Identifying Vacant and Uninhabitable Properties in Philadelphia, PA, using Publicly Available Data

DS 670 Capstone Winter 2021

Akshay Srivastava
Prianka Ball
Reina Carissa

The case in urban vacancy

Vacant storefronts along Second Avenue between 60th and 90th streets, Upper East Side, NY— a rate of more than two per block. (Nick Garber/Patch).

Source: https://patch.com/new-york/upper-east-side-nyc/vacancy-crisis-empty-storefronts-blanket-upper-east-side

What are Vacant Properties?

- Abandoned, boarded-up buildings
- Unused lots that attract trash and debris
- Residential, commercial, and industrial buildings that exhibit one or both of the following traits:
    - Poses a threat to public safety (meeting the definition of a public nuisance), or
    - The owners or managers neglect the fundamental duties of property ownership (e.g., they fail to pay taxes or utility bills, default on mortgages, or carry liens against the property.)
- Vacant or under-performing commercial properties known as greyfields (such as under-leased shopping malls and strip commercial properties)
- Neglected industrial properties with environmental contamination known as brownfields.

What Makes a Property Uninhabitable?

Non-Functioning Cooling and Heating

Rental properties that do not have a working air conditioner or furnace, depending on the season, might be deemed uninhabitable by the laws of many states.

Structural, Plumbing, and Electrical

Almost every state requires rental properties to be free of structural issues, have running water, and have electricity and/or gas

Mold, Mildew, and Water Leaks

Many forms of mold and mildew can be dangerous for humans and animals to be around. This falls into the category of environmental hazards.

Why are Vacant and Abandoned Properties a Problem?

Vacancy is associated with a number of concerns among communities

- Undermining their neighbors' **quality of life**
- Diminishing the **value** of nearby properties
- Imposing **fiscal burdens** on the city
- Need intervention from city governments to keep them from becoming **crime magnets, fire hazards, or dumping grounds**

Urban vacancy in Philadelphia, PA

Philadelphia is currently the 6th largest city in the United States, based on a population of 1.6 million according to the World Population.

Urban vacancy in Philadelphia

- Growing manufacturing economy that spurred population growth (Schilling & Hodgson, 2013).
- By the mid-20th century, Philadelphia lost over half of its industrial sector, and experienced a significant population loss.
- A study found that houses within 150 feet of a vacant or abandoned property in Philadelphia experienced a net loss of **$7,627** in value (Bass et al. 2005).

$20 million

Maintenance cost for **~40,000** estimated vacant lots in Philadelphia, PA.

Many of these lots have become sites for **illegal dumping** for soiled mattresses, abandoned cars, and household trash (Loesch, 2020).

Challenges in addressing vacant lands in Philadelphia

- Not all vacant property owners take the necessary steps to protect and care for their property.
- Many types of vacant property are not measured except when people walk or drive block by block to count them.
- While national source exists that projects the number and location of vacant lots in a particular location, the data is largely outdated (Mallach, 2018).

Objective

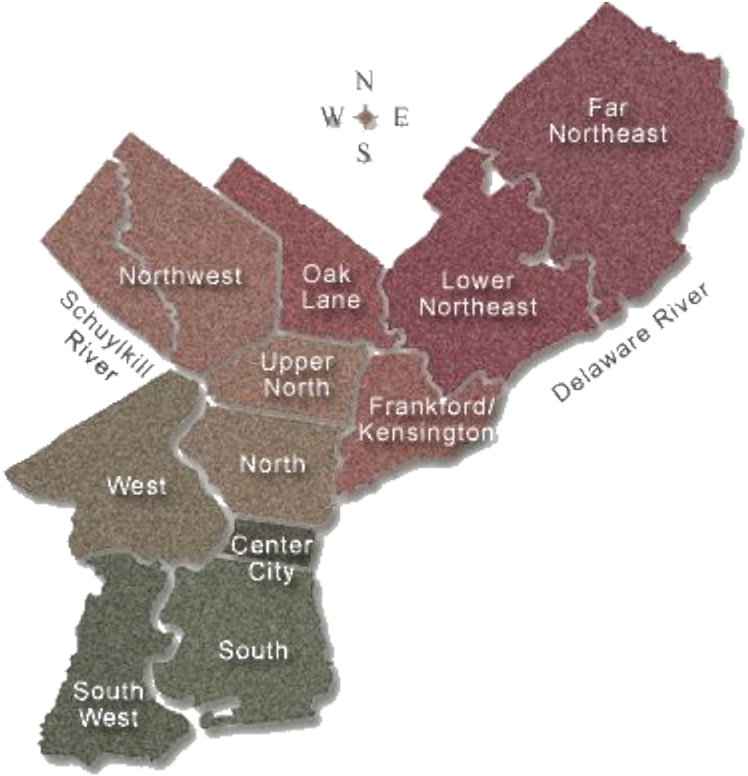
Need more research to make evidence-based decisions to:

- Rehabilitate or demolish
- Have a judicial or administrative foreclosure process
- Convert a brownfield to an affordable housing development or a green space
- Pursue smart growth or smart decline.

Our Approach

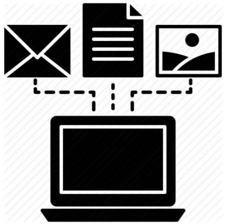
Using machine learning models

- Accurately identify lots that are in danger of becoming vacant.
- Explore data that have been historically connected to vacant lots.
- Improve and revitalize disinvested urban neighborhoods in Philadelphia.



