Lab 06

This lab will be done in Python. For the lab to be complete, all questions must be answered.

Examining customer spending

We'll be using the lab06_customers.csv data set for this lab. The data set covers the demographic characteristics of some customers and the amount they spent over the past year at an online retailer.

1. Import the CSV file and print out a few rows to examine the data.
2. Compute descriptive statistics. This includes computing percentages of each category of the categorical variables. The function `value_counts(normalize=True)` will provide percentages. If you don't set `normalize=True`, that function will provide counts by category.
3. Look at the distributions of the continuous variables. If income is not normal, choose an appropriate transformation.
4. Compute average spending by race using `groupby()`.
5. Compute average spending by sex.
6. We have reason to believe that groups of Hispanic Men and Black Women spend differently from each other. Fit a model to test these hypotheses. This is easiest to do using R-style coding for the interactions. See this link and scroll to about midway on the page to see an example.
    a. Look at the overall p-value for the interaction using `anova_lm()` and Type III sums of squares. Is it significant?
    b. Report p-values for this comparison and adjusted R^2 from the model.
    c. Create a pivot table of mean spend by race and sex. This link shows how (scroll to the part on adding columns).
    d. Did the findings support your hypothesis? Explain.
    e. Hint: Make sure the reference categories for sex and race is one of the comparison groups.
7. We also believe that education might interact with race and gender. Fit another model testing a 3-way interaction with schoolYears. Is it significant? What is the adjusted R^2 value?
8. Three-way interactions can be hard to explain. We will often visualize the components of them. Do an interaction plot with spend as the response, schoolYears on the x-axis, and sex as the trace using `statsmodels` `interaction plot`.
9. Of course, income and age probably play a factor in spending. Add those to the model and refit. Are they significant and how do you interpret their coefficients? What is the new R^2?
10. What did you learn about customer spending? Just provide a quick summary.