An Analysis of the linkages between Exercise and Property in King County

Pierre Augustamar
Matthew Peters
Derrick Priebe

Main Hypotheses

- Do exercise routes affect property values?
  - Why: To understand whether there is a strong case with regards to property value for community development of more formal exercise routes.

- What is behind what routes people choose for exercise?
  - Why: From a runner’s perspective, try to identify main characteristics that can be identified that would accurately align with chosen routes.

Related Work

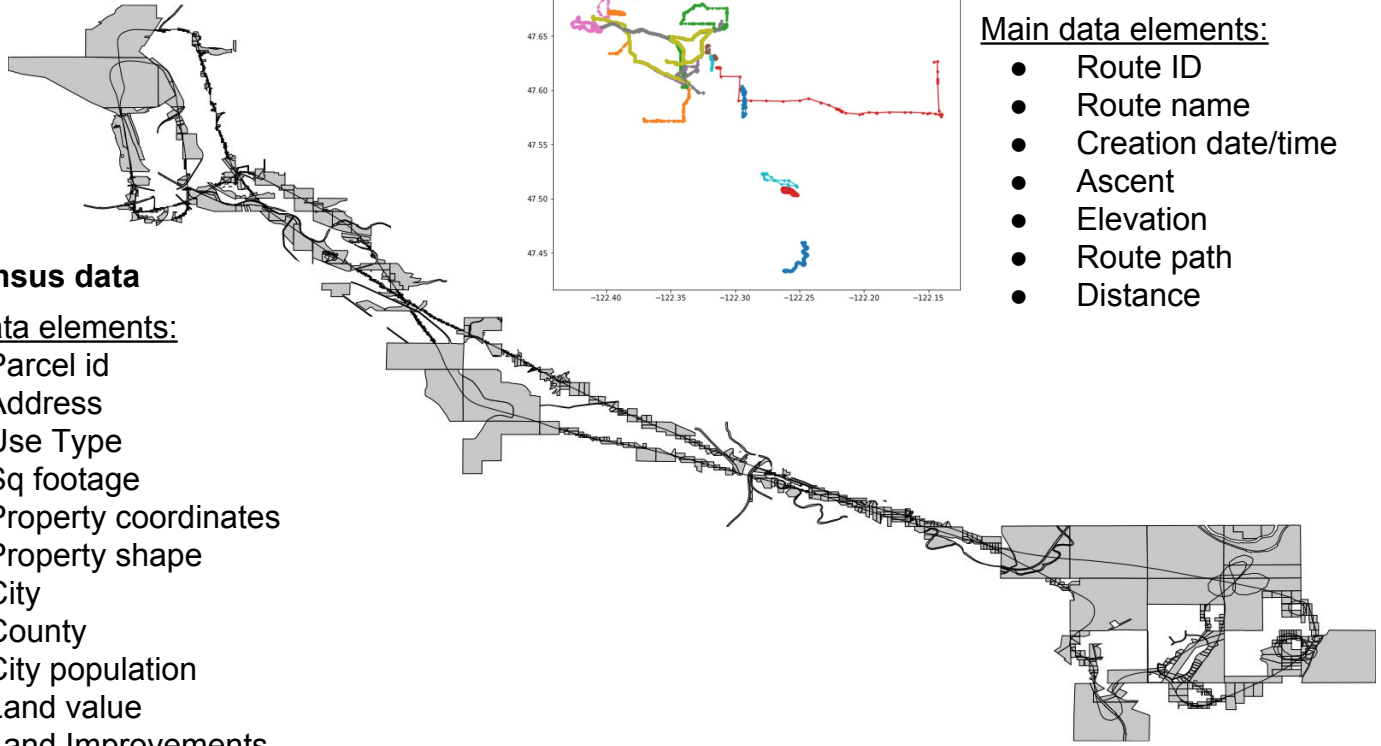
- The **paper "Not Just a Walk in the Park: Methodological Improvements for Determining Environmental Justice Implications of Park Access in New York City of the Promotion of Physical Activity"** takes a focused look at a single city to analyze availability of park to various socioeconomic and ethnic sub-groups. While we have academic interests that are similar, the scope and approach of our study is quite different. This may be a relevant read simply to provide academic and methodological insights, although it is not likely to have anything in common with means of research. As the paper notes, "This study is designed to shed light on the"unpatterned inequities" of park distributions identified in previous studies of New York City park access. (Miyake et al)".

