A. Individual Proposals

1. Chandan Nayak
Data titles: NOAA Global Historical Climatology Network – Daily measurements dataset
Data sources (URLs): Link
Size: 15GB, 230mn records. Data is in compressed csv files
Reasons: NOAA has compiled a time series of important meteorological records for over 180 locations around the world. The dataset goes back more than a hundred years but with more observations since the 1950s.
Possible analytic goals: time series analysis of temperature over time and validate if the trend is increasing

2. Xinming Wang
Data titles: eCommerce behavior data from Cosmetics Shop
Data sources (URLs): Link
Size: 2.43GB in total, 3.52mn records
Reasons: With this dataset we can mimic the data scientist in an ecommerce company, use these data to catch the trend, optimize the page layout, improve the purchase process, do targeted advertising and targeted promotions.
Possible analytic goals: 1, The most popular category/ brand trend 2, The price range/ time period in which most purchase behaviors happen 3, Retention rates of “View - Cart - Purchase” process 4, Define & label customers by their view frequency and purchase frequency

3. Yufeng Xing
Data titles: Dataset PUBG Match Deaths and Statistics
Data sources (URLs): Link
Size: 20+GB in total, 70+m records
Reasons: Easy for EDA, Few Features, Accurate Data
Possible analytic goals: Player's favorite weapons, Player's death place, Weapon K/D rates, Player's strategy in this game, etc.

4. Kris Knapp
Data titles: Open Payments at CMS.gov (Centers for Medicare & Medicaid Services)
Data sources (URLs): Link
Size: each year is between 380MB - 808MB / between 5.8mil - 11.7mil rows
total for all 7 years 2014-2020 is approximately 4.5GB / 73mil rows
Reasons: 75 columns provides a lot of information and directions we can look at the data from. Most of the data appears to be clean, although some columns have a lot of missing data.
Possible analytic goals: compare payment levels between states or other geographic regions, or examine trends in payment levels over time (2014-2020).

B. Reasons & Goals

After discussion, we finally decided to choose dataset 2 about eCommerce based on the following reasons,

- Structured data from a real company: Good selection of numerical and categorical data types from a real world e-commerce website.
- Possible a more practical data analysis compared to other datasets
- The dataset has entries from the holiday season in Nov to Dec as well as regular months.
- Detailed customer behavior included: The user actions (view, add to cart, checkout etc) are recorded as well.