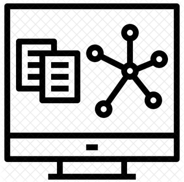
Methodology

Data collection



Exploratory data analysis (EDA)

Feature engineering, preprocessing, and modeling



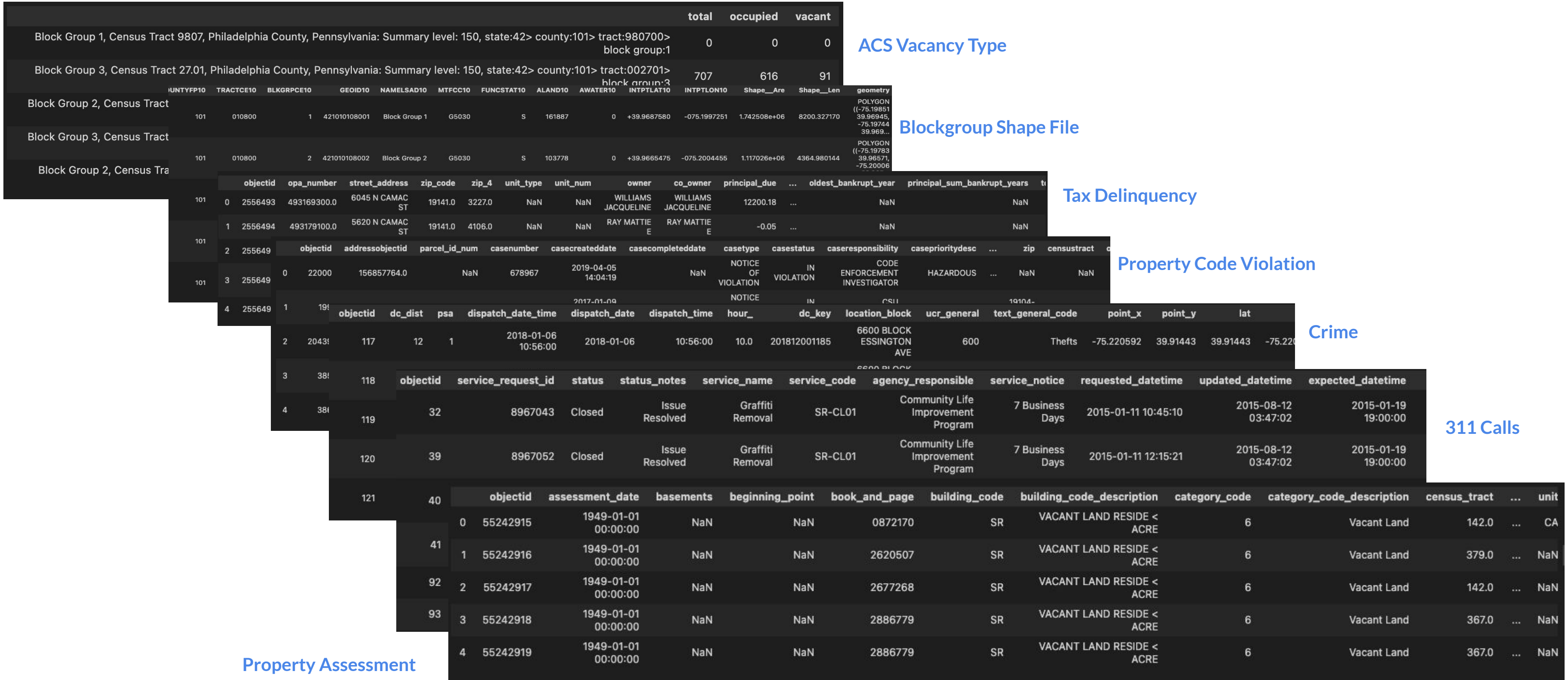
Data collection

Resources

- Philadelphia shapefiles
  - Census block groups
  - Zip code
- American Community Survey
  - Occupancy status
  - Vacancy status
  - Population
- Philadelphia open data program
  - Property assessment
  - Crime Data
  - 311 Data
  - Property tax delinquency
  - Property code violations

Most of the datasets are between 2015 and 2021

Data Screenshot



ACS Vacancy Type

	total	occupied	vacant
Block Group 1, Census Tract 9807, Philadelphia County, Pennsylvania: Summary level: 150, state:42> county:101> tract:980700> block group:1	0	0	0
Block Group 3, Census Tract 27.01, Philadelphia County, Pennsylvania: Summary level: 150, state:42> county:101> tract:002701> block group:3	707	616	91
Block Group 2, Census Tract			
Block Group 3, Census Tract			
Block Group 2, Census Tra			

Blockgroup Shape File

COUNTYFP10	TRACTCE10	BLKGRPCE10	GEOID10	NAMELSAD10	MTFCC10	FUNCSTAT10	ALAND10	AWATER10	INTPTLAT10	INTPTLON10	Shape__Are	Shape__Len	geometry
101	010800	1	421010108001	Block Group 1	G5030	S	161887	0	+39.9687580	-075.1997251	1.742508e+06	8200.327170	POLYGON ((-75.19851 39.96945, -75.19744 39.969...
101	010800	2	421010108002	Block Group 2	G5030	S	103778	0	+39.9665475	-075.2004455	1.117026e+06	4364.980144	POLYGON ((-75.19783 39.96571, -75.20006
101													
101													
101													

