UCLA Extension Data Science Intensive

Instructor: William Yu

Project 4

Submit your result via Canvas using R script and a screen shot of the Tableau chart.

A. Census Data Analysis

- In class (D04e_census.R), we learned how to access Census data directly and efficiently through R library: tidycensus.
- Let's use "get_acs" function to get data for the number of residents (above 25 years old) who has a bachelor's degree by county in the U.S. The variable's code is B15003_022. Meanwhile, let's also get the number of total adult residents by county (B15003_001).
- Get the median household income by county (B19013_001).
- Merge the above two dataset together by its geographic ID: GEOID.
- Calculate the percentage of college degree (over the total adult population) by county and add it as a new variable in the data.
- Using ggplot in R, plot the correlation between the college degree percentage (in x-axis) and the log median household income (in y-axis).
- Run a simple regression (x=college degree %, y=log median household income). From the regression result, what story can you tell?
- From the chart, it seems there are some abnormal observations in the correlation. Can you identify them? (Hint: Not about outliers) It would be even better if you can show them in a different color in the chart.

B. What Factors Contribute to Higher Adoption of Electric Vehicle (EV)?

- In the dataset dmv_zip.xlsx, we have the variable (p_beph) in Column D representing the % of vehicles of battery electric and plug-in hybrid (EV) over the total vehicles by zip codes in California.
- Use Tableau to visualize the p_beph. Hint: I show my chart below for your reference.
- Via the chart, we got some ideas. We want to know if certain variables will associate/explain the variation of EV adoption rate by zip code in California. Some candidate variables are e.g. income, education (CHCI), and age demographic.
- Use Census (ACS) data to get those variables and merge them to dmv_zip. And calculate CHCI by zip code level.
- Run a multivariate regression and answer the question: do these variables associate/explain the EV adoption rate? Such as the following:

  *P_beph = a + b1*CHCI +b2*Income + b3*Age Distribution*
- Use a simple R plot function to plot the correlation between the median household income (in x-axis) and the EV adoption % (in y-axis).
- Hint: the following code will help you to begin. Calculate CHCI as you did in the previous project. Construct/use a variable for the percentage for Age 25 to 44.

```
mincome =  read_excel("ACS_17_5YR_S1903.xlsx")    # Median household income from Census
```

```
education =  read_excel("ACS_17_5YR_S1501.xlsx")   # Education data from Census
dmv = read_excel("dmv_zip.xlsx")                   # California DMV data
income.17 = mincome[,c(2,8)]
colnames(income.17) = c("id","income")
income.17$income = as.numeric(gsub("-","NA",income.17$income))
ed.17 = education[,c(2,64,78,90,102,114,126,138,150,184,220,256,292)]
colnames(ed.17)=c("id","pop17","below9","g912","hs","scollege","associate","bachelor","higher","p2534","p3544","p4564","p65a")
str(ed.17)   # p2534: percentage of age 25 to 34 …p65a: percentage of age above 65
ed.17[,3:9]=sapply(ed.17[,3:9],as.numeric)
```

C. Machine Learning Models Summary and Cross-Validation

- In Project 3, we analyzed the breast cancer data. Let's analyze the data with more machine learning models by using similar functions in D04f_iris.R and D05a_mlmodels.R.
- With the full sample, use "rpart" method (which is a tree model) to plot a tree chart for the classification (i.e. using "fancyRpartPlot")
- Using 10-fold cross-validation and analyze the following models: LDA, tree, KNN, Bayesian GLM, SVM, Random Forest, and XGBoosting.
- To make sure we have the same result, let's have the same random procedure: set.seed(99).
- List the mean of accuracy and kappa, identify the best model to predict the breast cancer.
- Briefly explain your result.

Figure. Percentage of EV (both battery electric and plug-in hybrid) of total cars by zip code

Green color represents higher percentage; darker green means further higher percentage