(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.

Part A: Rudimentary Analysis

1)

```
ai_data_df.info()
```

```
<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
```

The given dataset contains 5000 rows and does not contain any nulls. The columns in the dataset are: Year, Major, University, Time and Order

2) Next, we look at the possible values each of these columns can take and the frequency in which they occur.

   a) Year column:

```
Frequency for Year :
Year 3    2719
Year 2    2273
Year 1       5
Year 4       3
Name: Year, dtype: int64
```

We see that almost the entire dataset consists of orders from Year 2 and Year 3 students.

b) Major column:

```
Frequency for Major :
Chemistry                   640
Biology                     635
Astronomy                   619
Physics                     610
Mathematics                 582
Economics                   511
Business Administration     334
Political Science           309
Marketing                   239
Anthropology                146
Finance                     135
Psychology                   76
Accounting                   62
Sociology                    31
International Business       29
Music                        21
Mechanical Engineering       11
Philosophy                    4
Fine Arts                     3
Civil Engineering             3
Name: Major, dtype: int64
```

The data consists of a total of 20 majors (as seen in descending order of number of orders above).

- The top 10 Majors correspond to 92.5% of the total orders.
- The top 4 Majors (Chemistry, Biology, Astronomy & Physics) make up 50% of the total number of orders.

c) University column:

```
Frequency for University :
Butler University                                          1614
Indiana State University                                   1309
Ball State University                                      1085
Indiana University-Purdue University Indianapolis (IUPUI)   682
University of Notre Dame                                   144
University of Evansville                                   143
Indiana University Bloomington                              12
Valparaiso University                                        9
Purdue University                                            1
DePauw University                                            1
Name: University, dtype: int64
```

- Butler University students are our biggest customers (32.3%)
- Butler University, Indiana State University & Ball State University students are responsible for a little over 80% of the total orders.

d) Time column:

```
Frequency for Time :
13    1316
12    1314
14     883
11     857
15     282
10     247
16      49
9       40
8        8
17       4
Name: Time, dtype: int64
```

- By listing down the distinct values of the Time column in ascending order (8,9,10,11,12,13,14,15,16,17), we observe that our Food truck operates from 8 am to 5 pm.

- **11 am to 2pm** is the time window that sees the most sales **(~87% of the orders**).

e) Orders column:

```
Frequency for Order :
Sugar Cream Pie                                                  512
Indiana Pork Chili                                               510
Cornbread Hush Puppies                                           510
Sweet Potato Fries                                               508
Ultimate Grilled Cheese Sandwich (with bacon and tomato)         503
Indiana Buffalo Chicken Tacos (3 tacos)                          496
Indiana Corn on the Cob (brushed with garlic butter)             495
Breaded Pork Tenderloin Sandwich                                 494
Fried Catfish Basket                                             490
Hoosier BBQ Pulled Pork Sandwich                                 482
Name: Order, dtype: int64
```

