New York State Bridge Analysis

Scripting for Data Analysis – IST652

Professor Landowski

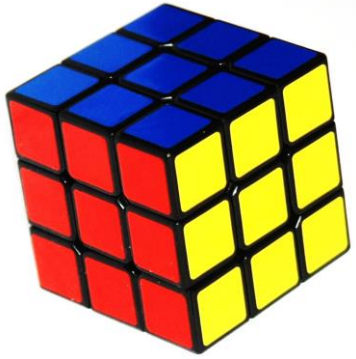
Brian Hogan & Katie Poole

New York Bridge State Analysis

Analysis performed to distinguish which bridges and counties priority condition rating. Needing most attention a function of funding, location, and traffic use.

Complicated

Simple



- Data Sets (2017 & 2018):
  - NYS Bridge Data
  - Population by County
  - Car Traffic Flow
- Selected variables: condition; geocodes; traffic volumes
  - Priority rating formula includes bridge condition rate and AADT
- Data consolidation & spatial viewpoints in QGIS software
- Project focus: priority rating, location, and developing an understanding of population proximity to bridge condition

✓Resulting data set: ~20,000 records x 42 variables

References: https://www.dot.ny.gov/divisions/engineering/structures/repository/manuals/inventory/rc01_june06.pdf

Data Sources & Data Processing

Emphasis was placed on reading data in Python, working with data frames, and joining tables for road map visualization in QGIS software.

Munging…

- For QGIS shapefile converted to a csv w 18 fields; NAs were removed; used data types
- Data joined by county name; word county removed; names renamed to lower case; 10 digits
- 600+ Tweets pulled across a 10 day period across ~15 Twitter accounts

New York State Bridge Data (QGIS) (main data set)

- Main source of bridge statistics. Data library 23 fields: ID, lat, long, status, priority rating, etc
- https://www.dot.ny.gov/divisions/engineering/structures/repository/manuals/inventory/rc01_june06.pdf

New York State Average Annualized Daily Traffic (AADT)

- Total volume of vehicle traffic by road by year divided by 365 days
- https://www.dot.ny.gov/divisions/engineering/applications/traffic-data-viewer/tdv-definitions#AADT

New York State Population by County (and/or Department of Transportation Vehicle Ownership Stats)

- Population by county metrics may be useful for building new variables assessing ratios of populations, vehicle ownership counts, traffic throughput statistics, and similar.
- https://www.labor.ny.gov/stats/nys/statewide-population-data.shtm

Twitter / RSS /

- Team would like to gather text chatter from local news on road traffic status.

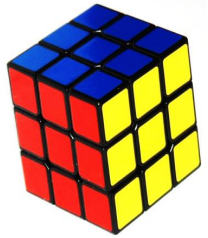
Bridge, population, and traffic data reside in well maintained tables (nominal cleaning)

Utilized a *good, bad, ugly* data classification approach

Metrics

Complicate

Simple



Analysis Questions

- Is there a correlation that can be seen visually between the county population and the number of low-quality bridges?
- What are the top 10 highest priority bridges?
- Which counties do we identify as our Good, Bad and Ugly counties?
  - High, medium and low percent bridge priorities
- Overall, how many bridges are ranked high medium and low priority?
- Which county has the highest count of high priority bridges?
- What is the sentiment analysis of Twitter traffic tweets?
- What factors help predict a bridge’s condition rating?

