Journal of Statistics Education

Rail Trails and Property Values: Is There an Association?

Ella Hartenian
Smith College

Nicholas J. Horton
Amherst College

Journal of Statistics Education Volume 23, Number 3 (2015)
www.amstat.org/publications/jse/v23n2/horton.pdf

Key Words: biking score, linear parks, greenways, multiple regression, real estate, walking score, Zillow.com

Abstract

The Rail Trail and Property Values dataset includes information on a set of $n = 104$ homes which sold in Northampton, Massachusetts in 2007. The dataset provides house information (square footage, acreage, number of bedrooms, etc.), price estimates (from Zillow.com) at four time points, location, distance from a rail trail in the community, biking score, and walking score. The dataset is amenable to use with exploratory data analysis in introductory courses, intermediate courses with a focus on visualization and multivariate relationships as well as advanced courses that utilize repeated measures regression models and more sophisticated graphics.

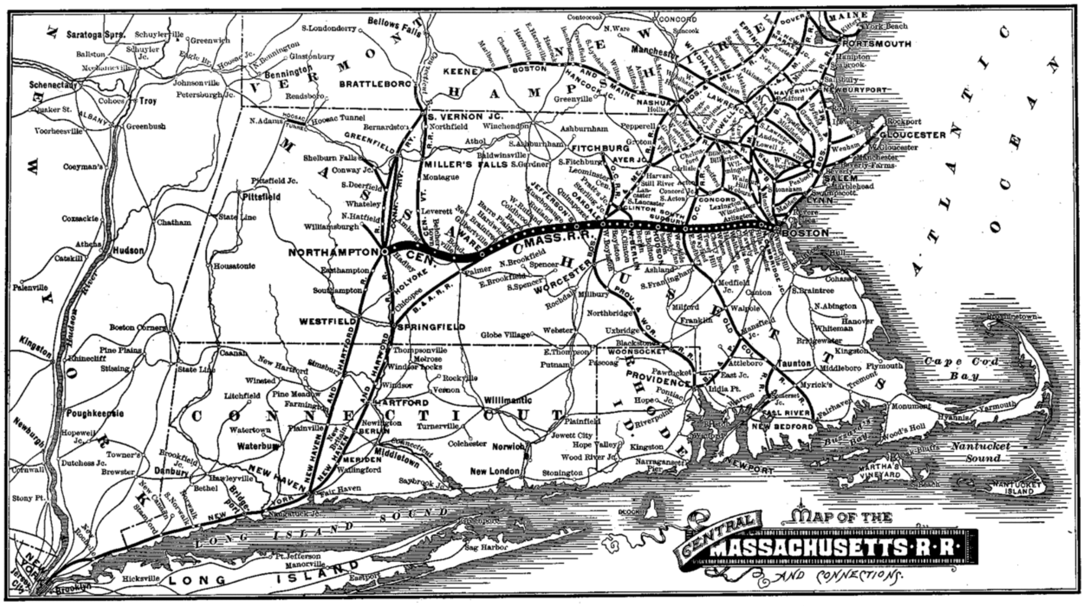


Figure 1: Map of the Central Massachusetts Railroad network from 1888.

1. Introduction

Railroads transported many of the goods and people in the United States from the middle of the 18th century until the early 20th century. In the 1870s, there were more than a dozen competing railroad systems in Southern New England alone, with nearly every medium to larger size town connected to one or more of these networks. As an example, Figure 1 displays a map of the extensive and interconnected Central Massachusetts Railroad system in 1888 and its connections to other companies in Massachusetts.

However, as automobiles became more affordable and road infrastructure expanded, train travel decreased dramatically. Only major lines remained intact while a huge number of branch lines were removed from service and tracks reclaimed for scrap. In Massachusetts alone thousands of miles of these corridors were formally abandoned and many of these paths were lost to development or reuse. The passage of the Railroad Revitalization and Regulatory Reform Act in 1976 and the National Trails System Act in 1978 led to funding and technical assistance to preserve the corridors ("rail-banking"). This legislation spurred the conversion of many of these old track beds to multi-use trails.

In 1986, an initial proposal was brought forward by then-Governor John Ashcroft of Missouri to convert 185 miles of unused railroad bed into what became the Katy Trail, one of the first rail trails in the US. That same year, the Rails-to-Trails Conservancy (http://www.railstotrails.org) was established and began advocating nationally for trail development. By 1989, 200 rail trails were open across the country and the Rails-to-Trails Conservancy had over 7,000 members. Now there are 20,000 miles of trail and tens of millions of users each year, with thousands of additional conversions underway (Rails to Trails Conservancy 2015). Throughout the last decades of the 20th century, the "rails to trails" movement gained increasing visibility within the burgeoning environmental movement; the trails simultaneously provided an alternative to auto-centered transportation as well as a way for people to enjoy nature without contributing to carbon emissions.

A secondary benefit of the trails has been to spur exercise. More than 40 percent of American adults do not engage in leisure-time physical activity and a similar number are overweight or obese (Centers for Disease Control and Prevention 2010). As such, the existence of rail trails responds to a public health need for exercise on safe, scenic and accessible paths. Because many trails link the centers of cities and towns, rail trails facilitate exercise through daily activities like commuting or grocery shopping. One observational study in Indiana found that 70% of users of six different paths statewide reported that they exercised more as a direct result of the trail (Eppley Institute for Parks & Public Lands 2001).

Advocates of these trails have also suggested that they have an economic benefit for the communities where they are located. Proximity to parks and greenspaces has thought to be associated with an increase in property values. However rail trails (also referred to as linear parks) are not necessarily analogous to traditional parks. Rail trails are long corridors of paved trail, which may or may not be surrounded by greenery, unlike the typically larger and consistently verdant landscape of parks. Additionally, rail trails provide easy access to locations along the trail, a feature more analogous to proximity to a subway line than a park. There are few studies empirically linking rail trails with economic activity or changes in home prices although one could imagine increases in property values related to easy access to the trail and the amenities along the trail, or as a result of increased economic activity by trail users (American Trails 2011).

Two studies have looked specifically at house prices and rail trails with a hedonic price model. This econometric model is used when a good (e.g., a house price) can be broken up into discrete components and a market value can be attributed to each component. Karedeniz (2008) at the University of Cincinnati examined home values along the Little Miami Scenic Trail in southwest Ohio and found that for houses within 10,000 feet of the trail, proximity to the trail positively impacts property values. Lindsey et al. (2004) used a hedonic price model to measure the impact of greenways on property values in Indianapolis, Indiana. They found that the association between price and trails varied from trail to trail. While homes near the Monon Trail sold for about 11% more than houses more than half a mile away, there was little or no difference in price for homes near other trails in Indianapolis.

Anecdotal evidence also suggests that rail trails are an increasingly sought-after amenity. The National Association of Home Builders found that nearly 36% of 2,000 recent home-buyers said multi-use trails were "important" or "very important" in their choice of a home (American Trails 2011). This is also consistent with other evidence such as the inclusion of a "Bike Path" field on the New England Multiple Listing Service and the number of hits on the housing section of Craigslist for "Rail trail", suggesting that proximity to trails is an increasingly important amenity for buyers and renters.

2. Goal

This dataset could be used in a first course in statistics which includes exploratory data analysis, visualization or modeling. This dataset is also appropriate for a second or more advanced course in statistics that includes modules on multiple regression modeling, data manipulation, mapping or repeated measures. It can be used to assess the association of various factors with change in property values to shed light on relationships of interest in terms of urban planning and sustainable development. The dataset provides material for a discussion on confounding variables, categorical versus continuous variables, and the process of scraping data from the internet. Students can look at basic descriptive statistics and compare houses of similar structures, then assess whether there are predictors of change in price of houses. Even in this simple analysis, after controlling for baseline price, houses closer to the trail appreciated more than houses further away between 1998 and 2014.

