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

Q1

Question: When predicting the 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?

Variables Involved: price (continuous), country of origin (categorical), organic (binary), weight (continuous), state (categorical), supermarket_type (categorical)
Cleaning: Missing values will be dropped, and country of origin, state, and supermarket_type will be "dummied" using get_dummies.
Modeling/Computation: A Train/Test split with 80/20 split will be used. Continuous variables will be z scored. Six linear regression models will be fit to predict price. One model will use all predictors, and the remaining 5 will use all but 1 predictor (e.g. all predictors except organic). R2 scores will be pulled for both train and test set and compared.
Graphs: A bar chart showing the train/test R2 scores for the various models. The predictor that improves the R2 the most will be made into another graph (scatter or boxplot) to show the relationship of it with price.

Brief Discussion of why analysis is effective at answering question: This analysis is effective because it compares a full model (all predictors) with models that are each missing 1 predictor in order to see how different the R2 values are when that specific predictor is removed. The plots will help visualize the numeric results, as well as demonstrate the relationship between the strongest predictor and the outcome (price).

(For Q2 and Q3, blank structure is shown, yours should be filled out like in Q1)

Q2
Question:
Variables Involved:
Cleaning:
Modeling/Computation:
Graphs:

Brief Discussion of why analysis is effective at answering question:

Q3
Question:
Variables Involved:
Cleaning:
Modeling/Computation:
Graphs:

Brief Discussion of why analysis is effective at answering question: