Data Analysis Pipeline (J. Erik Jarvis, Oscar Bailon, Anya Lauria, Ahmad Sadeed)

- Data set
  - Wholesale customers Data Set: https://archive.ics.uci.edu/ml/datasets/Wholesale+customers
  - Attributes:
    - Discrete
      - Channel: the type of customer purchasing wholesale products (Retail or Restaurant)
      - Region: location of customer (Lisbon, Oporto, Other)
    - Continuous
      - Fresh: annual spending on fresh products
      - Milk: annual spending on milk products
      - Grocery: annual spending on grocery products
      - Frozen: annual spending on frozen products
      - Detergents/Paper: anl. spending on detergent or paper products
      - Delicatessen: annual spending on delicatessen products
  - Other analyses mentioned by UCI:
    - https://econpapers.repec.org/bookchap/wsiwschap/9789814696357_5f0008.htm
    - https://arxiv.org/pdf/1211.0437.pdf
- Data Exploration and Visualization
  - We plan to generate several graphs in exploring our data:
    - Create several 2-dimensional graphs for each of the variables
    - A few 3d graphs
  - For each of these graphs we plan to create two versions:
    - One with the labels based off the channel these goods were purchased in
    - One with labels based off the geographic location of purchase
  - Calculate the mean, standard deviation, minimum and maximum for each variable by label (the data set info has already given us all these values for the general data set)
- Research question:
  - Can we identify market segments via k-means clustering? Do these segments correspond with the retail location/channel?
- Analysis
  - Clustering; we will remove labels and ignore them up until we want to compare the labels to the clusters we have developed.
  - We will be clustering with different sets of variables in both 2 and 3 dimensions in order to identify patterns in the data.
  - By clustering, hopefully we can visualize the different segments of the wholesale market.
- Planned report
  - Model experiments will be visualized for comparative analysis and confounding variables will be discussed in sections of association. The report will go through a process of data wrangling, algorithm implementations, explanation and exploration of visualized results (EDA), possible discussion of ethics and conclusive information fairly extrapolated from results, constructed through subplots.

This should be a bulleted outline of your project. This should be no more than a single page. Make sure to include at least the following:

- A description of your dataset
    - What attributes are included in your dataset? What do they represent?
    - Provide a link to your dataset and mention any analyses that were previously performed on your data.
- How will you explore and visualize your data?
- Your group's research question
- Your planned analysis
    - What type of analysis? Any cross-validation? How will this model help answer your research question?
- Your planned report
    - How will you report your results? Will you report just a single model or something else?