DataEng S23: Project Assignment 3

Data Integration

Due date: May 28, 2023 @10pm PT

Congratulations! By now you have a working, end-to-end data pipeline. Unfortunately, it does not have enough data to properly implement our Data Scientist's visualization. To fill out information such as "route ID" you need to access another source of data and build a new pipeline to integrate it with your initial pipeline. Here are your steps:

A. access the stop event data
B. build a new pipeline for the stop event data
C. integrate the stop event data with the breadcrumb data
D. testing

A. Stop Event Data

Access TriMet "Stop Event" data at this URL: http://www.psudataeng.com:8000/getStopEvents As with the previous data source, this data set gives all TriMet vehicle stop events for a single day of operation.

B. New Pipeline

Your job is to build a new pipeline that operates just like the previous one, including use of Kafka, automation, validation and loading.

C. Integrate Stop Events with Bread Crumbs

The two pipelines (BreadCrumb pipeline and StopEvent pipeline) must update the values in the Trip table such that all of the columns of both tables are filled correctly.

[5/20/2022] Alternatively, it would be OK to load the StopEvent data into a separate table and then use SQL views to integrate the two datasets.

D. Visualization

MapboxGL is a data visualization tool that allows you to view your breadcrumb data and display it on a map. Your job is to integrate this tool with your database tables so that you can query the breadcrumb and trip data in your database server, transform to geoJSON format and display the resulting map visualization. To get started, see this guide.

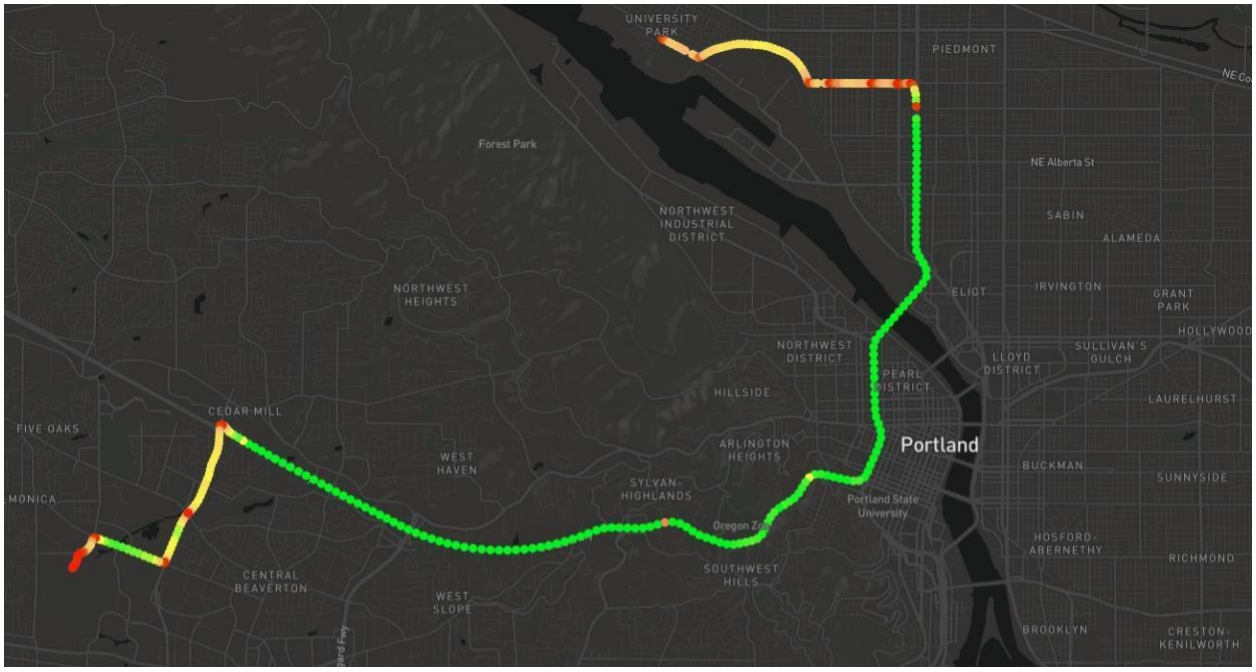
[5/20/2022] Alternatively, you may use an alternative visualization tool (such as folium) to create the required visualizations. We do not provide any guides for doing it, but you are free to do so if you prefer. The submitted visualizations must be equivalent or superior to the visualizations produced by the provided MapboxGL based visualization tool.