  Source: Miyake KK, Maroko AR, Grady KL, Maantay JA, Arno PS. Not Just a Walk in the Park: Methodological Improvements for Determining Environmental Justice Implications of Park Access in New York City for the Promotion of Physical Activity. *Cities and the environment*. 2010;3(1):1-17.

- A **paper, "City, Culture and Society", by Lawson and Fadare** uses "simple random sampling of household heads" (Lawson, Fadare) in three distinct neighborhoods of Eti Osa, Nigeria. Here, our methodologies and score vary greatly. As noted in the review: "This paper considers the effects of socio-economic status as a determinant of urban health outcomes. Issues examined include housing and environmental conditions as well as socio-economic characteristics such as age, gender, income and household size. Furthermore, health seeking behaviour was investigated and these include expenditure on health as well as health and nutritional habits."

  Source: Taibat Lawanson, Samson Fadare, 2014, Elsevier Ltd, City, Culture and Society, Volume 6, Issue 1, Pages 43-52, *Environment and health disparities in urban communities: Focus on Eti Osa, Nigeria*

Data Sources



MapMyRun

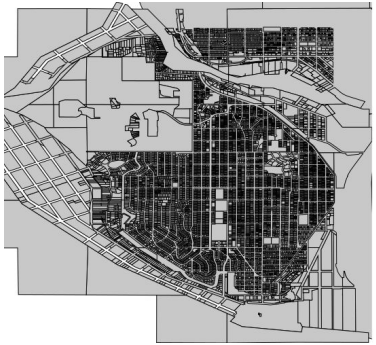
API for exercise data

Main data elements:

- Route ID
- Route name
- Creation date/time
- Ascent
- Elevation
- Route path
- Distance

