Net Economic impact: Introduction of Street Car in Downtown Cincinnati

Exploratory Data Analysis, Predictive Analysis and Forecast

K. Rajesh Jagannath

Foundations of Data Science
Mentor: Anirban Ghosh
08/05/2016

Table of Contents

Net Economic impact: Introduction of Street Car in Downtown Cincinnati 1

Introduction 1

- **Objective** 1
- **Motivation** 1
- **Data Sources** 2

Extraction, Transformation and Loading of Data 3

Feature Extraction 4

Exploratory Data Analysis 6

- **Conclusion** 7

Net Economic impact: Introduction of Street Car in Downtown Cincinnati

Introduction

The Cincinnati Streetcar is a modern streetcar system designed to link major employment centers in downtown and uptown, connecting through Cincinnati's historic Over-the-Rhine neighborhood.

It will operate 18 hours a day, 365 days a year.

Objective

The study's goal is to analyze and predict the "net positive effect" on the economy in the buffer zone around the streetcar route by selecting meaningful features from various data sets.

Motivation

The City of Cincinnati is the client. Downtown is Cincinnati's largest employment center, with approximately 70,000 people working in the area every day. It has been proven in cities from Atlanta to Seattle that fixed rails in the ground with thousands of potential riders **draw new storefronts and businesses**, as well as **housing**. These new businesses provide employment opportunity and **boost a city's tax revenue.** Also, here may have been inconveniences to the neighborhood, during the construction phase. Hence, there are two camps of opinion -

- One opinion insists that the introduction of the streetcar is disruptive to the neighborhood (crowding, transient population, noise), and
- The other opinion is that it provides easy access to business, shops, dining and commuting to work and home and draws new business, expansion of storefronts, revenue from ridership, permit fees, **property tax** and restaurant license fee.

Three buffer zones around the streetcar route were established as shown below.

- CORE: The area shown in Red color is the CORE Buffer zone. The Streetcar runs through the center of this area along a North South corridor.
- CENTER: The area shown in Magenta color is the designated CENTER Buffer zone
- EDGE: The area shown in Green color is the EDGE Buffer zone