We first describe the dataset in detail including how the home value data was obtained, how the distance between each house and a rail trail entrance was calculated, and how other variables were scraped from the internet. We then consider how to undertake exploratory analysis and modeling, including use of repeated measures regression to assess whether there is differential price growth for properties nearer to the rail trail. We conclude that homes within a half mile of the rail trail tend to appreciate more than homes further away, and suggest several other possible avenues for analysis.

3. Dataset

3.1 Background

Data were collected on house sales in Northampton, Massachusetts, a city of more than 28,500 people located along the Connecticut River approximately 100 miles west of Boston, Massachusetts and 80 miles north of New Haven, Connecticut. The city, founded in 1654, served as the nexus of train travel between Boston and New Haven in the nineteenth century until passenger service was discontinued in the 1920's and freight service largely abandoned in the 1960's.

The dataset was collected using information from Northampton because the conversion of a nearly 3 mile long section of the Williamsburg Branch of the New Haven railroad into one of the oldest municipally operated rail trails in 1984 allows us to assess possible changes over time between 1998–2014. This natural experiment allows us to compare the change of house prices for houses close to the trail versus those that are more distant. Figure 2 displays a map of this trail (labeled Francis Ryan Northampton Bikeway). A number of other rail trail projects followed in and around Northampton, including the Norwottuck Rail trail (1992, ∼ 10 miles), the Manhan Rail trail (2003, ∼ 5 miles) and the 2011 linkage between the Northampton and Easthampton trail networks.

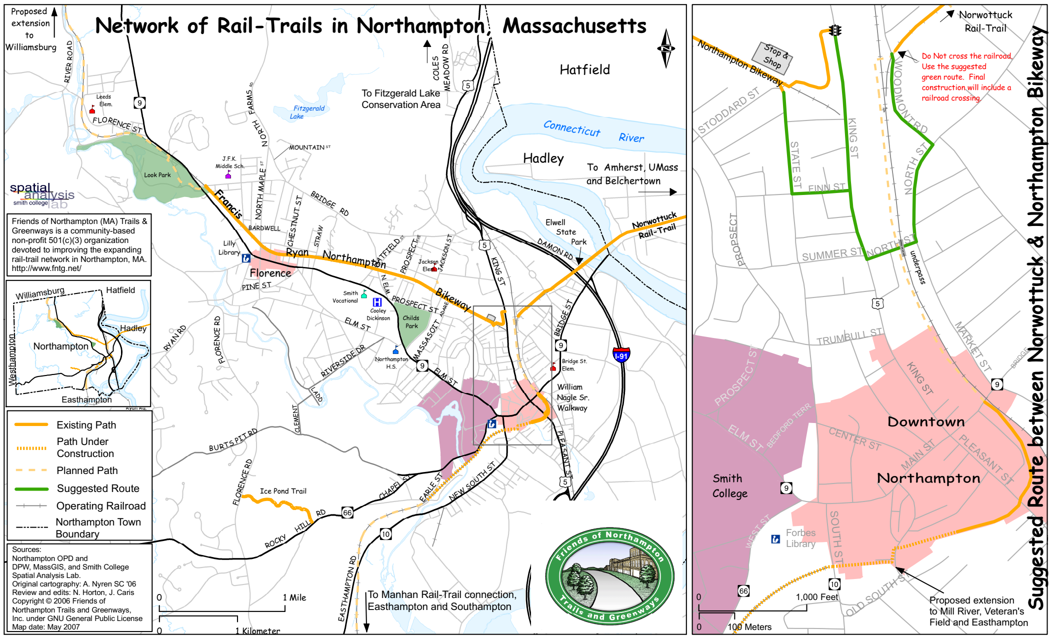
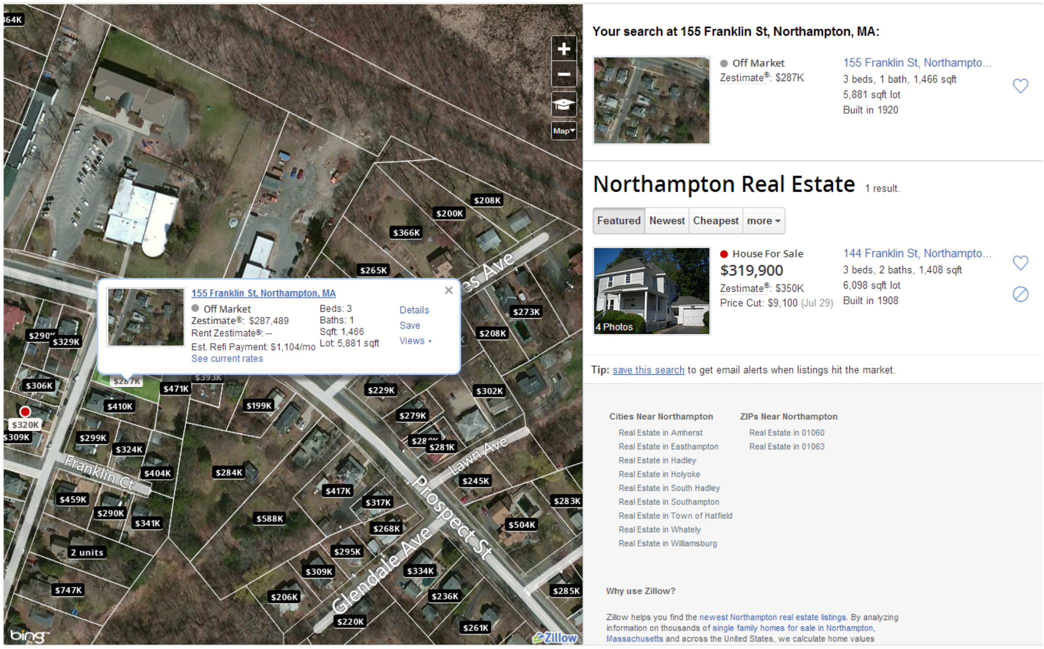

Network of Rail-Trails in Northampton, Massachusetts
Proposed extension to Williamsburg
RIVER ROAD
Leeds Elem.
FLORENCE ST
Look Park
spatial analysis lab
smith college
Friends of Northampton (MA) Trails & Greenways is a community-based non-profit 501(c)(3) organization devoted to improving the expanding rail-trail network in Northampton, MA. http://www.fntg.net/
Williamsburg
Hatfield
Hadley
Northampton
Westhampton
Easthampton
Existing Path
Path Under Construction
Planned Path
Suggested Route
Operating Railroad
Northampton Town Boundary
Sources: Northampton OPD and DPW, MassGIS, and Smith College Spatial Analysis Lab. Original cartography: A. Nyren SC '06 Review and edits: N. Horton, J. Caris Copyright © 2006 Friends of Northampton Trails and Greenways, Inc. under GNU General Public License Map date: May 2007
0
1 Mile
0
1 Kilometer
To Fitzgerald Lake Conservation Area
Fitzgerald Lake
Hatfield
Connecticut River
Hadley
To Amherst, UMass and Belchertown
Elwell State Park
Norwottuck Rail-Trail
Francis Ryan Northampton Bikeway
Florence
Lilly Library
J.F.K. Middle Sch.
Smith Vocational
Cooley Dickinson
Childs Park
Jackson Elem.
Northampton H.S.
Bridge St. Elem.
William Nagle Sr. Walkway
Ice Pond Trail
KING ST
ELM ST
BRIDGE ST
PLEASANT ST
DAMON RD
CHAPEL ST
EARLE ST
NEW SOUTH ST
EASTHAMPTON RD
ROCKY HILL RD
BURTS PIT RD
FLORENCE RD
RYAN RD
RIVERSIDE DR
PROSPECT ST
MASSASOIT ST
PINE ST
MOUNTAIN ST
NORTH MAPLE ST
BRIDGE RD
To Manhan Rail-Trail connection, Easthampton and Southampton
Friends of Northampton Trails and Greenways
Suggested Route between Norwottuck & Northampton Bikeway
Norwottuck Rail-Trail
Northampton Bikeway
Stop & Shop
Do Not cross the railroad. Use the suggested green route. Final construction will include a railroad crossing.
STODDARD ST
STATE ST
KING ST
FINN ST
WOODMONT RD
NORTH ST
PROPSECT
SUMMER ST
NORTH ST
underpass
TRUMBULL ST
MARKET ST
BRIDGE
PROSPECT ST
ELM ST
BEDFORD TERR
CENTER ST
MAIN ST
PLEASANT ST
Downtown
Northampton
Smith College
WEST ST
Forbes Library
SOUTH ST
OLD SOUTH ST
0
1,000 Feet
0
100 Meters
Proposed extension to Mill River, Veteran's Field and Easthampton