King County

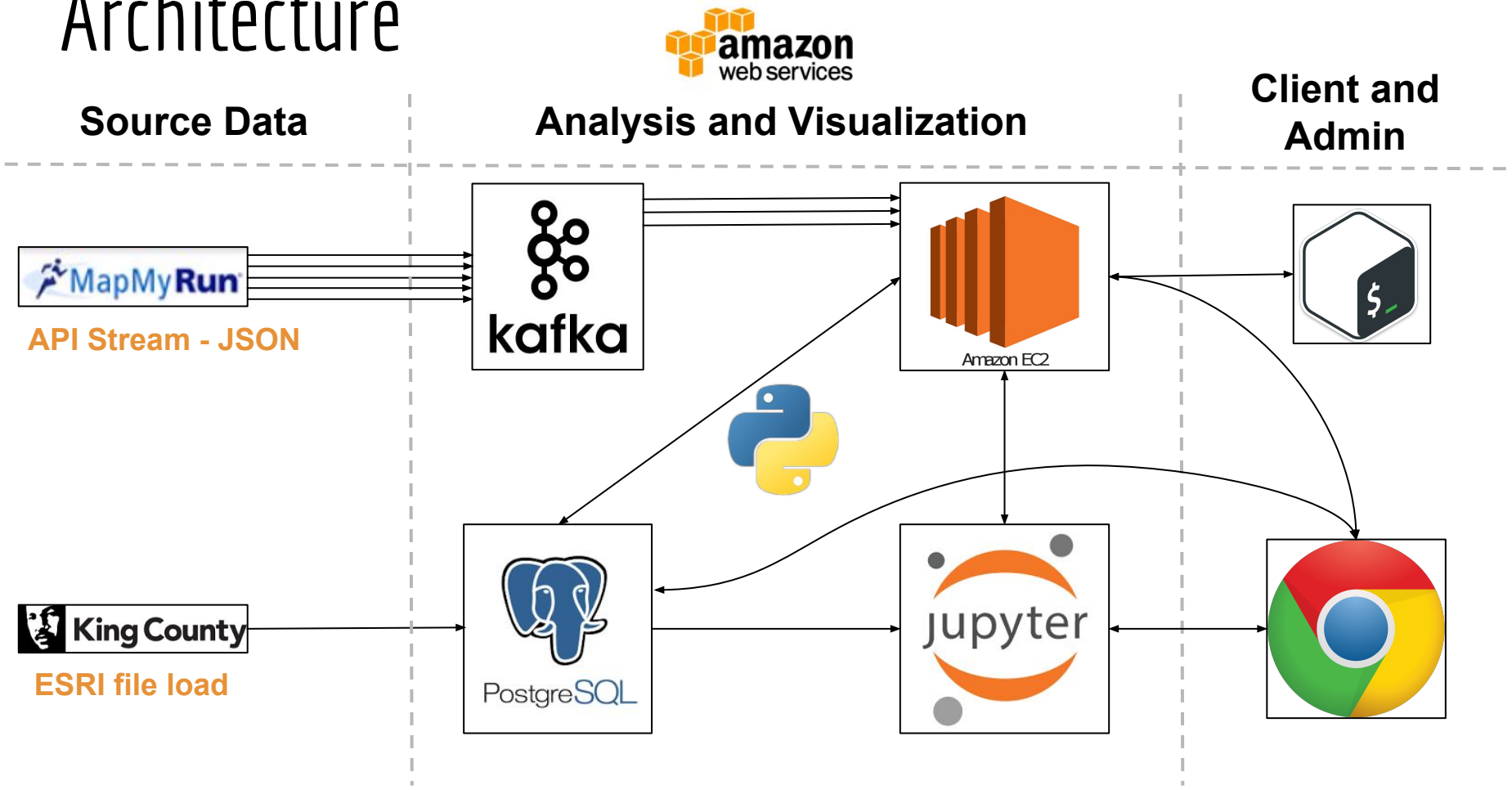
GIS Property Information & census data



Main data elements:

- Parcel id
- Address
- Use Type
- Sq footage
- Property coordinates
- Property shape
- City
- County
- City population
- Land value
- Land Improvements

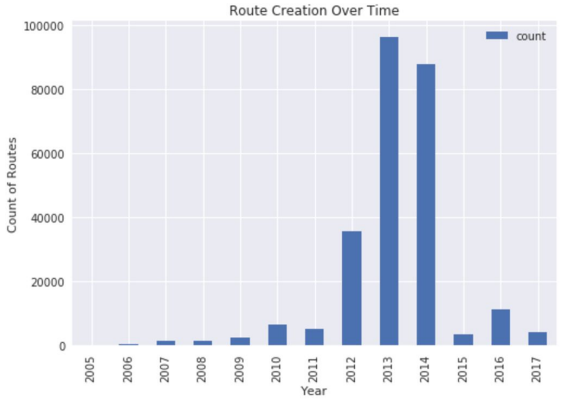
Architecture

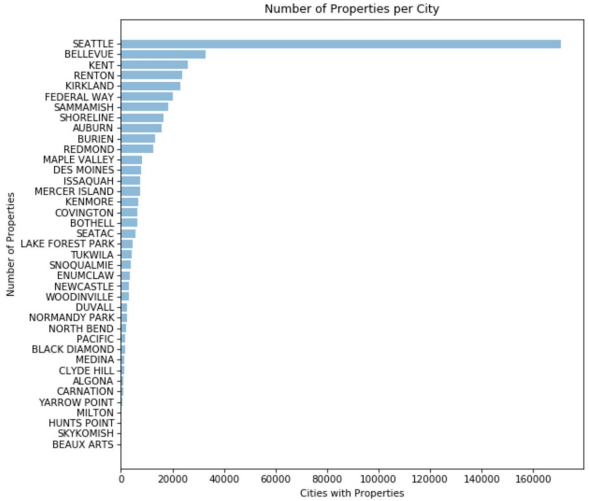


General Data Observations

- Route/Exercise Data Notes:
  - Inclusive of King County only
  - Accessed MapMyRun API via JSON in May 2017 over 2 week period
  - Not fully representative of a complete dataset (stopped to complete project)
  - Route creation covers years 2005 - 2017

- Property Data Notes:
  - Property data inclusive of King County only
  - Properties exist in 40 cities across the county
  - Accessed data from ESRI files from King County website
  - Includes all properties recognized by King County