Tax Delinquency

	objectid	opa_number	street_address	zip_code	zip_4	unit_type	unit_num	owner	co_owner	principal_due	...	oldest_bankrupt_year	principal_sum_bankrupt_years
0	2556493	493169300.0	6045 N CAMAC ST	19141.0	3227.0	NaN	NaN	WILLIAMS JACQUELINE	WILLIAMS JACQUELINE	12200.18	...	NaN	NaN
1	2556494	493179100.0	5620 N CAMAC ST	19141.0	4106.0	NaN	NaN	RAY MATTIE E	RAY MATTIE E	-0.05	...	NaN	NaN
2	255649												
3	255649												
4	255649												

Property Code Violation

	objectid	addressobjectid	parcel_id_num	casenumber	casecreateddate	casecompleteddate	casetype	casestatus	caseresponsibility	caseprioritydesc	...	zip	censustract
0	22000	156857764.0	NaN	678967	2019-04-05 14:04:19	NaN	NOTICE OF VIOLATION	IN VIOLATION	CODE ENFORCEMENT INVESTIGATOR	HAZARDOUS	...	NaN	NaN
1	19				2017-01-09		NOTICE	IN	CSU			19104-	
2	2043												
3	38												
4	38												

Crime

	objectid	dc_dist	psa	dispatch_date_time	dispatch_date	dispatch_time	hour_	dc_key	location_block	ucr_general	text_general_code	point_x	point_y	lat
117	12	1		2018-01-06 10:56:00	2018-01-06	10:56:00	10.0	201812001185	6600 BLOCK ESSINGTON AVE	600	Thefts	-75.220592	39.91443	39.91443
118									6600 BLOCK					
119														
120														
121														

311 Calls

	objectid	service_request_id	status	status_notes	service_name	service_code	agency_responsible	service_notice	requested_datetime	updated_datetime	expected_datetime
	32	8967043	Closed	Issue Resolved	Graffiti Removal	SR-CL01	Community Life Improvement Program	7 Business Days	2015-01-11 10:45:10	2015-08-12 03:47:02	2015-01-19 19:00:00
	39	8967052	Closed	Issue Resolved	Graffiti Removal	SR-CL01	Community Life Improvement Program	7 Business Days	2015-01-11 12:15:21	2015-08-12 03:47:02	2015-01-19 19:00:00
	40										
	41										
	92										
	93										

	objectid	assessment_date	basements	beginning_point	book_and_page	building_code	building_code_description	category_code	category_code_description	census_tract	...	unit
0	55242915	1949-01-01 00:00:00	NaN	NaN	0872170	SR	VACANT LAND RESIDE < ACRE	6	Vacant Land	142.0	...	CA
1	55242916	1949-01-01 00:00:00	NaN	NaN	2620507	SR	VACANT LAND RESIDE < ACRE	6	Vacant Land	379.0	...	NaN
2	55242917	1949-01-01 00:00:00	NaN	NaN	2677268	SR	VACANT LAND RESIDE < ACRE	6	Vacant Land	142.0	...	NaN
3	55242918	1949-01-01 00:00:00	NaN	NaN	2886779	SR	VACANT LAND RESIDE < ACRE	6	Vacant Land	367.0	...	NaN
4	55242919	1949-01-01 00:00:00	NaN	NaN	2886779	SR	VACANT LAND RESIDE < ACRE	6	Vacant Land	367.0	...	NaN

Property Assessment

Data Challenges

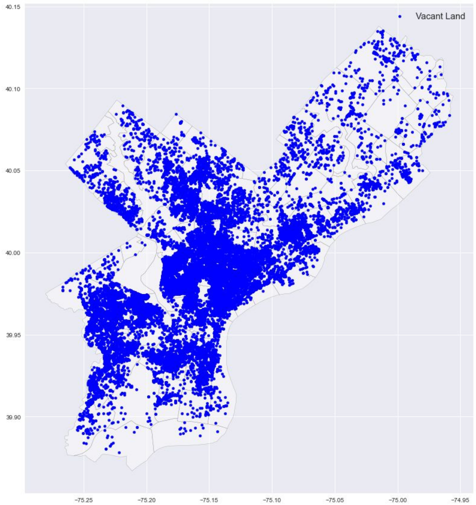
- First time working with geospatial data.
- Joining different types of datasets: geospatial data with regular datasets
- Very large number of columns.
- Large number of columns with null values.
- Project required intensive research to understand all data descriptions. This was needed to understand how to preprocess and extract dataset.
- Long time to finish queries.
- Difficult to draw clear conclusions by just looking at maps.
- Hard to understand effect of different variables in each location.
- Unbalanced labelled dataset in final model. Final labelled dataset had only 7% labelled vacant lots.
- A lot of skewed data.