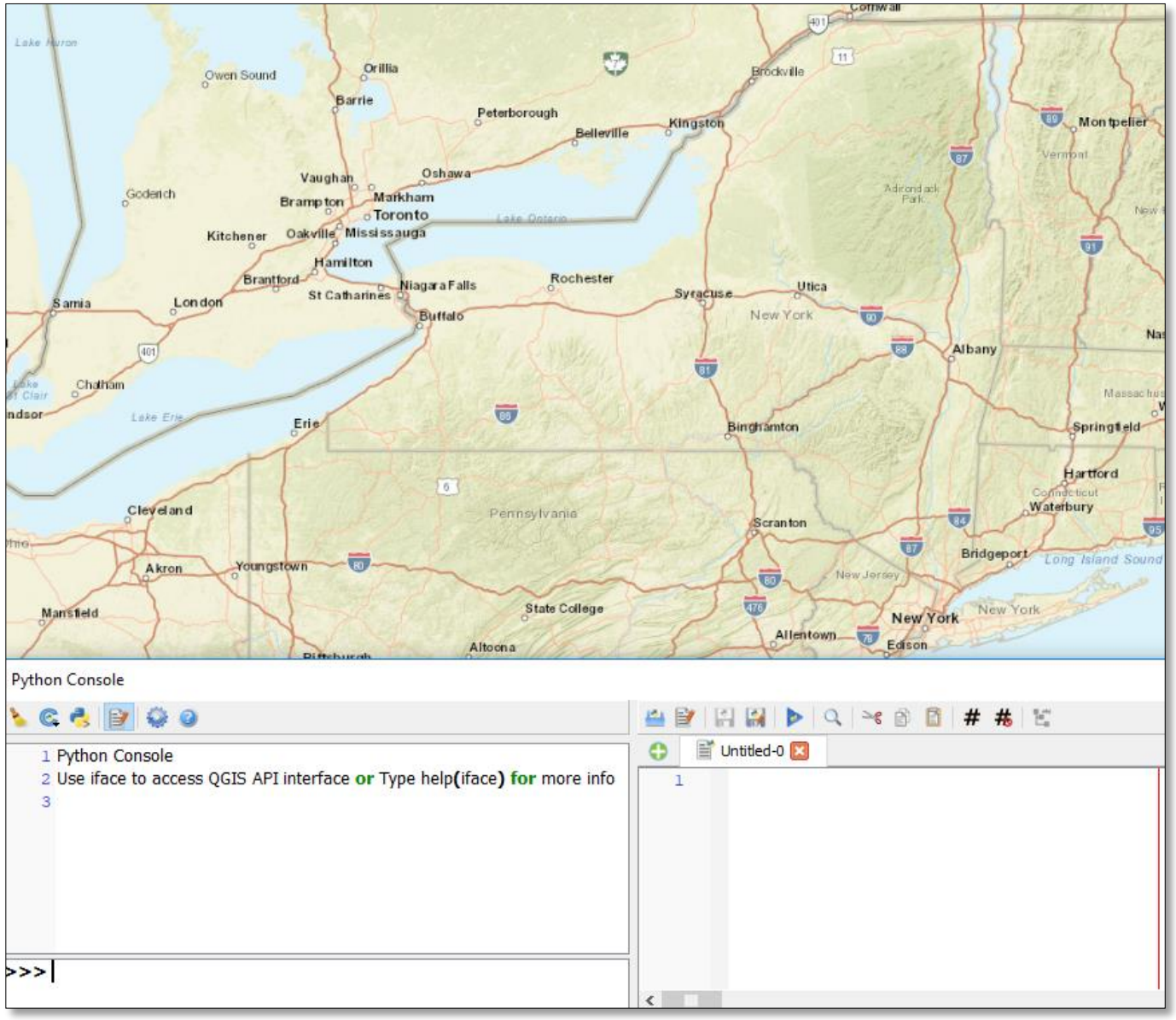
Description of Program

1. Preprocess
2. Bridge Analysis
3. Twitter Analysis
4. Prediction

- **Python cleanup for QGIS Analysis**
- **Data scaling (population, AADT rating)**
- **Bridge shapefile creation**
- **Traffic flow visualization**
- **AADT – annualized traffic - analysis**
- **Spatial data join by bridge**
- **Bridge & traffic flow finalization**
- **Tweet results over 15 day period**
- **Good, bad ugly name binning but had issue with LOTS of misc. characters**
- **Sentiment analysis**
- **Bridge condition rating & prediction work**

Why QGIS and Analysis Questions

- QGIS is an opensource GIS Software, where geospatial data can be created, edited, visualized and analyzed.
- PyQGIS is a Python environment in QGIS, can process and analyze geospatial data
- Could've used geopandas instead, however more powerful visualization capabilities
- For fun and work related



