Part B

PCA analysis was conducted on the universities data set with categorical variables removed and instances with missing numeric measurements removed. This is because PCA analysis only works on numeric data due to the fact orthogonals are being calculated. Also, the data needed to be normalized before the PCA analysis could be performed. This is because PCA is trying to find the combinations that capture the most variance, if the data is not on the same scale then the variables with the largest scale will dominate the first component which may or may not be appropriate.

The PCA analysis revealed that the variance in the data set is primarily accounted for by two principle components. These two components account for over 55% of the variability. Also, from looking at the elbow lot, 3 principle components would be the correct number to select. The first, PC1, is primarily characterized by having high in-state and out-of-state tuition and the new students being from the top 10% or top 25% of their graduating class. Also, PC1, has have a low part time undergraduate student enrollment, low personal cost, and a low student to faculty ratio. The second principle component, has high in-state and out-of-state tuition and no other positive correlations. However, it has a low number of applications received and accepted as well as a low number of students enrolled. The third, PC3, is high in additional fees, number of faculty with PHDs, student faculty ratio, and the graduation rate. The third principle component is very low in book costs, personal costs, room, and board.

From these correlations, I would expect the first principal component to represent elite institutions given the quality of student, tuition rates, and lack of part time students. The second principal is smaller private institutions given the high in-state and out-of-state tuition and low number of applications. The plot of PC1 versus PC2 confirms these cluster representations.

Part C

The business question I am trying to answer is which counties a political party should concentrate their resources for the 2020 election given the data from the 2016 election in order to maximize their effectiveness. I have data from Kaggle that contains information about characteristics of all counties in the US and how they voted in the 2016 elections.  It contains information such as education level, ethnicity, income, etc.  I want to use this information to do a PCA analysis to see how many “clusters” account for the different voting groups. Also, I want to do a logistic regression to see if I can predict, given a person’s characteristics, how they are going to vote.