## STT 811 Project Proposal

**Team:**

Kevin Patel

Siva Ram Kottapalli

**Dataset:**

Brazilian E-Commerce Public Dataset by Olist

**Link:**https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?datasetId=55151

This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Overall dataset size is ~45 mb and below are the data tables corresponding to the dataset:

 olist\_customers\_dataset.csv

 olist\_geolocation\_dataset.csv

 olist\_order\_items\_dataset.csv

 olist\_order\_payments\_dataset.csv

 olist\_order\_reviews\_dataset.csv

 olist\_orders\_dataset.csv

 olist\_products\_dataset.csv

 olist\_sellers\_dataset.csv

 product\_category\_name\_translation.csv

**Project Agenda:**

As part of the project, we want to focus on Customer segmentation based on behavior. For this purpose, we may need to use SQL operations on the tables as well as ML modeling and our area of interest is **Clustering**.

We will be using Python for our Project coding purpose.