New York State Population

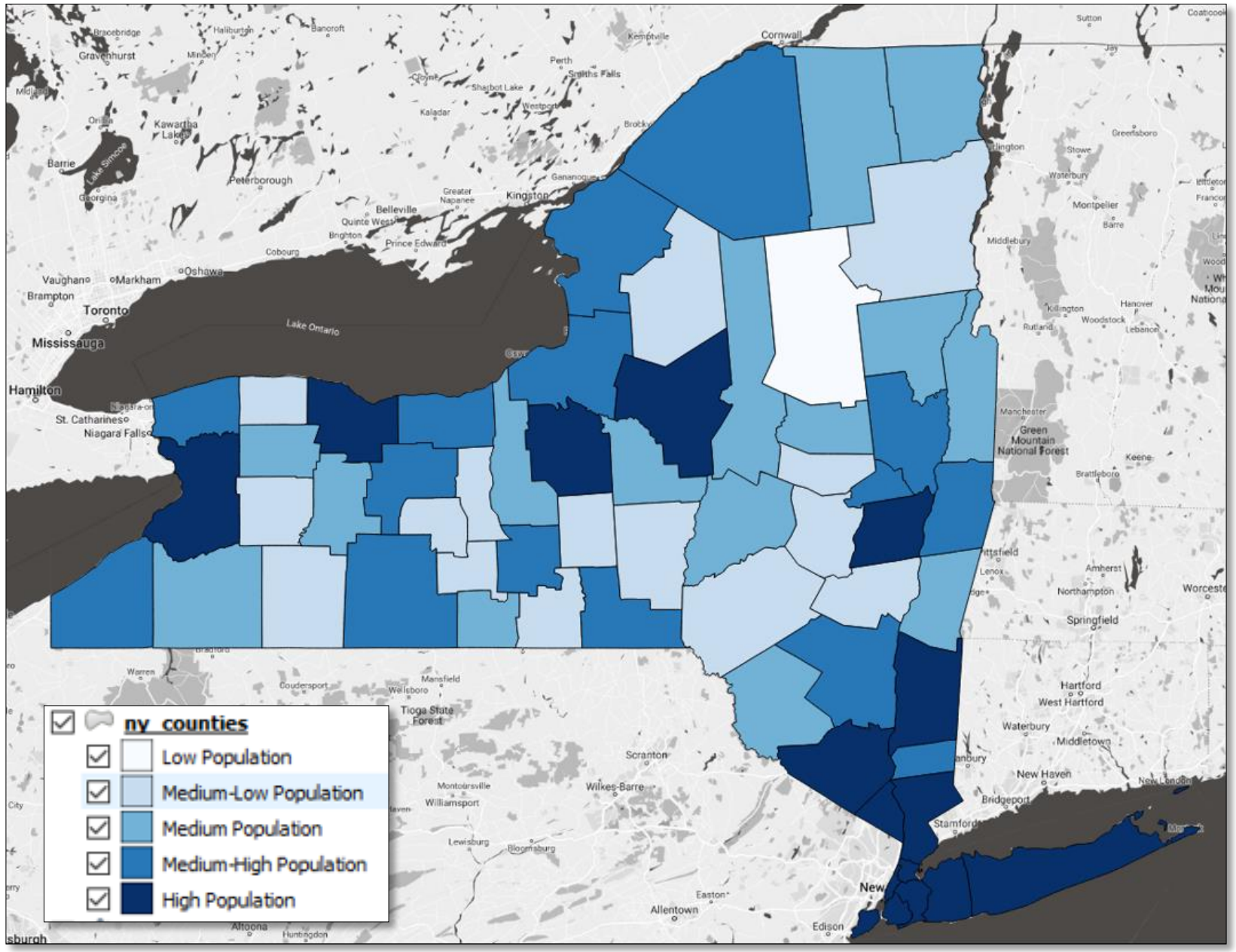
In order to produce this visualization, a table join was performed between the county shapefile and population csv

- Joined on county name data

Perhaps there is a correlation between population and bridge condition?