Submission

Make a copy of this document and update it to include the following visualizations. For each visualization extract from your database a list of (latitude, longitude, speed) tuples and then use the provided visualization code (see Section D above) to display bus speeds at all of the corresponding geographic coordinates. So, for example, if you are asked to visualize a “trip”, then you must query your database to find all of the (latitude, longitude, speed) tuples for that trip, and then display a map showing the recorded/calculated bus speed at each (latitude,longitude) location.

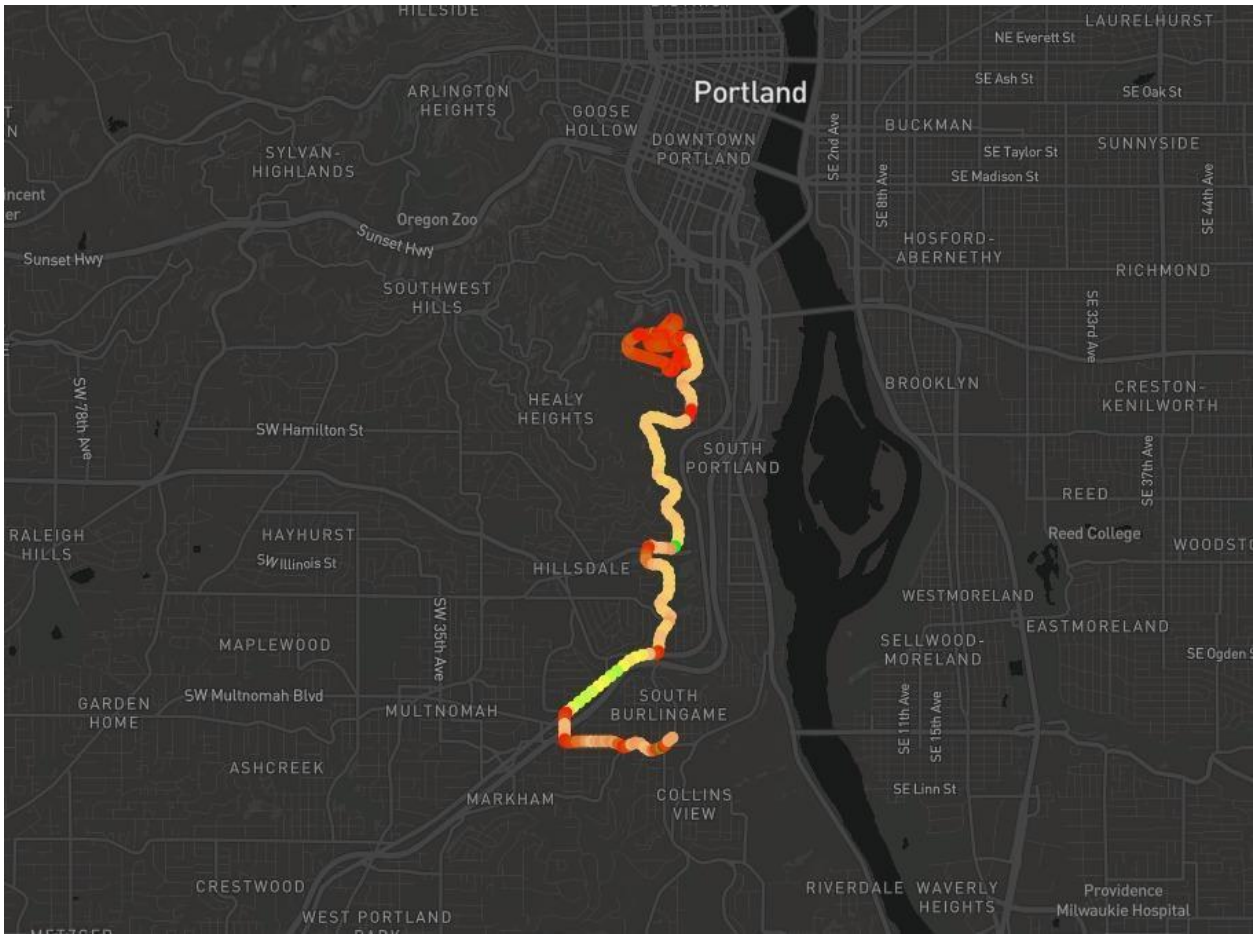
No need to produce software that neatly displays trips, routes, dates, times, etc. onto the visualization itself. Instead, just paste a screen capture of the map-based speed visualization into your submission document and then include a text description of the contents of the visualization. For example, text like this: “Bus Speeds for all outbound trips of route 72 between 9am and 11am on Wednesday, February 15, 2023.”

