XTERNSHIP – AI Question

Samyak Kashyap Shah | shahsamy@iu.edu

1. EDA

- Firstly we load the dataset and to exploratory data analysis on the given dataset, using pandas module.

```
[1]: import pandas as pd
```

```
[2]: df = pd.read_csv('XTern 2024 Artificial Intelegence Data Set - Xtern_TrainData.csv')
```

```
[3]: df.head()
```

[3]:

	Year	Major	University	Time	Order
0	Year 2	Physics	Indiana State University	12	Fried Catfish Basket
1	Year 3	Chemistry	Ball State University	14	Sugar Cream Pie
2	Year 3	Chemistry	Butler University	12	Indiana Pork Chili
3	Year 2	Biology	Indiana State University	11	Fried Catfish Basket
4	Year 3	Business Administration	Butler University	12	Indiana Corn on the Cob (brushed with garlic b...

```
[4]: df.info()
```

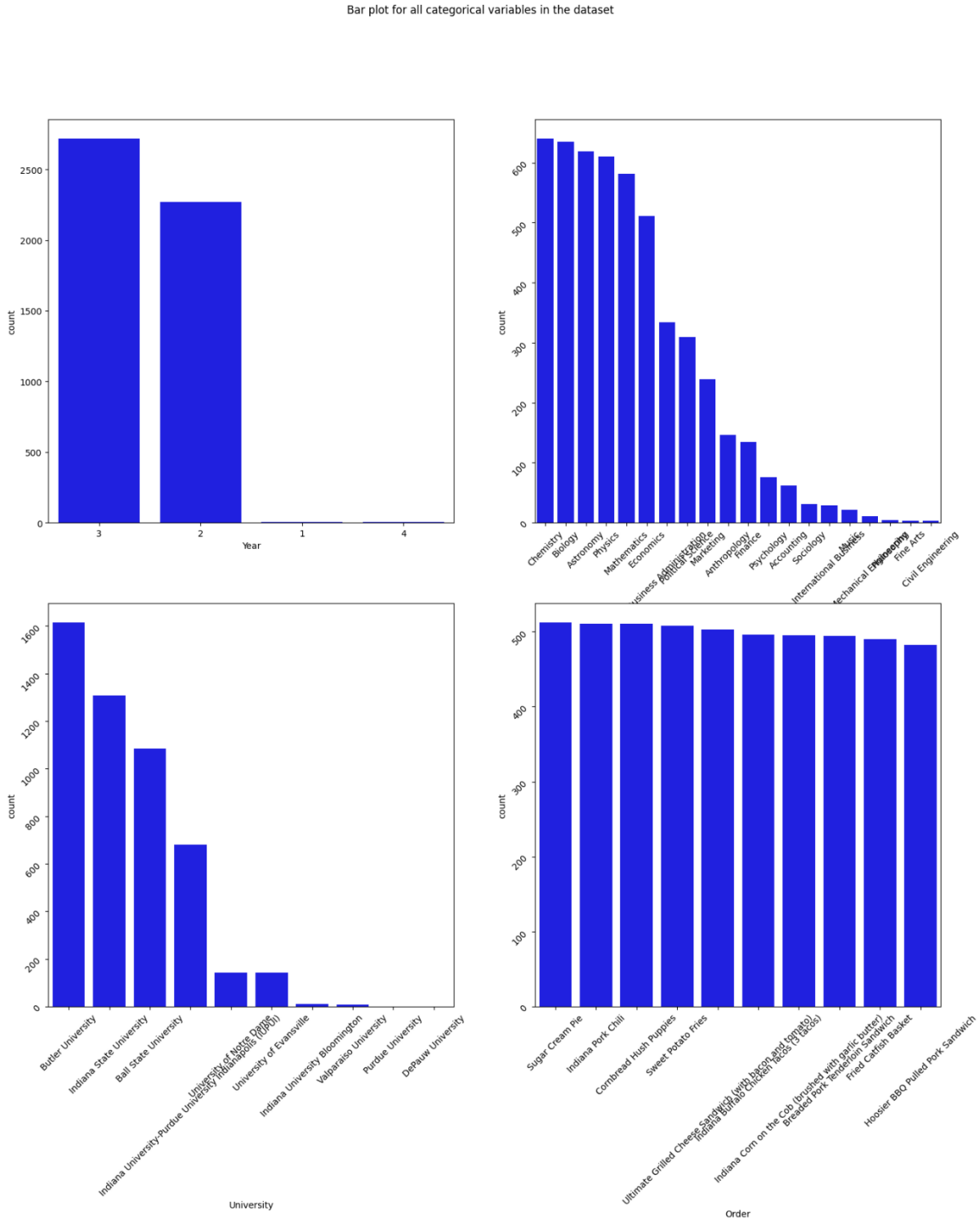
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Year        5000 non-null   object
 1   Major       5000 non-null   object
 2   University  5000 non-null   object
 3   Time        5000 non-null   int64
 4   Order       5000 non-null   object
dtypes: int64(1), object(4)
memory usage: 195.4+ KB
```

- Here we get to understand the data and number of inputs and outputs. For the dataset we are provided with contains, four dependent variables viz. Year, Major, University and Time. Since these variables serve as important factors for predicting outputs i.e. 'Order'.

[12]:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Year	5000.0	NaN	NaN	NaN	2.544	0.501313	1.0	2.0	3.0	3.0	4.0
Major	5000	20	Chemistry	640	NaN	NaN	NaN	NaN	NaN	NaN	NaN
University	5000	10	Butler University	1614	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Time	5000.0	NaN	NaN	NaN	12.5282	1.357193	8.0	12.0	13.0	13.0	17.0
Order	5000	10	Sugar Cream Pie	512	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We got to know from the dataset that most of the orders come from Year 3 students. Also, the major that orders the most is Chemistry, and Butler University sees the greatest number of students that order. Sugar Cream pie stays the most ordered food.



The above image shows us a distribution of the categories and their values.

2. When we consider data collection, storage and data biases in the context of developing a food order prediction model, it's important to take into account, various ethical, business and technical implications.
    a. Ethical implications: It is important to collect data with the consent of the individual and securely store it to avoid data leakage, which may inadvertently affect the individual. The participants should be informed regarding the usage of their data and the users should be informed regarding the usage of the prediction model trained on their dataset.
    b. Business Implications: Inaccurate predictions can cause customer dissatisfaction and they may receive recommendations that do no align with their preferences, which could harm the user experience. This could lead to reduction in brand image and improved models can help build better image.
    c. Technical Implications: Data collection and preprocessing must ensure data to be of a good quality. Inaccurate or incomplete data can negatively impact the model's performance. The model should be scalable as the user base can grow and handling the data should be easier at later stages.
3. Model

- I went through the sklearn library to check for best fit model for prediction. Usually, content based filtering/collaborative based filtering are the best recommender systems. However for easier workflow, I chose RandomForestClassifier as my model. I tested out other models as well, like DecisionTree and Nearest Neighbors, however I got the highest accuracy using RandomForestClassifier.
- After training and saving the model, I checked the score, however the accuracy I got wasn't that much and the model can be tweaked to achieve more accuracy

4. I would take in mind the following factors:

- Data quality and availability.
- The business idea behind the project
- Feasibility to achieve the results
- Complexity of the model.
- Scalability of the model