PyQGIS used to join the two tables, bucketed the data and visualize

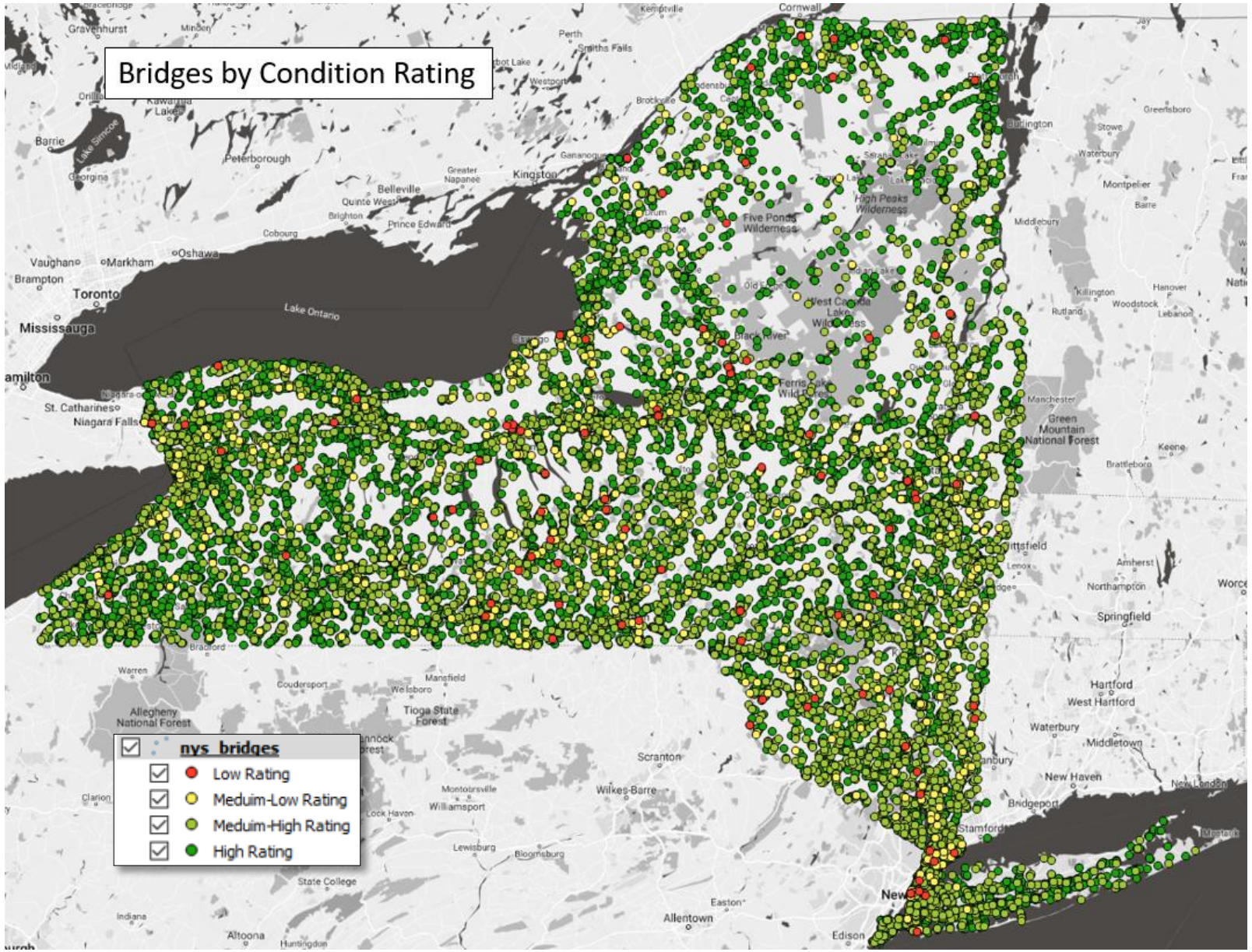
The county and population data will be used later for further analysis



Visual: Bridges by Condition Rating – Density Plot

Working with the Bridge Data:

- Downloads as a Geodatabase file
- Had to be cleaned
  - NA's removed
  - Data types adjusted
- Exported as a csv
- Plotted in QGIS via coordinates
- Bucketed data
  - Used "equal breaks"
  - Buckets were 0-1.75, 1.76-3.5, 3.51-5.25, 5.26-7
- Visualized in QGIS
  - Density Plot
  - Condition Rating
- Looking at Condition Data alone, not many low rated bridges

