(25%) Given the data set, do a quick exploratory data analysis to get a feel for the distributions and biases of the data. Report any visualizations and findings used and suggest any other impactful business use cases for that data.

The first thing I did was starting off with some basic analysis of the data, allowing me to find any key outliers or biases of the data. To do this, I used Python and Scikit Learn for its simplicity and flexibility with various models. From Google Sheets, I downloaded the data as a .csv file as that is the most compatible file for data like this. To get the data specifically and plot it, I used various Python libraries. I used Pandas to read in and get all of the data, matplotlib to calculate plots for all of my data, and Seaborn to generate the viewable plots. I then exported all of the plots using the PdfPages library.

To get an overall synopsis of how the data categories stacked up against each other, I also plotted the distributions of each of the categories. We can see that for the Majors, there were a very large number of Chemistry and Biology Majors while barely any Civil Engineering or Fine Arts Majors. We can see another large skew with the Universities, with many Butler and IU University students and barely and Purdue or DePauw students. However, it seems that all of the items are near equally popular as they all hover around 500 orders per item. Meanwhile, when we take a look at the years, we can see that an overwhelming majority of entries are from Year 2 and Year 3 students, with an insignificant number of students from Year 1 and Year 4. These graphs can be found in *Distributions.pdf*

With the number orders being what we wanted to predict, it made the most sense to plot the other categories (Major, Time, University, and Year) against the number of orders each one received. Each of the graphs can be found in the *Order_By_{Catagory}.pdf* files. With this data, some clear correlations could be determined between the various categories. For example, there didn't seem to be any orders that the Year 2 and Year 3's equally liked. Most of them seemed to be only favored by one or the other. Mathematics majors also seemed to heavily favor the *Ultimate Grilled Cheese Sandwich (with bacon and tomato*) over any other option.

(30%) Consider implications of data collection, storage, and data biases you would consider relevant here considering Data Ethics, Business Outcomes, and Technical Implications

1. **Discuss Ethical implications of these factors**
2. **Discuss Business outcome implications of these factors**
3. **Discuss Technical implications of these factors**

Ethical Implications:

Privacy and Consent: Collecting customer data without their explicit consent can raise ethical concerns. It's important to inform customers about data collection and ensure that they willingly provide their information for predictive modeling.

Data Bias: Biases can be introduced in the dataset based on where the data is collected from. In this case, university affiliation could introduce biases related to socio-economic status, culture, and preferences. Using such biased data can lead to unfair or discriminatory predictions.

Fairness: The model should be designed to provide fair predictions across different demographic groups. Failing to do so may result in unfairly targeting or excluding specific customer segments based on their university, major, or academic year. Because there are so few Year 1 and Year 4 students in the data set, the model would not be accurate to predict the orders for Year 1 and Year 4 students.

Transparency: It's essential to ensure that the data collection process and model are transparent. Customers should know how their data is being used and how the model makes predictions. Lack of transparency can erode trust and lead to our app gaining a lot of negative publicity.

Business Outcome Implications:

Customer Trust: Ethical data collection and modeling practices can build trust with customers. They are more likely to engage with the business when they feel their data is handled with care and respect.

Reputation Management: If data breaches or unethical practices come to light, it can damage the reputation of the business. Negative publicity can have long-term consequences.

Legal Compliance: Non-compliance with data privacy regulations can lead to legal repercussions, including fines. Ensuring ethical data practices helps the business stay within legal boundaries.

Market Segmentation and Targeting: Biased data can lead to ineffective targeting. A better understanding of biases can help the business refine its marketing strategies and better serve its customer base.

Technical Implications:

Data Quality: Biased or incomplete data can lead to inaccurate predictions. Ensuring data quality through data cleaning and validation is essential for model performance.

Unwieldy Data: We need to make sure that the data is properly formatted so that our models are able to take in all the data and train with them.

Bias Mitigation: Implementing bias mitigation techniques in the model, such as re-sampling or re-weighting, can help reduce biases and make predictions fairer.

Model Explainability: To address transparency and ethical concerns, using interpretable models or model explanation techniques can help users understand how predictions are made. Using many models and picking the best one can be best to have the most accurate predictions.

