Data Sources

We combine several data sources to do our analysis. Information about individual rides, including the GPS trace, the rider, and timestamp were provided by Ride Report. Weather data were collected from Weather Underground's archive of the KPDX weather station and a Portland Fire Bureau station.

Our goal in this chapter is to discuss this data and what considerations we should have in mind before exploring it in depth. This includes how and by whom the data was collected, who and what is this data actually representative of, and what samples were taken of the data.

Some of these considerations, such as the limited demographics represented in the Ride Report data, pose serious limitations to how our inferences can be generalized. Others, such the large number of missing responses in the Ride Report data, motivate the analysis we are doing in this thesis. Finally, there are other considerations which we will acknowledge here, but addressing them is out of the scope of this thesis. This data set contains an abunance of potential research questions, only a fraction of which could be reasonably addressed in one thesis.

Ride Report

Ride Report's data is the focus of this paper. Knock Software created the app to collect large amounts of information about urban cyclists' routes and experiences on those routes. The hope is that this information will be valuable to city planners.[1]

Ride Report's approach to crowdsourcing this data is particularly important to understand. The app automates every piece of the data collection process except for the rating given by the rider. Thus, the app casts aside nuanced and (somewhat) reliable human input in favor of increasing sample size (i.e. one could imagine a similar app where users have more control over how the route is recorded, have the ability to rate on a more fine-grained scale, and are given more direction in what they are rating for.) This leads to some potential issues that we need to address in our models.

Before we get into the potential issues in the data collection, let's examine the data collection process itself. When installed on a persons phone, the Ride Report app attempts to automatically detect when the user starts riding their bicycle, based acceleromoter data, when a user leaves a familiar Wi-Fi network, and some other pieces of information. When the app detects the start of a ride, it starts recording a GPS trace. At the end of the users ride, the app detects them getting of their bike (in a similar process to how it detected the start of a ride) and prompts them to give a rating of the ride. The ride data is saved then, even if the user does not provide a rating.

This automatic detection of when a ride starts and stops leads to two related and common errors in the dataset: first, one ride is often split into two or more rides at points, such as at a stoplight or a train crossing, where a cyclist stops for an extended period of time; second, car rides are sometimes misclassified as bicycle rides and vice-versa (car rides are not rated.) The app does allow riders to correct the misclassification, but there is currently no way to join split rides back together (Knock is working on changing that, though.)

The app only recently became publicly availible and has undergone significant changes in the course of it's life. In particular, while the ratings have always been binary, the labels have changed at various points in time. For a while the rating labels were "Stressful" and "Chill", while now they are labelled "Recommend" and "Avoid" (see Figure 1.) Other fundamentals of the data collection process, have remained very constant, however.

The data colletion method itself has some problems, but there also may be some biases in the population of riders using the app. The app is only availible on iOS, so only iPhone owners could use this application which may imply a bias toward riders of higher socioeconomic status. At the time of the start of the thesis,

[1] Knock's other project is making a cheaper bicycle counter for cities to monitor traffic flow, again intended to be sold to cities wishing to improve bike infrastructure.

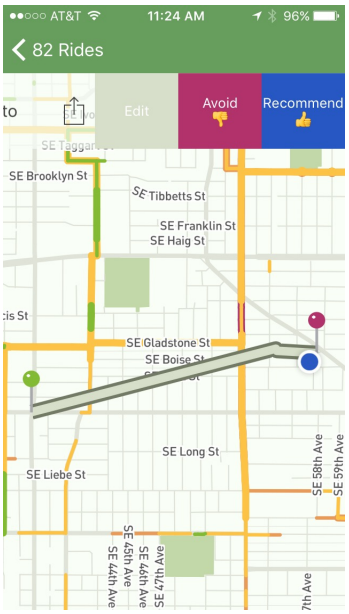

Figure 1: The Ride Report app's interface has changed significantly between versions, including the rating text displayed after a ride. This is the current version as of Februrary 2015.

the app was in private beta, meaning only people who actively sought out using the app were able to use it. Now the application is public and on the Apple App Store, making it more widely availible. So many of the earlier rides may be people within the developer's personal network. Unfortunately, it's hard to make any solid conclusions about the users of the app because Ride Report doesn't collect any demographic data about their riders.

One other issue with the Ride Report data guided our analysis: privacy. Because the data involves timestamps and GPS locations of people's commutes, the data is very sensitive: one could easily infer someone's home and workplace based on their most commons routes. In fact, this data is protected by an end-user license agreement (EULA) which prevents sharing of data, without the explicit permission of those involved. This presented a logistic challenge: how were we to do inference and data exploration without access to the data?

By agreement with Knock Software, identifying data must be kept private. With permission from five riders, Knock was able to give us a small subset consisting of all the rides from those five riders, to be kept confidential. That is the data set we used for prototyping models and some basic exploratory analysis. Knock also agreed to allow us to run models fitting scripts on larger samples of their dataset, as long as they were performed on their computers, with no identifying data leaving their system.

While at first this set up seems like an inconvenience, it actually has some advantages. One of the pitfalls of having an entire data set, especially a high dimensional one, is that in performing exploratory analyses it is often too easy to find spurious "statistically significant" results. Instead, we must come up with our models before running them, greatly limiting the choices we can make in the garden of forking paths.