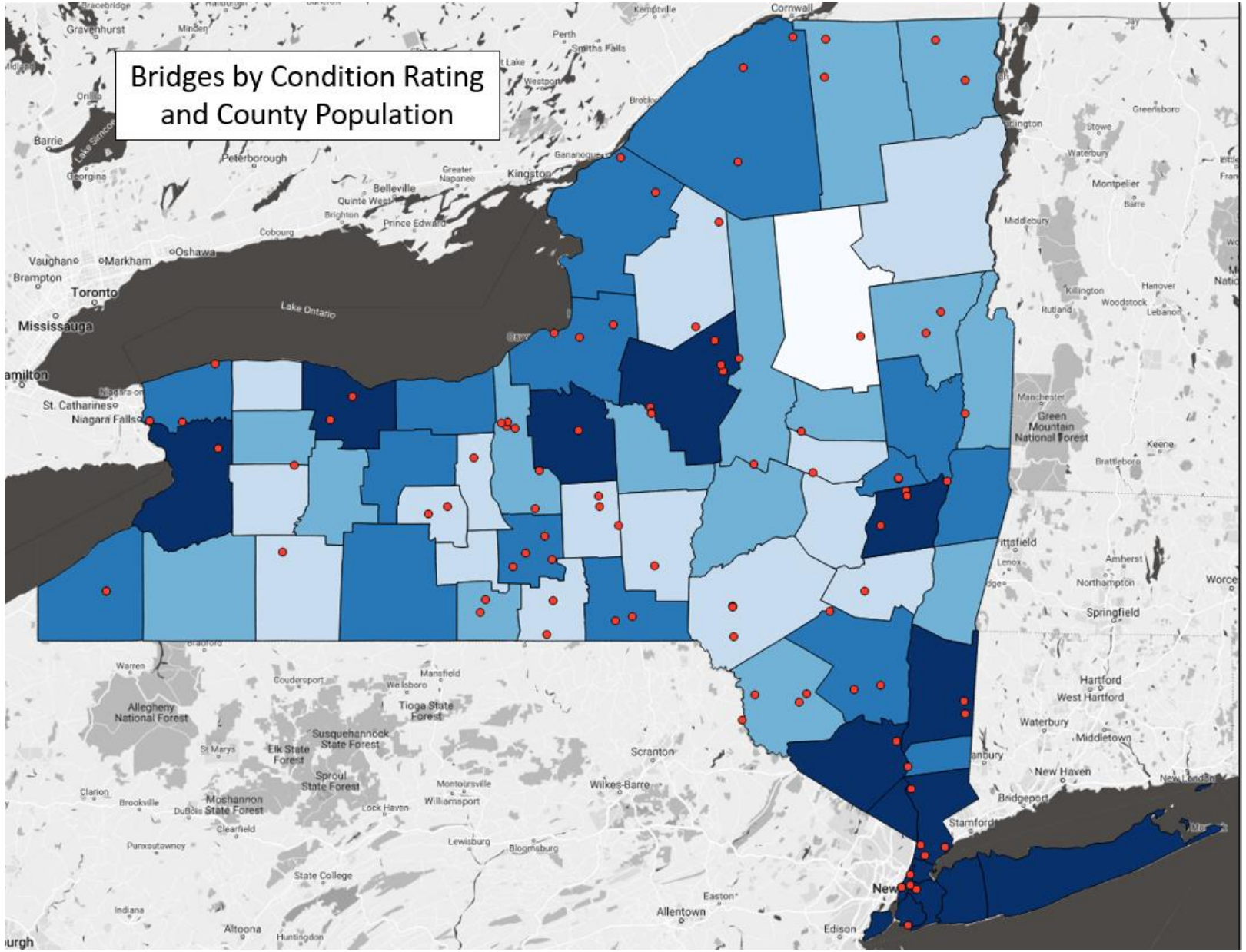


Correlation between Population and Low-quality Bridges?

Filtered out the lowest-quality bridges based upon Condition Rating

Other than New York City area, doesn't visually appear that highly populated areas indicate more low-quality bridges.

