



# **Data Mining**

## **CSCI-6401-01**

### **Phase 4 – Data Exploration**

#### **Team Name**

Technocrats

#### **Team Members**

Sravani Thoomuganti (Team Lead) – [sthoo1@unh.newhaven.edu](mailto:sthoo1@unh.newhaven.edu)  
Kamal Siddharth Teki – [kteki1@unh.newhaven.edu](mailto:kteki1@unh.newhaven.edu)  
Sai Teja Gunda – [sgund9@unh.newhaven.edu](mailto:sgund9@unh.newhaven.edu)

#### **Company Name**

DoorDash

## Datasets

<https://www.kaggle.com/code/deyashinichakravorty/door-dash-data-analysis/notebook>

<https://platform.stratascratch.com/data-projects/delivery-duration-prediction?tabname=solution>

## About the Datasets

The datasets we chose date back two years and come from a trusted domain. The DoorDash data analysis datasets will be used by us. Since we obtained the data from the official Kaggle website, which has historical stock data, we can guarantee that it is accurate to the nth degree. We also got dataset of delivery duration prediction which we also use in our research. We have taken this from StrataScratch website.

## Research Question

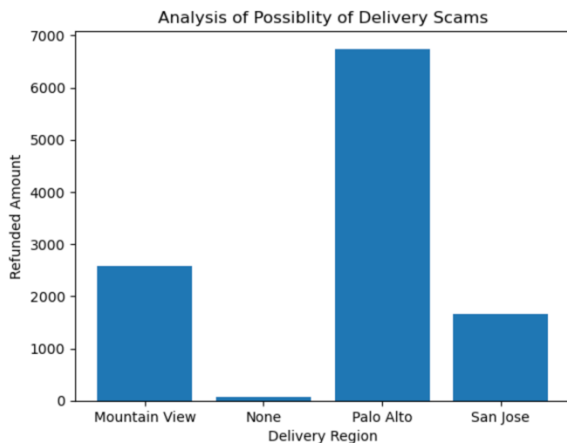
Ways to identify fraudulent activity on the DoorDash platform, such as delivery scams?

## Exploration Techniques

- Bar Chart
- Histogram
- Scatter Plot

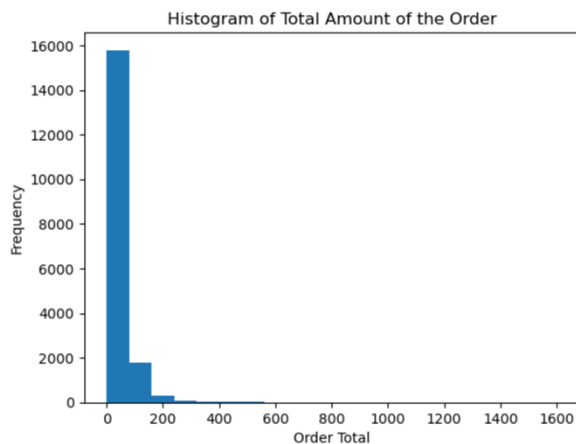
Initially we took the dataset and the raw data into consideration and then we have made an analysis on the required data sections that are required for finding a solution to our research question. Our main aim is to lay path to find the delivery scams going on in DoorDash. So for this, we have used some data exploration techniques picked from Bi-variate exploration techniques like bar chart, histogram and scatterplot. Which might be categorical and numerical.

## Bar Chart



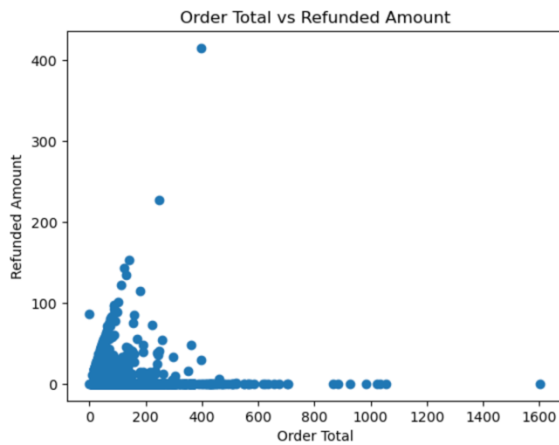
Here we have generated this bar chart using matplotlib functions in jupyter notebook. If you can see here after getting the raw data we have chosen x-axis as delivery region and y-axis as refunded amount to get a better idea on from which regions there is a possibility of delivery scams via unauthorized or invalid refunds reported from accounts that are either new or newly created.

## Histogram



Here in this histogram we have used the analysis to get an idea on who is placing unusual orders which are more than most of the orders that are placed and the frequency of the orders that have been placed from the account which is plotted on the data of total price of the order.

## Scatter Plot



Here in the scatter plot we have made an analysis on Order total vs Refunded amount because to identify the unusual orders placed and also the unusual activities that involve high order total and most frequent refunds placed from the account which can be the possible reason for delivery scams.

## Github Repository Link

<https://github.com/kamalsiddharthteki/Team-Technocarts>