Analyzing City Tax Data *with* Data Science Toolkit

Bill Krekeler

St Louis RUG 2017.12.14

Who, What, Why

- Hypothesis: *Unjustified* Property Tax Increases

- How to prove when homes only really have defined value when sold?
  - Zillow and Realtor.com estimates seem unrealistic
  - Realtor.com uses comparable properties miles from my house
  - City residents know sales $ varies by zip code, neighborhood, street, block, & whim

- Talk focuses more on basic how than the analysis/insight itself
  - Data
  - Tools

Tools/What is needed

- The data is in what? Mdb-setup or other access db export utility
- Data Science Toolkit
  - Where: http://www.datasciencetoolkit.org/
  - Support: https://github.com/petewarden/dstk
  - Why: Zipcode translation & Geolocation
- Vagrant + VirtualBox
- R

```
install.packages('RDSTK')
library(RDSTK)
```

RDSTK-package {RDSTK} R Documentation

RDSTK: A R wrapper for the Data Science Toolkit API

Description

This package contains several functions that provide direct access to the Data Science Toolkit API. See www.datasciencetoolkit.org for an overview of the API. The package is an attempt to R-ify calls to this API.

By default the packages accesses the API at www.datasciencetoolkit.org. Alternatively, because it is possible to clone the DSTK service on a local machine, you can point the package to an alternate API using `options("RDSTK_api_base"="http://localhost:8080")`.

DSTK Setup – Web vs Local VMs

- If http://www.datasciencetoolkit.org/ inaccessible how local?
  1) AWS EC2
  2) PreciseBox → update error apt-add-repository not present; old ubuntu version
  3) Roll your own issue with Ruby and other versions colliding
  4) Find a valid pre-built box
     - 0.51 file from *braddd* post in https:\\github.com\petewarden\dstk\issues\57
     - Alt: https:\\vagrantcloud.com\dmofot\boxes\dstk\versions\0.51\providers\virtualbox.box
- Local VM = independent, secure, scalable
- https://github.com/petewarden/dstk

DSTK Setup – Local VMs

- Vagrant via URL

```
$ vagrant init dmofot/dstk          # this downloads vagrant box and creates the Vagrantfile
```

- Vagrant via local file

```
$ wget -0 dstk_0.51.box https:\\vagrantcloud.com\dmofot\boxes\dstk\versions\0.51\providers\virtualbox.box
$ vagrant box add dstk dstk_0.51.box
$ vagrant init    # this create Vagrantfile
```

- Manual configuration:
  - edit Vagrantfile to uncomment lines 25 and 29 about network access

```
config.vm.network "forwarded_port", guest: 80, host: 8080
config.vm.network "private_network", ip: "192.168.33.10"
```

- Use the VM

```
$ vagrant up            # this starts VM server
$ vagrant ssh           # to connect directly in terminal
$ vagrant halt          # to stop (or could “su; shutdown -h now” inside the ssh terminal: username=vagrant; pswd=vagrant
```

DSTK Setup – Local VMs – Vagrant Mess

- DSTK Boxes are 20+ GB, exporting to Virtual Box 60+ GB
  - Copy made for vagrant (x2 if use vagrant to download)
  - Export made for VirtualBox (not explicitly necessary)
- Where did it put the files? [ref: https://stackoverflow.com/questions/10155708/where-does-vagrant-download-its-box-files-to]
  - during download: ~/.vagrant.d/tmp/boxXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  - post download boxes in <user_path>/.vagrant.d/boxes
- Vagrant Cleanup [ref: https://superuser.com/questions/1113532/vagrant-remove-one-of-many-boxes-with-different-versions]

```
$ vagrant box list
 data-science-toolbox/dst (virtualbox, 0.2.1)
 dstk                     (virtualbox, 0)
 dstk_0.51.box            (virtualbox, 0)
 precise64                (virtualbox, 0)
$ vagrant box remove precise64 --box-version 0
$ vagrant box remove dstk --box-version 0
```

DSTK Interface

- Insert *fancy* dynamic demo here of web service and RSDTK
- http:\\localhost:8080