Figure 2: Map of the Northampton Rail trail, dating from 1984, connecting from King Street to Bridge Street/Look Park (2.7 miles).

3.2 House Prices

We began with the set of all homes sold in Northampton in 2006 and 2007, and used data from Zillow to ascertain house prices. Zillow.com is a real estate website that provides information on residential property across the country. Houses (including those currently for sale and those not for sale) are listed on the website. "Zestimates" of home value are based either on list price or, when the house isn't for sale, an algorithm is used to estimate the price which takes into account the home's characteristics as well as the local housing market. In addition to the "Zestimate" is a Value Range (which corresponds to a 70% confidence interval, http://www.zillow.com/zestimate).

Searching for a house by address on Zillow.com produces a bird's eye view of the neighborhood with price estimates for that address and surrounding houses. Figure 3 displays the Zillow information for one of the homes in our dataset, while Figure 4 displays this home's Zillow estimate over time, along with other houses in its zip code and Northampton as a whole.


155 Franklin St, Northampton, MA

● Off Market
Zestimate®: $287,489
Rent Zestimate®: --
Est. Refi Payment: $1,104/mo
See current rates

Beds: 3
Baths: 1
Sqft: 1,466
Lot: 5,881 sqft

Your search at 155 Franklin St, Northampton, MA:

● Off Market
Zestimate®: $287K

155 Franklin St, Northampto...
3 beds, 1 bath, 1,466 sqft
5,881 sqft lot
Built in 1920

Northampton Real Estate 1 result.

● House For Sale
$319,900
Zestimate®: $350K
Price Cut: $9,100 (Jul 29)

144 Franklin St, Northampto...
3 beds, 2 baths, 1,408 sqft
6,098 sqft lot
Built in 1908

Tip: save this search to get email alerts when listings hit the market.

Cities Near Northampton
Real Estate in Amherst
Real Estate in Easthampton
Real Estate in Hadley
Real Estate in Holyoke
Real Estate in South Hadley
Real Estate in Southampton
Real Estate in Town of Hatfield
Real Estate in Whately
Real Estate in Williamsburg

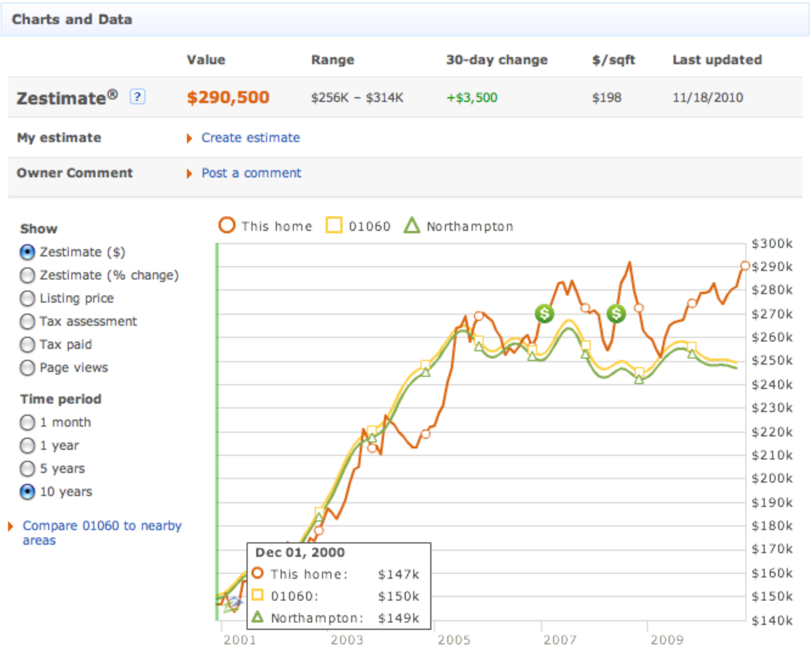
ZIPs Near Northampton
Real Estate in 01060
Real Estate in 01063

Why use Zillow?

Zillow helps you find the newest Northampton real estate listings. By analyzing information on thousands of single family homes for sale in Northampton, Massachusetts and across the United States, we calculate home values

Figure 3: An image of the Zillow page for one home in our study, showing a bird's eye view of the neighborhood as well as providing general housing information including price and square footage.

Photos, basic information and price trends are available for many houses. The accuracy of the information is dependent upon the location of the home; only some counties publicly release housing attribute information. In counties that don't, Zillow uses market trends,



Charts and Data

	Value	Range	30-day change	$/sqft	Last updated
Zestimate®	$290,500	$256K – $314K	+$3,500	$198	11/18/2010
My estimate	Create estimate				
Owner Comment	Post a comment				

Show
● Zestimate ($)
○ Zestimate (% change)
○ Listing price
○ Tax assessment
○ Tax paid
○ Page views

Time period
○ 1 month
○ 1 year
○ 5 years
● 10 years

Compare 01060 to nearby areas

Figure 4: The ten-year price data for a house in our study, taken in 2010. The trends for the house, zip code and town are displayed. The money symbols represent dates the house was sold.

user-entered information and list/selling price when available. As of June 2014, Zillow claimed that the 70.8% of their "Zestimates" in the Boston, Massachusetts area (1.4 million homes) were within 10% of the selling price, with a median error of 5.9%.

Zillow has created an application programming interface (API) to facilitate automated access to their database ("data scraping"). Users need to register with Zillow and receive a ZillowID which allows a finite number of database accesses per day for free. The Zillow package in R (available from http://www.omegahat.org/Zillow) allows the user to find the Zillow estimate for a property specified by a street address, find information about the house (e.g. number of bedrooms and lot size) as well as find comparable properties. More information about the Zillow API can be found at http://www.zillow.com/wikipages/Zillow-API.

We used Zillow's estimated price in 2007 along with the 10-year history of home values to extract the estimated price of each of the homes in 1998 and then additionally collected the price estimate in 2011 and 2014, yielding four estimated prices for each home.

To allow comparisons of the house prices over time, we adjusted for the Consumer Price Index using a Bureau of Labor Statistics calculator (http://data.bls.gov/cgi-bin/cpicalc.pl) to allow all prices to be expressed in 2014 dollars. As an example, a US dollar in 1998 would be worth $1.46 in 2014 equivalents.

> *Helpful Hint: Chapter 14 of Utts (2005) (reading the economic news) provides an excellent introduction to price indexing and inflation.*

3.3 House Characteristics

Home characteristics and amenities were obtained through the generous collaboration of Craig Della Penna through the Multiple Listing Service for homes on the market in 2007. This included which zip code (01060 or 01062), household area (in thousands of square feet), number of acres, number of garages, number of bedrooms, and number of full bathrooms. The number of garages was recoded to a dichotomous variable (having any garage versus no garage) while the number of bedrooms was categorized as 1-2, 3, or 4+ bedrooms.