Visualization 1. A visualization of speeds for a single trip for any bus route that crosses the US-26 tunnel. You choose the day, time and route for your selected trip. To find a trip that traverses this tunnel, consider finding a trip that includes breadcrumb sensor points within this bounding box: [(45.506022, -122.711662), (45.516636, -122.700316)]. Any bus trip that includes breadcrumb points within that box either drove across the tunnel or teleported across!



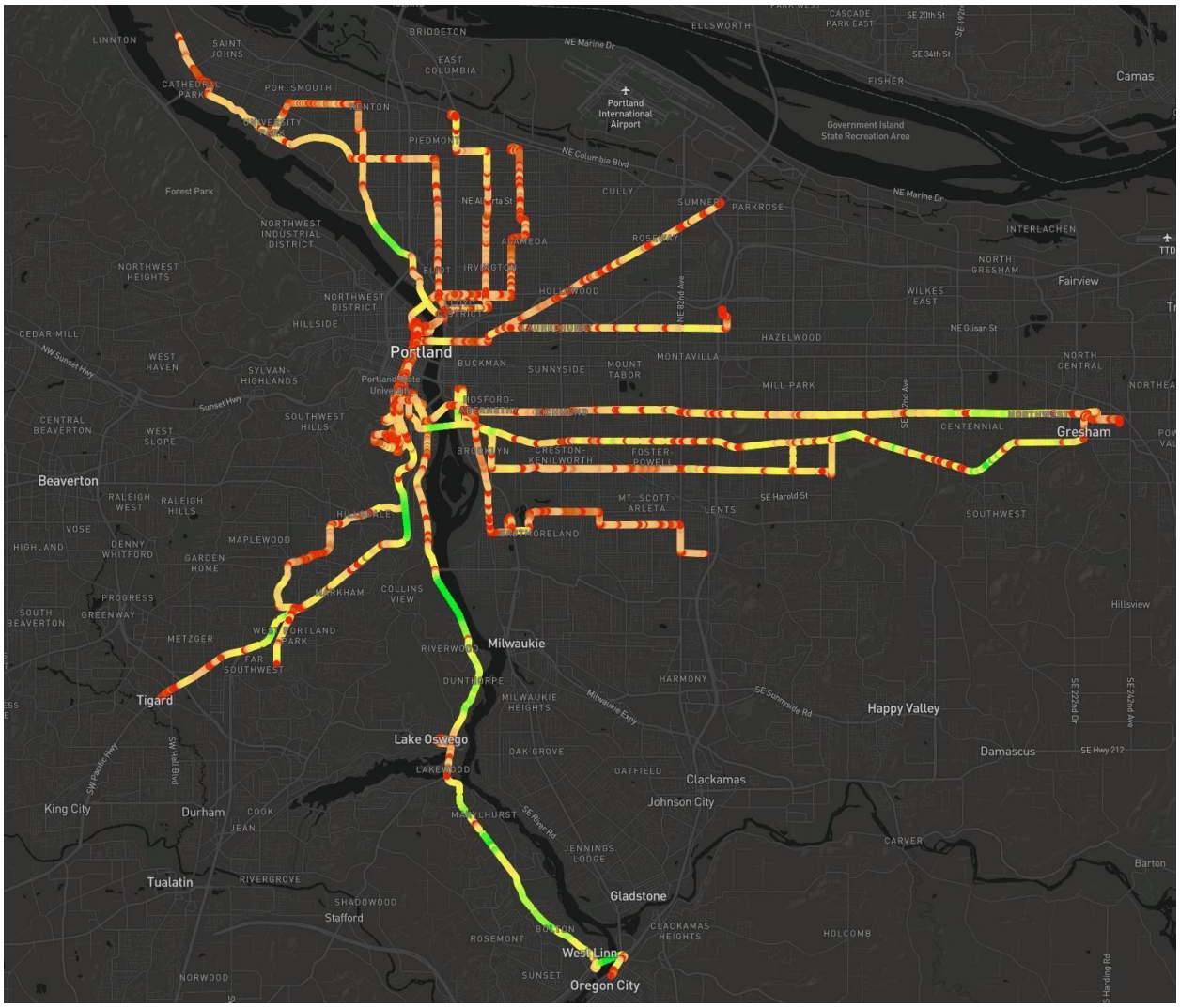
Bus Speeds for outbound trip 238333463 between 10pm and 10:30pm on Sunday, January 15, 2023.

