Project Process Book

Data Visualization (Fall 2025) group 6:
- Will Watkins
- Bryson Newman
- Eden Sharp

I. Overview and Motivation

Primary Goal:

The primary goal of this project is to analyze trends in customer orders from FoodPanda across five cities in Pakistan. The aim is to uncover any potential correlations and insights related to customer orders, preferences, and behaviors.

Motivation:

Food delivery services have been growing rapidly in Pakistan since post COVID. Being able to understand how different cities engage with the platform could provide valuable insights for both FoodPanda and local businesses, improving service and efficiency.

Questions:

- What specific trends in customer orders can be identified across the five cities?

II. Data

Source:

The dataset from FoodPanda was found on Kaggle and can be found below: https://www.kaggle.com/datasets/zubairamuti/foodpanda-review-dataset

Cleanup:

Since the idea was to create a scatter plot, rendering every point on the plot would make the graph unintelligible. So the scatter plot filtered out a random 25% of the points to be rendered. This is why the points change every time the page is refreshed.

III. Chronological Process Entries

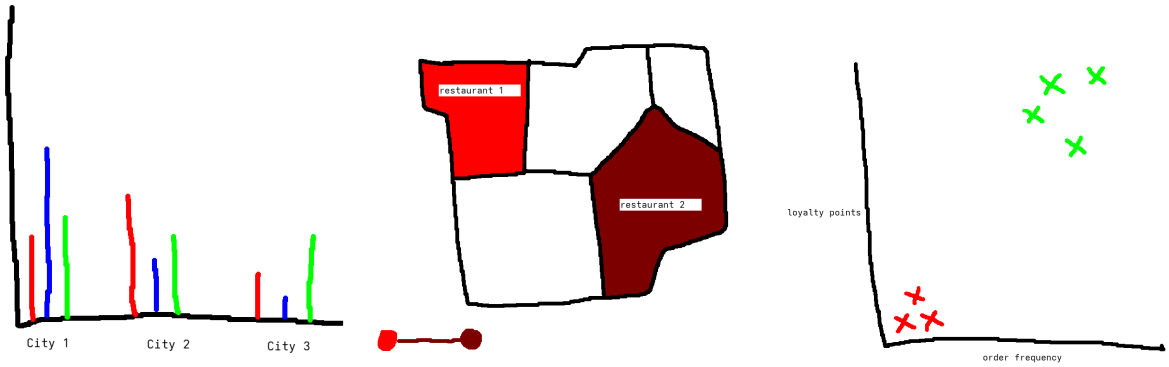
Dataset Collection:

We started out in our trying to find a fitting dataset we might like to try and visualize, and ended up looking in kaggle and found our FoodPanda dataset for online orders, which looked superficially promising to find insights in which weren't obvious in

the kaggle overview. We submitted that as our dataset proposal and got a thumbs up from it, probably because at the time it was very hard to tell its lack of potential.

Project Proposal:

For this milestone we attempted to make a concept art of a layout that we thought would do a good job at communicating some insights in the FoodPanda dataset. We ended up with these sketches which were examples of ideas we discussed:



These concepts were of course what we thought a simplification of our data might look like idealistically.

Project Prototype:

We ended up scrapping the idea of the map and instead replaced it with a pie chart. We wanted to try and give better context for the points on the scatter plot and show off what different categories of food they could fall under. We added a rudimentary interaction where when you click on a city you can see the associated points on the scatterplot. For our final submission, we want to discard the pie chart in favor of a distribution chart or something similar. We're still figuring out which visualization tools we want to utilize to properly display everything, but we at least have a foundation at this point.