Weather Data

Slippery roads and formidable winds are no fun for anyone balancing on a two-wheeled vehicle. Weather is, then, one of the most obvious family of predictors for ride rating, at least intuitively. We use the time of a ride to join in data about the weather conditions during the ride, including

- the temperature,
- whether and how much it is raining,

- whether the roads are wet or have puddles,
- wind and gust speed.

We include the first two, temperature and percipitation, to account for rider comfort. A swealtering, frigid, or stormy day could make an unpleasant experience for a bicyclist and thus could lead to more negative ratings.

On the other hand, we include the last two, wet road and gust speed, as factors that impact safety. During and after storms, puddles often accumulate in bike lanes before the center of the road, pushing cyclists into lanes shared by cars, which are often more dangerous.

Gust speeds impact the aerodynamics of a ride, which are particular important for bicyclists. It's one of the main reasons cyclists care about getting into lower (and more aerodynamic) rider positions. Thus, high wind or gust speeds may affect rider rating.

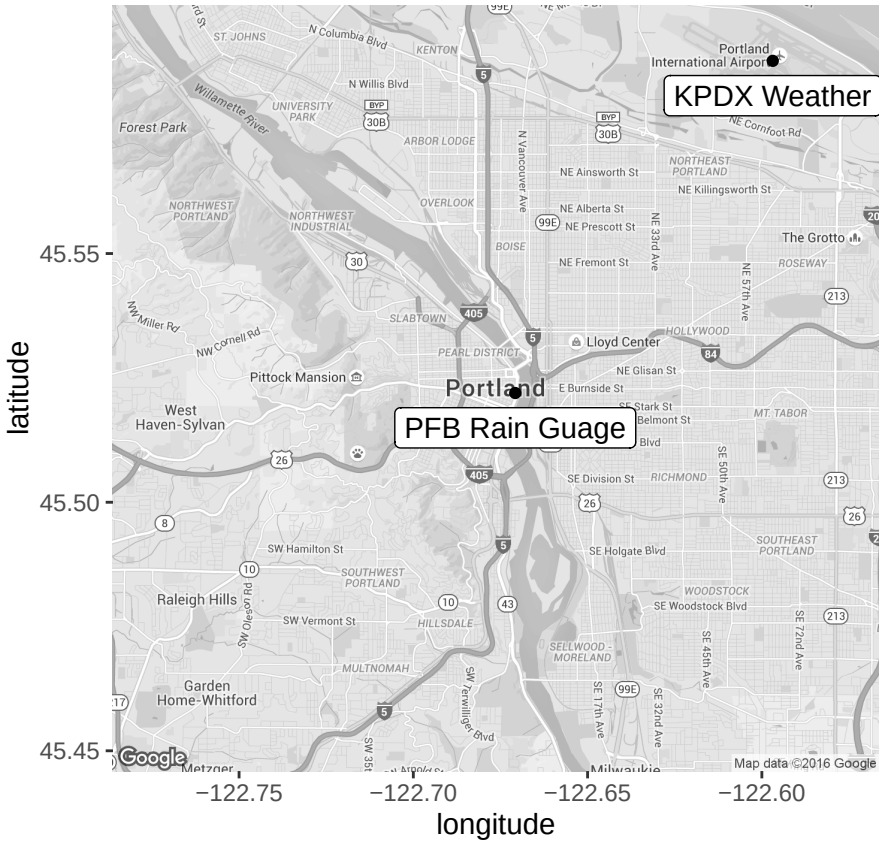

Figure 2: Positions of weather data collection sites. Daily weather information was collected at the KPDX weather station at Portland International Airport. Hourly percipitation data was collected at the Portland Fire Bureau's rain guage in downtown Portland.

We are limiting our study to rides in Portland, Oregon. Given this, we can first assume that it may be reasonable to expect that riders are used to the same climate, and thus have somewhat similar responses to weather. This also makes it reasonable to use data from one nearby weather station, rather than attempting to collect from several stations and creating a spatial model for weather.

For daily summaries of weather conditions, we used weather history from the KPDX weather station at Portland International Airport downloaded from Weather Underground. From this we were able to get daily weather data, including

- Average, minimum, maximum temperature for the day.

- Total percipitation.
- Mean wind speed, as well as gust speed (speed of brief, strong winds.)

We also got hourly rainfall data from a data stream at the Portland Fire Bureau Rain Gage at 55 SW Ash St., which is just about the geographic center of Portland. This just gives raw uncorrected rain guage data, but gives us a fine grain look at how much rain there has been recently.

For daily weather data, such as temperature highs and average wind speed, we use information from the KPDX weather station. This weather station is the best calibrated in the area. It is further from the geographic center of the rides we are examining, but because the weather is daily summary statistics, we don't expect closer weather stations to be much more informative. Figure Figure 2 shows the geographic positions of these two stations.

Road Data

To get road feature data, such as bike lanes and intersection types, we had to bring in other data. More on this later. (We may use Open Street Map data and/or data from civicapps.org/datasets)