- List of built-in tool references: http://localhost:8080/about

City Property Data

- Zillow
  - https://www.zillow.com/research/data/
  - Data aggregation at too high a level: zipcode, neighborhood

```
$ cd ~/projects/data_housing/zillow_20170813/
$ grep 'a href' ../zillow_research_data.html | grep csv | sed 's/a href=\"/\n/g' | sed 's/\">/\n/g' | grep http | xargs wget
```

- St Louis City
  - The home of many city data sets: https://www.stlouis-mo.gov/data/
  - The real data: http://data.stlouis-mo.gov/downloads.cfm
    - Tax, sales, type data stored in zipped Access Database files *.mdb

```
# example export to text
$ ../../mdb_table_export.sh bincode | /bin/sh
$ ../../mdb_table_export.sh PrclCode | /bin/sh
$ ../../mdb_table_export.sh prclsale | /bin/sh
$ ../../mdb_table_export.sh prcl | /bin/sh
```

```
#!/bin/bash
# script = mdb_table_export.sh
tableName=$1; mdb-tables $tableName.mdb | sed 's/ /\n/g' | perl -ne
'chop($_); if ($_ =~ /[a-zA-Z]/ ) { print "mdb-export '"$tableName"'.mdb $_
> '"$tableName"'_$_.txt;\n"}'
```

Decoder Ring

- Property Lookup: https://www.stlouis-mo.gov/government/departments/assessor/records-mapping/index.cfm
- 5403 Lisette Ave = parcel 60410002700
  - 6041 = block
  - 270 = parcel
  - Note: RDSTK::coordinates2politics(lat,long) indicates Princeton Heights
- Note: house arbitrarily chosen.

Decoder Ring (Partial)

- Note questions marks on assorted annotations

prcl_BldgCom.txt	commerical building type ... lists apartments, types
prcl_BldgResImp.txt	improvement codes by cityblock and parcel
prcl_BldgRes.txt	**residential building type, basement finish, exterior w all type, living area, bathrooms, half baths, ac, attic, garage, year
prcl_BldgSect.txt	is this specific to large buildings? Has building information but not for single family or small multi-family blocks? story heights are > 8 generally
prcl_CxPrclCnBlk10.txt	reference the census block, use w ith census data file (zcta_tract_rel_10.txt) to lookup address zip code
prcl_PrclAddrLRMS.txt	seems similar to prcl_PrclAddr.txt
prcl_PrclAddr.txt	**look up cityBlock, parcel id by address, also use PrimaryAddr flag to determine rental property
	rank = 32 is residential, 49 is commercial or land? 8 is light industry or w arehouse? 53 is w arehouse?
	AddrType = 3 is residential, 2 is commercial, A is non-existant? center of street, R = mixed use apartment restaurant bottom
	RecNum = 1 if single, 2 if double, 3 if triple
	RecNumLRMS generally matches RecNum but may differ if original double now single, vice versa
	NLC ??
	Parity = E and O, meaning unclear
prcl_PrclAsmt.txt	assessed values by: block, parcel, ow ner, class?, land value, building assessed, assessed total=new assessed value, billland=old land value, billimprove = old building assessment, bill total = old assessed total, aprland?, costaprimprove?, asrlanduse?
prcl_PrclAttr.txt	some kind of attributes by block and parcel, attrtype =HPRA and attrnum = 12
prcl_PrclDate.txt	just the database update date
prcl_PrclImp.txt	does not list: unclear, header = CityBlock,Parcel,Ow nerCode,ImpNum,ImpCode,ImpMeasure
prcl_PrclREAR.txt	show s tax values back to 2009, last update is the day someone pays. isprior must indicate w hich taxes are paid in full and closed out
prcl_Prcl.txt	
prclsale_CdCityBlockPartSrc.t	explains update? types: A, C, L, P, U
prclsale_CdSaleType.txt	explains sales code type
prclsale_HistPrclSale.txt	lists sales prices historical by parcel, our house is not listed? w arning it imports dates w ith 'DATE so the dates are not sortable by date, many more fields than prclsale below
prclsale_PrclSale.txt	lists sales prices by parcel, does show our house, also show s LRMS Sale Date column means w hat? Land Resource Management system?

Interesting Plots and Examples

- Insert awesome demo
  - test_city_data.r
  - house_cost_estimate_report2-sales.rmd / html

Insight

- Know your audience