Exploratory data analysis (EDA)

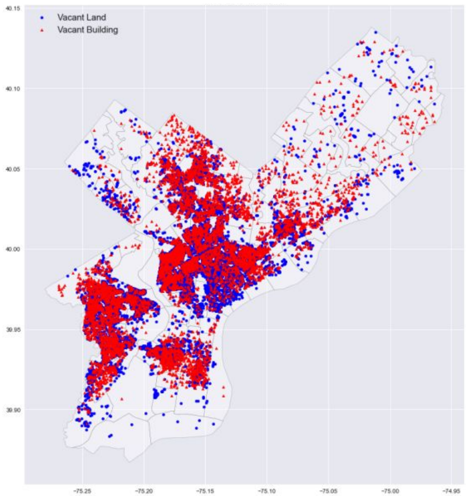
Important Libraries

- Geopandas: geospatial data
- CensusData: census data
- Sklearn: ML model and feature engineering
- Seaborn
- Matplotlib
- Shapely
- Pandas
- Numpy

Vacant Lots in the City of Philadelphia



Current Vacant Lots (source: Property Assessment)



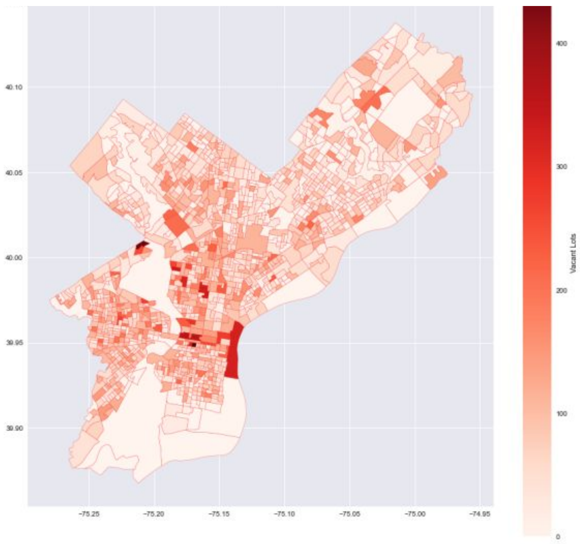
Predicted Vacant Lots (source: Predicted Vacant Places by City of Philadelphia)

Vacant places are more concentrated in the middle.

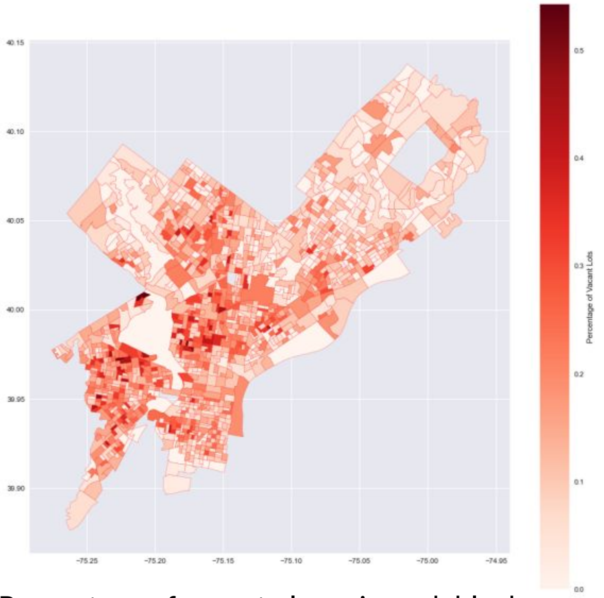
Property Assessment

- An average house size is around 1,000 sq ft and has a radius of 6m
- Total livable area and total area are very skewed. It did not say anywhere what metric it was using. We are guessing it was square feet
- Most of assessment was from 2021 which shows that the data is quite recent
- Vacant lot has less market value than non-vacant lots
- Vacant lot has higher mean depth meaning they were more away from the roads. The depth is measured from the principal street back to the rear property line or secondary street
- Property assessment from year 2015 - 2021 reveals that market value of properties has increased over the years

American Community Survey (ACS) Occupancy



Number of vacant places in each block group



Percentage of vacant places in each block group

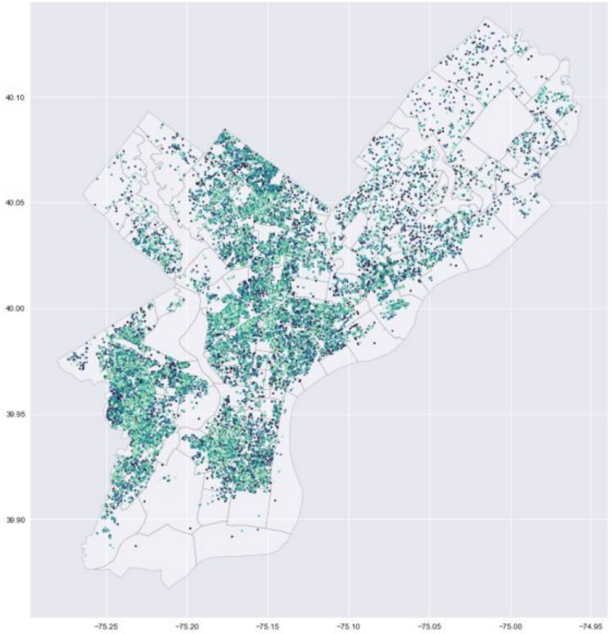
- Distribution of percentage of vacant lot is different from number of vacant lots.
- Median percentage of vacant lots in each block group is around ~12%

American Community Survey (ACS) Vacancy Type

- "Other" type of vacant lot is the highest percentage of vacant lots when looking at all types of vacant places
- There are around 6.7% of "other" vacant places when compared to the total housing in the area.

