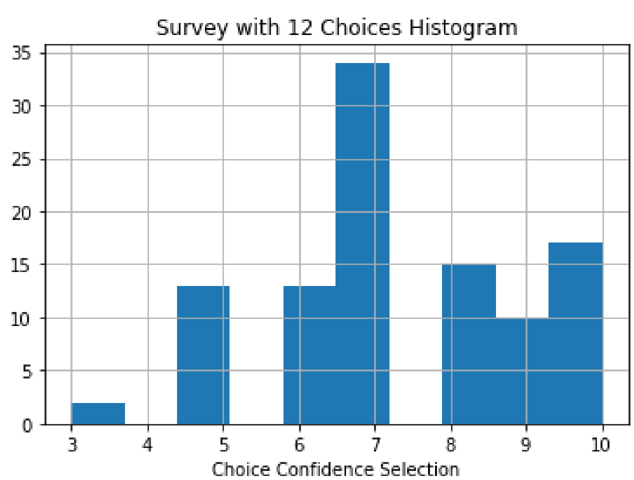
As for the correlation matrices, the surveys with three choices and twelve choices have a negative correlation. The three-choice survey shows the most negative correlation, showing that those that had high confidence in their choice did not necessarily find it easy to choose their menu item and vice versa. The seven-choice survey and twelve-choice survey are both near 0, showing almost no correlation between choice confidence and choice easiness. This goes against the idea that those who are the most confident would have the easiest time choosing a menu item, while those who are least confidence seemed not to struggle.

The “Food Selection” bar chart that we attached to our analysis showed all of the data collected through the three surveys. The salmon, burger, and fried chicken were the most popular.

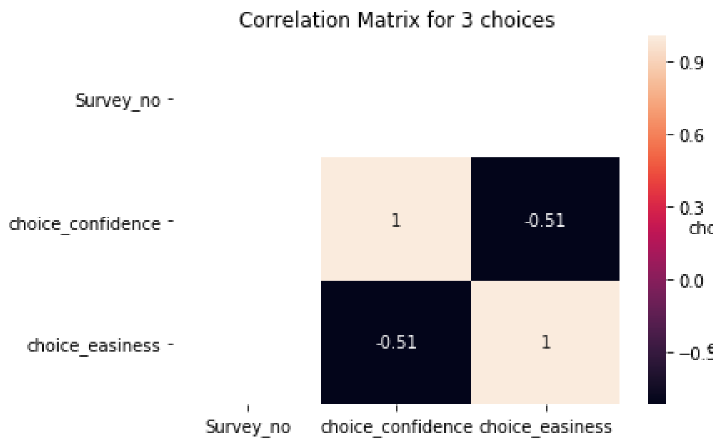
The boxplot chart that we attached shows that there’s a much larger range in confidence in Survey 3 and that surveys 1 and 2 had the same mean. The three-choice survey had the least range.