Visualization 2. All outbound trips that occurred on route 65 on any Friday (you choose which Friday) between the hours of 4pm and 6pm.



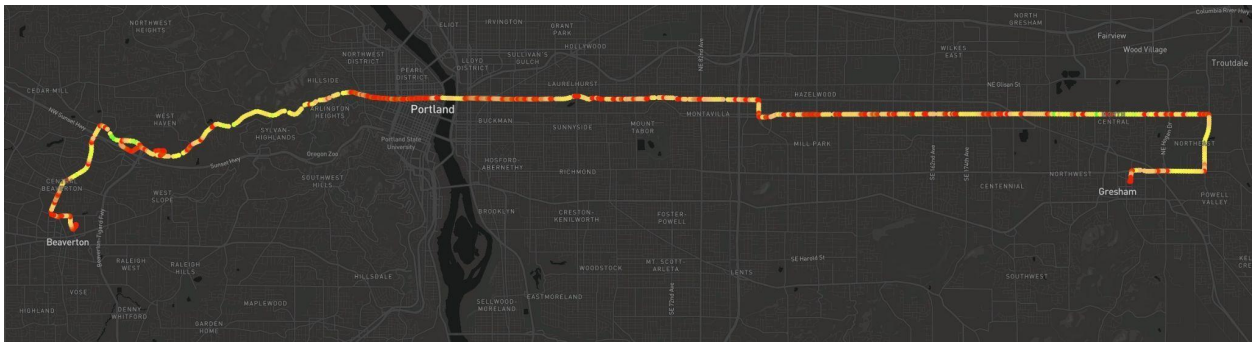
Bus Speeds for all outbound trips of route 65 between 4pm and 6pm on Monday, January 30, 2023.

Visualization 3. All trips that travel to and from PSU campus on any Sunday morning (you choose which Sunday) between 9am and 11am.



