Meeting Summary

Date: Monday, 5/10/21, 630p

Materials: Available on github at [here](https://github.com/justinehly/olist/tree/main/meeting%20notes-presentations)

We reviewed the Powerpoint deck that covered

1. CRISP-DM framework
2. Grading Rubric for Project 1

Notes:

* Overall Business Objective Suggestions based on what the EDA to date shows us
  + Helping sellers sell more products/ increase their customer base (loyalty)
    - Different parts of listing affect sales
  + Helping Olist reduce shipping costs by evaluating freight costs and timing
    - Freight costs
    - Delivery estimates vs actual delivery
* Data Merge
* List of Pre-Merge Items
  + Need to merge products + translations as *pdf*
    - impute 2 missing translations as products\_eng
    - remove original category
    - rename english category to shorten name to product\_category\_english
  + side note: there are 19,015 duplicate zip code prefixes in the geolocation dataset, merging it with the customers df increases the the df to 15,083,733 records and it takes up over 1.2G memory before we even start to merge with other df's -I expect the same thing might happen with with the sellers df – so this was tabled
* Merge order for the dataset

1. **DF1:** inner merge customers + orders as df
2. **DF2:** right merge df + reviews as df2
3. **DF3:** right merge df2 + payments as df3
4. **DF4**: right merge df3 + items as df4
5. **DF5:** left merge df4 + pdf as df5
6. **DF6:** left merge df5 + sellers as df6
7. **Olist:** set olist = df6

* Article is in the github that details Brazilian e-commerce shopping habits during the same time period our dataset covers 2016-2018 and that is very helpful in explaining sales trends by month and possibly by day as well as why shipping costs may be such an important factor in predicting consumer outcomes, pricing or other things
* Simple Stats: Need to run simple statistics and look for anomalies as well for the write-up
* Visualize Attributes: need 5 visuals for the most important attributes, interpretation and why that visual was chosen
* Explore Joint Attributes: Explain interesting relationships
* Explore Attributes and Class: identify and explain interesting relationships between features and the class we are trying to predict
* New Features:
  + John: Purchase Month Name
  + John: Purchase Day Name
  + Helene: Create a *product\_vol* variable by combining *length \* width \* height* (length is misspelled in the dataset, need to fix that)
  + Other variables that may come up as needed
* Exceptional Work:
  + Run and interpret PCA
  + LASSO/ LDA-QDA?