> *Alternative Application: Variables such as the number of bedrooms or the number of bathrooms could be considered either categorical or quantitative. Some discussion of modeling tradeoffs may be appropriate to discuss.*

3.4 Distance from Rail Trail

The “My Places” feature of Google Maps was used to calculate the distance from each house to the rail trail. This application allows users to trace lines on a map from one location (a home) to another location (a rail trail entrance) and to identify the length of the line in feet. Following the shortest combination of roads and sidewalks from each home to the nearest trail-entrance point, we collected data on the proximity to the trail. Figure 5 displays an image showing the distance calculation for two of the homes in our study using Google’s “My Places” application, where the blue line indicates the shortest link on roads and sidewalks, with the green line indicating the trail.

This dataset can be used to assess whether homes closer to the original Northampton rail trail appreciated more between 1998 to 2014 relative to more distant houses for the set of all houses in Northampton which sold in 2007 (before the most recent “Great Recession”). This task was greatly facilitated due to the emergence of a web-based repository of

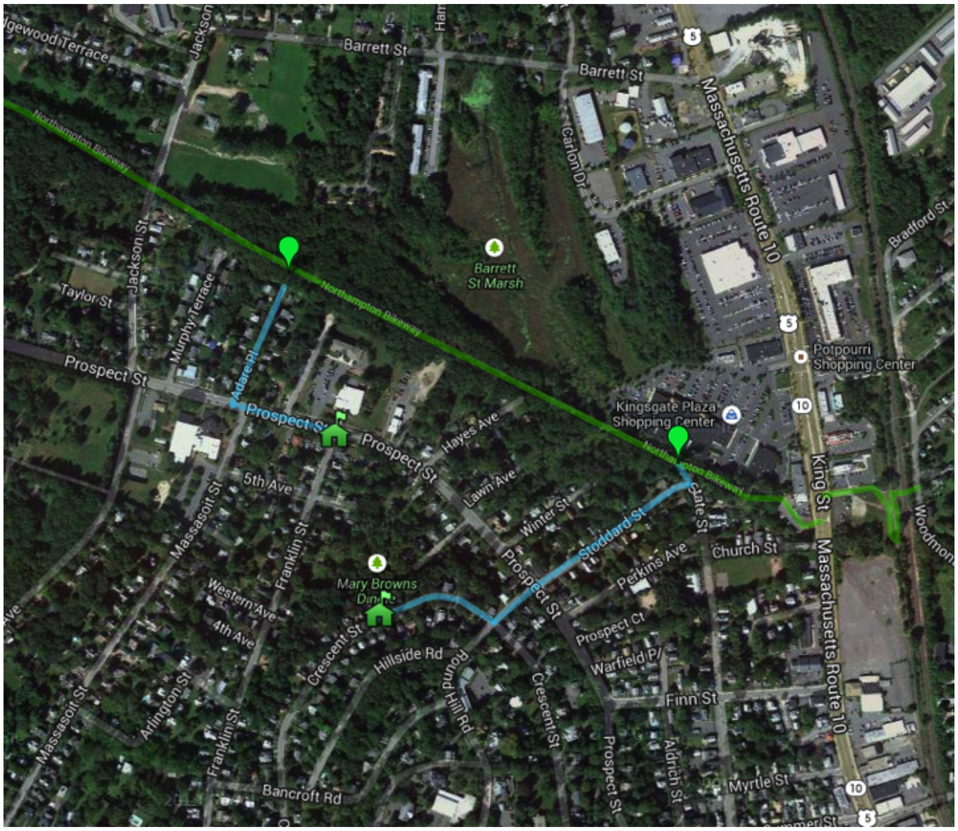


Figure 5: Two houses in our study are linked to the Northampton rail trail by the shortest route, (traced in blue), to two trail access points represented by green flags. Google indicates the length of these lines in feet in a sidebar.

information about residential properties (including home characteristics, sale history and "Zestimates" of house prices) throughout the US.

We divided our sample into two groups based on their proximity to the trail, using an arbitrary but useful metric to describe an average resident's ability to access the paths by bike or on foot. Our first group ($n = 404$) was comprised of houses within a half mile of the trail, equivalent to a walk of 10 minutes at 3 mph. We called this group "Closer." The second group consisted of any house further away than one half mile (equivalent to more than a ten-minute walk, $n = 64$). We considered the residents of these houses as less able to easily access the trails (this group is called "Farther away").

3.5 Neighborhood Characteristics

Measures of walkability and bikeability were scraped from the website http://www.walkscore.com. These scores range from 0–100 with higher scores being better. A walkscore in the

90's is described as a "Walker's paradise: daily errands do not require a car", while a bikescore in the 90's is described as "Biker's paradise: flat as a pancake, excellent bike lanes."

3.6 Creation of Analytic Sample

We restricted our analysis to homes with at most 0.56 acre in total property size, since this was the largest lot size for homes near the rail trail (ten homes were dropped). In addition, two homes were eliminated from the analyses because they appreciated more than $500,000 between 1998 and 2007 and were identified respectively as a "massive fixer-upper" and a "major league make over" (personal communication, realtor Craig Della Penna). This left us with an analytic dataset of $n = 104$ homes.

> *Helpful Hint: The identification and handling of outliers is an important topic for any statistical analysis.*

The following files can be downloaded from the *JSE* website:

- The original dataset,
  http://www.amstat.org/publications/jse/v23n2/horton/dataset_in_wide_format.csv;
- tall dataset,
  http://www.amstat.org/publications/jse/v23n2/horton/dataset_in_tall_format.csv;
- codebook,
  http://www.amstat.org/publications/jse/v23n2/horton/documentation.docx;
- and R Markdown code used to generate figures and run analyses,
  http://www.amstat.org/publications/pubdump/railtrails.zip.

The code in R (R Core Team 2015) leverages routines from the `mosaic` package to facilitate the teaching of statistics (Pruim et al. 2015), functions from the `dplyr` (Wickham and Francois 2015; Horton et al. in press) and `tidyr` (Wickham 2014) packages to undertake data manipulation, mapping routines from the `ggmap` package (Kahle and Wickham 2015) as well as the `knitr` package for reproducible analysis (Xie 2015). More information on the use of R Markdown in introductory statistics can be found in (Baumer et al. 2014).

4. Analyses

Here we undertake a series of analyses using these data, including exploratory analysis of single variables and bivariate relationships, mapping, unadjusted and adjusted comparisons of price changes by group, data transformations, and more sophisticated approaches using repeated measures models.

4.1 Exploratory Analysis

Figure 6 displays a plot matrix of 2007 price and home characteristics (including distance from the rail trail). The diagonal entries of Figure 6 provide density plots of a subset of the study variables (for continuous measures including price [in thousands of dollars], acreage, the actual number of bedrooms, and square footage [in thousands of feet]) and barcharts for the categorical variables (distance group). The lower triangular entries of this figure provide scatterplots for combinations of continuous measures or stacked dotplots for combinations involving categorical variables. The upper triangular entries display correlations (for combinations of continuous measures) or side by side boxplots (for combinations of continuous and categorical variables).