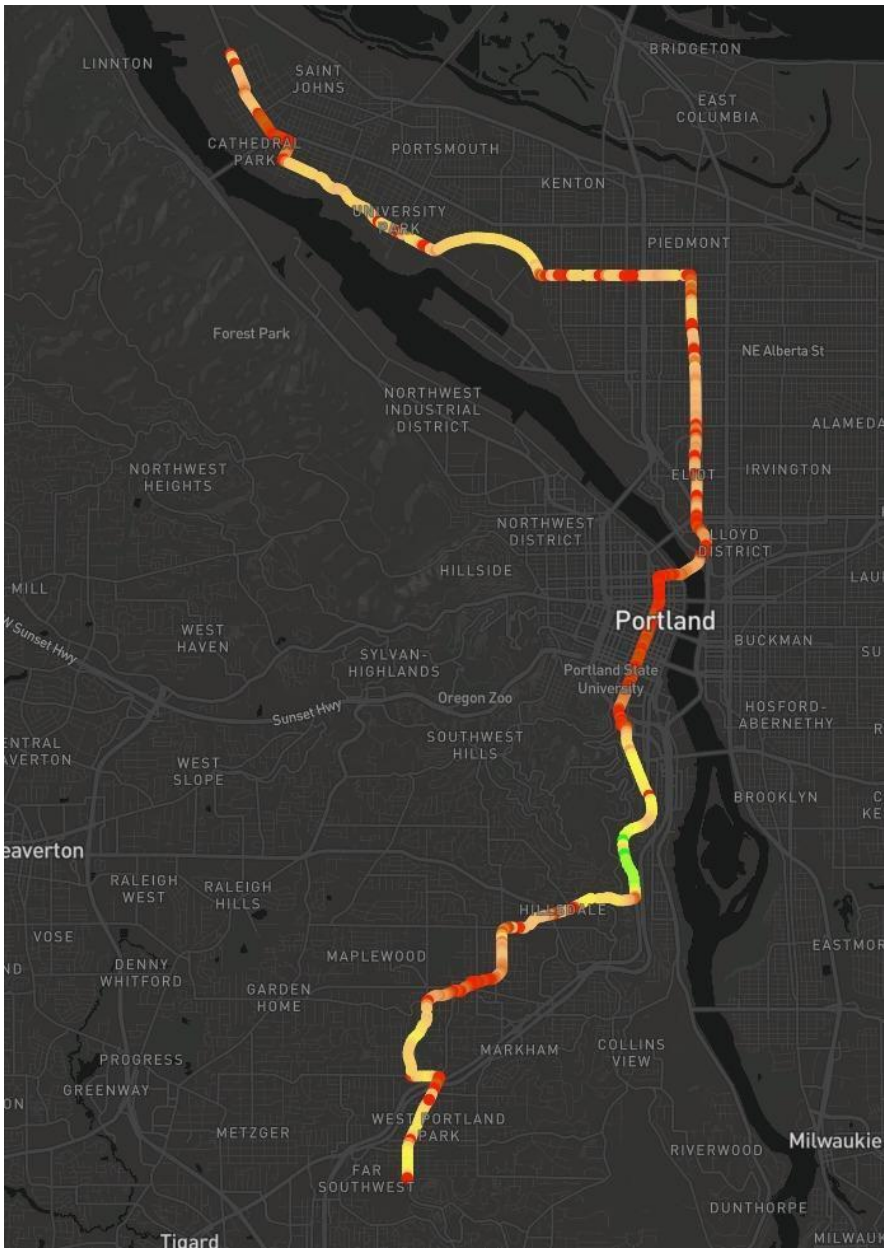
Bus Speeds for all outbound trips traveling to and from PSU campus between 9am and 11am on Sunday, January 15, 2023.

Visualization 4. The longest (as measured by time) trip in your entire data set. Indicate the date, route #, and the trip ID of the trip along with a visualization showing the entire trip.



