**Case Study: Leisure activities of adults**

Using the “Adult Participation In Selected Leisure Activities By Frequency: 2018 [As Of Fall]” data as produced by the MRI Survey of the American Consumer and provided by the ProQuest Statistical Abstract of the U.S. (<https://statabs-proquest-com.ezproxy.bellevue.edu/sa/docview.html?table-no=1269&acc-no=C7095-1.26&year=2020&z=187A84E2AB438865E1A89BF8BE7405A9A7BC5F95>), I will observe the varieties of activities in which US adults participate on a regular basis and the relationships between them. In particular, I am focusing solely on the year 2018 (the data set provides information on 2010 through 2018), the total numbers (as opposed to the percentages), and I have added a source of metadata which classifies each activity as either a local activity or a remote one (i.e., an activity that can be performed close to the participant’s residence or one which requires travel to another location). The data approximates totals (in thousands) of the number of adults who participated at least once in the activity in a specified period of 12 months, as well as views the totals of those who participated in the activity two or more times a week, once a week, two or more times a month (but less than once a week), and once a month.

The study that I was going to conduct would have been primarily observational in order to see how the distribution of leisure is laid out over the course of the year of 2018; however, as actual applications progressed, this developed into one where I used the data to try and predict whether an unknown activity was “Local” or “Remote” based on the values of the participation. Additionally, since each activity is only “Local” or “Remote”, the information for the associated activities would be left out in order to maintain a true estimator off of the training data.

Variables:

* Activity
* Year
* Locality
* Participated in the last 12 months
* Weekly Participation (multiple)
* Weekly Participation (once)
* Monthly Participation (multiple)
* Monthly Participation (once)

**Process for Dimensionality and Feature Reduction**

* Transformation of quantitative data through logarithms in order to attempt to produce a more normalized output
* Conversion of the categorical variable Locality to a numerical format

**Model Selection and Evaluation**

I chose to use a logistic regression model due to how strongly quantitative the data was – since I believed that the information on the numbers of participants would help indicate whether the activity was Local or Remote, I felt that using only the logarithmic numbers (for their near normal distribution) would give the best results. As I did not want to use Activity as a predictor, I had no categorical variables incorporated into the model. After running the model (following a very similar path to that used by the Titanic tutorial), the end result did show some rather good results for the test data; however, I wonder if the data set was too small and if it would have produced different results if I had included the data for all of the years available in the set (2010 – 2018) instead of just 2018.

**Original Case Study Analysis**

Per my own observation in the last part, I redid my entire analysis using all of the associated data by adding the other years to see how that would affect the results. In addition to merging the data from each years’ tab and adding a “Year” column to differentiate the data, I shifted the data values from the approximate population totals previously used to the percentages also provided with the initial data set in order to bring the information from the differing years into balance. Per the attached files, the “table1269” file is the original file downloaded from the site above, the “Leisure data 2018” was the data used for the first 3 parts of the case study, and the “Leisure data full” is the cleaned version used for this final part.

**Final Observations and Thoughts**

As mentioned in the previous paragraph, I chose to use percentages in this final attempt due to the fact that the estimated total population values of each year varied, so the percentage values helped to equalize the totals. I considered including “Year” in the analysis but decided against it because the concept of locality is constant irrespective of time. Overall, the results were fairly good, with the estimations by the model predicting remote activities with higher accuracy than local.