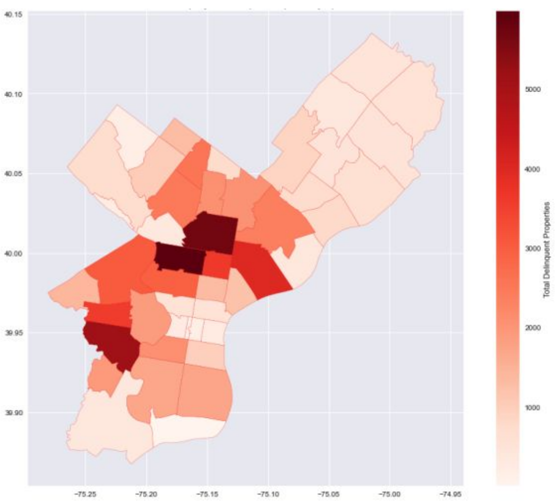
"Other" Vacant type on block groups

Property Tax Delinquency

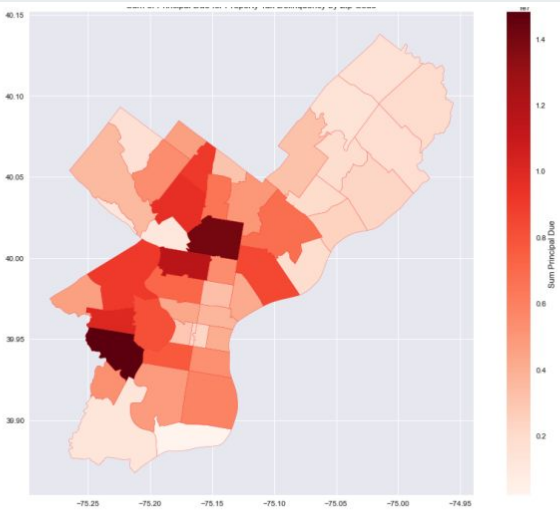


Property tax delinquency of principal value more than $2K on neighborhood.

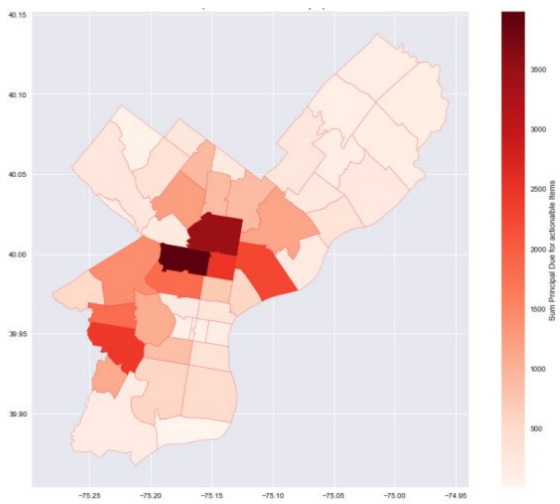
- An account is delinquent when Real Estate Tax is still unpaid on January 1 the following year the tax was due.
- Median for principal value is around $2K.
- Principal due data is skewed.



Number of Property Tax Delinquent Properties by ZIP code



Sum of Principal Due for Property Tax Delinquency by ZIP code



Sum of Principal Due Actionable by ZIP code

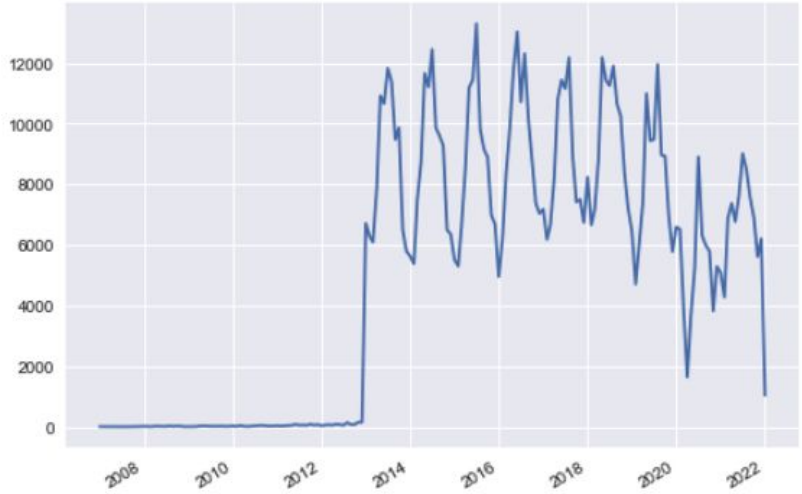
- Actionable: the city is actively working to collect these accounts
- Non-actionable: the city can't do anything further or they are barred from collection.
- Accounts that are in payment agreement, bankruptcy, or overdue but not yet delinquent are considered "not actionable". Payment agreement is one of the way the city collect debts.
- Sheriff sale and Sequestration are actionable

More on Property Tax Delinquency

- Most of the principal due is owed for 1-4 years
- The fifth higher principal due is owed for 25 years
- When looking at median principal due, a lot of taxes are owed for 18, 23, 22 and 27 years. Principal value is skewed so median is a better measure.
- Most of the delinquent properties have taxes that are owed for 1-6 year. A lot of properties have taxes owed for 25 years.
- Most of the delinquent properties are **residential**, not commercial.
- **89%** of residential places are delinquent properties.
- Most of the places were assessed in 2021
- Principal due is the most for houses and vacant land. However, the median and mean principal due is not high for house and vacant lots. Utility buildings have the highest median principal due.

