- 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.
    - There is a high distribution of Year 2 and Year 3 data. There is very little data for Year 1 or Year 4 students. The decreased amount of data could decrease the accuracy of a model's prediction for Year 1 and Year 4 students. Within Year 2, the most popular order is Indiana Pork Chili and within Year 3, the most popular orders are Cornbread Hush Puppies, Indiana Buffalo Chicken Tacos, and Sugar Cream Pie. The food truck could give out freshman and senior discounts to attract more students from Year 1 and Year 4.
    - The most popular majors are Chemistry, Biology, and Astronomy. Chemistry's and Biology's most popular order is Indiana Pork Chili and Astronomy's is Ultimate Grilled Cheese Sandwich.
    - The most popular universities are Butler, Indiana State, and Ball State. There is very little data for Purdue, Depauw, and Valparaiso. Butler's most popular order is Indiana Pork Chili, Indiana State's is Hoosier BBQ Pulled Pork Sandwich, and Ball State's is Indiana Corn on the Cob. The food truck could try marketing to other universities to increase sales from them. This approach may be futile due to the distance from the food truck and the other universities. Students are not usually willing to drive a long distance just to go to a food truck so the franchise could deploy closer food trucks to those far-away universities.
    - The distribution of times is centered around 12:30 pm. The further away from lunch time, the lower the number of orders. The morning and evening distributions are roughly similar.
    - The distribution of orders for each item is roughly uniform. The orders range from around 470 to 510. The most popular order is Sugar Cream Pie and the least popular is Hoosier BBQ Pulled Pork Sandwich. The food trucks should account for this accordingly and purchase slightly more ingredients for popular orders and less for the unpopular ones.
- Consider implications of data collection, storage, and data biases you would consider relevant here considering Data Ethics, Business Outcomes, and Technical Implications
    - Discuss Ethical implications of these factors
        - The data collection and storage seem to be ethical as the student information is anonymized and there is no risk of leaking personal information.
    - Discuss Business outcome implications of these factors
        - The business outcome is greatly affected by biases in the data. The lack of data on Year 1 and Year 4 students will make predictions on them less accurate and increase the likelihood of giving those customers discounts, decreasing revenue. Any factor that could contribute to a decrease in accuracy of the model will have an effect on the business as it will decrease revenue.
    - Discuss Technical implications of these factors

        - The technical implications are not very severe. The model training doesn't seem to require too much time or computational power.
- Build a model to predict a customer's 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
    - Outline your process for model selection, training and testing. Including data preparation.
        - I selected the KNearestNeighbors model because of its simplicity and effectiveness in classification problems.
        - I first created a pandas DataFrame of the dataset. I used a OneHotEncoder to transform the categorical features into a numerical version. Similarly, I then used a LabelEncoder to transform the categorical responses into a numerical version.
        - I used Grid Search Cross Validation to test different KNN models with K ranging from 1 to 30. I used classification accuracy as the metric for the cross validation. I used 10 fold cross validation in the Grid Search.
        - The best KNN model was with K = 12 with an accuracy of 61.84%.
- 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?
    - I think more data needs to be collected on the lacking factors before an accurate model can be deployed. Once there is more data collected, the model should be more accurate for the lacking factors like Year 1 and Year 2 students. Once deployed, this model would save money for the franchise and effort for the employees.