CSE701 Project 1 Proposal

Simple Taxi Fare Prediction System Based on Telematics

Yunfei Ma

Student Number: 400390380

September 25, 2021

Project Background

With the fast adoption of vehicle telematics for fleet management, tons of data are generated by taxi companies. The telematics dataset are generally very big and it is important to extract critical information from the large dataset. Therefore, my project is about building a simple taxi fleet fare prediction system to predict the fare for the driver based on travel time and miles.

Dataset

The dataset I will be using is from City of Chicago Taxi Trip dataset from June 1 to July 1. I would like to create the system based on some of the variables in the dataset:

Table 1: Chicago Taxi OD data

Variable Name	Data Type	Number of Observations
Taxi ID	Character	332263
Timestamp	DateTime	332263
Trip Seconds	Numerical	332263
Trip Mile	Numerical	332263
Fare	Numerical	332263
Tip	Numerical	332263
Tolls	Numerical	332263
Trip Total	Numerical	332263
Payment Type	Character	332263

Tentative Project Details

Goal: To build a linear multiple regression model for taxi fare prediction system. **When the user enter the trip time and trip in miles, a predicted taxi fare base on the model will appear**. The variable including but may not limited to:

- Dependent variable Y: The taxi Fare

- Independent variable X1: The trip mileage

- Independent variable X2: the trip seconds

$$Y = a + b_1 * X_1 + b_2 * X_2$$

Based on the multiple regression (with 2 variable). I will be using the computational techniques described in "Regression with Two Independent Variable". For the two variables case:

$$b_1 = \frac{(\sum X_2^2)(\sum X_1 Y) - (\sum X_1 X_2)(\sum X_2 Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1 X_2)^2}$$
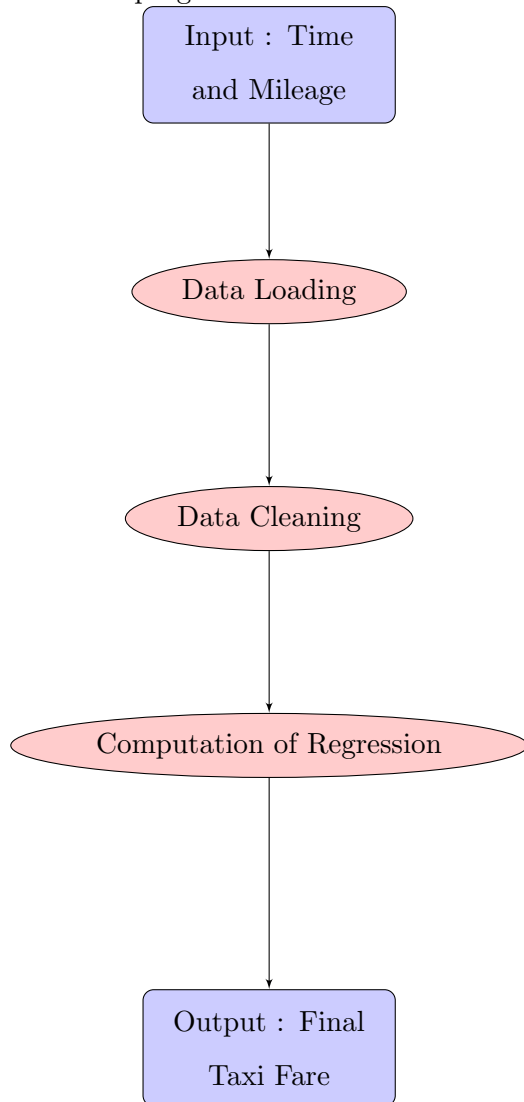
and

$$b_2 = \frac{(\sum X_1^2)(\sum X_2 Y) - (\sum X_1 X_2)(\sum X_1 Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1 X_2)^2}$$

The constant

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

The final program should be like:

```
Input : Time
and Mileage

        |
        v

   Data Loading

        |
        v

   Data Cleaning

        |
        v

Computation of Regression

        |
        v

Output : Final
Taxi Fare
```

The output would look like:

Taxi trip with time xxx and xxx has the expected fare (excluding tips) of xxx.

Extra Possible Part for Project

This part maybe include into the project to boost the computation side of program, the extra part will be more on statistical analysis.

- calculate R squared of the regression

- Create a simple Monte Carlo simulation based on the regression variance and risk factor to predict the fare.

Project Expectation

If approved, the final project may vary in minor details but shall not affect the overall goal mentioned in the proposal. If major changes are needed, I shall inform and discuss further with Professor Barak Shoshany and make another proposal.