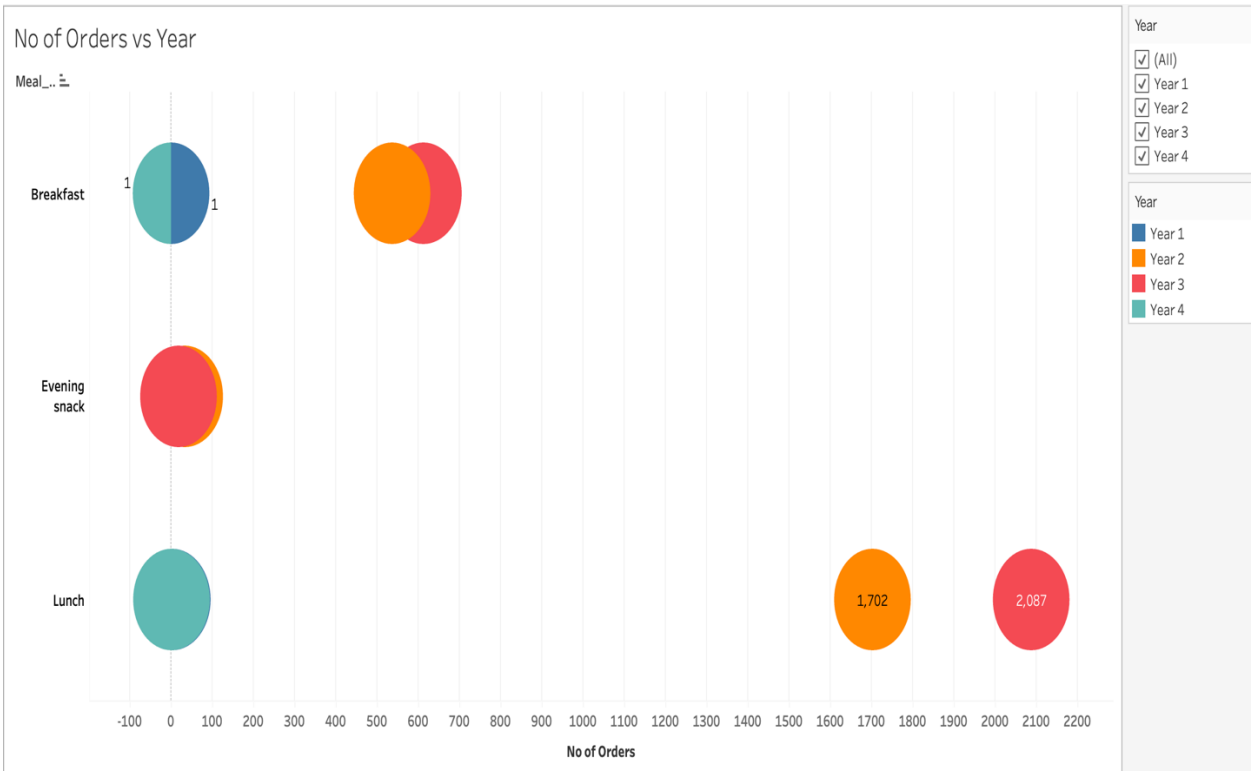
- There are 10 items on the menu.
- All the items on the menu are sold roughly in the same numbers. No item is a clear favorite.

Part B: Intermediate Level Analysis

1) We divide the working hours of the food truck into three parts as follows:

   a) 8 am to 11 am (8, 9, 10,11): Breakfast