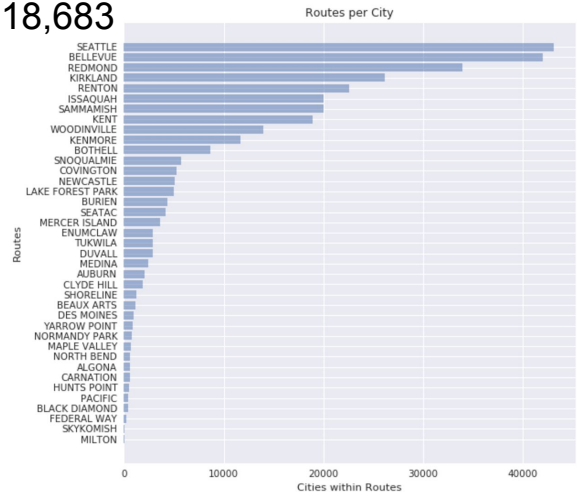


General Data Observations

- Number of routes by type
  - **Total:** 255,716

route_type	count
Run	103887
Walk	88712
Bike	34438
Unknown	28679

- Number of routes that intersect cities
  - **Total:** 318,683

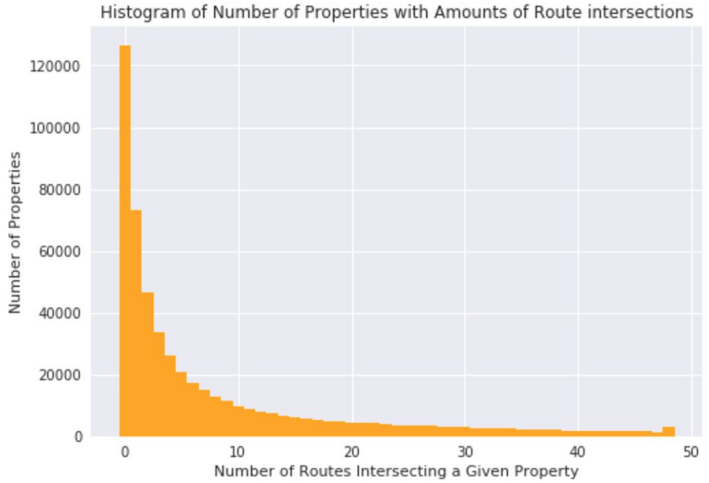


- Number and total value of properties
  - **Total:** 610,232; **Median:** 437,000

present_use_class	count
Residential	564025
Commercial	25879
Unknown	13936
Public	3602
Utility	2128
Nature	662

present_use_class	median
Public	1334300.0
Commercial	1006600.0
Residential	436000.0
Utility	119000.0
Nature	10000.0
Unknown	0.0

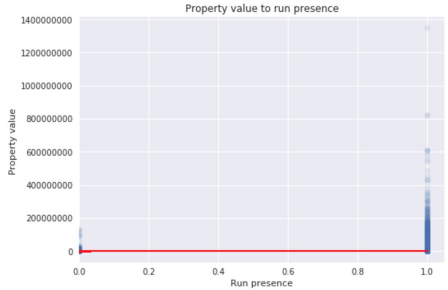
- Properties intersected by routes

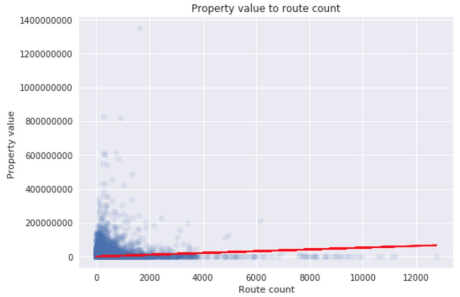


Exercise Routes and Property Values

- Correlation plot
  - Total property value vs Route count:
    - Slope = 585,275; Rsquared = .04; Pvalue = <.05
  - Total property value vs Route presence:
    - Slope = 5,234; Rsquared = .01 ; Pvalue = <.05

- Regression
  - Total property value -- Use class, lot sq ft, route count, route presence, lot sq ft vs city sq ft
  - Coefficients - route presence, route count, lotsqft had the highest positive correlations; commercial and public properties also had the highest correlations of use class
  - Rsquared = .23
  - Pvalue = <.05





