**PA 541: Homework # 1**

**Due: February 8th at 3:00pm ~ Submit via Blackboard**

**Topics: Data Wrangling and Regression Basics**

**NAME:**

\*\*You can answer each of these questions in a word or similar text editor. For those of you that know R Markdown, that is a great option as well. As needed by the question, please provide the R code along with the relevant output. I should be able to see from your code how you arrived at the answer. Please only paste in the relevant code and output. Note, it is better to copy and paste from the R console (script + output) within RStudio (the lower left pane). When pasting, choose to keep source formatting. [70 pts]

**PART ONE**

**Data Wrangling. Read in the ‘fatalities’ data. The data come from the US Department of Transportation Fatal Accident Reporting System. Total vehicle miles traveled annually by state was obtained from the Department of Transportation. Personal income was obtained from the US Bureau of Economic Analysis, and the unemployment rate was obtained from the US Bureau of Labor Statistics. You can read in the data using following: *fatal = read\_csv("Fatalities.csv")***

**Names and descriptions of those variables are below.**

* **state -** state.
* **year -** year.
* **spirits -** Spirits consumption.
* **unemp -** Unemployment rate.
* **income -** Per capita personal income in 1987 dollars.
* **emppop -** Employment/population ratio.
* **beertax** - Tax on case of beer.
* **baptist** - Percent of southern baptist.
* **mormon -** Percent of mormon.
* **drinkage -** Minimum legal drinking age.
* **dry -**  Percent residing in “dry” countries.
* **youngdrivers -** Percent of drivers aged 15--24.
* **miles -** Average miles per driver.
* **breath -** Preliminary breath test law?
* **jail -** Mandatory jail sentence?
* **service-** Mandatory community service?
* **fatal -** Number of vehicle fatalities.
* **nfatal -** Number of night-time vehicle fatalities.
* **sfatal -** Number of single vehicle fatalities.
* **fatal1517 -** Number of vehicle fatalities, 15--17 year olds.
* **nfatal1517 -**  Number of night-time vehicle fatalities, 15--17 year olds.
* **fatal1820 -**  Number of vehicle fatalities, 18--20 year olds.
* **nfatal1820 -** Number of night-time vehicle fatalities, 18--20 year olds.
* **fatal2124 -** Number of vehicle fatalities, 21--24 year olds.
* **nfatal2124 -** Number of night-time vehicle fatalities, 21--24 year olds.
* **afatal -** Number of alcohol-involved vehicle fatalities.
* **pop -** Population.
* **pop1517 -**  Population, 15--17 year olds.
* **pop1820 -**  Population, 18--20 year olds.
* **pop2124 -** Population, 21--24 year olds.
* **milestot -** Total vehicle miles (millions).
* **unempus -** US unemployment rate.
* **emppopus -**  US employment/population ratio.
* **gsp -** GSP rate of change.

**QUESTION 1 (12 pts)**

**The following will require you to use the tools and verbs we learned during week 1 to wrangle data. The results of these tasks will produce a tibble. You only need to copy and paste the tibble itself (what R reports) and not all of the variables or observations (i.e., don’t print out the whole dataset).**

**a. First, let’s select a handful of variables to focus on and remove the others. Create a new dataset, call it fatality2, that contains only the following variables: fatal, state, year, spirits, unemp, income, dry, pop, and miles. Use this dataset for all steps below. (2pts)**

**b. For each year available in the dataset (i.e., 1982 – 1988), how many total fatalities were there in each of those years? (2pts)**

**c. Which state had the largest number of fatalities in 1982? (2pts)**

**d. Which states in which years had more than 1,000 fatalities and more than 20% of its population residing in dry counties. (2pts)**

**e. What is the average number of fatalities in each state? (2pts)**

**QUESTION 2 (8 pts)**

**Create a new variable, ‘fatal.cat’ that breaks the continuous variable fatal down into three categories: (i) 0 - 300, (ii) >300 - 1000, (iii) >1000. Please label the categories “low”, “mid”, “high”. Set this new variable to be a factor.**

**What is the mean of miles in each of the fatal categories?**

**PART TWO**

**Regression. For part 2, let’s limit the fatality2 data from above to only the year 1987. So, to begin part 2, create this new dataset and call it fatality3.**

**QUESTION 3 (6 pts)**

**Using the newly created fatality3 dataset, test the correlation between miles and fatal. What are your findings (i.e., what is the size of the correlation and is it significant)?**

**QUESTION 4 (12 pts)**

**Create a new population variable, that is population in 100,000s. Call the new variable pop\_100k. Run a simple linear regression predicting fatal from pop.100k. (a) Interpret the estimates of the slope and intercept coefficients in the context of the problem. (b) What is the percentage of variation in fatal explained by pop\_100k? (c) Predict the number of fatalities in a state if the population was 8 million.**

**QUESTION 5 (8 pts)**

**Which state has the largest negative residual in our model from question 5? Which state has the largest positive residual? Tell me what these large positive and large negative residuals mean within the context of our data and model.**

**QUESTION 6 (12 pts)**

**Fit another regression model with fatal as the dependent variable and pop\_100k, miles, and dry as the predictors. (a) What percentage of the variation in the dependent variable is explained by the predictors? (b) Ignoring whether the predictor is significant or not, interpret the coefficient estimates for each predictor. Be specific when discussing the relationship. (c) How do we interpret the p-value for dry? (d) By how much did our R-squared increase from our initial model that only included pop\_100k as a predictor?**

**QUESTION 7 (12 pts)**

**Run the following two models and compare the difference in the size and direction of the coefficient on *miles*. What is happening here? Can we trust the estimate of the effect of miles in the first model?**