Final analysis to consider Traffic and Condition Rating

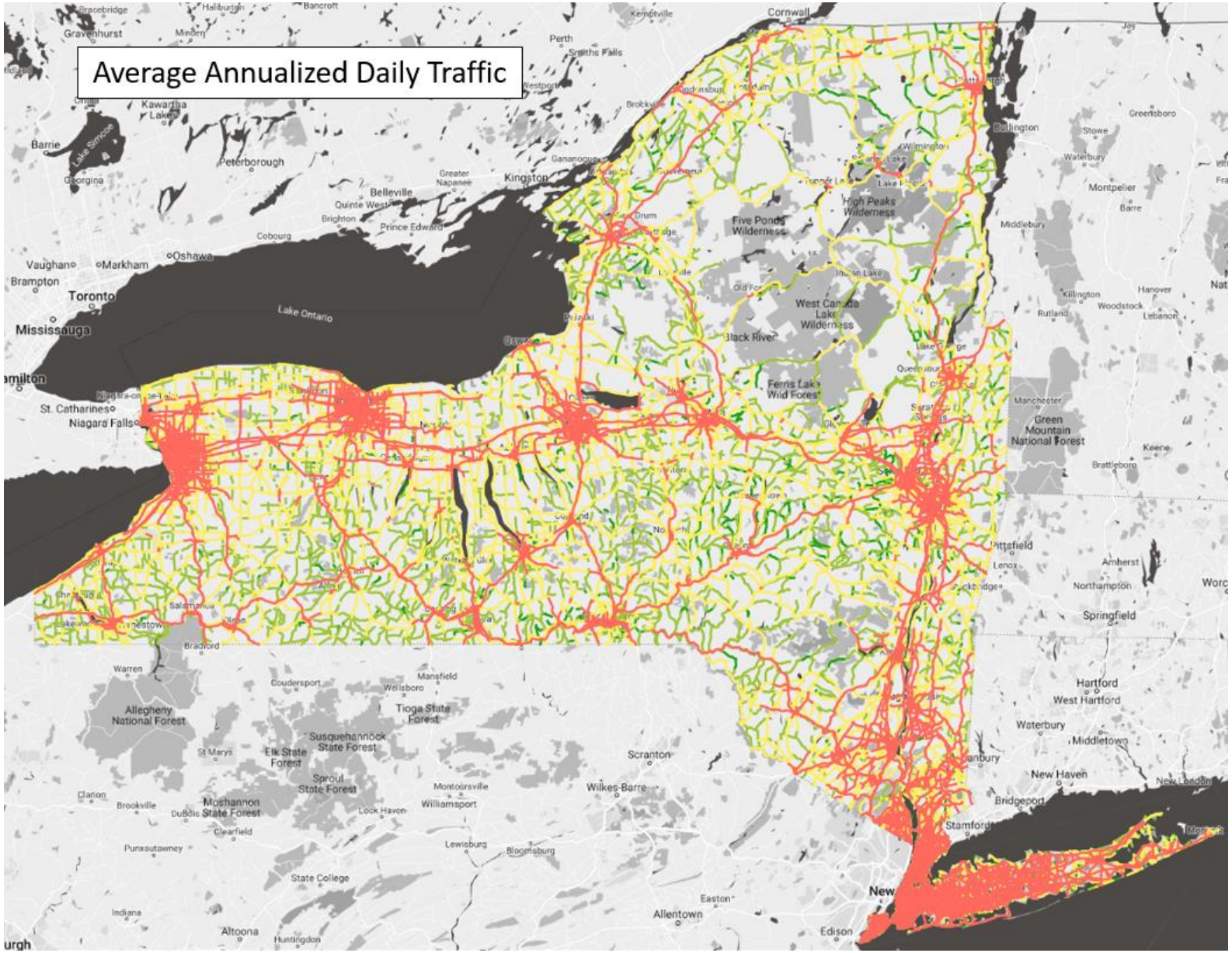


AADT – Average Annualized Daily Traffic

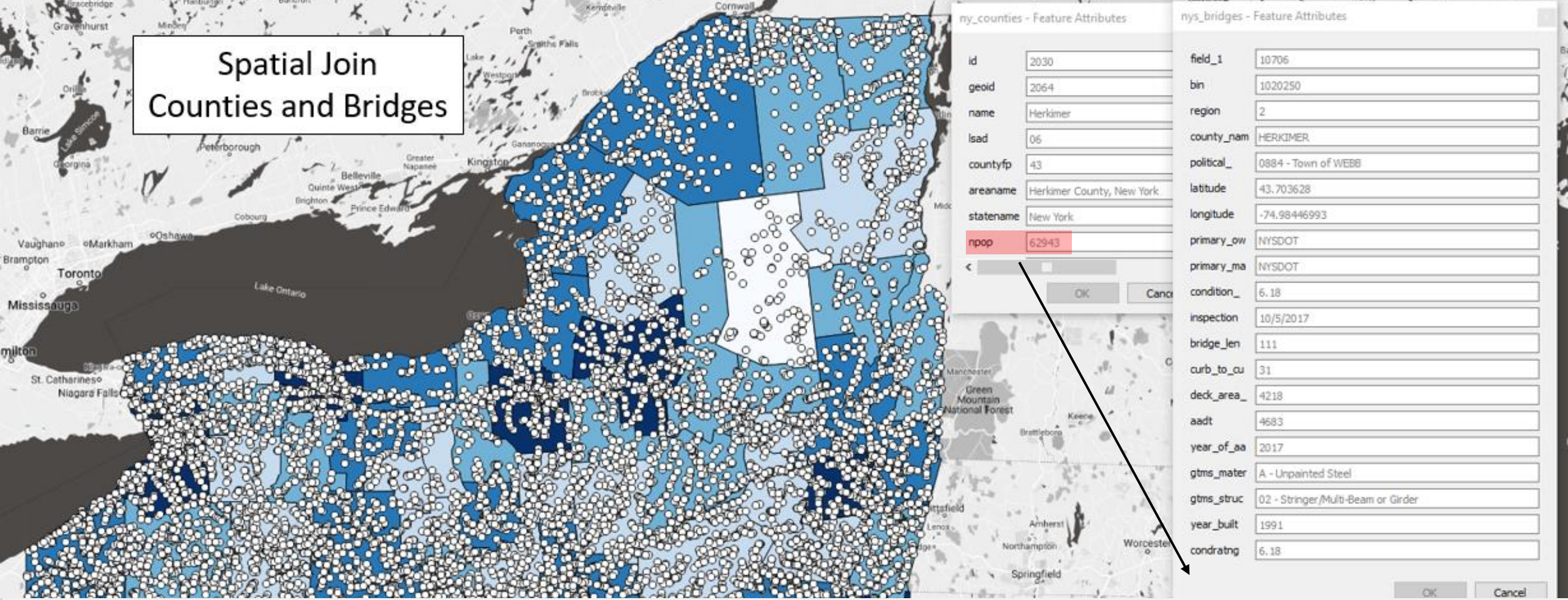
The other important component of the analysis: AADT