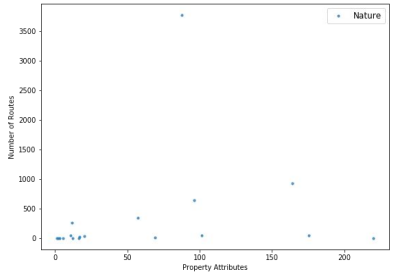
Routes chosen - Linear Regression

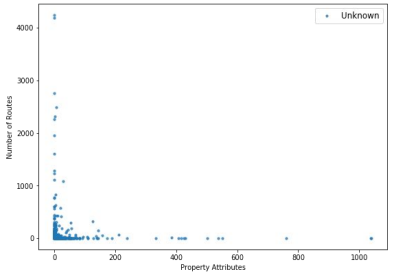
OLS Regression Results

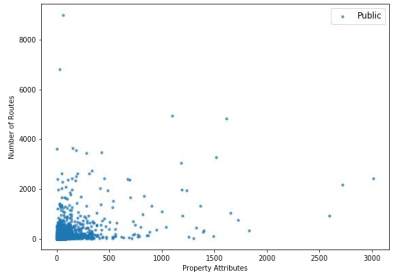
Dep. Variable:	route_count	R-squared:	0.058
Model:	OLS	Adj. R-squared:	0.058
Method:	Least Squares	F-statistic:	9591.
Date:	Mon, 29 May 2017	Prob (F-statistic):	0.00
Time:	06:52:24	Log-Likelihood:	-2.8469e+06
No. Observations:	467323	AIC:	5.694e+06
Df Residuals:	467319	BIC:	5.694e+06
Df Model:	3		
Covariance Type:	nonrobust		

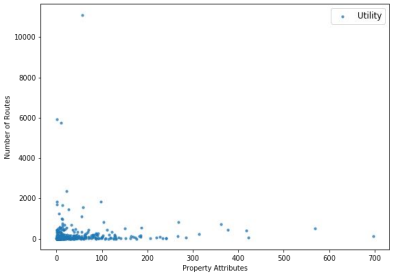
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	11.2978	0.082	137.240	0.000	11.136	11.459
present_use_class[T.Nature]	-0.0358	0.000	-137.240	0.000	-0.036	-0.035
present_use_class[T.Public]	-0.0109	7.93e-05	-137.240	0.000	-0.011	-0.011
present_use_class[T.Residential]	11.1818	0.081	137.240	0.000	11.022	11.342
present_use_class[T.Unknown]	0.0147	0.000	137.240	0.000	0.015	0.015
present_use_class[T.Utility]	0.0018	1.29e-05	137.240	0.000	0.002	0.002
lotsqft	0.0001	1.83e-06	72.175	0.000	0.000	0.000
apprlndval	9.761e-06	1.03e-07	95.104	0.000	9.56e-06	9.96e-06
appr_impr	3.035e-06	4.08e-08	74.350	0.000	2.95e-06	3.11e-06
appr_impr:apprlndval	-4.265e-14	4.09e-16	-104.350	0.000	-4.34e-14	-4.18e-14
appr_val_vs_city_appr	0.0007	4.74e-06	137.240	0.000	0.001	0.001
lotsqft_vs_citysqft	0.0004	2.75e-06	137.240	0.000	0.000	0.000

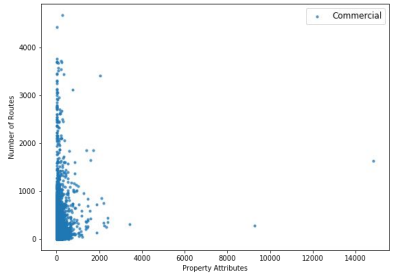
Omnibus:	997991.603	Durbin-Watson:	1.295
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14389962681.341
Skew:	18.449	Prob(JB):	0.00
Kurtosis:	861.868	Cond. No.	1.64e+18

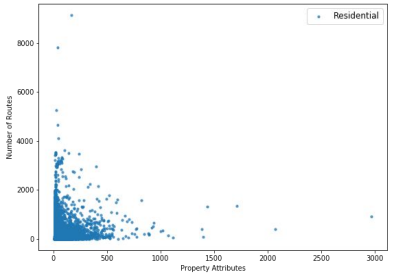












Routes Chosen - T-test

```
#filter property with no routes
propertyNoRoutes = df.query('route_count == 0').dropna()
#calculate the average enrollment for treated group
print(propertyNoRoutes.total_appr_val.mean())

#filter property with routes
propertyWithRoutes = df.query('route_count > 0').dropna()
#calculate the average enrollment for controlled group
print(propertyWithRoutes.total_appr_val.mean())
```

```
427669.0338925823
1095981.753910244
```

```
from scipy.stats import ttest_ind
#calculate t-test for routes and noroutes property
t, p = ttest_ind(propertyWithRoutes.total_appr_val, propertyNoRoutes.total_appr_val)
print(t, p)
```

```
29.5652927288 6.28841032737e-192
```

Summary Analysis This t-est generates a p-value that's vastly less than 0.05, thus we reject the null hypothesis that there is no difference between properties built arount running routes and those built without running routes around.

