# Data Analytics for Smart Cities & Transportation

## Final Project Guidelines

*Assignments are due **at the beginning of the class** on the due date, unless otherwise specified. Late submission will not be accepted. Please ensure the guidelines for submissions are followed to receive full credit.*

This final project involves the application of the data analysis concepts on relevant transportation datasets. You must select one (1) of the options listed below. The datasets are made available to you on CANVAS. *You may work in pairs (no more than 2 members) for this project and if you decide to do so, please notify your instructor as soon as possible.*

### Option 1

You are provided with data extracted from the 1998 national household travel survey conducted in The Netherlands (same data you used in Assignment 1). This sample provides the number of weekly trips undertaken by 1631 households (the corresponding variable in the dataset is called ntrips). The dataset includes the following variables:

| Variable name | Description |
|---|---|
| hhsize | Number of persons living in the household |
| nchlt12 | Number of children < 12 years of age in the household |
| nchgt12 | Number of children >=12 years of age in the household |
| nworker | Number of workers in the household |
| nstudent | Number of students in the household |
| ncar | Number of cars owned by the household |
| income | Household income |
| resloc | Residential location (1=household resides in city, 2=household resides in suburb, and 3=household resides in rural area) |

Please estimate at least 4 appropriate statistical models to predict the number of weekly trips and determine the best fit model.

Please include the null hypothesis you are testing when you examine whether the coefficient of a specific variable is statistically significant.

## Option 2

You are given 151 observations of a travel survey collected in State College Pennsylvania (same data as in assignment #1). All of the households in the sample are making the morning commute to work. They are all departing from the same origin (a large residential complex in the suburbs) and going to work in the Central Business District. They have the choice of three alternate routes; 1) a four-lane arterial (speed limit = 35mph, 2 lanes each direction), 2) a two-lane rural road (speed limit = 35mph, 1 lane each direction) and 3) a limited access four-lane freeway (speed limit = 55mph, 2 lanes each direction).

Please estimate at least **two** models of route choice (i.e., the likelihood of an individual traveler taking one of the three routes). Please include the t-statistics and justify the sign of your variables. The dataset includes the following variables:

| Variable Name | Description |
|---|---|
| x1 | Route chosen, rows: 1 - arterial, 2 - rural road, 3 - freeway |
| x2 | Arterial row indicator; 1 for arterial row, 0 for others |
| x3 | Rural row indicator; 1 for rural row, 0 for others |
| x4 | Freeway row indicator; 1 for freeway row, 0 for others |
| x5 | Traffic flow rate |
| x6 | Number of traffic signals |
| x7 | Distance in tenths of miles |
| x8 | Seat belts: 1 - if wear, 0 - if not |
| x9 | Number of passengers in car |
| x10 | Driver age in years: 1 - 18 to 23, 2 - 24 to 29, 3 - 30 to 39, 4 - 40 to 49, 5 - 50 and above |
| x11 | Gender: 1 - male, 0 - female |
| x12 | Marital status: 1 - single, 0 - married |
| x13 | Number of children |
| x14 | Annual income: 1 - less than 20000, 2 - 20000 to 29999, 3 - 30000 to 39999, 4 - 40000 to 49999, 5 - more than 50000 |
| x15 | Model year of car (e.g. 86 = 1986) |
| x16 | Origin of car: 1 - domestic, 0 - foreign |
| x17 | Fuel efficiency in miles per gallon |

## Option 3

You are given 204 observations from a travel survey conducted in the Seattle metropolitan area. The purpose of the survey was to study the number of times (per week) commuters' changed their departure time on their work-to-home trip to avoid traffic congestion. The data are non-negative integers with the mean approximately equal to the variance thus, the data are well suited to the Poisson regression approach. Remember in a Poisson regression, you are estimating a parameter vector $\beta$ such that:

$$\lambda = EXP(\beta X)$$

where λ is the Poisson parameter that in this case is the expected number of departure changes per week.

Provide the results of at least two of your best model specifications. Include t-statistics and justify the sign of your variables.

In addition, run a negative binomial model on your two best models and ensure that the overdispersion parameter is not significantly different from zero.

The dataset includes the following variables:

| Variable Number | Explanation |
|---|---|
| x1 | Household number |
| x2 | Do you ever delay work-to-home departure to avoid traffic congestion? 1 - yes, 0-no |
| x3 | If sometimes delay, on average how many minutes do you delay? |
| x4 | If sometimes delay, do you 1- perform additional work, 2 - engage in non-work activities, or 3 - do both? |
| x5 | if sometimes delay, how many times have you delayed in the past week? |
| x6 | Mode of transportation used 1- car SOV, 2 - carpool, 3 - vanpool, 4 - bus, 5 - other |
| x7 | Primary route (work-to-home) 1 - I90, 2 - I5, 3 - SR520, 4 - I405, 5 - Other |
| x8 | Do you generally encounter traffic congestion on your work-to-home trio? 1 - yes, 0-no |
| x9 | Age: 1-(<25), 2-(26-30), 3-(31-35), 4-(36-40), 5-(41-45), 6-(46-50), 7-(>50) |
| x10 | Gender: 1 - male, 0 - female |
| x11 | Number of cars in household |
| x12 | Number of children in household |
| x13 | Annual income: 1 - less than 20000, 2 - 20000 to 29999, 3 - 30000 to 39999, 4 - 40000 to 49999, 5- 50000 to 59999 6 - more than 60000 |
| x14 | Do you have flexible work hours? 1-yes, 0-no |
| x15 | Distance from work to home (in miles) |
| x16 | Face LOS D or worse? 1 - yes, 0 - no |
| x17 | Ratio of actual travel time to free-flow travel time |
| x18 | Population of work zone |
| x19 | Retail employment in work zone |
| x20 | Service employment in work zone |
| x21 | Size of work zone in acres |

Dr. Sravani Vadlamani
Florida Polytechnic University

## Final Report

Please submit a professional report (PDF version only) on CANVAS. Please ensure to follow these guidelines:

1. Your report should look professional and include an introduction, methodology, results and conclusion sections.
2. The outputs (tables and graphs) from utilizing the statistical software should be cleaned up for improved legibility and comprehension. Figures and tables using snipping tool are NOT acceptable.
3. Your write up should be at most 15 pages (including figures, tables and R code). This is not very long and you should be concise and to the point.
4. Please include your R Code in the appendix. Do not provide me a link to your Rnotebook.
5. When submitting any work, your objective is to communicate information to the reader (in this case, your instructor) in a Clear, Concise, Complete, Careful, and Courteous manner (5 C's of good writing). If your work does not possess the "5C" qualities and/or does not adhere to the guidelines specified below, you will definitely lose A LOT OF credit even though you may have the correct answer(s).

## Submission

1. The final report is due by 12/08/19 11:59 PM EST.
2. Please upload your report via CANVAS.

## Project Presentation

1. The presentations will be in class on 12/03/19.
2. You have 10 minutes to share your project and summarize your findings.
3. The presentation should include 5-10 slides summarizing your progress.