Test 2

Instruction

- The duration of this test is **2.5 hours**.
- You are free to use Internet except that no communication is allowed throughout the test. Any communication even if it is not relevant to the test will be treated as cheating.
- All the questions must be answered via coding. No manual answer is accepted.
- **DO NOT** print out the full data frame in your final output. Use *head*() to print out a few lines if necessary.
- SAVE your RMD file regularly to avoid loss of work due to computer problem.
- Name your file in the following format: [Session Name]_[Your Name]. For example, SA1_Liu Qizhang.
- Submit your RMD file at the end of the test. Then, additional 10 Mins will be given for you to knit your solution and submit your final version in HTML format. 5 marks will be deducted if you could not submit your HTML file in time.

Please download *journeys.csv* and *stations.csv* from LumiNUS and load them into R as data frames named *journeys* and *stations* respectively. Answer the following questions based on these two data sets.

Q1. Create a column in *journeys* named *Start.Time* that stores the complete date time of the starting time of each journey. For example, a record with Start.Year = 17, Start.Month = 9, Start.Date = 19, Start.Hour = 17, and Start.Minute = 26 should have Start.Time = "2017-09-19 17:26:00". Convert this column into a date format.

Similarly, create a column named *End.Time* to store the complete date time of the ending time of each journey.

Then create a column named *Duration* that is the duration of the journey in **seconds** calculated using *Start.Time* and *End.Time*. (**10 marks**)

Q2. There is already a column named *Journey.Duration* in the data frame, which is the actual journey duration accurate to seconds. Unfortunately, as the data does not provide start time in seconds and end time in seconds, *Duration* calculated in Q1 may not equal to *Journey.Duration* exactly. Please come up with a reasonable way to validate that *Duration* is indeed close to *Journey.Duration* and thus your creation of *Start.Time* and *End.Time* in Q1 are correct. (**5 marks**)

Q3. Develop a visualisation to compare the bike sharing demand across different days in a week. What is your conclusion? (**15 marks**)

(Note: think carefully before coding.)

Q4. In *stations* data frame, create columns *Latitude* and *Longitude* based on *Location* column.

(**5 marks**)

Q5. It is reasonable to assume that bike sharing demand patterns differ between weekdays and weekends. For example, weekday demands may be driven by working people to and from their offices, while weekend demands may be driven by leisure needs. Follow the steps below to validate whether this assumption is true. (**25 marks**)

a) Break the start times of the journeys into half-an-hour time slots. For example, if start time >= 09:00:00 and start time < 09:30:00, then it belongs to time slot *09:00-09:30* (to simplify the notation, we omit seconds); if start time >= 09:30:00 and < 10:00:00, then it belongs to time slot *09:30-10:00*. Create a new column named *Timeslot* in *journeys* to store time slot of each start time.
b) Create a new column named *Is.Weekend* in *journeys* to indicate if the start time falls in a weekend.
c) Plot a chart to show how bike sharing demands change over different time slots on weekdays and weekends respectively. Does the visualisation support the assumption? What is your conclusion?

Q6. One key decision that bike sharing company need to make is the capacity at each station. Over capacity is waste of resources and under capacity means loss of opportunity. In *stations*, each station has a capacity associate with it, which is the current capacity planned by the company. Please develop a visualisation based on data in *journeys* and *stations* to relate total number of journeys starting at each station VS the station capacity. What is your conclusion? (**15 marks**)

Q7. Use ggmap or Leaflet to plot the top 10 busiest starting stations and the top 10 busiest ending stations during weekdays. You plot must clearly distinguish three scenarios: a station that is only one of the busiest starting stations, a station that is only one of the busiest ending stations, and a station that is among both busiest starting stations and busiest ending stations (if any such case existing).

Note: if you do not know how to do Q4, you may manually change the stations.csv file using Excel to get longitude and latitude and then reload the stations data into R. (**20 marks**)

Q8. Ask your own question(s) and develop visualisation to answer them. Please note that I value insights in your solution more than number of charts or technical skills in this question.

(**25 marks**)