James Doe
CPSC 392

Variables

- `price` (in USD) of avocados per pound
- `country_origin:` country where avocado was grown
- `organic`: 1 if organic, 0 if not
- `weight (in grams):` weight of avocado
- `supermarket_type`: type of supermarket (categorical) chain, local, farmers market, service station
- `county`: county avocado ended up in
- `state`: state avocado ended up in
- `distance_traveled`: distance traveled from origin to destination
- `color_R`: Red value for image taken of avocado skin
- `color_G`: Green value for image taken of avocado skin
- `color_B`: Blue value for image taken of avocado skin

1. (**Supervised Model**) When predicting price paid for avocados, which predictor (country of origin, organic, weight, state, supermarket_type) improves the R2 the most when compared to a model with all other variables except itself?
2. Which counties have the highest prices avocados and are counties similar to the counties near them?
3. (**Dimensionality Reduction**) When comparing a model using PCA on all the continuous variables (other than weight) in the dataset and retaining enough PCs to keep 90% of the variance, to a model using all the continuous variables (other than weight), how much of a difference is there in mean absolute error when predicting avocado weight?
4. When looking at the region the avocados were *sourced* from, which region has the smallest variance in price?
5. (**Clustering**) When considering price, weight, and distance traveled, what clusters emerge and what characterizes those clusters?
6. (**Supervised Model**) Looking at the coefficients, which variables (of country of origin, organic, weight, state, supermarket_type, region_of_origin, distance_traveled) have the strongest relationship on price?
7. Is avocado color when picked (measured by the R,G,B values from a sample of the avocado skin) related to the distance traveled?