   b) 12 pm to 3pm (12, 13, 14, 15): Lunch
   c) 4 pm to 5 pm (16, 17): Evening snack

   On visualizing, we observe that Year 3 students place **significantly more lunch orders** than Year 2 students.



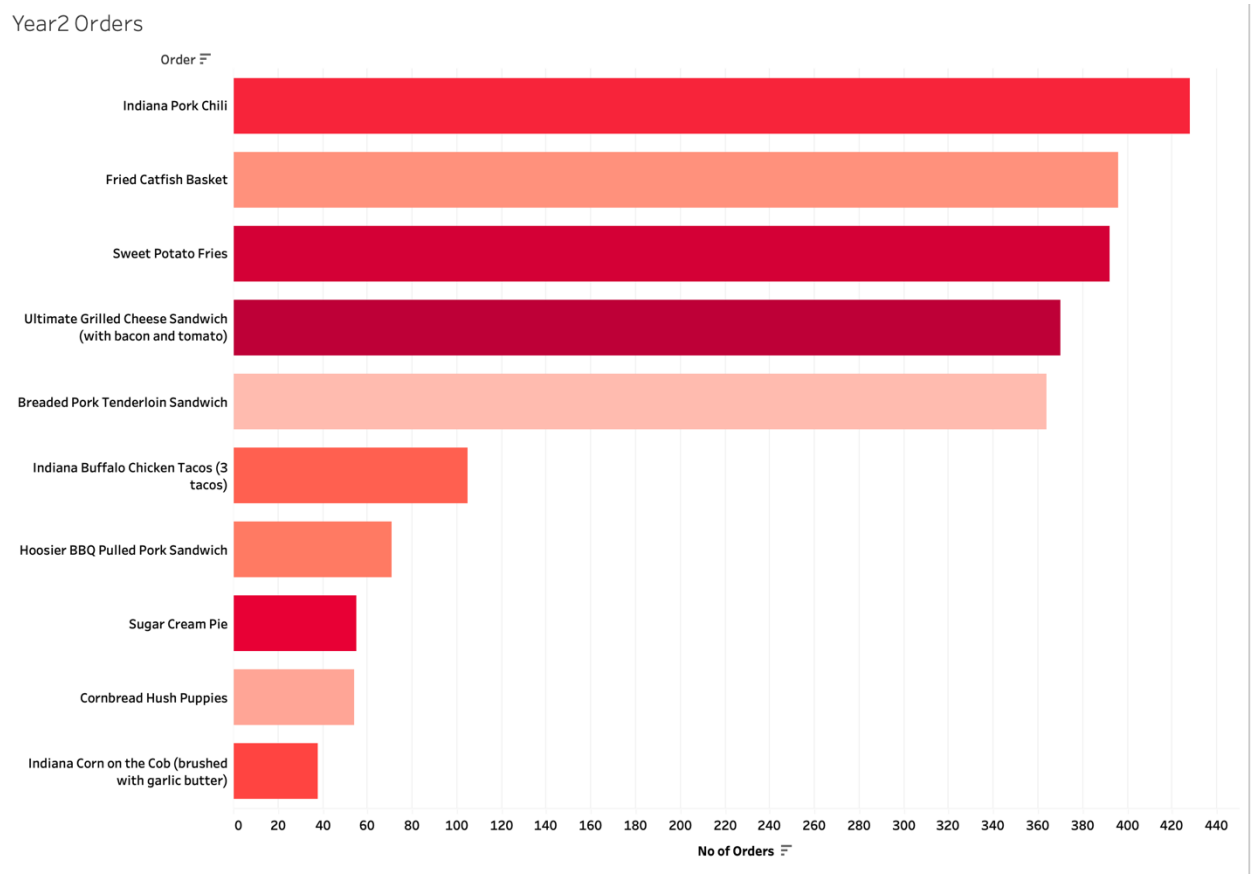
2) What foods do students of Year 2 and 3 prefer?

