Five essential skills for transportation data science

The transportation industry is ravenous to hire data scientists; no PhD required

Transportation touches the lives of nearly every living person every day. The public sector entities that build and operate roads, highways, and other public transportation networks are ravenous to hire data scientists with the skills to make sense of their voluminous data. The needs are abundant and the application areas are not particularly complex. Functional proficiency with just the five skills listed below is sufficient to be able to answer many of the questions that, for example, U.S. state departments of transportation grapple with daily.

In fact, the industry is a perfect entry-point for new data scientists.

Interested in applying your skills in an arena where you can have a broad and lasting impact? Here are five essential skills you’ll use as a data scientist supporting a transportation agency:

1. Data transformation. Calculating sums and averages by group; transforming between long and wide format; log transformations.
2. GDAL or other libraries for manipulating spatial data, including a working knowledge of spatial projections, and GIS concepts like buffering and spatial joins.
3. Decision Trees
4. Plotting
5. Count Regression Models

This article illustrates the application of these skills through an applied investigation of the relationship between roadway lighting and traffic safety in Pennsylvania using R. The statistical programming language R provides an excellent workbench for many transportation questions, in part because of the rich ecosystem of community packages for R. Python’s SciPy is an excellent alternative (though with an arguably steeper learning curve).

The Setup

Suppose you’re an analyst for the Pennsylvania Department of Transportation (PennDOT). The agency is weighing investing more money into highway lighting versus other safety improvements such as installing guardrails. Your charge is to help answer the question: “what impact does street lighting have on crashes?”

Data Preparation: GIS + Data Transformation

Data Transformations

Data analysis is almost always proceeded by data preparation to transform source data into a format suitable for analysis. This includes operations like deriving new attributes (such as grouping ranges into bins), joining tables, aggregation, and reshaping (e.g. long-to-wide transformation). R’s data.table and dplyer and Python’s Pandas libraries are data transformation multi-tools; learn one well, and develop sufficient familiarity with the others to be able to perform simple operations.

In this article, I use R’s data.table package, which I prefer for its speed and concise syntax. Those new to programming may prefer dplyr, which has a more verbose but human-readable syntax.

[data transformation source code]

GIS

Because transportation describes the movement of people and goods through space, transportation data is often spatial. If you’re going to work with transportation data, you’re going to need a basic GIS skillset. At a minimum, you’ll need to be familiar with basic GIS operations like buffering, dissolving, and joining, as well as reprojecting spatial data between Web Mercator, StatePlane, and UTM projections.

In R, the `sf` package provides an interface for using the powerful GDAL spatial library, allowing you to use the same general functions in R that may be familiar from working with PostGIS or Python.

Projections are a way of mapping spatial information on a round planet onto a flat computer screen. Projections define locations in space, by defining distance in units from a fixed point of reference (e.g. the equator). Web maps typically display data with distances measured in degrees of latitude and degrees of longitude, using a projection called Web Mercator (EPSG:4326). Most state datasets are created in a StatePlane projection (a state may have one or more regional StatePlane projections), locations are identified feet (x, y) from a defined point of reference. Another useful projection is UTM, which divides the world into zones and identifies points based on meters (x, y) from a known point of reference. E.g. Pennsylvania is mostly within UTM Zone ??N, (EPSG:326??).

A degree of longitude is widest at the equator and narrows toward the poles, so it typically not used for most GIS operations like creating buffers. To measure distances, first reproject into a StatePlane or UTM projection. When finished, reproject back into Web Mercator to show your spatial data as a web map.

(Note: both data.table and sf are extensions of R’s base data.frame object. data.table objects and sf objects often do not play well together.)

Decision Trees.

Decision trees are useful tools for identifying structural relationships in data. Their principal use is to classify observations into two or more classes by deriving a rule set (a task for which many other alternatives offer superior performance), but their method for defining rulesets can reveal much about a dataset, and the ability to visualize a decision tree classifer’s “under the hood” logic in a way that is interpretable by anyone makes a decision tree an extremely useful diagnostic tool. I use Decision Trees both to learn about my datasets (which attributes are predictive of an outcome, for example), and will also use a Decision Tree to fill in missing attribute values.

Plotting

Data visualization enables significant quantities of information to be presented in a way in which a human can immediately apply our powerful spatial reasoning and pattern recognition abilities to a far greater set of information than we could otherwise hold in our mind at once. `ggplot2` is R’s workhorse data visualization library.

Count Models

Regression is a highly useful statistical tool for identifying reliable relationships in data. You’re most likely familiar with Ordinary Least Squares linear regression, but regression is, in fact, a family of models. Expanding your repertoire to include count models (e.g. counting crashes) has many applications in a transportation context. [Reference] Even if you obtain a good fit from a linear model, if your Guass-Markov assumptions aren’t met, any inference you draw from the model is likely to be invalid.

The data generation process and distribution of your data

Close with a pitch for the importance

Nearly every living person interacts with and is impacted by the quality and effectiveness of our transportation systems, whether through its success—mobility and economic opportunity—or transportations many challenges (degraded air quality; nearly one in one hundred Americans born today will die in an automobile accident; congestion; dangerous cities).

For data transformation and management, the R community seems evenly split between two competing packages: `data.table` and `dplyr`. Both packages make it easy to aggregate and transform data. In general, `data.table` has a slight edge in terms of speed. data.table’s syntax is extremely concise, making it more difficult to learn but ultimately faster to write (e.g. `DT[, c := a + b]` versus `df … mutate blah blah blah`). `data.table` also ships with a number of useful functions that are highly useful for transportation.

Because the distance between a degree of longitude is different at the equator (~69 miles) than at the poles (~0 miles), and, except at the equator, a degree of latitude is not equal to a degree of longitude, distances are typically measures in feet or meters.

Projections, here’s what you need to know:

Web maps are created in Web Mercator (EPSG:4326) projection, with locations identified by degrees of latitude from the equator and longitude from the prime meridian.

If you are measuring distances, you will almost always need to reproject into a StatePlane or UTM projection.