Routes Chosen - ROOT-MEAN-SQUARE Error

```
# Generate our predictions for the training set.
predictions = model.predict(train[columns])

# Compute error between our test predictions and the actual values.
print(compute_rmse(predictions, train[target]))
```

```
5519571.414
```

```
# Generate our predictions for the test set.
predictions = model.predict(test[columns])

# Compute error between our test predictions and the actual values.
print(compute_rmse(predictions, test[target]))
```

```
6914200.9208
```

In our model, we have found that the RMSE for the test data is slightly higher than the RMSE in the training data. Basically, we have a model that tests well in sample, but may not have a strong predictive value when tested out of sample.

Conclusion

- Existence of a correlation between the value of a home and the availability of routes running in the area
- Our data rejected the null hypothesis
- Model possible use:
  - Advertisers targeting ads to home buyers who are also fitness enthusiasts
  - Affect how municipalities design building code
  - Policy based on a set of machine learning that utilizes this model.
  - Add into Zillow's Zestimate algorithm

Challenges & Future Consideration

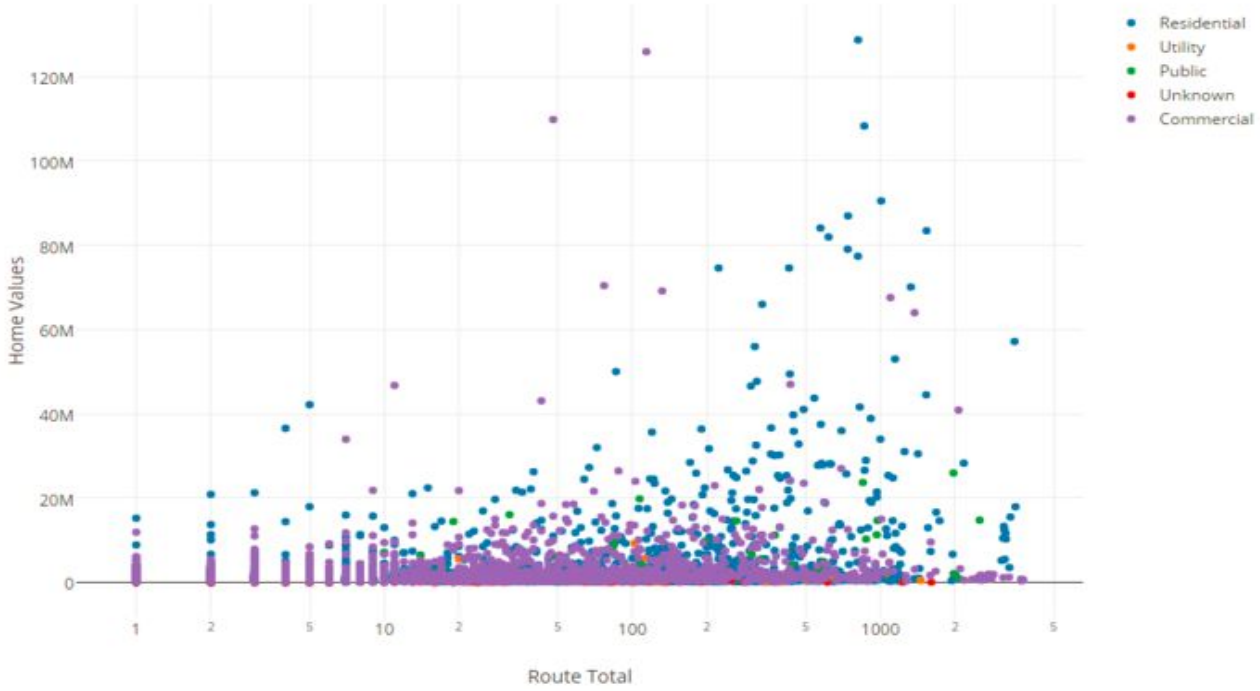
- Challenges
  - Geolocation transformations can be difficult
  - Processing power for many tasks required cloud processing
  - Data large for standard dataframe creation techniques
  - Tuning ETL processes for speed and de-duplication (services frequently require restarting)

- Future Considerations
  - Need to access more complete exercise dataset
  - Integrate more independent variables in order to find better correlations
  - Year-over-year analysis
  - Route proximity to parcel, distance traversed over property, property sequence

End

- Thank you for your time
- Questions?

Additional - Graphics



Additional - Graphics