For our analysis, bridges on high traffic roads are considered more important than bridges on low traffic roads

Imported a shapefile from the New York State GIS Clearinghouse, bucketed AADT level to create visualization



Spatial Join Counties and Bridges



joinresult :: Features total: 16637, filtered: 16637, selected: 0

	primary_ow	primary_ma	condition_	inspection	bridge_len	curb_to_cu	deck_area_	aadt	year_of_aa	gtms_mater	gtms_struc	year_built	condratng	id	geoid	name	lsad	countyfp	areaname	statename	npop
1	NYSDOT	12 - State - Subc...	7	9/27/2017	102	34	3860	162016	2017	5 - Prestresse...	05 - Box Beam o...	2005	7.000000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759
2	NYSDOT	12 - State - Subc...	6.48	4/30/2018	77	160.8	12936	142609	2011	5 - Prestresse...	05 - Box Beam o...	2014	6.480000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759
3	NYSDOT	12 - State - Subc...	7	7/11/2018	102	34	3862	126055	2011	5 - Prestresse...	05 - Box Beam o...	2005	7.000000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759
4	NYSDOT	12 - State - Subc...	5.34	4/12/2018	68	168	12009	126055	2011	5 - Prestresse...	05 - Box Beam o...	2014	5.340000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759
5	NYSDOT	12 - State - Subc...	5.66	11/3/2017	21	40	5376	107537	2017	2 - Concrete (...	19 - Culvert	1968	5.660000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759
6	NYSDOT	12 - State - Subc...	6.53	1/12/2018	207	50	11600	97460	2011	3 - Steel	02 - Stringer/Mul...	1962	6.530000000...	2182	2214	Richmond	06	85	Richmond County, New York	New York	4759

