Visual Analytics 690V – Homework 2

* Ajay Shaan Shanmugam

[ashanmugam@umass.edu](mailto:ashanmugam@umass.edu)

SPIRE: 31133955

**Datasets used:**

Wholesale customers data - Missing Values.csv

Wholesale customers data - No Missing Values.csv

**Preferred Visualizations:**

Scatterplots

**Interactive tools used:**

Hover, Lasso, Box zoom, Wheel zoom, Pan

The notebook contains a visualization of values in the dataset followed by visualizations of comparisons between actual and imputed values of missing data in the dataset.

The following 6 techniques are used to visualize these comparisons:

1. Ignoring tuples with missing values

All records with even one missing value are removed.

1. Using a global constant

A global constant (-1) is used to fill in missing values.

1. Using the attribute mean

The mean of an attribute is used to fill up missing values in it.

1. Filling in the missing values manually

Involves eyeballing the data and choosing appropriate values to fill in missing data.

1. Filling in most probable values (Using KNN ML Model)

The K nearest neighbors of a data point are determined using a similarity metric (Euclidian distance) and the values in their target column are averaged. K-cross validation is performed to determine the best K value and also nominal columns such as Channel and Region are ignored for training the model.

1. Filling in most probable values (Using Linear Regression model)

A line is determined that best fits the data and this model is then used for estimating probable values of missing data.

**Note**:

The Bokeh user guide was used for reference.