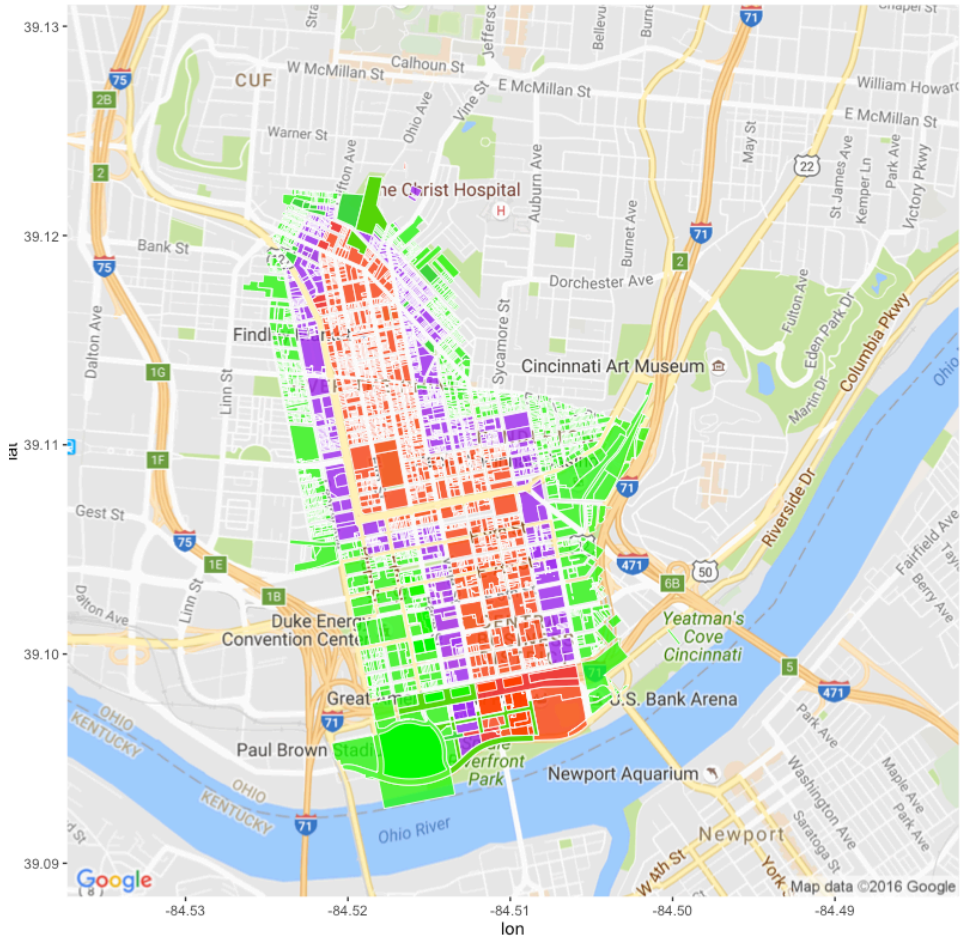


Table 1. ggplot of the downtown area under study illustrating the Buffer zones CORE, CENTER and EDGE around the Street Car route

Data and Sources

Data Sources

Data source is Open Data Cincinnati and Cincinnati Area Geographic Information Systems (CAGIS), City of Cincinnati, OH.

1. **Buffer Area Parcels: There are three .csv files** that with an observation for each parcel in the three buffer zones under study.
    - StreetCarParcels_CORE.csv
    - StreetCarParcels_CENTER.csv
    - StreetCarParcels_EDGE.csv

Column name	Example Data	Description
PARCELID	7500010007	Unique id to identify parcels
EXLUCODE	C	Existing Land use Code e.g. Commercial
ADDRNO	1208	Address, street and type of street
ADDRST	SYCAMORE	
ADDRSF	ST	

Table 2: .csv files are used to identify the parcel id. of the three areas around the Street Car - Core, Center and Edge Buffer zones

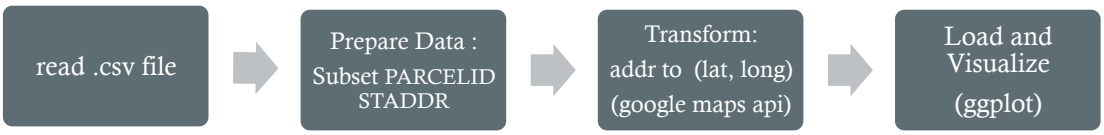
2. **Assessors Tax Information 2007-2015: The Assessors Office** provided data for 9 years in Fixed Width Format in 9 files.
    - taxinfo2007.txt
    - taxinfo2008.txt
    - taxinfo2009.txt
    - taxinfo2010.txt
    - taxinfo2011.txt
    - taxinfo2012.txt
    - taxinfo2013.txt
    - taxinfo2014.txt
    - taxinfo2015.txt

Column Name	Example Data	Description
PARCEL_ID	10001000100	Unique id for a parcel
LOC_STREET		Location e.g for 2327 Sussex Ave., Cincinnati Oh
LOC_HOUSE_NO	2327	
LOC_ST_DESC	SUSSEX	
LOC_ST_IND	AV	
LOC_ST_DIR		
VALID_SALE	Y	Yes or No
NUM_PARCEL	3	Number of Parcels
MKT_LAND_VAL	23000	Value of the Land
MKT_IMPR_VAL	140570	Market value of the Land
MKT_TOTAL	163570	Mkt. Total Val
ACRES	0.246	Acreage of the building
SALE_AMOUNT	116000	Sale Amount
SALE_DATE	20121129	Sale date in YYYYMMDD format
NEW_CONSTR	N	Newly constructed building
ANNUAL_TAXES	3693.14	Annual Taxes Assessed
TAXES_PAID	3693.14	Annual Taxes Paid
DELQ_TAXES	6088.56	Delinquent taxes
FORECL_FLAG	Y	Tax Foreclosure Flag

