IST 652 – Scripting for Data Analysis
School of Information Studies
Syracuse University
Fall / 2021

Final Exam – Due December 15/2021

Use the initial commands listed in the *Final_Exam_Fall2021.ipynb* file to get access to the following datasets:

- Bus Ridership Activity for the years 2019 and 2020 in route SY36 for the city of Syracuse, NY
- Daily weather conditions (temperature, snow, precipitation) for the year 2019 in the city of Syracuse, NY

Note: You will need to upload the *Final_Exam_Fall2021.ipynb* notebook to your Jupyter environment. Please upload it to a location that is inside the *library* folder so that your work doesn't get lost when the server for your environment goes down.

Using the SY36 bus route datasets and the daily weather conditions data perform the data analysis tasks listed at the end of this document. Perform as many tasks and subtasks as needed to get up to 100 points. If the tasks you complete exceed 100 points, and they are all done correctly, your maximum grade will be capped at 100 points.

Each task has a point value for completing it. There are sub-tasks associated with each task that provide ADDITIONAL points. You can complete a sub-task ONLY IF you have completed the main task to which it is associated.

To get all the points for a task or sub-task, in addition to performing the required data analysis mentioned in the task/subtask you must also provide clear documentation of the steps you used in the analysis and, if applicable, meaningful graphs that show your results.

Final exam submission:

Submit your notebook in Notebook format (.ipynb file) and in PDF format. You can use the "Download as" options in the "File" menu of your Jupyter Hub environment to download your notebook in these formats. Submit your files via Blackboard before the deadline.

<u>List of tasks and subtasks</u>

Task 0 (10 points): Mention and summarize two key results from any of the final project presentations of a team that is not your own OR one result each from two different project presentations that are not your own.

Task 1 (30 points): For the year of 2019, determine the number of passengers that board the bus (PASSENGERS_ON) at a particular STOP_ID per day. Use this data to understand how the changes in weather affect the ridership at your selected Bus Stop. Select a bus stop with a daily annual average of at least 5 passengers using it (This means that for any service day of the year, on average, at least 5 passengers boarded the bus from that bus stop).

- **Subtask 1.1** (+10 points): Group the activity at the selected bus stop per month and compare against the average temperature for that month
- **Subtask 1.2** (+10 points): Compare the activity between 2 or more bus stops over each month of the year
- **Subtask 1.3** (+10 points): Determine the 5 bus stops that provide the highest average number of daily passengers during the year

Task 2 (30 points): For the years of 2019 and 2020, determine the number of passengers that board the bus (PASSENGERS_ON) at a particular STOP_ID per week. Select a bus stop where you would expect a high number of users (i.e. Near a shopping mall, school, hospital, etc). Compare the ridership activity between the two years and mention any hypothesis as to why changes could have taken place.

- **Subtask 2.1. (+5 points per hypothesis):** Provide evidence to support that one of your hypothesis is likely true. They evidence could be in the form or newspaper articles, government declarations, etc.
- **Subtask 2.2**. (+10 points): Determine the 10 bus stops that provide the highest average number of weekly passengers for each year. Provide a brief comment on any differences that you find interesting.

Task 3 (15 points): For any of the years 2019 or 2020, compute the distance traveled by the buses serving the route per day. Use the data in the SEGMENT_MILES column for this purpose.

- **Subtask 3.1** (+10 points): Break down the miles travelled by the buses per month. Assuming an efficiency of 7 miles per gallon and a price of US$ 3 per gallon, compute how much it costs to run the buses per month if they use diesel fuel.
- **Subtask 3.2** (+5 points): Determine if there are any anomalies (outliers) in the values of daily distance travelled by the bus. Give a possible hypothesis for why they may have occurred.

Task 4 (20 points): For any of the years 2019 or 2020, study the differences between the time of arrival (TIME_ACTUAL_ARRIVE) and the scheduled bus time (TIME_SCHEDULED) to determine the hours of the day, or the weeks of the year (or some other time period) where significant deviations between the arrival time and the scheduled time appear. Deviations of more than 10 minutes can be considered significant but you can change this limit if it makes sense for your analysis

- **Subtask 4.1** (+10 points): (For 2019 ONLY) Investigate if there is any relationship between weather conditions (temperature, wind, snow) with the deviations between time of arrival and scheduled bus time. Justify your analysis with graphs and documentation.
- **Subtask 4.2** (+15 points): Investigate if the VEHICLE_NUMBER or OPERATOR_ID have any influence in the deviations between time of arrival and scheduled bus time. Justify your analysis with graphs and documentation.