Utkarsh Nigam, DATS 6401, Summer 2020

Progress Report

## **Impact of Holiday Season 2019 on US Online Retail Sales**

E-commerce sales data for the time range January-2019 to April-2020 was downloaded from Kaggle. Raw data is in the form of clickstream i.e., every user event has been tracked with time stamp and necessary session details. Firstly, I aggregated the monthly data into one data set and then, labeled the data as Non-Holiday Season and Holiday Season for the months January – October and November – December respectively. Moreover, I observed that Holiday Season data has different campaign specific attributes, which may reveal a lot about the user behavior with respect to the campaign.

I researched and created a list of [7 Hypotheses](https://github.com/utkarsh-nigam/DATS-6401-Final/blob/master/Final%20Project_Hypothesis.xlsx), and processed data based on the requirements of these hypotheses. Since the raw data size was around 50 GB, I used python to do the preprocessing ([script link](https://github.com/utkarsh-nigam/DATS-6401-Final/blob/master/Preprocessing_FinalProject.py)). Post the preprocessing, data is now ready to be uploaded and analyzed ([processed data](https://github.com/utkarsh-nigam/DATS-6401-Final/tree/master/Data)). I used tableau for the initial exploratory data analysis wherein key trends were observed such as:

* Mobile device usage outnumbers Desktop and Tablet
* Grattol is the most holiday sensitive brand
* Peak sale days generate 15% of the revenue during the holiday season
* California and Texas are the top revenue generating states

I have started to build the visualizations in d3.js with respect to the hypotheses and storyline I wish to present. Post that, I will be working on the construction of the web page in HTML, including the CSS styling. The final project deliverable will consist of one fixed side navigation web page with scrolling page anchors.