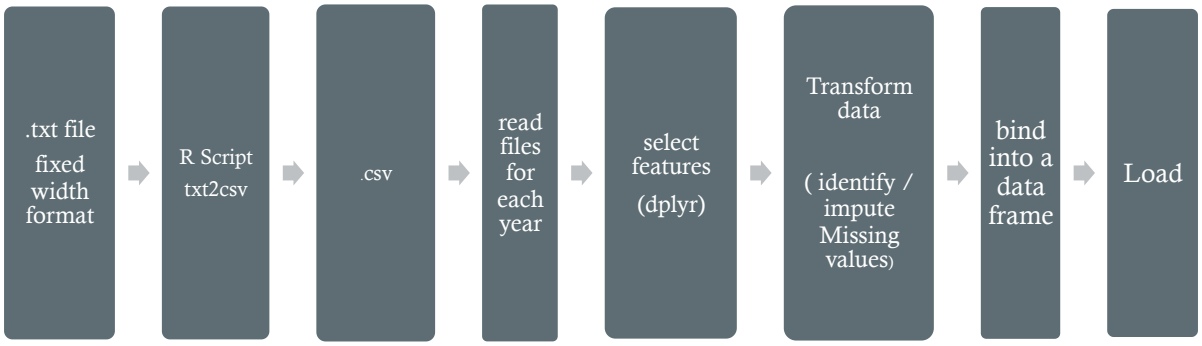
Table 3. Features selected from Property Tax Information from years 2007 - 2015

Extraction, Transformation and Loading of Data

1. **Buffer zone under study**: The three buffer-zone parcel files were in .csv format. It was read in using read_csv. Features to obtain street address and parcel id were selected. The Street address was used to geocode the data to obtain longitude and latitude of the parcel. The data was visualized for exploratory analysis. There are 900-1700 observations in each file. The file size is about 1.2 MB.



2. **Property Tax information 2007-2015**: The original datasets were provided in fixed width format. An R script converted it to .csv file. The problem here was each of the groups of years 2007, 2008 and 2009-2014 and 2015 had different column widths. The field width was clearly documented. There are 300,000 observations for each year. Files are about 150MB to 260MB in size for reach year.



Feature Extraction

Yearly Tax Data from Years 2007-2015 is available from the Auditors office. From that data-set a few features have been identified for selection. These selections are indicative of economic growth – Market Value, Assessed taxes, Revenue from Taxes paid, Sales data , Foreclosure Data and New Construction Flag

Sl	VARIABLE	Description
1	**PARCELID**	*A Unique identifier of the parcel*
2	**LOC_STREET**	*Address for plotting on ggplot or other package to identify spatial correlation*
3	**LOC_HOUSE_NO**	*Street Address Location + Latitude and Longitude*
4	**LOC_ST_DESC**	
5	**LOC_ST_IND**	
6	**LOC_ST_DIR**	
7	**cent_long**	
8	**cent_lat**	
9	**EXLU_CODE**	*Existing Land Use code*
10	**MKT_LAND_VAL**	*Market Value of land, Improvements, and Total*
12	**MKT_IMPR_VAL**	
13	**MKT_TOTAL**	
14	**ANNUAL_TAXES**	*Net Prop Tax revenue: Annual Taxes assessed, Taxes actually Paid, Delinquent Taxes and Tax Foreclosure*
15	**TAXES_ PAID**	
16	**DELQ_TAXES**	
17	**FORECL_FLAG**	
18	**ACRE**	*Acreage to compute Property Value / sq. ft.*
19	**SALE_AMT**	*Sales data of Property : Amount, Sale Date, New Construction*
20	**VALID_SALE**	
21	**SALE_DATE**	
22	**NEW_CONSTR**	

Table 4. There are several features available in the data set for years 2007-2015. The features in the table above have been selected and are indicators of Market value of the parcel, Annual taxes, Acre-age, Sales Data. These are representative of the net economic effect.

PARCELID	2007	2008	2009	2010	2011	2012	2013	2014	2015	DummyVars
Over 300,000 observations/year										

Table 5. Parcel id uniquely identifies an observation. Each year from 2007-2015 has a subset of features shown in Table 4. A set of Dummy variables will be used to identify CENTER, CORE and EDGE Buffer parcels. This data is not tidy data and will need to be transformed using tidyr into Table 6. Below.

