(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.

Based on the exploratory data analysis of the dataset, it was found that the most popular food items among students are the Sugar Cream Pie and the Indiana Pork Chili, both exceeding 500 sales. Conversely, the Fried Catfish Basket and Hoosier BBQ Pulled Pork Sandwich were less favored, with less than 500 sales each. A notable trend emerged when examining orders by major; different majors exhibited distinct food preferences. For instance, Biology majors showed a preference for Sugar Cream Pie and Indiana Pork Chili, while Physics majors ordered more Cornbread Hush Puppies, and Economics majors had a penchant for the Ultimate Grilled Cheese Sandwich with bacon and tomatoes. Similarly, food preferences varied by university. Students from Butler University ordered more Sugar Cream Pie and Indiana Pork Chili, whereas Indiana State University students ordered more Cornbread Hush Puppies and Sweet Potato Fries. These findings suggest that both the major and university of a student can influence their food preferences. This information could be leveraged to tailor menu offerings to different groups of students, potentially increasing sales. Furthermore, this data could be utilized to optimize inventory management by predicting which items will be most popular among different groups of students and to tailor marketing efforts towards specific majors or universities based on their food preferences. However, it is crucial to consider potential biases in the data, such as overrepresentation of certain majors or universities, as this could skew the results.

(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 considerations play a critical role in ensuring that data collection involves informed consent and the anonymization of personally identifiable information to protect student privacy. Additionally it is important to recognize any biases, in the data collection process to ensure unbiased and accurate results are obtained. The business implications of the insights derived from this data are substantial, as they can inform menu customization, potentially leading to increased sales and improved customer satisfaction. However, the reliability of these insights hinges on the quality of the data; biased or incomplete data can lead to erroneous conclusions and adversely affect business decisions. From a technical perspective, prioritizing secure and reliable data storage is imperative, along with maintaining data quality by addressing missing or inconsistent data. Additionally, the choice of appropriate tools and technologies for data analysis can significantly impact the efficiency and accuracy of the derived insights.

(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.**

For this task, I chose the RandomForestClassifier as my model, given its robust ensemble learning approach and the capability to handle a substantial number of features, making it well-suited for our dataset. Initially, I loaded the data from a Google Sheet into a pandas DataFrame, separating the target variable 'Order' from the rest of the data. Using an OrdinalEncoder, I encoded the categorical variables into numerical values and split the data into training and testing sets. I then trained the model on the training set using the fit method of the RandomForestClassifier. Next, I evaluated the model's performance on the testing set, printing a classification report that highlighted essential metrics like precision, recall, and F1-score. Finally, I pickled the trained model, as well as the scalar and encoder used in data preparation, to ensure their availability for future applications.

(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?

The model's current accuracy stands at 65%, marking a significant improvement from the initial 45% achieved with Logistic Regression. Although not perfect, this is a promising start, and there is potential for further enhancement through additional tuning or data. From a business standpoint, the model could have a substantial impact as it can predict with 65% accuracy what a student would order based on their information. This could facilitate better decision-making and boost efficiency by providing insights into which universities could generate the most sales. However, it's crucial to consider the data management aspect. The Google Sheet needs to be updated with every sale to maintain accuracy, necessitating regular retraining of the model to incorporate new data. These considerations are vital in determining the suitability of this course of action.