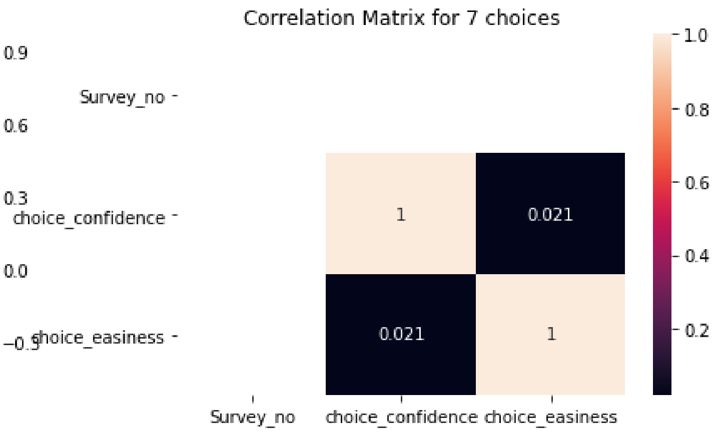


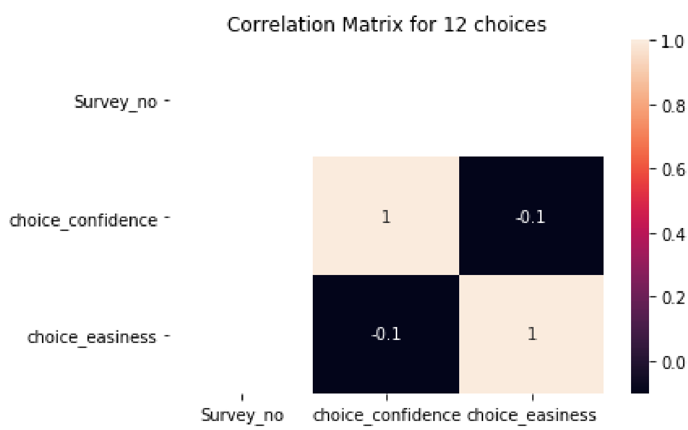


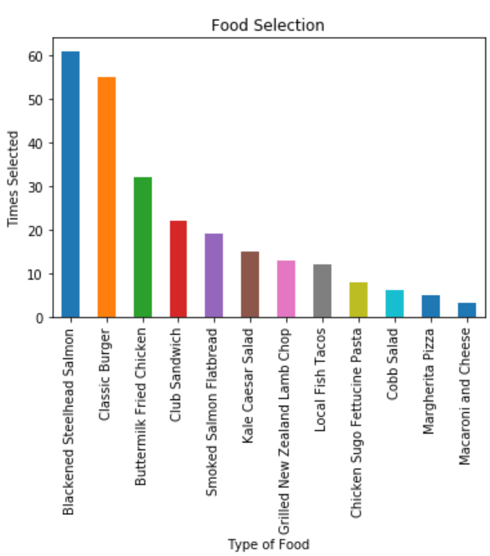


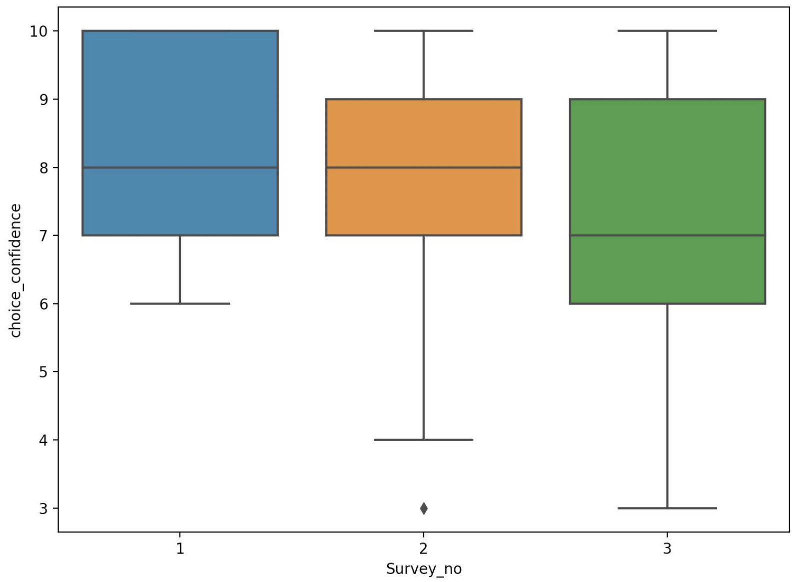
,


Correlation Matrix for 3 choices
Survey_no
choice_confidence
choice_easiness
1
-0.51
-0.51
1
Survey_no
choice_confidence
choice_easiness
0.9
0.6
0.3
0.0
−0.3


Correlation Matrix for 7 choices
Survey_no
choice_confidence
choice_easiness
1
0.021
0.021
1
Survey_no
choice_confidence
choice_easiness
1.0
0.8
0.6
0.4
0.2


Correlation Matrix for 12 choices
Survey_no
choice_confidence
choice_easiness
1
-0.1
-0.1
1
Survey_no
choice_confidence
choice_easiness
1.0
0.8
0.6
0.4
0.2
0.0


Food Selection
Times Selected
60
50
40
30
20
10
0
Blackened Steelhead Salmon
Classic Burger
Buttermilk Fried Chicken
Club Sandwich
Smoked Salmon Flatbread
Kale Caesar Salad
Grilled New Zealand Lamb Chop
Local Fish Tacos
Chicken Sugo Fettucine Pasta
Cobb Salad
Margherita Pizza
Macaroni and Cheese
Type of Food



Random Forest Classifier:

We one-hot encoded every possible selection for the food in the surveys. This allowed us to perform additional prediction classification with more variables as now each food selection was its own explanatory variable. We dropped the columns “Timestamp”, “age”, “date”, “time”, “hour”, “allergy_No”, and “num_choices” due to various problems such as converting time/date formats to int/float data types and age and allergy having only one value throughout the dataset. After converting all of the encoded data types to floats, we were ready to perform random forest classification. Without taking number of choices into consideration as a variable, keeping all of the data in the master file, we see an accuracy of classification for the target variable “Choice Confidence” of 23.81%. Considering there are 11 possible options for this target variable, our model performs better than randomly predicting 1/11 which would get you an accuracy of 9.1%. It’s interesting to note that some of the choices such as “0”, “1”, “2”, and “4” didn’t have any selections from the surveys. Maybe in future iterations of this study, we could group ranges of confidence from 0-3, 4-6, 7-10 to get a better idea of where people stand. When we switched the target variable from Choice Confidence to Choice Easiness, the accuracy was much better and we got 28.57%. This could be another interesting thing to focus on if continuing to run this study. Those two questions are essentially the same just different diction, so we could analyze diction in how we ask people how confident they are in their selections and see if the language itself in how the questions are presented to people correlate with their selections. After tuning some of the hyperparameters in the Random Forest classifier, we were able to achieve much greater prediction accuracy in our model. By changing max_features from 10 to 4, and doubling the n_estimators to produce more trees in the final majority voting decision in the random forest, the accuracy for the target variable of Choice Confidence increased from 23.81% to 35.29% and the accuracy for Choice Easiness target variable increased from 28.57% to 37.25%. This could be due to only one of the food choices being an option as max_features at 10 doesn't really make much sense in the context of our data. Also increasing n_estimators creates more decision trees and with a larger sample size of trees playing a part in the final majority voting mechanism, this should definitely produce greater accuracy. Obviously, in future iterations, it would definitely help to have a much larger sample size and maybe even feature engineer additional explanatory variables to look at.

Random Forest (Choice Confidence Target)
No tuning:

```
Random Forest Accuracy (Choice Confidence Target Variable):
0.23809523809523808
[[0 0 0 0 1 0 0]
 [0 0 0 1 1 0 0]
 [0 1 0 5 1 0 0]
 [1 2 0 4 2 1 4]
 [1 1 1 7 2 0 3]
 [0 2 1 1 3 0 0]
 [0 1 2 1 3 1 9]]
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           5       0.00      0.00      0.00         2
           6       0.00      0.00      0.00         7
           7       0.21      0.29      0.24        14
           8       0.15      0.13      0.14        15
           9       0.00      0.00      0.00         7
          10       0.56      0.53      0.55        17

   micro avg       0.24      0.24      0.24        63
   macro avg       0.13      0.14      0.13        63
weighted avg       0.24      0.24      0.24        63
```

Random Forest (Choice Easiness Target)
No Tuning:

```
Random Forest Accuracy (Choice Easiness Target Variable):
0.2857142857142857
[[ 0  0  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  2  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  4  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  1  0  0  0]
 [ 0  0  0  1  0 14  0  0  2  0  0]
 [ 0  0  0  0  0  8  0  0  0  1  1]
 [ 0  0  0  0  0  5  0  1  0  0  0]
 [ 2  0  2  0  0  7  1  0  0  1  0]
 [ 0  0  0  0  0  2  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  0  0  0  3]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.00      0.00      0.00         3
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         4
           4       0.00      0.00      0.00         2
           5       0.33      0.82      0.47        17
           6       0.00      0.00      0.00        10
           7       0.25      0.17      0.20         6
           8       0.00      0.00      0.00        13
           9       0.00      0.00      0.00         2
          10       0.75      0.75      0.75         4

   micro avg       0.29      0.29      0.29        63
   macro avg       0.12      0.16      0.13        63
weighted avg       0.16      0.29      0.19        63
```

Random Forest (Choice Confidence Target)
Tuned Parameters:

```
Random Forest Accuracy (Choice Confidence Target Variable):
0.35294117647058826
[[ 0  0  0  0  0  0  0]
 [ 0  0  0  1  0  0  0]
 [ 0  1  0  2  3  0  1]
 [ 1  2  0  4  2  1  4]
 [ 1  1  1  3  4  0  3]
 [ 0  1  0  0  2  0  0]
 [ 0  0  2  1  0  0 10]]
```

Random Forest (Choice Easiness Target)

```
Random Forest Accuracy (Choice Easiness Target Variable):
0.37254901960784315
[[ 0  0  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  1  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  2  0  1  1  0  0]
 [ 0  0  0  0  0  1  0  1  0  0  0]
 [ 0  0  0  1  0 14  1  0  1  0  0]
 [ 0  0  0  0  0  5  1  0  1  0  1]
 [ 0  0  0  0  0  3  0  0  0  0  0]
 [ 2  0  1  0  0  3  1  1  1  1  0]
 [ 0  0  0  0  0  1  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  3]]
```

Final Thoughts and Future Considerations

We had a fair amount of difficulty due to the limitations in our data collections. In the future, we would have liked to collect more demographic data for the survey, most notably gender and hunger level at the time of taking this survey. This would have given us more factors to analyze, and we could’ve tried to do some more sentiment analysis. We also would’ve liked to include more age groups and more respondents.