PARCELID	YEAR	LOCATION	EXLU	MKT VALUE	TAXES	SALE	DUMMY VARIABLES			
	2007									
	2007									
	2008									
	2008									
	2015									
	2015									

Table 6: The same data frame in Tidy Data Frame.

PARCELID	YEAR	LOC_STREET	LOC_HOUSE_NO	LOC_ST_DESC	LOC_ST_IND	LOC_ST_DIR	cent_long	cent_lat	EXLU_CODE	MKT_LAND_VAL	MI

Table 7. Microsoft Excel object of the data-frame shown in Table 6

Exploratory Data Analysis

Plotting the position - Longitude, Latitude vs. Existing Land Use Code, vsualizes the **expected 2-D distribution** of the parcels concentrated in the CORE, CENTER and EDGE zones.

Figure 1. Scatterplot CORE Buffer Zone

Figure 2. Scatterplot CENTER Buffer Zone

Figure 3. Scatterplot EDGE Buffer Zone

Within the Buffer Zones, we find that the distribution with respect to Existing Land Use is **not uniform**. The distribution is skewed towards Multi-family, Mixed Used, Vacant, Commercial and Public/Semi-public parcels. Also, There are too many parcels classified as vacant lots – that needs investigation.

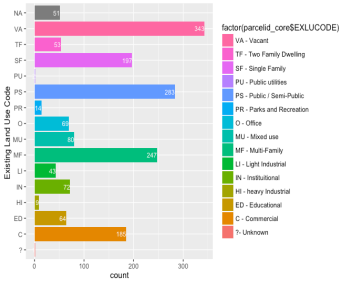


Figure 4. Histogram of parcels in CORE buffer zone.

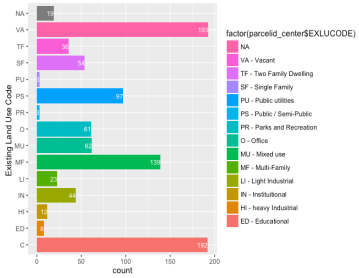


Figure 5. Histogram of parcels in CENTER buffer zone.

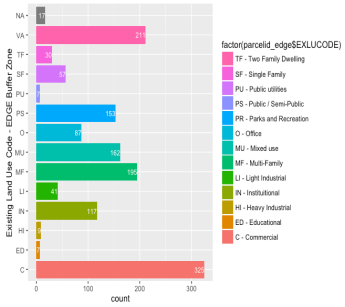


Figure 6. Histogram of parcels in the EDGE buffer zone

Performing 2D Kernel Density plot, we find the areas of high Market Land Value to be centered on the Buffer Zone. There **is an unanticipated concentrated distribution in the center of the Downtown** in all the three plots. This is indicative of **a problematic geocoding or the street addresses in the data are not correct**. In the scatterplot, Figure 1, Figure 2 and Figure 3, this problem is masked because the points are over-lapping each other in a single point in the center

of the downtown. However, a 2-D Kernel Density Map, reveals an unusually high concentration of observations in areas **not expected** to be in the CENTER and EDGE buffer zones.

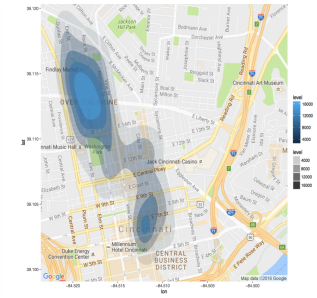


Figure 7. 2-D Kernel Density plot of CORE

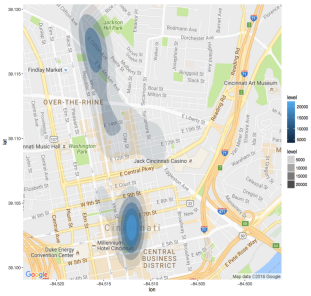


Figure 8. 2-D Kernel Density plot of CENTER: High density of observations near Central Business District is not expected