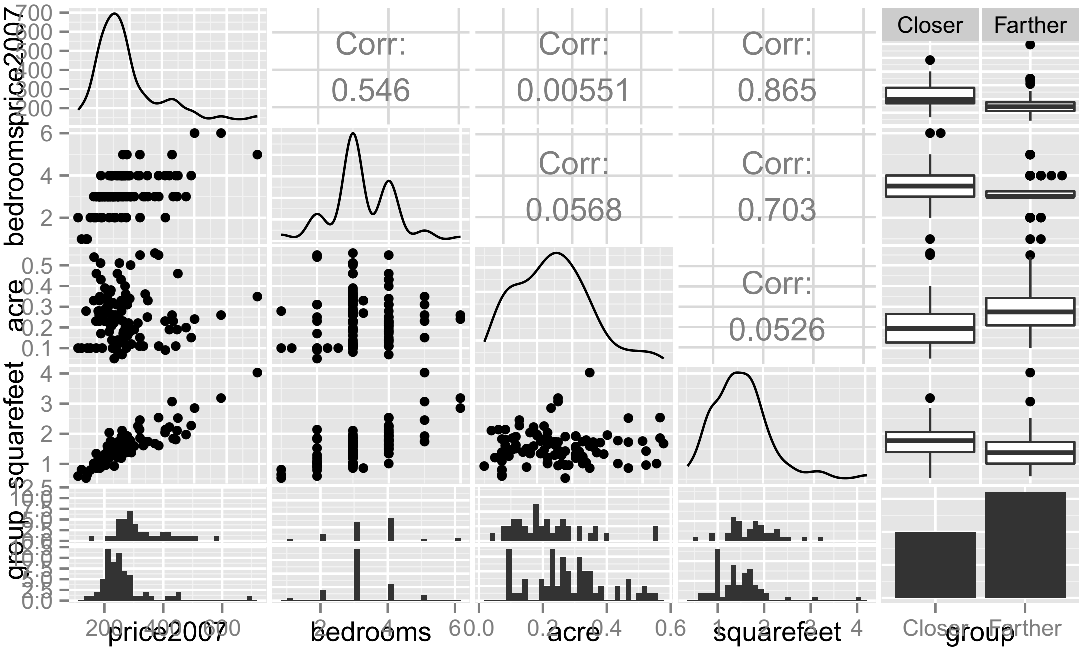


Figure 6: Plot matrix of study variables. The diagonal displays the univariate distribution of each variable, while the off-diagonals provide bivariate relationships.

The distribution of price and square footage are all skewed with long tails to the right. There is a strong correlation between price and square footage, as well as an indication that the number of houses with 1, 5, or 6 bedrooms is sparse [which led to its use as a categorical variable with levels 1-2, 3, 4+]. Price appears to be a nonlinear function of the number of bedrooms and that the number of bedrooms is positively associated with the square footage. In terms of the distance from the rail trail, houses closer to the rail trail tended to be more expensive in 2007, have less acreage, have more bedrooms and be larger.

> *Alternative Application: The* `squarefeet` *variable is skewed with a long right tail. It might be worthwhile to consider other ways to incorporate this into a regression (e.g., transformation or categorization).*

Figure 7 displays a plot matrix of change in price from 1998 to 2014, price in 1998, plus neighborhood characteristics and distance from the rail trail.

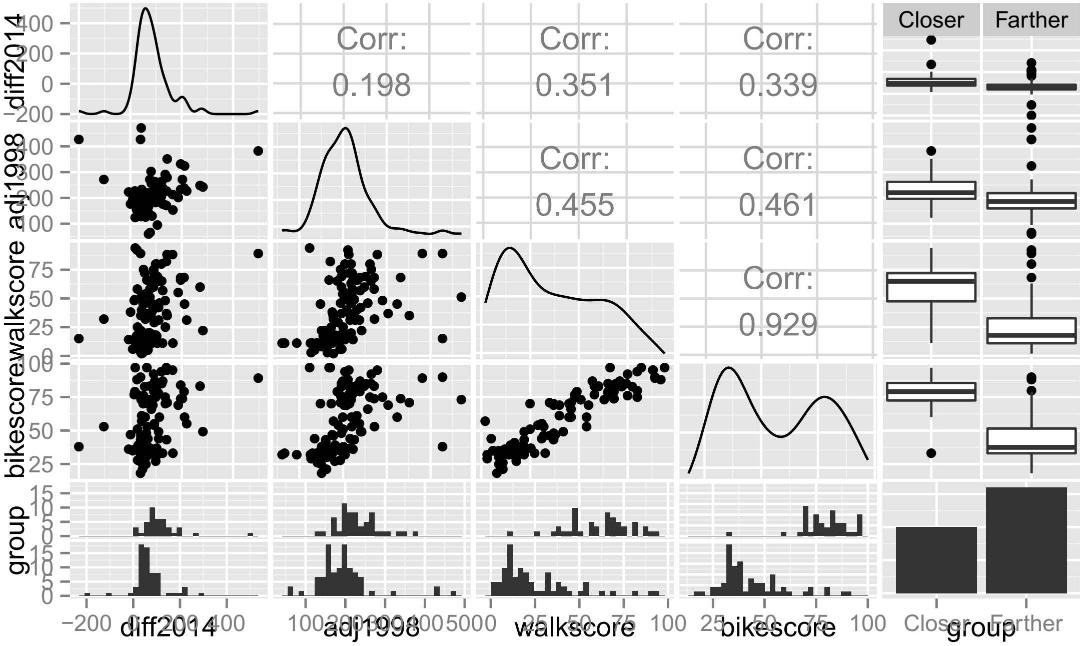


Figure 7: Plot matrix of other variables. The diagonal displays the univariate distribution of each variable, while the off-diagonals provide bivariate relationships.

The distribution of differences in prices and adjusted 1998 price [both in thousands of dollars] have heavy tails, but are roughly symmetric. The distribution of walkscores is almost triangular (many with low scores), while the bike scores are somewhat bimodal (with almost no overlap between the bike scores by distance group). There is a strong correlation between the two measures of walkability and bikeability, with positive correlations between price and score.

It's straightforward to calculate summary statistics by group as well as generate other graphical representations. As an example, consider the following distribution of price change (in thousands of dollars) from 1998 to 2014 by distance group. Table 1 displays a variety of summary statistics by group, while Figure 8 displays overlaid density plots.

Table 1: Summary statistics of difference in adjusted price from 1998 to 2014 (in thousands of dollars) by distance group

	Group	min	Q1	median	Q3	max	mean	sd	n	missing
1	Closer	15.42	75.42	94.50	136.47	497.82	114.02	81.87	40	0
2	Farther Away	-199.87	40.38	59.35	83.72	282.47	66.10	67.37	64	0

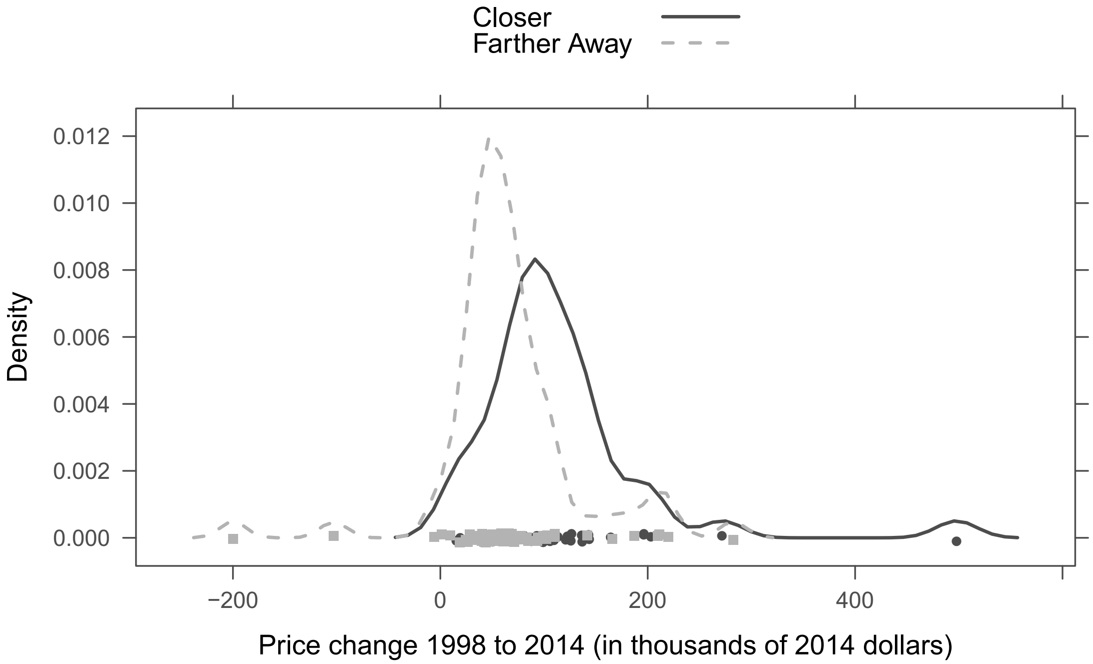


Figure 8: Density plot of price change from 1998 to 2014 (in thousands of 2014 dollars).