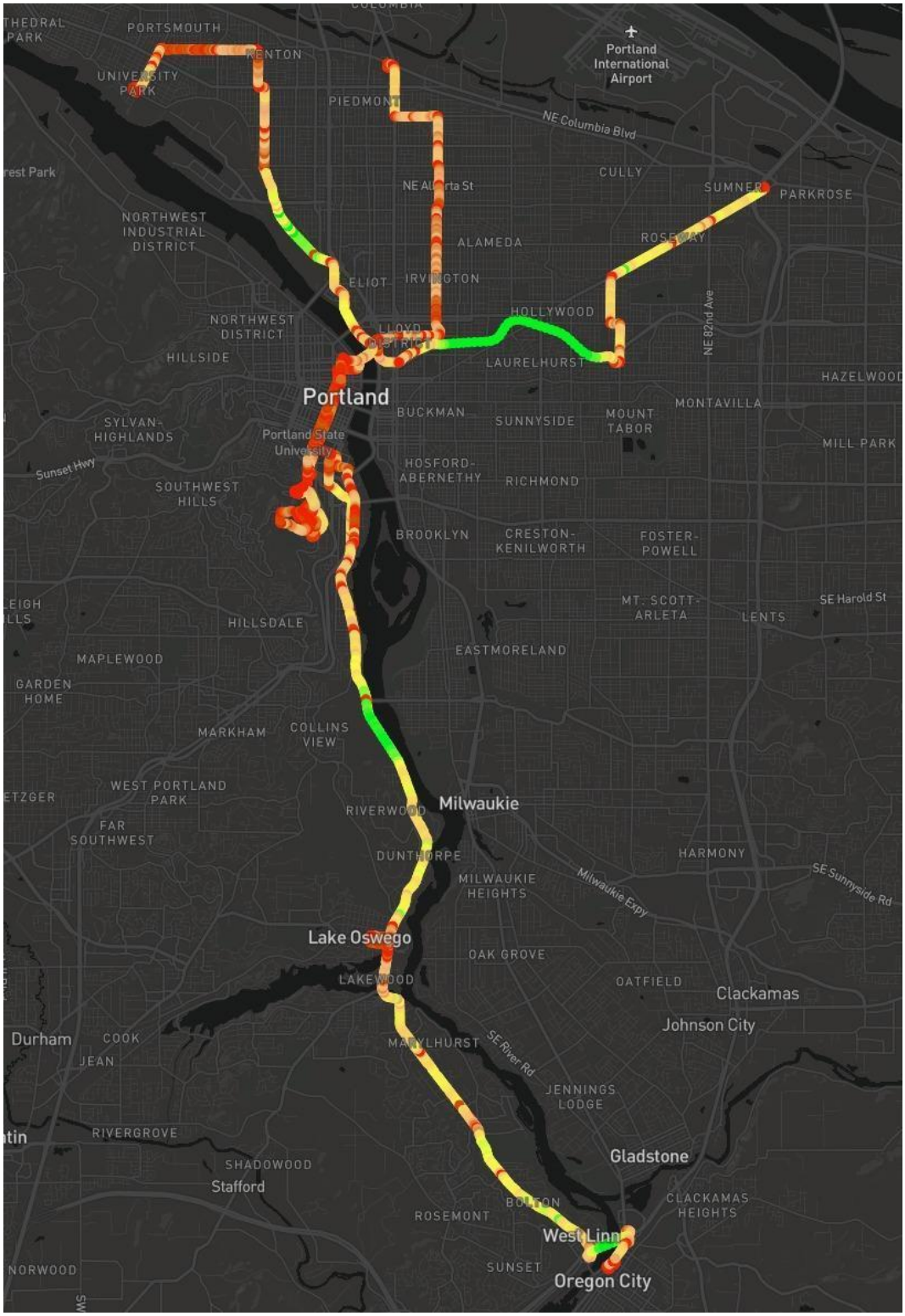
Bus Speeds for trip 237812768 of route between 15:25:53 and 17:52:54 on Friday, January 13, 2023.