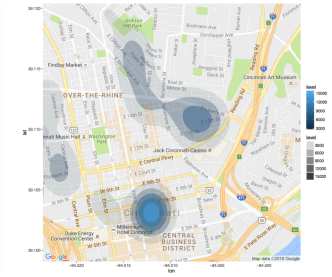


Figure 9. 2D Kernel Density plots of EDGE: High density of observations near Central Business District is not expected

---

The data sets were analyzed further with CAGIS and a cleaner data set was obtained

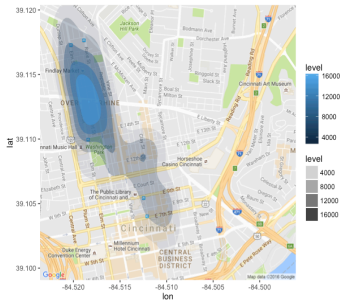


Figure 10. 2-D Kernel Density plot of the CORE parcels :AFTER – Clean Dataset provided by CAGIS

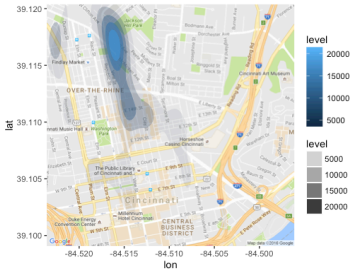


Figure 11. 2-D Kernel Density plot of the CENTER parcels : AFTER - Clean data provided by CAGIS - Central Business District are no longer there

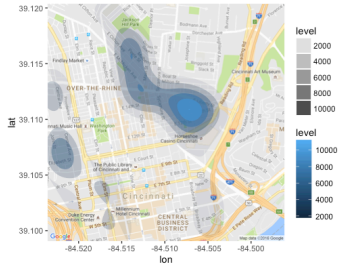


Figure 12. 2-D Kernel Density plots of the EDGE : After data was cleaned, observations are in line with expectation

Conclusion

- Data 1
  - 3 .csv files qualifying CORE, CENTER and EDGE buffer zone were analyzed
  - Their dimensions were 1713x8, 946x8 and 1418x8 respectively
  - Top 5 Land Use in each buffer zone were: Commercial, Multi-Family, Mixed Use, Semi-Public and Vacant
    - Examples of Vacant lots were Parking lots which is to be expected near Commercial centers
  - Parcels with 0 Annual Taxes value were analyzed further. They correspond to one of the several parcels owned by the same owner. The taxes are assessed on one parcel only for billing convenience and others are marked 0.
  - Density map indicated some geocoded co-ordinates are not spatially situated in the buffer zones as expected.
    - For example, in Fig. 8 and Fig. 9, there is a high density of observations Near the Central Business District which seems to be present in all 3 buffer zone
    - Some of these observations do not have complete addresses for Google Maps Geocoding API to provide accurate longitude and latitude co-ordinates
  - Further analysis of the data set with the client, CAGIS, indicated that condominium parcels are also not correctly treated in the data-set provided
  - More accurate data-set was requested
    - Fig. 10, Fig. 11, Fig. 12 illustrates a better distribution of the parcels in the expected buffer zones
    - In particular, the high density of observations near the Central Business District in Fig. 7, 8 and 9, prior to clean up is no longer observed
    - This paves way for future work
  - Instead of using Google Maps geocoding, it was decided to obtain longitude and latitude co-ordinates from CAGIS directly

- Data 2
    - Tax information for 9 years was provided in fixed width format and converted to csv. File size for each year's data set is of the order of 150 MB.
    - Each year's data set has 300000 observations.
    - 18 features indicating net positive economic effect were identified and feature extraction was performed
    - Annual Taxes was identified as dependent variable for Time series forecast
    - A data frame and its corresponding tidy-form has been proposed for future work

Bibliography

References

1. https://dev.socrata.com/foundry/data.cincinnati-oh.gov/emnx-rw6d
2. http://www.cincinnati.com/story/news/2016/05/05/streetcar-nation-kc-opens-friday-cincy-next/83874740/