We note a number of outlying values, which can and should be identified. Table 2 displays those with values less than -190 and more than 250.

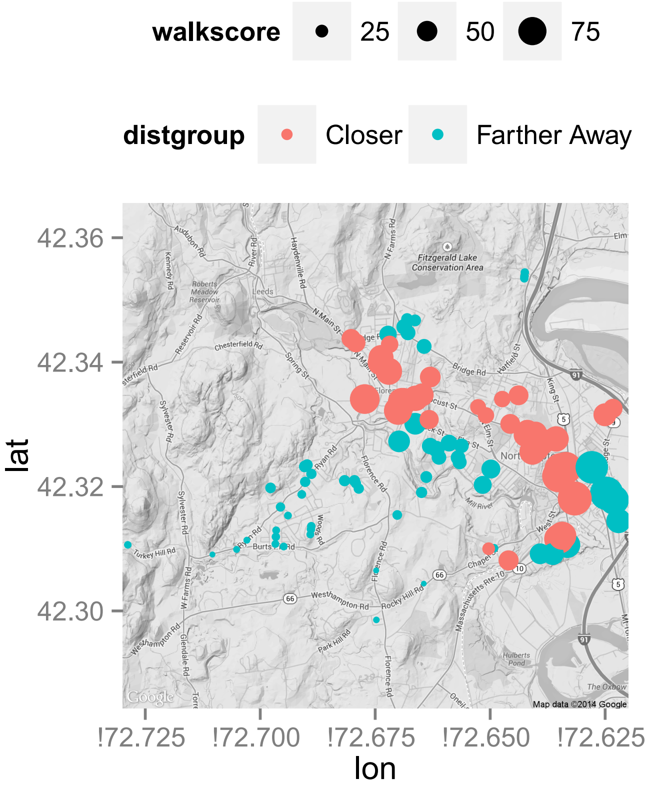


Figure 9: Map of walking score by location (with different colors for distance from rail trail)

4.2 Mapping

Spatial relationships can be very important, and it is straightforward to incorporate the geolocations (latitude and longitude) of the houses to try to discern patterns.

> *Helpful Hint: Nicolas Christou of UCLA has created materials to facilitate the teaching of spatial data analysis for introductory statistics courses: see* http://www.stat.ucla.edu/~spatial *for details.*

Table 2: Outlying observations of price changes (in thousands of dollars)

	streetno	streetname	price1998	adj1998	price2014	diff2014
1	497	Burts Pit Rd	292.50	427.55	227.68	-199.87
2	248	Prospect Street	170.00	248.49	520.15	271.66
3	14	Liberty Street	166.00	242.64	525.11	282.47
4	40	Trumbull Road	261.00	381.50	879.33	497.82

Figure 9 displays the distribution of walking scores (arranged by size) by location (with different colors for the groups denoting distance from the rail trail). The houses that are closer tend to have higher walking scores.

4.3 Comparisons of Houses by Distance

Our goal was to create groups of houses with similar characteristics that influence price thereby trying to reconstruct a randomized comparison. We undertook this adjustment by controlling for potential confounding factors (number of bedrooms, acreage and square footage) using multiple regression modeling.

As demonstrated in Figures 6 and 7, the houses in the two distance groups were not equivalent at the start of the study.

> *Helpful Hint: One of the most important conditions for drawing causal conclusions from a statistical study is randomization of subjects to treatment groups. Obviously, houses were not randomized to be near or far from the rail trail. When comparing houses close to the trail to houses farther away, there may be other factors that differ between the houses beyond proximity to the rail trail that account for the differences in price. The potential for confounding is an important topic to reinforce for students (see Kaplan (2012) or chapters 4–8 of Vittinghoff et al. (2012) for a comprehensive and readable introduction for instructors).*

4.4 Comparing Distance Groups

Table 3 compares the distribution of the house characteristics. Using the Wilcoxon rank sum test test (for continuous characteristics: number of acres and square feet) and Fisher's exact test (for categorical characteristics: number of bathrooms, bedrooms and garage spaces) we see that there are significant differences between the distance groups in terms of the number of bedrooms ($p = 0.024$), existing of a garage ($p = 0.028$), acreage ($p = 0.005$), square footage ($p = 0.0006$), and location ($p = 0.0009$). Homes near the rail tend to have more bedrooms, a garage, less acreage, a larger interior, and be in Northampton.

Table 3: Comparison of house characteristics. Fisher's exact test used for comparisons of categorical variables (number of bathrooms, bedrooms and existence of garage) while the Wilcoxon rank sum test test was used for continuous variables (number of acres and square feet).

House characteristic	Closer (n=40)	Farther away (n=64)	p-value
Number of full bathrooms (0-4)	$\bar{X}$ = 1.55	$\bar{X}$ = 1.39	$p = 0.51$
Percent bedrooms (1-2/3/4+)	15%/35%/50%	16%/59%/25%	$p = 0.024$
Percent with garage	65%	42%	$p = 0.028$
Acreage	$\bar{X}$ = 0.22	$\bar{X}$ = 0.28	$p = 0.005$
Household area (sf)	$\bar{X}$ = 1,755	$\bar{X}$ = 1,449	$p = 0.0006$
Percentage zip 01060	63%	28%	$p = 0.0009$

4.5 Multiple Regression Modeling

Our goal was to create groups of houses with similar characteristics that influence price thereby trying to reconstruct the hypothetical randomized comparison. A naive approach would be to simply compare house prices in 2014. We can fit a simple regression model (equivalent to an equal variance two sample t-test). Table 4 displays the results from this naive model.

Table 4: Unadjusted comparison of difference in 2014 price by distance group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	114.018	11.583	9.84	0.0000
distgroupFarther Away	-47.922	14.765	-3.25	0.0016

We observe that homes that are farther away tend to appreciate about $4,792 less than those closer to the rail trail. This result is statistically significant ($p = 0.0016$). A major limitation of this approach is that it assumes that the houses were comparable with the exception of the distance from the rail trail.

We can fit a more complex regression model that also controls for baseline price. Table 5 displays the revised results.

Table 5: Comparison of difference in 2014 price by distance group (adjusted for baseline price

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	80.188	28.251	2.84	0.0055
adj1998	0.147	0.112	1.31	0.1925
distgroupFarther Away	-42.759	15.230	-2.81	0.0060

We observe that after controlling for baseline price, homes that are farther away tend to appreciate about $4,276 less than those closer to the rail trail. This result is statistically significant ($p = 0.006$) but somewhat attenuated.

We fit a more complex regression model that controls for baseline price and other house characteristics that were observed to be different between the groups in Table 3. Table 6 displays the revised results.

