**Regression Project**

**QMB-6304 Analytical Methods for Business**

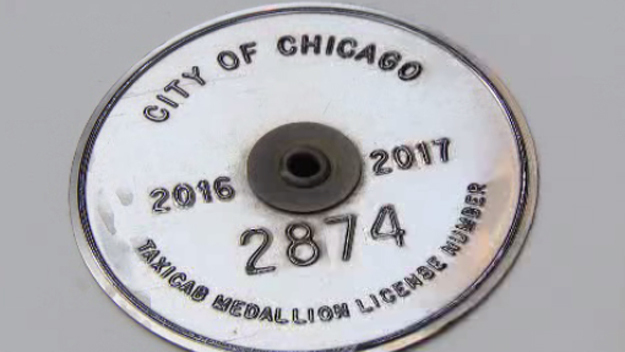
Write a simple R script to execute the following data preprocessing and statistical analysis. Show your R code, analytical output, and interpretations.

**Preprocessing**

1. Load the file “6304 Regression Project Data.csv” into R. This file contains information on 1,705,805 taxi cab trips in the City of Chicago during 2016. The data was taken from kaggle.com and modified. Variables in this data set are:
   1. taxi\_id: a unique identifier for each individual taxi cab.
   2. trip\_seconds: the number of seconds elapsed during the trip.
   3. trip\_miles: the number of miles logged during the trip.
   4. fare: the base fare charged to the customer for the trip.
   5. tips: the tip given by the customer to the driver for the trip.
   6. tolls: any surcharges for road or bridge tolls incurred during the trip.
   7. extras: charges for any incidentals requested by the customer.
   8. trip\_total: the total charge to the customer for the trip.
   9. payment\_type: the method of payment used by the customer. This includes cash, credit card, and several other methods of payment jointly classed as “other”.
2. Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 100 taxi trips from this master data set.
3. Using your judgment and the R tools you know, cleanse your random sample data of aberrant cases. Such cleansing cases is somewhat subjective, so explain your process and reasoning for identifying aberrancies in the data and removing them. State how many cases you are left with in your random sample after cleansing. This will be your primary data set for analysis.

**Analysis**

1. Using your cleansed sample data, provide summaries and density plots of each of the continuous variables in your data set with the exception of taxi\_id. Explain any apparent differences in the statistical distributions of these variables in your sample data.
2. Using the payment\_type factor variable and your cleansed sample data, provide a table of the number of cases in each level of payment\_type.
3. Construct an easily read and easily understood correlation matrix using all continuous variables except taxi\_id. Give a brief interpretation of the matrix understandable by a non-statistician.
4. Using fare as the dependent variable, build a regression model using trip\_seconds, trip\_miles, and payment\_type as potential independent variables. Evaluate the quality of fit of the model to your cleansed data. Explain the impact each independent variable in your model on the dependent variable, considering the 95% confidence interval on the beta coefficients.
5. Investigate relevant interactions and common independent variable transforms to determine if adding these to your model will result in a better model fit. Depending on your random data selection you may find it necessary to do some additional cleansing of your data in order to get a better model fit for the majority of data points.
6. Of the various combinations you ran in Step 5, report the model which provides what you deem as the “best fit” to your sample data. Explain why you selected this particular model and show the standard R regression output for the model. Evaluate and explain your model’s conformity to the LINE assumptions of regression.
7. Investigate and remove any data points deemed to have an inappropriately high leverage in determining the plot of the model. Rerun your model without these points and evaluate the quality of fit in this final regression model.
8. Return to the full data set of 1.7 million cases. Pull another sample of n=100 cases. (Be sure to use a new random number seed of the numerical portion of your U number plus 5.) To this data set apply the same cleansing procedures you used on your original sample data set. Referring to the model you developed in Step 6 above, apply that model to the new random set of data and evaluate how well the model fits this second data set.



Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is the “Regression Project” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.