Steven Brannum

BIA 6301 – ADM

Homework #2

**Part A**

1. Perform some preliminary EDA on the data and describe the sample. What do you learn about fatal automobile accidents from this exercise? How can you use that information?

Looking through the dataset you can see that the rows represent every individual involved in a fatal accident with no information to determine which party they were in the accident, that is to say, you can’t tell who was at-fault. This can and does lead to duplicate data across columns *State, Atmospheric Condition, Fatalities, Roadway, Crash Date, Crash Date.Year, Crash Date.Month, Crash Date.Day-Of-Week.* Running a *summary* lets us see that Wyoming followed by Arizona then Colorado have the highest number of people involved in fatal accidents. This does **not** tell us which has the most accidents because as was mentioned already the data seems to be duplicated across lines. We could attempt to get a state by state count by taking just the columns *State, Atmospheric Condition, Crash Date, Fatalities,* and *Roadway* and grouping them but since there is no identifier we could potentially group different accidents. It appears roughly 65% of the involved persons were reported as male and 90.5% were age sixteen or older. Also, interestingly it appears that the accidents tend to be high early in the year and taper off. January, February, and March have the highest number of people involved in accidents whereas October, November, and December have the least.

1. What steps will you take to prepare the data? Why?

I will begin by factoring states into regions since we may be confined to going nationwide by one region at a time. I will group the roadways into urban and rural to remove the large number of variables when creating dummy variables and also remove the three observations with “unknown” or “N” values. We will also remove the observations that have unreported or unknown genders.

Columns being dropped:

* *Atmospheric Condition* because from the EDA we can tell that the conditions were overwhelmingly clear during the accidents. That seemingly eliminates the thought of choosing a region with more mild weather patterns first.
* *Person.Type* because all persons involved in at-fault accidents are typically covered by the person’s insurance who was at-fault and there is no identifier for which person’s insurance paid out.
* *Age* also because in accidents, the insurance paying out covers all persons regardless of age.
* *Crash Date*.year because the data is all from 2011.
* *Crash Date.day-of-week* and *Crash Date.day-of-month* because I believe it is too granular for what we need.
* *Alcohol Result* because there is not enough completeness. We are missing Alcohol Result values for over 2000 observations.
* *Race* because it also lacks completeness with over 2000 observations missing values and could lead into unethical practices.
* *Drug Involvment* will be removed due to poor consistency and interpretability. The value of “Not Reported” is ambiguous.
* *Fatalities* will be removed because it is duplicated across indistinguishable observations.
* *Injury Severity* because it’s not relevant to the trends I am looking at.

1. What questions will you be able to answer with the cleaned data? What limitations will you face?

The main limitation is how the data was put together. Having both summary columns and individual observation columns causes duplicated and indistinguishable data, so clusters will more than likely include, if they have it, their corresponding observations. There are four columns being left in, *Region, Roadway* (urban or rural), *Gender*, and *Crash Month*. From this data I am hoping to determine where and when a certain gender is more likely to be involved in car accidents and also if in general a region or roadway area has more fatal car accidents during a given month or season.

**Part B**

1. How many types or groups of crashes are prevalent in the provided data set after you have conducted the analysis? How did you determine that? Describe the characteristics of each group.

When looking at two clusters we tended to see the split by gender with males trending towards accidents on rural roadways and females having accidents on urban roadways. With four clusters we saw again that the splits trended toward gender, but the rest of the data was inconclusive and mixed except for one cluster showing roughly half the male observations having zero involvement in accidents in the West region on Rural roadways.

1. Are there geographic differences in the crash types or fatalities? If yes, what were they? How might that affect the CEO’s decision?

When looking at clusters of both two and four we can see that Rural and Urban roadways both dominate separate clusters. Rural roadways accidents disproportionately appear in the South and Midwest while the Urban roadways appear to have much higher frequencies of fatal accidents in the West region. This makes sense as the South and Midwest are more rural areas. And the far West states tend to be predominately urban areas. One thought for the CEO would be to charge a higher premium to customers in the South that live in more Rural areas and charge customers in the West regions more that drive in Urban areas. This isn’t as clear cut due to not having population by region or how much of each region is rural vs. urban. If a region has more rural area than urban area and more people living in rural areas then logically there is more likelihood that rural areas will have more fatal accidents.

1. Can you make at least two or three recommendations for policy or pricing options to the CEO based on your analysis? These recommendations can range from suggestions on whether or not to offer automobile insurance to certain areas or drivers as well to suggesting pricing differentials for different customers. Explain your reasoning.

Firstly, if we look at just the total number of fatal accidents, we can see that the Northeast has the least number of accidents, also the least per state of the regions. Looking at just the Northeast region it would be wise to market towards females that live in rural areas as they tend to be the least involved party. My last recommendation would be to prepare to payout the most claims at the beginning of every year or to target marketing or rewards for safe driving the most during this time of the year.