Table 6: Comparison of difference in 2014 price by distance group (adjusted for baseline price and house characteristics)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14731.398	9025.596	1.63	0.1060
adj1998	-0.582	0.161	-3.62	0.0005
bedgroup3 beds	1.140	19.060	0.06	0.9524
bedgroup4+ beds	-14.427	22.888	-0.63	0.5300
garagegroupyes	17.564	14.440	1.22	0.2269
acre	-45.742	63.249	-0.72	0.4713
squarefeet	110.938	21.885	5.07	0.0000
zip	-13.832	8.510	-1.63	0.1074
distgroupFarther Away	-21.886	14.643	-1.49	0.1383

We observe that after controlling for baseline price and home characteristics, homes that are farther away tend to appreciate about $2,189 less than those closer to the rail trail, but that this result is not statistically significant ($p = 0.14$).

> *Alternative Application: the percentage change in adjusted price from 1998 to 2014 could be modeled. The variable* `pctchange` *has been created in this manner. Table 7 displays the results from this model.*

Table 7: Comparison of percent change in adjusted price by distance group (adjusted for baseline price)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	58.842	9.691	6.07	0.0000
adj1998	-0.080	0.044	-1.80	0.0746

Regression diagnostics are always important to undertake whenever a model is fit. It is straightforward to assess the normality of residuals from this linear model (see Figure 10). The assumption of normality is suspect here (with many extreme residuals).

4.6 Translating the Dataset from Wide to Tall Format

While the dataset is provided in two formats, all analyses to date have used the “wide” format with one row per house. As an example, let’s consider the data from 40 Trumbull Road (house number 97), which is displayed in Table 8.

```
> mplot(lm4, which=2)
[[1]]
```

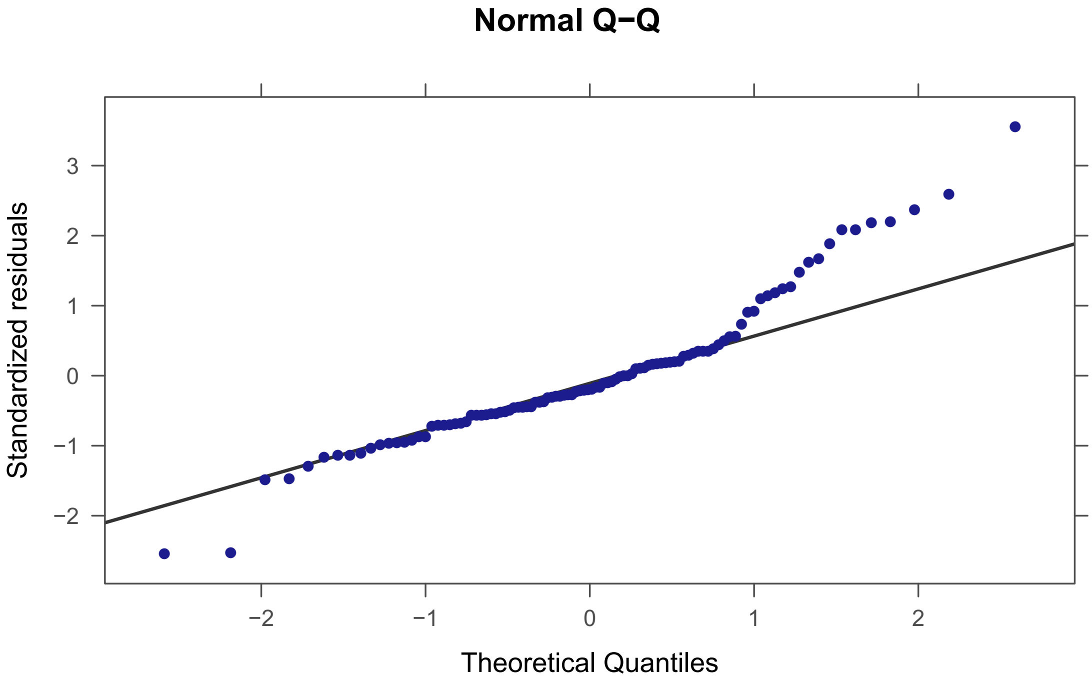


Figure 10: Distribution of residuals for alternative application model.

Table 8: Listing of a subset of information for 40 Trumbull Road in wide format

	streetno	streetname	adj1998	adj2007	adj2011	price2014	acre	distgroup
1	40	Trumbull Road	381.50	669.93	698.54	879.33	0.26	Closer

Some analyses require a structure with one observation per time period. We call this format “tall” since there will be four times as many rows. Table 9 displays the data for the 40 Trumbull home in this format.

Table 9: Listing of a subset of information for 40 Trumbull Road in tall format

	housenum	streetno	streetname	year	price	acre	distgroup
1	97	40	Trumbull Road	1998	381.50	0.26	Closer
2	97	40	Trumbull Road	2007	669.93	0.26	Closer
3	97	40	Trumbull Road	2011	698.54	0.26	Closer
4	97	40	Trumbull Road	2014	879.33	0.26	Closer

Having the data in tall format facilitates some plots. For example, we can display boxplots of adjusted price over time while stratified by distance from the rail trail and grouping of square footage (see Figure 11). We observe that the price increases tend to be larger for larger homes.

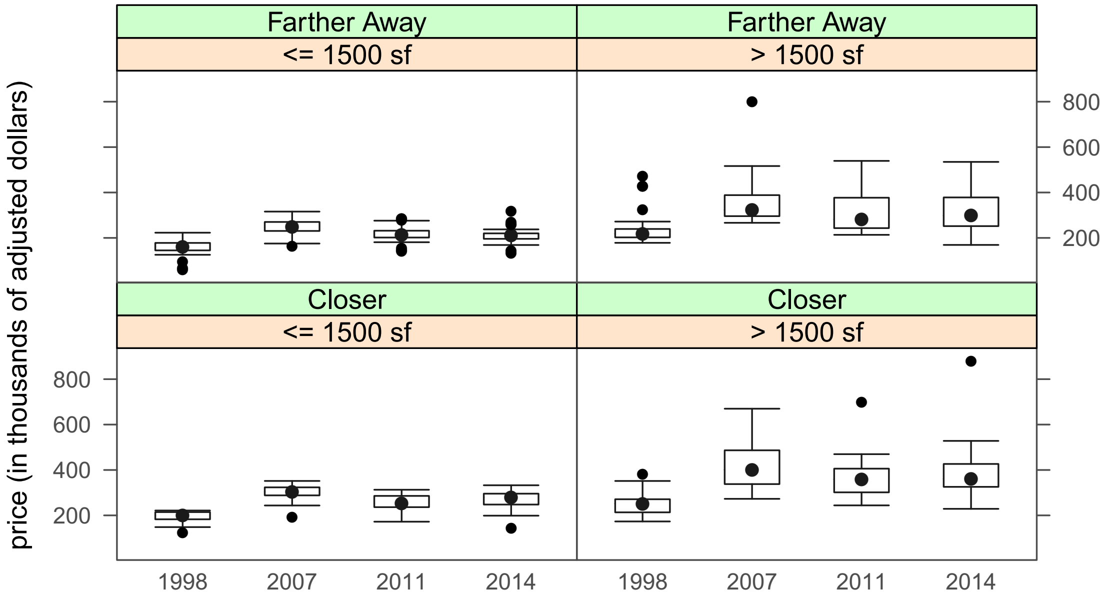


Figure 11: Distribution of adjusted price over time by distance from rail trail and grouping of square footage (less than 1500 square feet versus greater than or equal to 1500 square feet).