Visualization
5a,



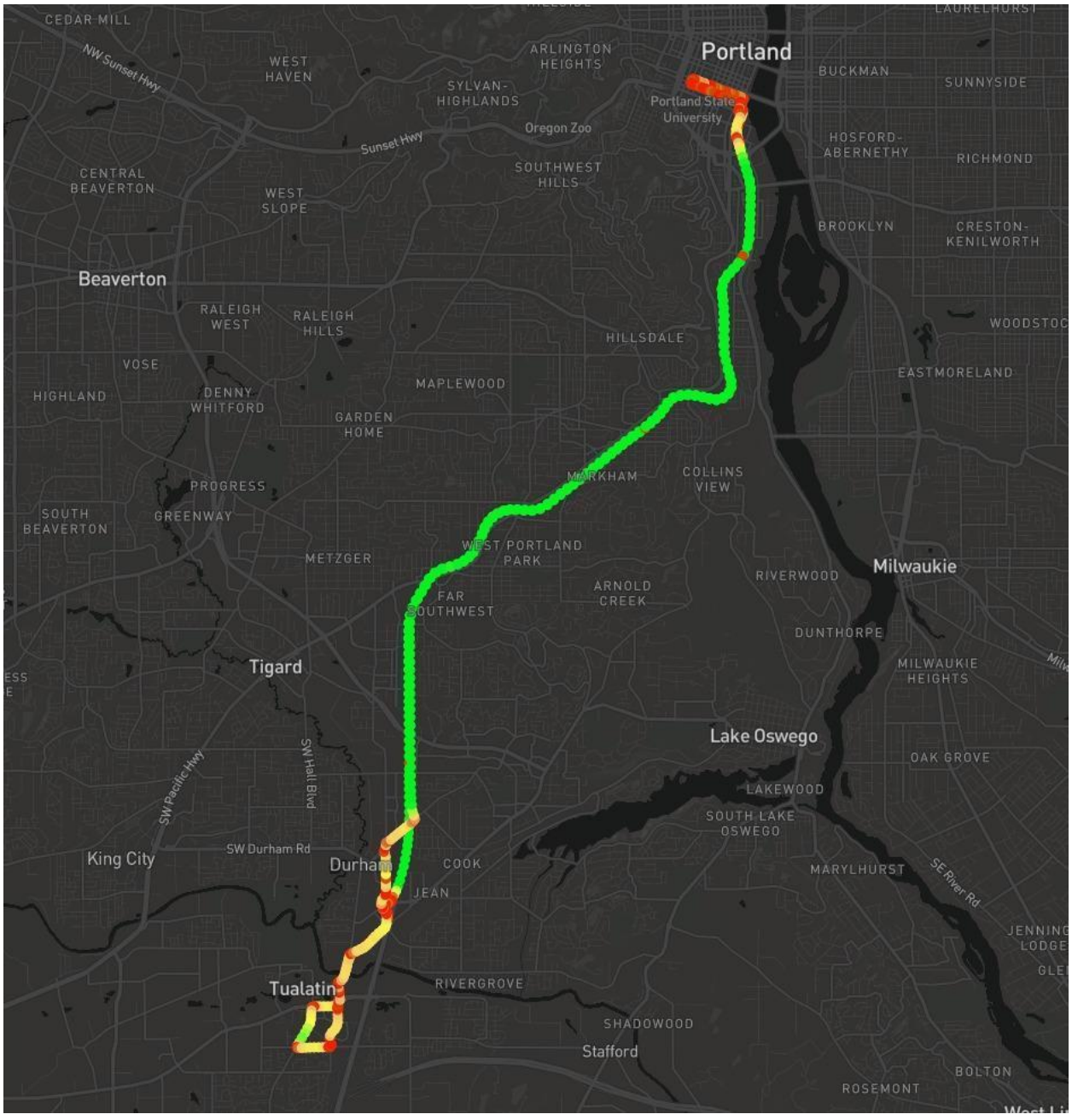
Bus Speeds for all outbound trips traveling to and from UoP campus between 9am and 11am on Sunday, January 15, 2023 as a contrast to visual 3 using UoP instead of PSU.

5b,



Bus Speeds for all outbound trips traveling to and from The Moda Center between 5pm and 8pm on Friday, January 13, 2023 simulating access to The Moda Center on a popular event night.

5c,



Bus Speeds for all outbound trips of route 96 between 10am and 12pm on Monday, January 30, 2023 to show the route I took to get to this class that day.

Your Code

Provide a reference to the repository where you store your code. If you are keeping it private then share it with Bruce and Mina.

https://github.com/asn4psu/data-wizards-project/tree/main/part_three