CB2203 Data-Driven Business Modeling
Assignment 1

Please save your files as Asm1-XXXXXXXX.xlsx/.pdf/.docx as appropriate, where XXXXXXXX is your student id number.
For Question 2, you may submit a single Excel file with organized layout and clearly labelled answers to different parts of the question.
Display all non-integer numeric values to 2 decimal places and time-based values in the hh:mm:ss format, if necessary.

Question 1 (40 marks)

Study the effect of different time periods of the day on the taxi tip using the dataset "TLC Trip Record Data" in Lectures 2 or 3 Excel file (worksheet "Jan1-4 2019 Data").

(a) Develop the least squares regression model to predict the average tip from four independent variables: area, number of riders, high tipper and pre-tip amount. Give the estimated regression equation. (3 marks)

(b) Comment on the goodness of fit of the model using the adjusted coefficient of determination. (3 marks)

(c) Test the significance of the overall model at 0.05 level of significance. State the null hypothesis, alternative hypothesis, test statistic, *p*-value, decision and final conclusion. (6 marks)

(d) Discuss the significance of each independent variable at 0.05 level of significance. (4 marks)

(e) For simplicity, split the 24-hour day into four time periods:
06:00:00 – <12:00:00, 12:00:00 – <18:00:00, 18:00:00 – <00:00:00 and 00:00:00 – <6:00:00.
Introduce *three* indicator variables to indicate whether or not the drop-off time of a trip falls into exactly one of these periods. (7 marks)

(f) Explore whether any time period in (e) is significant in predicting the average tip, given the presence of the four independent variables in (a).

(i) If one or more of these time periods is significant, suggest an alternative linear regression model that could improve the adjusted coefficient of determination in (b) and pass the tests in (c) – (d). (13 marks)

(ii) Referring to the alternative linear regression model in (f)(i), interpret the slope coefficient for each indicator variable of time period. (4 marks)

Question 2 (60 marks)

A cross-harbour tunnel has implemented a time-varying toll for arriving vehicles. Simulating vehicle arrivals to the automatic toll collection system can help estimate the revenue and vehicle waiting time during a morning peak period, given the past distribution of time between vehicle arrivals (or inter-arrival time) to a single lane at the entrance of the tunnel in Table 1. For simplicity, consider only two major types of vehicles arriving at the single lane. The distribution of private cars and goods vehicles are 70% and 30%, respectively.

Table 1: Inter-arrival time distribution (all vehicles)

Time between arrivals (seconds)	Probability
0.5	0.26
1	0.19
2	0.15
3	0.11
4	0.07
5	0.06
6	0.04
7	0.03
8	0.02
9	0.02
10	0.01
11	0.01
12	0.01
13	0.01
14	0.01

The service start time (end time, respectively) of a vehicle is assumed to be the time when the vehicle front (back, respectively) passes a fixed location (e.g., camera scanning the car plate). The length of private cars and goods vehicles are estimated to be 5 metres and 11 metres, respectively, based on the local standard car park sizes. The vehicle speed passing through the fixed location is assumed normally distributed with an average of 35 km per hour and a standard deviation of 5 km per hour for any vehicle. The service time can be treated as the time taken for the entire vehicle (from its front to end) to pass the fixed location. (Note: Time taken = Length/speed.) To ensure enough stopping time in an emergency, each vehicle in the lane should maintain a safe time gap of at least 2 seconds with the vehicle in front.

Assume the simulation starts at 7:00:00 (or 7 am) with no car at the entrance to the single lane. When the front car plate of a vehicle is detected at the fixed location (e.g., camera scanning car plate), the time is recorded and the charge is based on Table 2. For example, if a private car is detected at 7:29:59, the charge will be $20. If a private car is detected at 7:30:00, it will be charged $22. A goods vehicle pays a fixed charge of $50 at any time.

Table 2: Time-varying toll

Starting time (hh:mm:ss)	Private car charge	Goods vehicle charge
7:00:00	$20.00	
7:30:00	$22.00	
7:32:00	$24.00	
7:34:00	$26.00	
7:36:00	$28.00	
7:38:00	$30.00	
7:40:00	$32.00	
7:42:00	$34.00	
7:44:00	$36.00	$50.00
7:46:00	$38.00	
7:48:00	$40.00	
10:15:00	$38.00	
10:17:00	$36.00	
10:19:00	$34.00	
10:21:00	$32.00	
10:23:00	$30.00	
16:30:00	$32.00	

(a) Simulate 5000 vehicle arrivals using spreadsheet starting at 7:00:00. For each vehicle, show its arrival order, vehicle type, time between arrival of previous vehicle and itself, arrival time, service start time, service time, service end time, waiting time and charge.
(For the first vehicle, the arrival of the previous vehicle is assumed to be the simulation starting time.) (30 marks)

(b) Based on 1 replication, find the following performance measures:

(i) Revenue collected during the period 7:15:00 – 10:30:00 (5 marks)
(ii) Number of vehicles arriving between 7:15:00 and 10:30:00 (5 marks)
(iii) Mean waiting time for vehicles arriving between 7:15:00 and 10:30:00 (5 marks)
(iv) 90^{th} percentile waiting time for vehicles arriving between 7:15:00 and 10:30:00 (5 marks)

(c) Replicate the above simulation for 200 times using Data Table. For each replication, show the average performance measures in (b)(i), (ii), (iii) and (iv). (10 marks)

- End -