   We ignore Year 1 and Year 4 from this analysis as there are too few datapoints for Years 1 & 4.

   Year 2:
   Top 5 foods

```
ai_data_df['Order'][ai_data_df['Year']=='Year 2'].value_counts().head(5)
```

```
Indiana Pork Chili                                          428
Fried Catfish Basket                                        396
Sweet Potato Fries                                          392
Ultimate Grilled Cheese Sandwich (with bacon and tomato)    370
Breaded Pork Tenderloin Sandwich                            364
Name: Order, dtype: int64
```

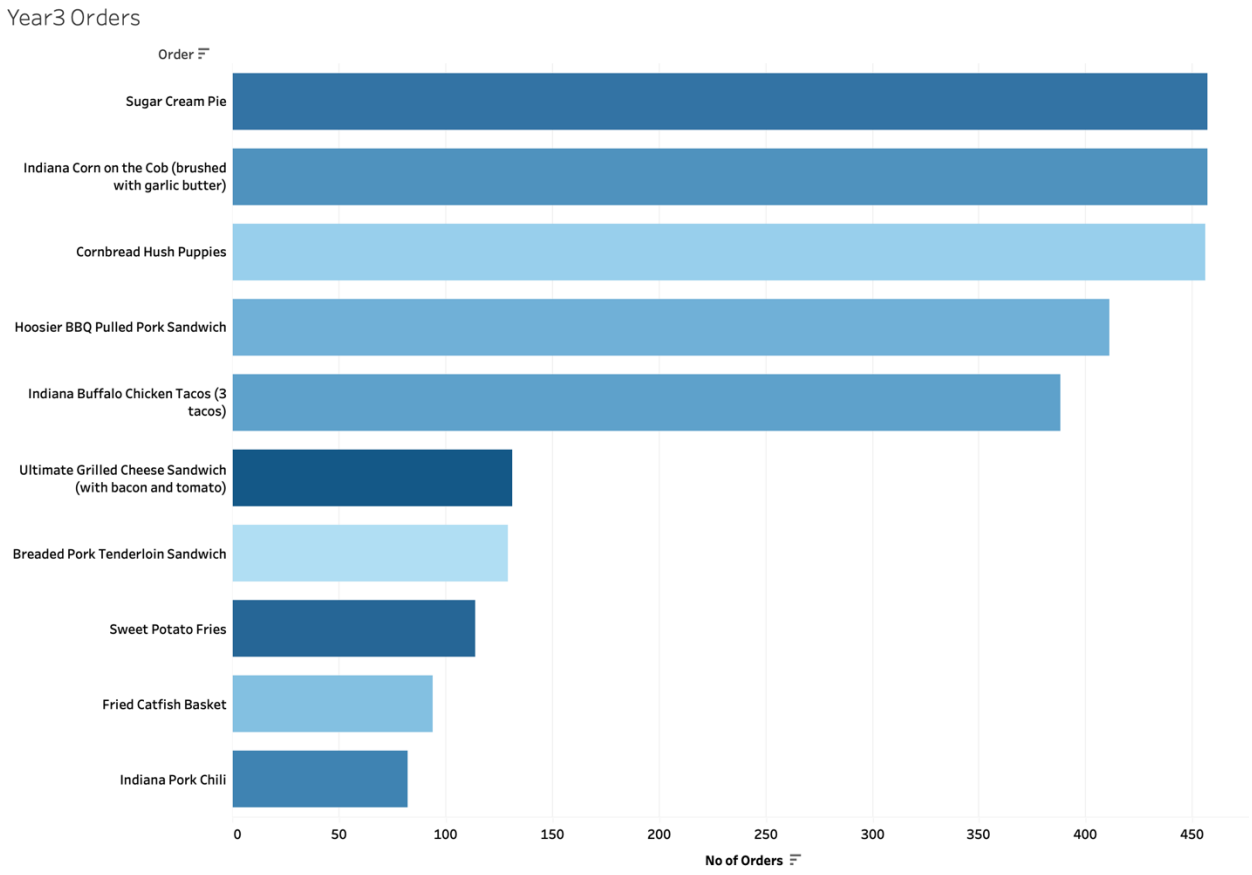


It is evident that there is a clear liking displayed by Year 2 students towards five specific items out of ten options. In fact, Year 2 students bought one of the top 5 foods 86% of the time. This is a clear bias.

We explore below, the food choices of Year 3 students in a similar fashion.

```
ai_data_df['Order'][ai_data_df['Year']=='Year 3'].value_counts().head(5)
```

```
Sugar Cream Pie                                          457
Indiana Corn on the Cob (brushed with garlic butter)     457
Cornbread Hush Puppies                                   456
Hoosier BBQ Pulled Pork Sandwich                         411
Indiana Buffalo Chicken Tacos (3 tacos)                  388
Name: Order, dtype: int64
```



Just like Year 2 students, Year 3 students also display a clear inclination towards five specific items out of the ten choices. They choose one item from the top 5 about 80% of the time!

3) Year 2 vs Year 3 food preferences:

We observe that the top 5 food choices of Year 2 students differ completely from that of Year 3 students.

Food priority Rank	Year 2 – Top 5 foods (% of all orders)	Year 3 - Top 5 foods (% of all orders)
1	Indiana Pork Chili (18.8%)	Sugar Cream Pie (16.8%)
2	Fried Catfish Basket (17.4%)	Indiana corn on the cob (with garlic butter) (16.8%)
3	Sweet Potato Fries (17.2%)	Cornbread Hush puppies(16.7%)
4	Ultimate Grilled Cheese Sandwich (with bacon and tomato) (16.2%)	Hoosier BBQ pulled pork sandwich (15.1%)
5	Breaded Pork Tenderloin Sandwich (16%)	Indiana Buffalo chicken Tacos (14.3%)

We observe that the top 5 food choices of Year 2 students and Year 3 students are mutually exclusive!
Thus, if the customer is a Year 2 student, there is a very high likelihood that the order is going to be from one of the top 5 preferences of Year 2 students. The same logic applies to Year 3 students as well.
Also, if the time of order falls between 12 and 4, it is more likely that a Year 3 student is placing the order (as they demonstrate a higher tendency to place lunch orders).

Part C: Advanced Analysis

We observe that all the columns in the dataset are categorical columns (the order time column is an ordinal categorical variable). Our aim is to quantify the correlation between these variables and the Ordered food.
Since all variables are categorical, we perform a Chi-square test for correlation.

I have added a new column “Meal_type” which takes one of the three values (Breakfast, Lunch & evening snack) depending on the order time.

On running the Chi-square analysis in python, we observe the following:

```
feature_vs_correlation
```

```
{'Year': 0.3813953575194632,
 'Major': 0.25143208517789967,
 'University': 0.25172121175512524,
 'Time': 0.25086443856455637,
 'Meal_type': 0.343501995393279}
```

The closer the correlation value to 1, the more correlated is the variable to the output class. We see that the ‘Year’ column is highly correlated with the output. This is corroborated by the differences in food preferences of Year 2 and Year 3 students as seen in the Intermediate Level Analysis section.
The engineered feature (Meal_type) is also highly correlated with the final output, thus increasing the model’s prediction accuracy.