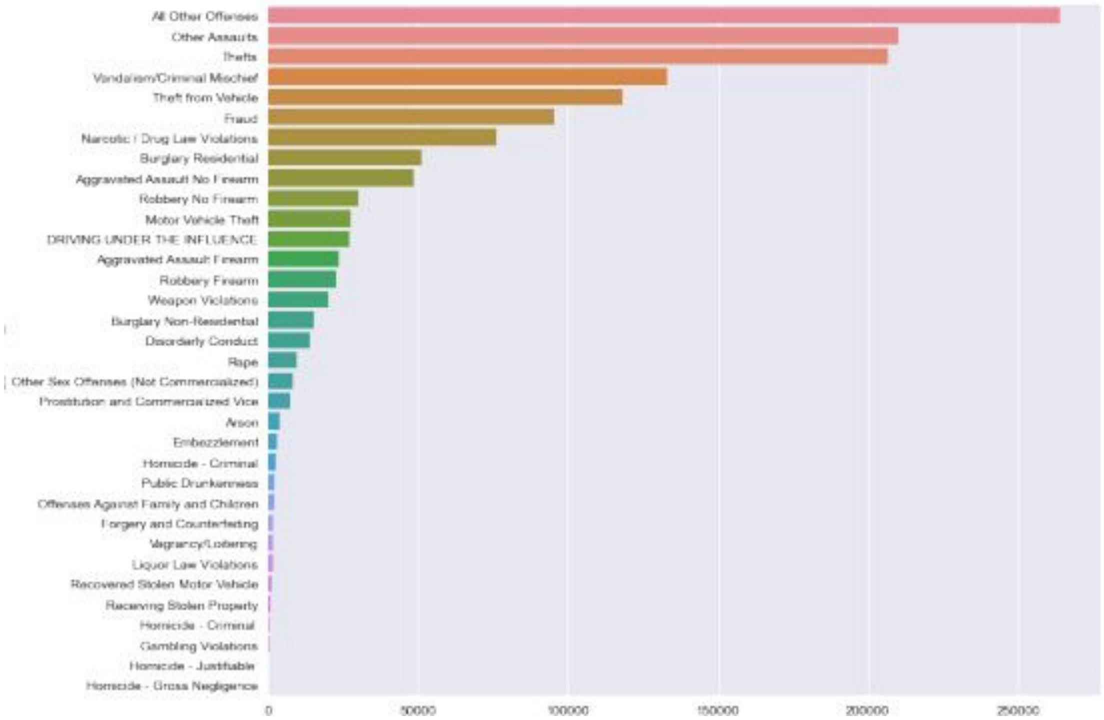
Property Code Violations

- There might be some seasonality in code violation
- There are around 2,000 violation types
- A lot of the violation code titles has vacant lot in the title.
- Most of the violation statuses were labeled as “complied” which shows that most property owners acted in accordance with the city government



Number of reported case violations year 2015 - 2021

Crime



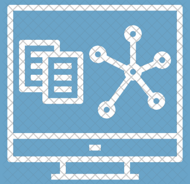
- Most of the crime are **all other offenses**
- Assaults is the second highest type of crime
- After grouping the number of crimes that happened within 50m of each parcel number, we noticed that the data is quite skewed.

311 Calls

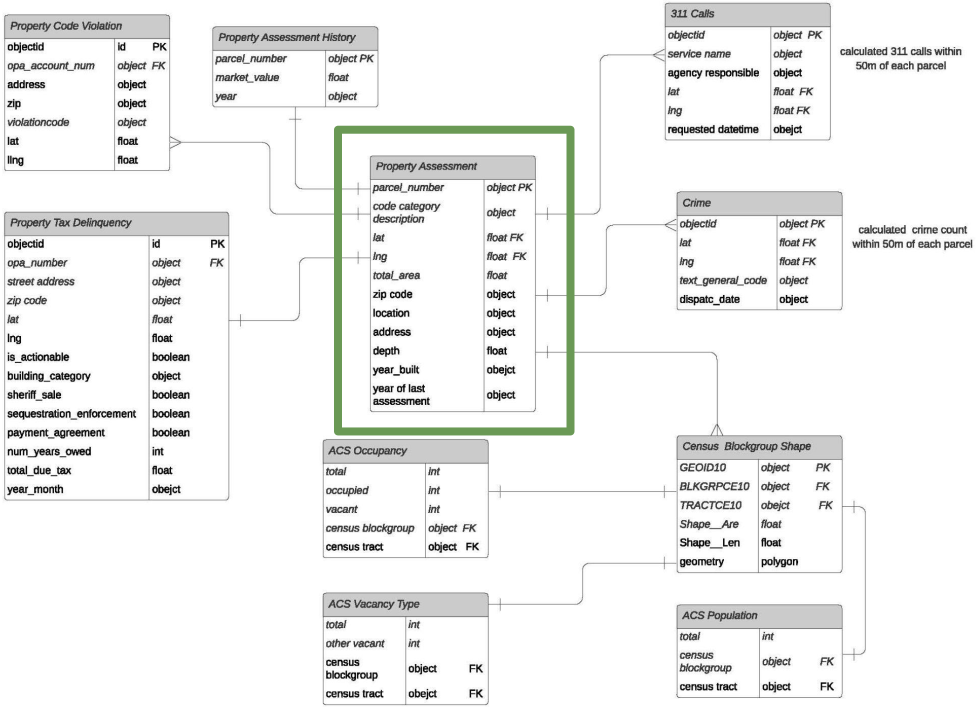
Volume of Service Requested from year 2019