IV. Exploratory Data Analysis

Initial Visualizations:

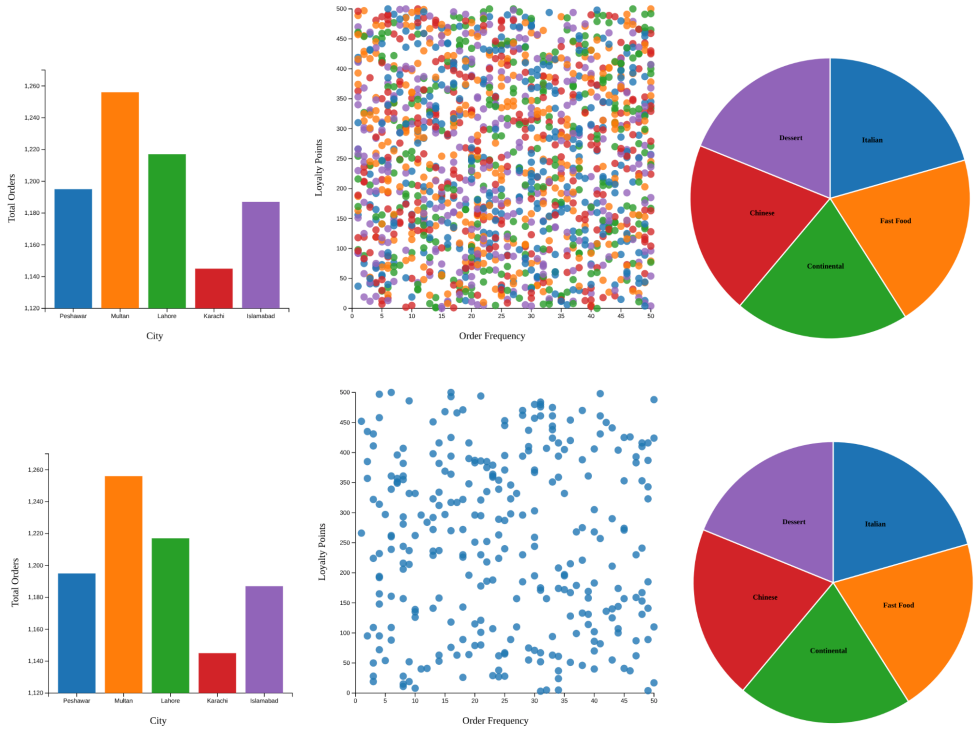
- Created the scatter plot to display the order frequency and loyalty points.
- Created bar chart to show volume of orders per city.
- Created a pie chart to show the distributions of the different genres of food.

Insights Gained:

This showed that our original idea of having a scatterplot to display all of our points would not work as well as we intended it to. Visualizing the entirety of our dataset proved to look extremely busy and ultimately doesn't bring any interesting trends to light. We will need to adjust our visualizations to have better interconnectivity and visual clarity.
Implementation:

Interactive Visualizations:

- Implemented a dashboard using D3js that allows users to filter data by city and food category.
- The scatter plot points are colored the same as the associated bar chart city.