As a rule, correlation coefficients below 0.3 denote low correlation. Thus, we can infer that the order time, university and Major do not significantly impact the order.

(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:

a) Privacy implications:

The classifier model in its current form is limited. It does not capture any personal/ sensitive information. The more relevant data (features) we capture, the better are our odds at making correct predictions. This necessitates capturing more data than the current scope. For instance, if the customer's name is captured, it becomes a unique identifier for the customer. Having this data could greatly enhance our odds of making a correct prediction of the food order. If the customer's locality (address) is captured, further analysis can be done on whether it is a relevant feature.

Care must be taken to ensure proper consent is obtained from the customer for capturing personal data. Customers should be made aware of means to withdraw their personal data from our systems at will. Strict data security measures must be established.

b) Data bias implications:

ML models are notorious for propagating the underlying biases. If there are race/ gender-based biases in the data, the same will reflect or even amplify in the outcome of the model.

c) Environmental implications:

AI models can optimize operations for cost-efficiency, but not necessarily for environmental sustainability. (Fun fact: ChatGPT consumes 500 ml of water for every 20-40 questions, to maintain server temperatures).

Business Implications:

The success of our ML classifier model is fully quantifiable. On each successful prediction, we save 10% of the order value, which we would lose otherwise. It helps to look at each transaction as a sale with an opportunity loss component. We aim to minimize the opportunity loss here. We trade off opportunity loss for getting higher engagement & in turn, higher sales volume.

It is difficult to quantify the exact increase in customer engagement that has resulted from the "10% off on wrong prediction" offer. Our efforts are best invested in increasing the prediction accuracy.

To capture the sales data, we need to link the billing software to a database in the backend. This comes at a cost.

If we want to scale up the model to multiple food trucks, the model's parameters would need to be tuned specifically to capture the trends in the locality of the given food truck. This would need multiple instances of the model to be implemented & run in real-time. This set up would require a server, which would increase the overall cost.

Also, if we choose to collect relevant personal information from the customers, data security systems need to be established. This costs too.

All these costs need to be quantified and evaluated to decide whether the whole idea is worthwhile.

Technical Implications:

For the idea of predicting orders to work continually, the underlying ML model must run frequently to stay in tune with changing preferences. Since ours is a food truck business, the scale of the operation can be enhanced significantly only by increasing the number of trucks. If the given data corresponds to, say, 30 days, we average about 167 orders per day (5000/30). In other words, 167 rows of data get added to our dataset each day. Also, the change in purchase trends isn't likely to change drastically overnight. Thus, the Classifier model doesn't require a real-time connection to the sales dataset and can be run on a need basis.

Let us try quantifying the revenue in terms of the success rate of the model.

For simplicity's sake, let the price of each food item be $X.

Let the model's Success rate be a

a = [No of correct predictions]/[Total no of orders]

Let the Total no of orders be n.

Revenue, R = anX + (1-a)(0.9 nX) {since we lose 10% on each wrong prediction}

R = (a + 0.9)nX – (0.9a)nx

R = (0.1a + 0.9)nX

When a = 0, R = 0.9nX {We make 10% less total revenue in exchange for increased customer engagement}

When a = 0.75 , R = 0.975nX {A model with 75% prediction accuracy would reduce our opportunity loss from 10% to only 2.5%}

When a = 0.9, R = 0.99nX {A model with 90% accuracy would reduce the opportunity loss to just 1% in exchange for increased customer engagement}

Also, to maximize their odds, the customers might choose deliberately to buy out of their set preferences. This will invariably add bias to the model

as we wouldn’t necessarily capture the real customer preferences, under a rigged circumstance such as this.

(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 ultimate measure of the success of this solution is cost-effectiveness. Expenses should be less than the potential increase in revenue obtained by increased customer engagement.

The expenses involved are as follows:

a) Technical labor cost
b) Database and server related costs
c) Opportunity loss up to 10% of the total revenue
   Out of the above three, only the opportunity loss is a variable expense.

   The difference in revenue before and after the implementation of the 10% off offer can be a measure of increased customer engagement. The solution is worthwhile if the increase in revenue is more than the net expenses incurred. (More details on minimizing opportunity loss in the Technical Implications section.)