- Big percentage of the lat and lng data was missing in this dataset
- A lot of the service names were “Information Request”.
- Most of the calls were for the Streets department
- 2018 and 2020 had highest number of 311 calls
- After grouping the number of 311 calls that happened within 50m of each parcel number, we noticed that the data is quite skewed.

Feature engineering, preprocessing, and modeling



Connecting Datasets



Important Data Cleaning and Preprocessing

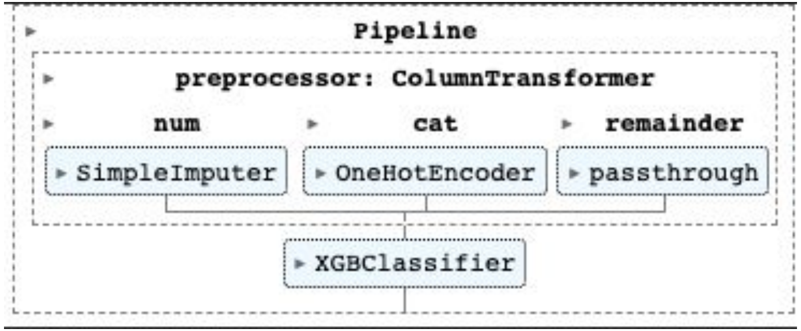
- Crime Dataset:
  - Dropped rows that had null values in lat and lng columns as this data was very important for us to join it with the rest of the datasets
  - Removed lat and lng values that look unusual or fell outside of Philadelphia region
- Property Assessment Dataset:
  - Dropped properties with unknown lat and lng data as this was very important to join other datasets
- 311 Call Dataset:
  - Removed "Information Request" service as it was not related to vacant lots. This service had the highest number of lat and lng as null values
  - Removed rows that did not have lat and lng data. Also, removed rows where the lat and lng looked unusual or fell outside of Philadelphia region.

Important Feature Engineering

- Percentage of "other" vacant lots of total places in each block group
- Population density in each block group
- Created variable to find out the number of time vacant lot related property violation appeared in the property
- Crime and 311 calls within 50m of each property. The data was broken down by last 6 month, last 3.5 year and full period
- Replace some null values with 0
- Replaced other null values with median
- Removed variables with high correlation
- "Category_code_description" variable from Property Assessment dataset was used to create labelled column Y for ML models
- Removed variables that were related to vacant lots other than the labelled column
- One Hot Encoding
- Power Transformation

Modeling trials

- Logistic Regression
  - No scaler or power transformation
  - Standard scaler
  - Quantile Transformation
  - Power Transformation
- AdaBoost
  - Logistic Regression using Power Transformation
  - Decision Tree
- SVC
  - No power transformation and balanced dataset class weight
  - Power transformation and balanced dataset class weight



Pipeline used on XGBoost Model

- XGBoost (Best Model)
  - Used default values
  - Specific values
- Random Forest Classifier
  - Numerical columns
  - Removing columns with high number of categories

Modeling Conclusion

- As we were using unbalanced dataset, we looked closely at precision and recall score of vacant lots to evaluate models.
- Tree Based Ensemble models tend to perform well
- XGBoost performed the best
- Logistic Regression did not perform well. It only performed well when we used Quantile Transformation or Power Transformation.
- Logistics regression when used on Adaboost did not perform well
- SVC only performed well when we used power transformation and balanced dataset class weight

Model Comparison

	Vacant Lots		
	precision	**recall**	**f1**
Logistic Regression (Power Transformation)	0.90	0.84	0.87
SVC	0.83	0.97	0.89
Adaboost (Logistic Regression)	0.86	0.45	0.59
Adaboost (Decision Tree)	0.95	0.88	0.91
Random Forest	0.98	0.89	0.93
XGBoost	0.97	0.96	0.96

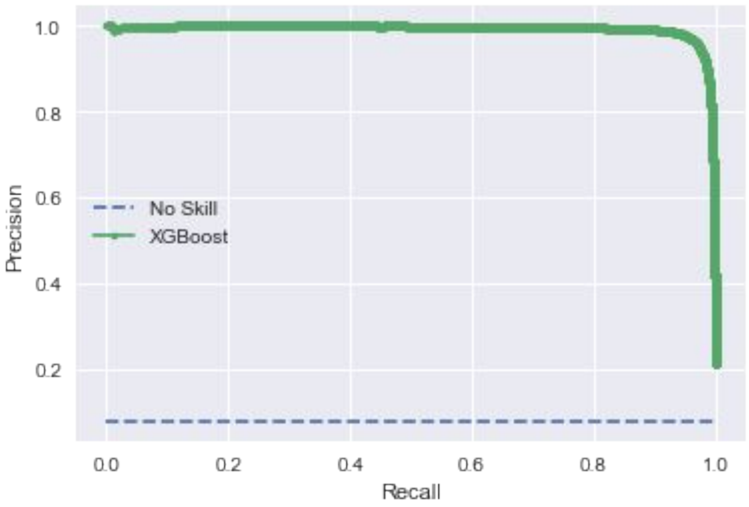
XGBoost (used some specific values) - BEST MODEL

```
model score on train: 0.998
model score on test: 0.994
[[84634   207]
 [  313  7027]]
                   precision    recall  f1-score   support

 Not Vacant Lots        1.00      1.00      1.00     84841
     Vacant Lots        0.97      0.96      0.96      7340

        accuracy                            0.99     92181
       macro avg        0.98      0.98      0.98     92181
    weighted avg        0.99      0.99      0.99     92181
```