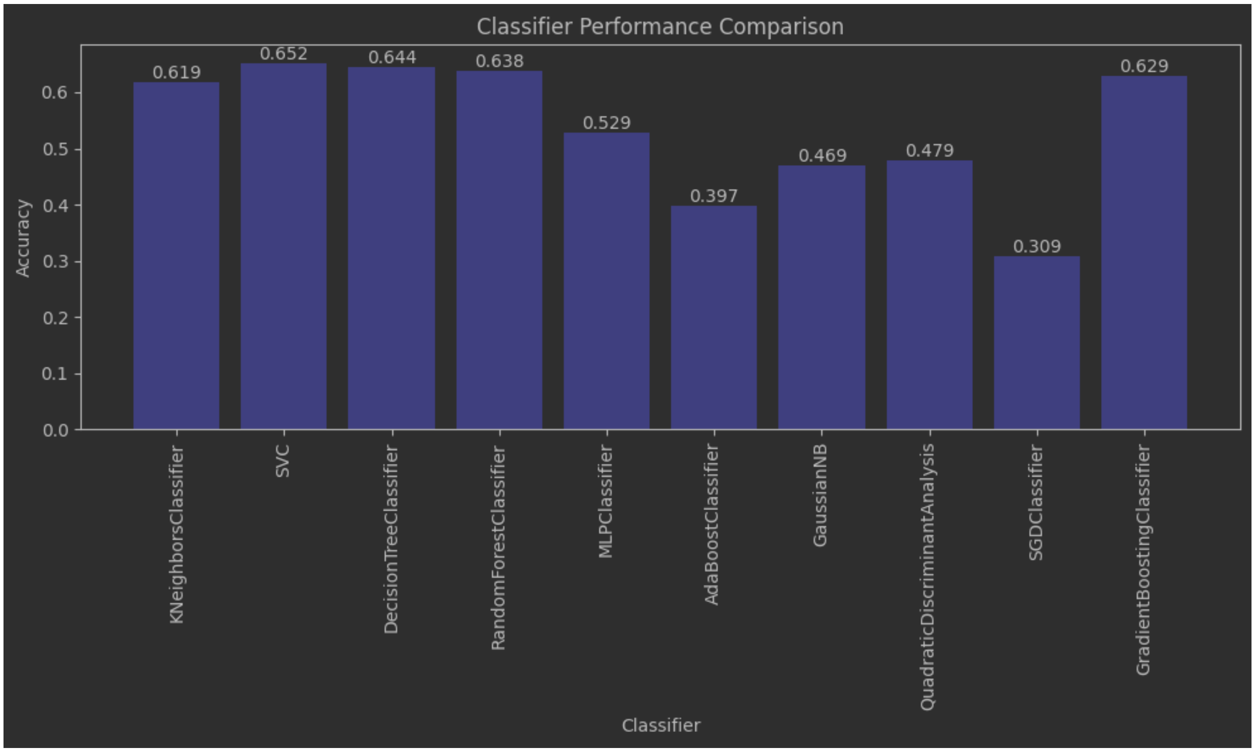
Data Security: Safeguarding customer data is crucial. Proper encryption and access controls should be in place to protect sensitive information.

Regulatory Compliance: The technical infrastructure should be designed to comply with data protection regulations such as GDPR or CCPA. This might include features like data anonymization or the ability to delete customer data upon request.

(35%) Build a model to predict a customers order from their available information. You will be graded largely on your intent and process when designing the model, performance is secondary. It is strongly suggested that you use SKLearn for this model as to not take too much time. You may use any kind implementation you would like though, but it must be pickelable and have a ".predict()" method similar to SKLearn

1. **Outline your process for model selection, training and testing. Including data preparation.**
2. **Design a function that prepares your data by loading the provided dataset and processes it into an appropriate machine readable format if necessary.**
3. **Design a function to train your model and pickle it.**
4. **Train and test your model. Submit any training, testing and model selection visuals or metrics.**
5. **Upload your work to GitHub and link the repository, make sure it is public.**

For my model selection, I decided to use many classifiers to see which one could perform the best. To do this, I imported many classifiers and created an array of imported classifiers to run through the data set with each one. Looking at the accuracies for each of the classifiers, it seems that the SVC (Scalable Linear Support Vector) classifier was the best performing one, following bu the Decision Tree Classifier. I believe that the Decision Tree Classifier would have been significantly more accurate if ScikitLearn's implementation supported categorical variables.



Process of Code

1. First, I load the data using Panda using read_csv
2. Then I preprocess all of the data, where I take in the categories using Label Encoder and put them all into their own arrays.
3. Load in all of the classifiers and set their parameters, with most of them being their default values.
4. For each classifier, I train and evaluate them, splitting the data set into 80% for training and 20% for testing.
5. After testing the classifiers, I take the accuracies from the test and list them
6. Then I take each of the accuracies and plot them into a graph

(10%) Given the work required to bring a solution like this to maturity and its performance, what considerations would you make to determine if this is a suitable course of action?

With machine learning models, it would be great to know what a potential customer would like to order. However, determining data to differentiate students from each other would be difficult as there are many ethical, business, and technical implications. It wouldn't be easy to sort students based on the inputs given in the data set as there would need to be cameras to determine what a student is or have them input the data themselves, which could lead to heavy bias.

To have this implemented on a large scale, it would be more suitable to use a model that reaches 80% accuracy, instead of around 60%. To do this, using a custom model other than the ones found on Scikit Learn would be a good idea. We would also need more variety in our data, as some majors and years were completely left out.