Appendix E: Data Analytics Process

Overview

We present you with our primary analytic goals and the analytic methods, testing processes, and challenges associated with each. There were three categories of analysis. The first was scooter deployments; how many and where were scooter put out each day. The second was trips, how many trips were taken based on where the trip began and basic stats for the trips including duration and distance. The third was to look at trips and deployments combined together and see how ratios of trips to deployments for different areas differed. This also allowed us to look at average trips per deployment in an area.

Deployment Analytic Goals

Availability is defined as any time when a vehicle is in a state that is available for rental. Deployment is defined as the first availability for each unique vehicle for each day. The City's compliance target for each company was a minimum of 100 unique vehicles deployed in the Eastern Pattern Area.

Fleet is defined as 683 unique vehicles per company deployed in a day. The City's compliance target for each company was 90-100% deployment, 615-683 vehicles. Our goal was to develop a dashboard that showed for each company how many vehicles were deployed in the Eastern Pattern Area and how many total vehicles where deployed to monitor our compliance our fleet and Eastern Pattern Area deployment metrics.

Deployment Analysis

Data granularity was at each duration of time each vehicle was available for rental. Each row represented a time period of availability for a unique vehicle. The deployment metric was developed by finding the earliest start datetime for the date for a unique vehicle. Looking at each vehicle for each day and finding the earliest availability time point gave us the measured deployments numbers.

GROUP BY vehicle ID AND Date AGGREGATE Min(Start Datetime) = [Min start per vehicle]

IF [Start Datetime] = [Min start per vehicle] THEN 1 ELSE null

Deployment Data Testing

We tested the deployment data by comparing within a given geography the total deployment number, total unique vehicles available, total vehicles involved in trips starting in that geography, and total trips starting in that geography.

Deployment Analysis Challenges

The data did include a field for placement reason, however all placement reason data was coded the same so we were unable to use this filed to differentiate deployments that were from a company placing a vehicle from a user ending a tip that began before midnight but ended after midnight. We were also not able to differentiate deployments that were from the company replacing one vehicle with another. This can cause fleet to be overcounted when looking at the fleet size from a daily unique vehicle count.

Trip Analytic Goals

A trip for analytic purposes was defined as any trip that was 60 seconds or longer in duration. Tips that were less than a minute were filtered out, which is consistent with BIKETOWN. Trip data measures were

at a per day level and included total trips, average trip distance, average minutes duration, and total distance.

Trip Analysis

Data granularity was to the trip level. Each row represented a trip and included trip measure fields for distance, duration, start & end location. We calculated mile per hour (MPH) speed of trips on an hourly basis and found a pattern of higher speeds from 4:00 am-6:00 am. We created a heat map to show trips by hour per day of the week and found a pattern of trips peaking in the afternoon. Looking at average trip distance by pattern area showed shorter trips in the inner city and longer distance trips in outer areas of the city.

Trip Data Testing

The speed analysis revealed some data quality issues with the distance and or duration data. For example, one trip had a distance recorded at 34 miles and a duration of 2 minutes giving it a 1,020 MPH speed. However, only about .04% of trips had outlier speeds above 20 MPH.

Trip Data Challenges

Creating route data was challenged by the quality of the location data. The location data was not precise enough to create detailed maps. To capture a representation of where scooters traveled on trips we grouped the trip data by street segment midpoint. If a scooter on a trip traveled within 100 feet of the street segment midpoint it was counted toward a trip that took that street segment. Street segment length variation and high variation of density of routes presented challenges in consistent and consolidated data visualization of routes. We found that a hexagonal binning approach worked the best to mitigate these issues.

Combined Trip and Deployment Data

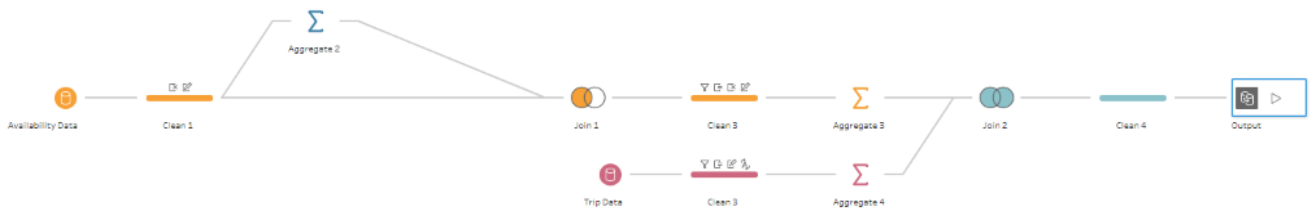
Combined Data Analytic Goal

The goal for combining the trip and deployment data was to be able to show the relationship between how changes in deployment might influence trip numbers.

Combined Data Analysis

Initially we tried combining the data through data blending, however the software we were using was not able to blend the trip and deployment data due to one of the calculation's in the deployment data not being supported for data blending. To combine the data successfully we grouped the availability data by date then by vehicle ID and for that grouped data we aggregated the start datetime to the earliest (the min of start datetime). This filtered for the deployment data. The deployment data was then joined back to the availability data on vehicle ID and start datetime to add in location information. With the deployment data now including the location information we grouped again by date and location and aggregated count of earliest start datetime and count of vehicle.

For the Trip data we grouped the data by date and trip start location and aggregated count of trip, average of distance, and average of duration. Having the deployment and trip data aggregated similarly we could then join them together on date and location fields. This gave us one set of data that included deployment and trip data that we could compare across days and locations.



We created a scatter plot to show the relationship between number of daily deployed vehicles in an area verses the daily number of trips taken in that area. We used a box plot chart to look at daily trips per deployed vehicle by location area.

Combined Data Testing

To evaluate the strength and significance of the relationship we used R and P values.

Combined Data Challenges

The challenge with combining these two data together was that the data had to be reshaped rows to columns in order to make the tables have the same granularity level to join them on date and area fields.