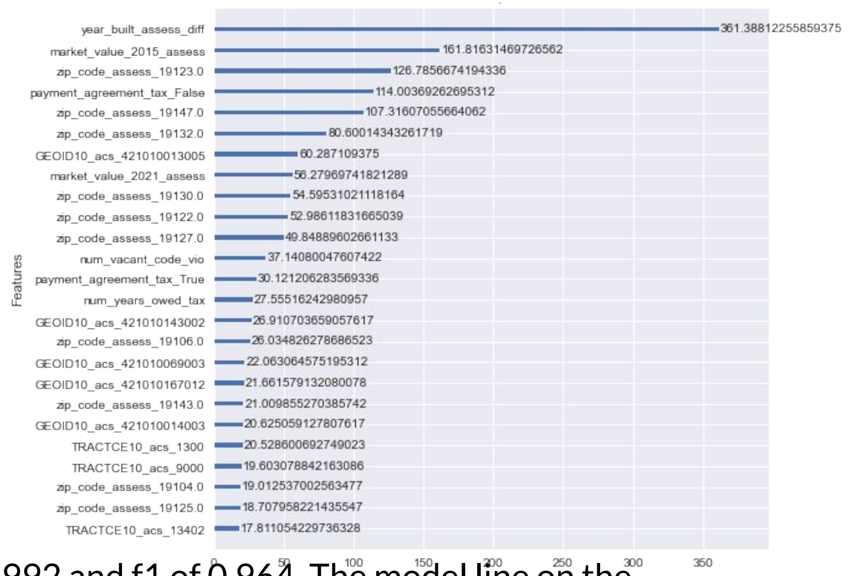
- Used values within XGBClassifier. Source: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
- Replaced null values with median
- Used one-hot encoding on category values

XGBoost Result

Precision-Recall Graph



Feature Importance(top 25)



- XGBoost model had a very high AUC of 0.992 and f1 of 0.964. The model line on the precision-recall graph has been way above the no-skill threshold.
- Year built of property was the most important feature that improved the accuracy of score the most. A lot of zip codes and GEOIDs were also important features

Discussion

Discussion

- Model runs very slow so would recommend using distributed processing system like Apache Spark in future
- Joining geospatial data using GeoPanda is very slow so would recommend using GIS software to make the process faster
- Future projects can include other census datasets like education, poverty, household income, race, poverty status.
- Project can also incorporate data from Openstreetmap.
- Test models on neural network.
- Compare model result with vacant property model created by City of Philadelphia Office of Innovation and Technology.
- Expand the research to other cities using the same methodology

Deployment

Deployment

Predicting Vacant Lots in Philadelphia

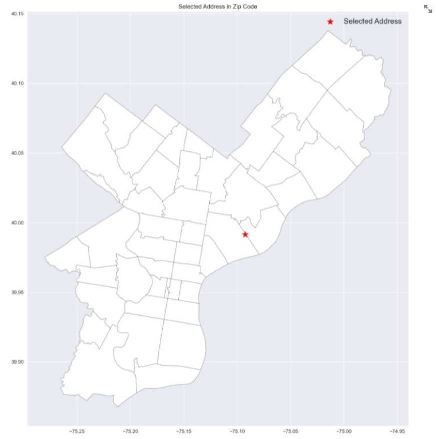
What is the address of the place? Please select one address only.

5228 F ST ×

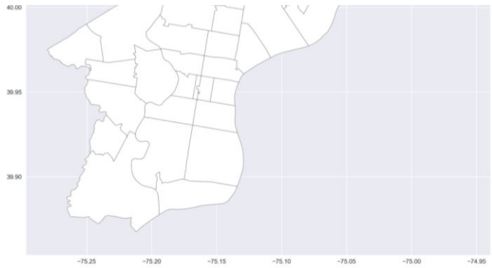
We found the address 5228 F ST 😊!

The following parcel numbers will be considered for the model:

	Parcel Number	Address	Zip Code
444041	351338100	5228 F ST	19124



Predict



Predict

Your address is NOT VACANT LOT

Top 10 reason impacting this result are:

	Reasons
0	Age of Property
1	Market Value of Property in 2015
2	Zip Code = 19123
3	Did not have tax payment agreement
4	Zip Code = 19147
5	Zip Code = 19132
6	Census GEOID = 421010013005
7	Market Value of Property in 2021
8	Zip Code = 19130
9	Zip Code = 19122

- Used Steamlit which easily turns python script into shareable web apps.
- Downloaded the best model using pickle
- Currently uses pre-existing joined file to run model.

Deployment: Video Link

Thank you