4.7 Repeated Measures Regression Modeling

A disadvantage of the previous analyses is that they do not incorporate all of the available price data. Because each house was measured at four different time points, it is not reasonable to assume that each price is an independent measurement. To account for these clustered observations, we used a generalized linear model for correlated data (Fitzmaurice et al. 2004). This model accounts for repeated measures on each house by estimating a 4 × 4 covariance matrix. If the mean model is correctly specified, the model yields unbiased estimates of the fixed effects parameters and their standard errors.

Helpful Hint: instructors in introductory courses (that generally do not cover repeated measures models) may choose to utilize our earlier approach where the increase in prices over time were assessed by picking two time points and controlling for the initial value, or just modeling the difference in prices, rather than use of the more complicated repeated measures model.

Alternative Application: A random effects model could also be fit (and is illus-

> *trated using the R Markdown file). Results are similar using this alternative approach.*

The model for the mean sales price includes the following variables: the number of bedrooms, acreage, square feet of the house, distance and year. We included house characteristics that significantly differed between distance groups in 2007. Because we hypothesized that the difference in prices by location might vary by year we included this interaction term. The interaction between time and distance was statistically significant (F(3, 402)=5.32, $p = 0.0013$), so the interaction was retained (see Table 10).

Table 10: Comparison of prices over time by distance group (adjusted for house characteristics) using repeated measures regression.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.5	15.4	2.76	0.006
zip01062	-12.2	9.9	-1.23	0.22
bedgroup3 beds	14.2	11.0	1.30	0.20
bedgroup4+ beds	3.5	13.3	0.26	0.79
garageYes	-4.9	8.3	-0.59	0.56
acre	41.1	36.4	1.13	0.26
squarefeet	102.4	9.0	11.4	< 0.0001
year2007	144.6	8.6	16.8	< 0.0001
year2011	96.0	8.8	10.9	< 0.0001
year2014	114.0	11.6	9.8	< 0.0001
distgroupFarther Away	-5.9	9.1	-0.65	0.52
Farther*2007	-41.7	10.9	-3.81	0.0002
Farther*2011	-32.6	11.3	-2.89	0.004
Farther*2014	-47.9	14.8	-3.25	0.001

We note that at the first observation time, there was not a statistically significant difference in prices ($p = 0.52$) but that the predicted difference in prices in 2014 between those closer away and farther away is $5,900 + $47,900 = $53,800. Use of the repeated measures has allowed us to more efficiently model these data.

5. Conclusion

This dataset can be used at many levels in the statistics curriculum, beginning with exploratory data analysis and informal inference. Further extensions allow use in data manipulation, mapping, and formal inference. A question of interest is whether it is possible to assess changes in house prices relative to their distance from rail trails, after accounting for other measured factors. Our analysis found that homes within one half mile of a well-established rail trail in Northampton, Massachusetts tended to appreciate more (or retain a higher value) during the period from 1998–2014 than homes farther than one half mile away from the original 1984 trail.

Other analyses are possible using this dataset, taking advantage of its relatively rich characteristics. These include assessing relationships between house characteristics, sales, and neighborhood characteristics (such as walkability).

Potential pitfall: It's always important to ensure that limitations of an analysis are clearly spelled out. Students should be encouraged to brainstorm possible issues. A starting point of possibilities is enumerated here: (1) We considered a small subset of houses out of the total number of houses in Northampton. These were the houses sold within a single community in a particular year; (2) The construction of additional rail trails in recent years means that many homes are now closer to the expanding network but were categorized in the "farther away" category in our analyses; (3) We used an arbitrary coding of distance from the rail trail; (4) Data scraping using Zillow may have non-negligible measurement error; (5) There may also be other unmeasured confounding factors of these houses that could potentially account for these results.

Acknowledgments

This work was supported by NSF grant 0920350 (Phase II: Building a Community around Modeling, Statistics, Computation, and Calculus). Craig Della Penna provided invaluable guidance and assistance, while Molly Johnson contributed to the data collection and analysis. We are also appreciative to the Editor, Associate Editor and anonymous reviewers for a number of suggestions which led to improvements in the manuscript.

REFERENCES

American Trails 2011, "Benefits of Trails and Greenways", Technical report. Available at http://www.americantrails.org/resources/benefits/homebuyers02.html, accessed March 8, 2015.

Baumer, B., Çetinkaya Rundel, M., Bray, A., Loi, L., and Horton, N. 2014, "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics," *Technology Innovations in Statistics Education* **8**(1). Available at http://escholarship.org/uc/item/90b2f5xh, accessed March 8, 2015.

Centers for Disease Control and Prevention 2010, "Vital Signs: State-Specific Obesity

Prevalence Among Adults—United States", 2009, *Morbidity and Mortality Weekly Report*. Available at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm59e0803a1.htm, accessed March 8, 2015.

Eppley Institute for Parks & Public Lands 2001, "Summary Report, Indiana Trails Study", Technical report. Available at http://www.in.gov/indot/files/z-CompleteDocument.pdf, accessed March 8, 2015.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. 2004, *Applied Longitudinal Analysis*, New York: Wiley.

Horton, N. J., Baumer, B. and Wickham, H. in press, "Setting the Stage for Data Science: Integration of Data Management Skills in Introductory and Second Courses in Statistics", *CHANCE* 28(2): 40–50. http://arxiv.org/abs/1502.00318, last accessed March 8, 2015.

Kahle, D. and Wickham, H. 2015, "ggmap: A Package for Spatial Visualization with Google Maps and OpenStreetMap," R package version 2.4. Available at http://CRAN.R-project.org/package=ggmap, last accessed March 8, 2015

Kaplan, D. 2012, *Statistical Modeling: A Fresh Approach* (2nd edition), http://www.mosaic-web.org/go/StatisticalModeling, accessed March 8, 2015.

Karedeniz, D. 2008, "The Impact of the Little Miami Scenic Trail on Single Family Residential Property Values", Technical report. Available at http://www.americantrails.org/resources/economics/littlemiamipropvalue.html, accessed March 8, 2015

Lindsey, G., Man, J., Payton, S., and Dickson, K. 2004, "Property Values, Recreation Values, and Urban Greenways", *Journal of Park and Recreation Administration*, 22(3), 69–90.

Pruim, R., Kaplan, D., and Horton, N.J. 2015, "mosaic: Project MOSAIC (mosaic-web.org) Statistics and Mathematics Teaching Utilities". R package version 0.9-2-2. Available at http://CRAN.R-project.org/package=mosaic, last accessed March 8, 2015

R Core Team 2015, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org, last accessed March 8, 2015.

Rails to Trails Conservancy 2015, "Rails-to-Trails in the Making," Technical Report. Available at http://www.railstotrails.org/about/history, accessed March 8, 2015.

Utts, J. 2005, *Seeing Through Statistics* (3rd edition), Cengage Learning.

Vittinghoff, E., Glidden, D., Shiboski, S., and McCulloch, C. 2012, *Regression Methods in Biostatistics* (2nd edition), New York: Springer.

Wickham, H. 2014, "Tidy Data," *Journal of Statistical Software* **59**(10). Available at http://www.jstatsoft.org/v59/i10/, last accessed March 8, 2015.

Wickham, H., and Francois, R. 2015, "dplyr: A Grammar of Data Manipulation," R package version 0.4.1. Available at http://cran.r-project.org/web/packages/dplyr, last accessed March 8, 2015.

Xie, Y. 2015, "knitr: A General-Purpose Package for Dynamic Report Generation in R," R package version 1.9. Available at http://CRAN.R-project.org/package=knitr, accessed March 8, 2015.

Ella Hartenian
Department of Biological Sciences
Smith College, Northampton, MA

Nicholas J. Horton
Dept of Mathematics and Statistics
Amherst College
AC#2239, PO Box 5000
Amherst, MA 01002-5000