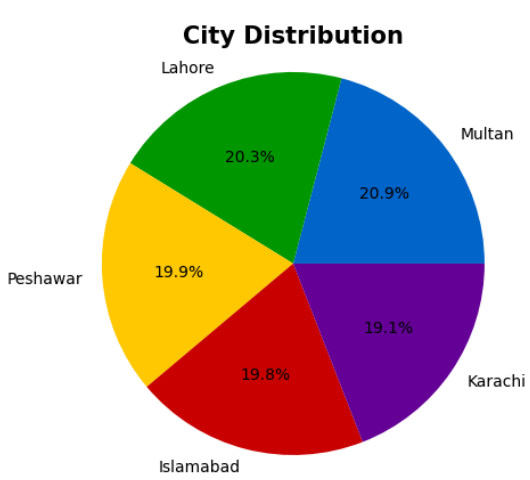


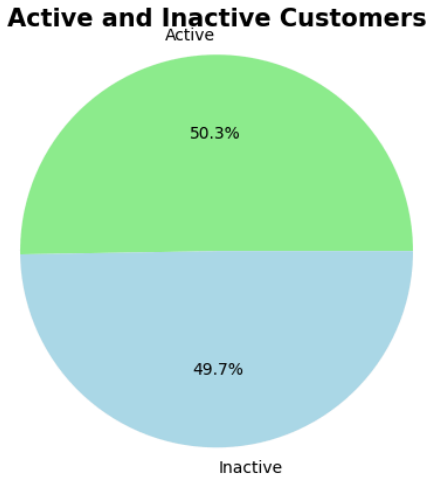
Evaluation

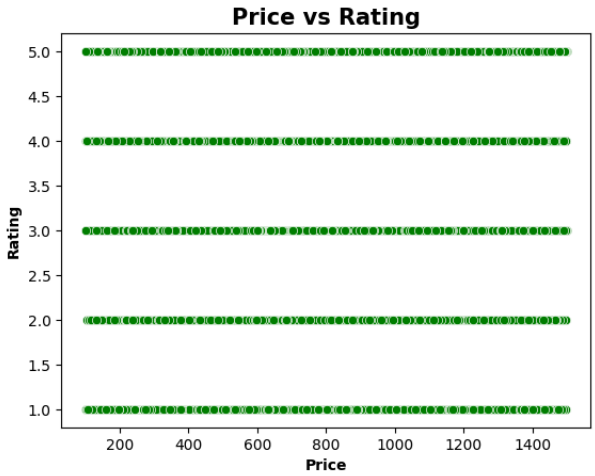
As of right now the visualizations did not prove to be effective in answering the initial questions. Although we attempted to pivot to align ourselves with provided feedback, we still need to work on finding a better correlation between the cities and another data point. The data currently still looks unstructured and uncorrelated. We will need to look deeper into the data and think about what other connections would be interesting to see and play with.

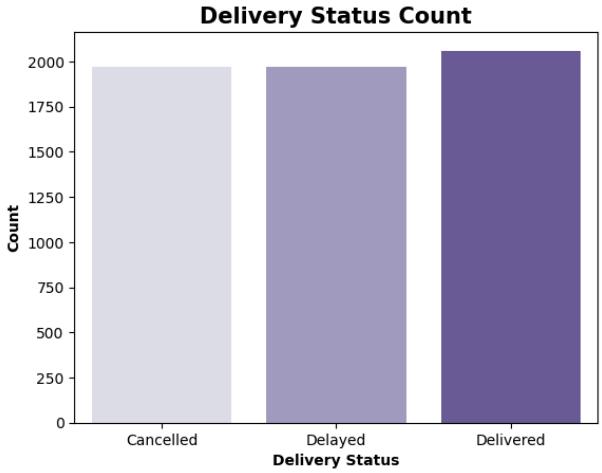
V. Issues with Dataset

In the process of discussing how to best proceed from here, we ran into a big issue: **unexpected uniformity**. Save for our bar chart, the rest of the data ended up looking extremely uniform with no trend. Even the bar chart only appeared non-uniform because we had zoomed the y axis to focus on the differences between the top of the bars. Upon further digging, the author of this dataset created some of their own visualizations that do a great job effectively showcasing the issue:









We found this to be a little suspicious, and we suspect that a random generator script or otherwise an AI model *could* have been used to generate or manipulate this dataset. Upon further research, a very interesting discovery was made. If you refer back to our bar chart of total orders per city, you'll notice that the city in the middle, Lahore, has a lesser order volume than the second city and not much more than the first city. This is interesting, as **Lahore has a greater population than the first two cities combined and doubled**. Peshawar and Multan both have roughly 2.5 million and 2.2 million people living in them respectively. Lahore has a population above **14 million** people. While sometimes data can debunk preconceived biases and go against basic intuition, we do not feel very comfortable saying this dataset is reliable. The author pumps out a

suspiciously large volume of datasets, the graphs are extremely uniform, and the data seemingly goes against verifiable metrics.
The only way we can think of that this data is from a real source would be through an algorithm that cherrypicks a subset of a dataset of millions of orders down to 6000 which tries to make (usually) all factors other than one hidden factor controlled for by the population distribution. This process would generate something like this, but it advertises itself to be used to train networks to learn many different aspects of the orders and customer habits, but that claim doesn't seem to be founded to us using the dataset.

VI. Development for presentations preceding final submission

Our interactive visualization was initially, even if the dataset was good, not up to par with the state of desired quality, in terms of design choices of the chart types selected, which were mostly as proof of concept implementations of visualizing in general. Our task was then to implement and/or modify better versions which feature more effective visualizations, at this point effective meaning such that it would be helpful if the dataset were useful in the first place.