Final Analysis – Find the lowest-quality Bridges

joinresult

Cleaned data

Normalized Data (to 0 to 1 scale)
Condition Rating: 0-7 scale
AADT: 0-276476

Summed Normalized Data to get Official Rating figure

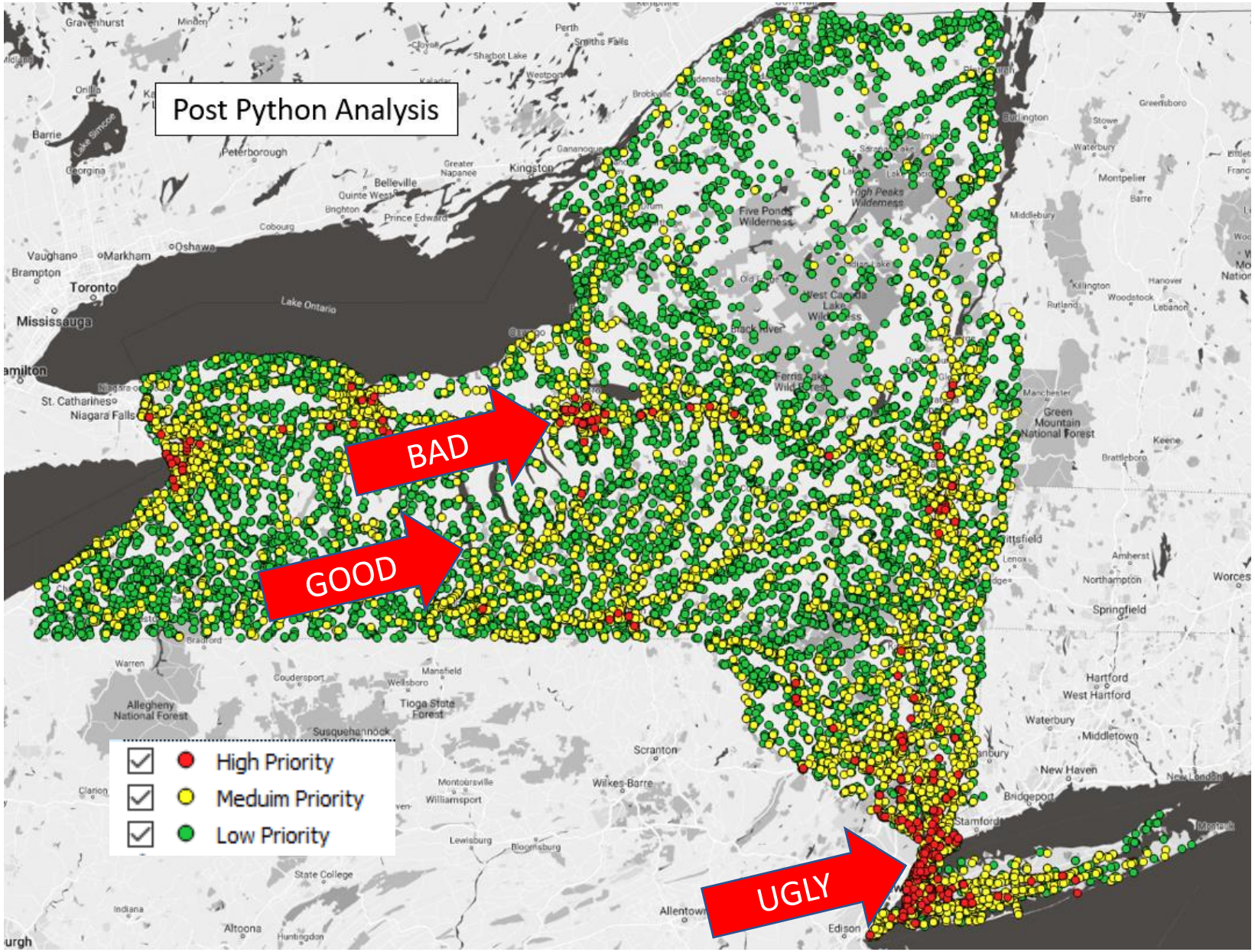
Bucketed and Categorized to High, Medium and Low Priority Bridges

Analysis_R...

Final Analysis Results – The Good, the Bad and the Ugly

The below table shows percent of High, Low and Medium priority Bridges

county_name	high	low	medium
NEW YORK	0.10	0.26	0.63
KINGS	0.09	0.33	0.57
BRONX	0.07	0.28	0.65
WESTCHESTER	0.06	0.31	0.63
ROCKLAND	0.06	0.33	0.61
ONONDAGA	0.05	0.43	0.52
QUEENS	0.05	0.27	0.68
RICHMOND	0.04	0.42	0.54
SENECA	0.03	0.76	0.20
NASSAU	0.03	0.31	0.66
SARATOGA	0.03	0.66	0.31
ORANGE	0.02	0.45	0.52
ALBANY	0.02	0.45	0.52
PUTNAM	0.02	0.33	0.65
ERIE	0.02	0.57	0.41
SUFFOLK	0.02	0.40	0.58
ONTARIO	0.02	0.70	0.29
RENSSELAER	0.02	0.60	0.38
CAYUGA	0.01	0.73	0.26
MONROE	0.01	0.51	0.48
MADISON	0.01	0.78	0.21
SULLIVAN	0.01	0.75	0.24
ULSTER	0.01	0.67	0.32
DELAWARE	0.00	0.84	0.16
FULTON	0.00	0.85	0.15
CHAUTAUQUA	0.00	0.86	0.14
WYOMING	0.00	0.88	0.12
ALLEGANY	0.00	0.89	0.11
ORLEANS	0.00	0.89	0.11
ST LAWRENCE	0.00	0.92	0.08
CATTARAUGUS	0.00	0.92	0.08
LEWIS	0.00	0.93	0.07
CLINTON	0.00	0.94	0.06
FRANKLIN	0.00	0.95	0.05
HAMILTON	0.00	0.95	0.05
YATES	0.00	0.96	0.04

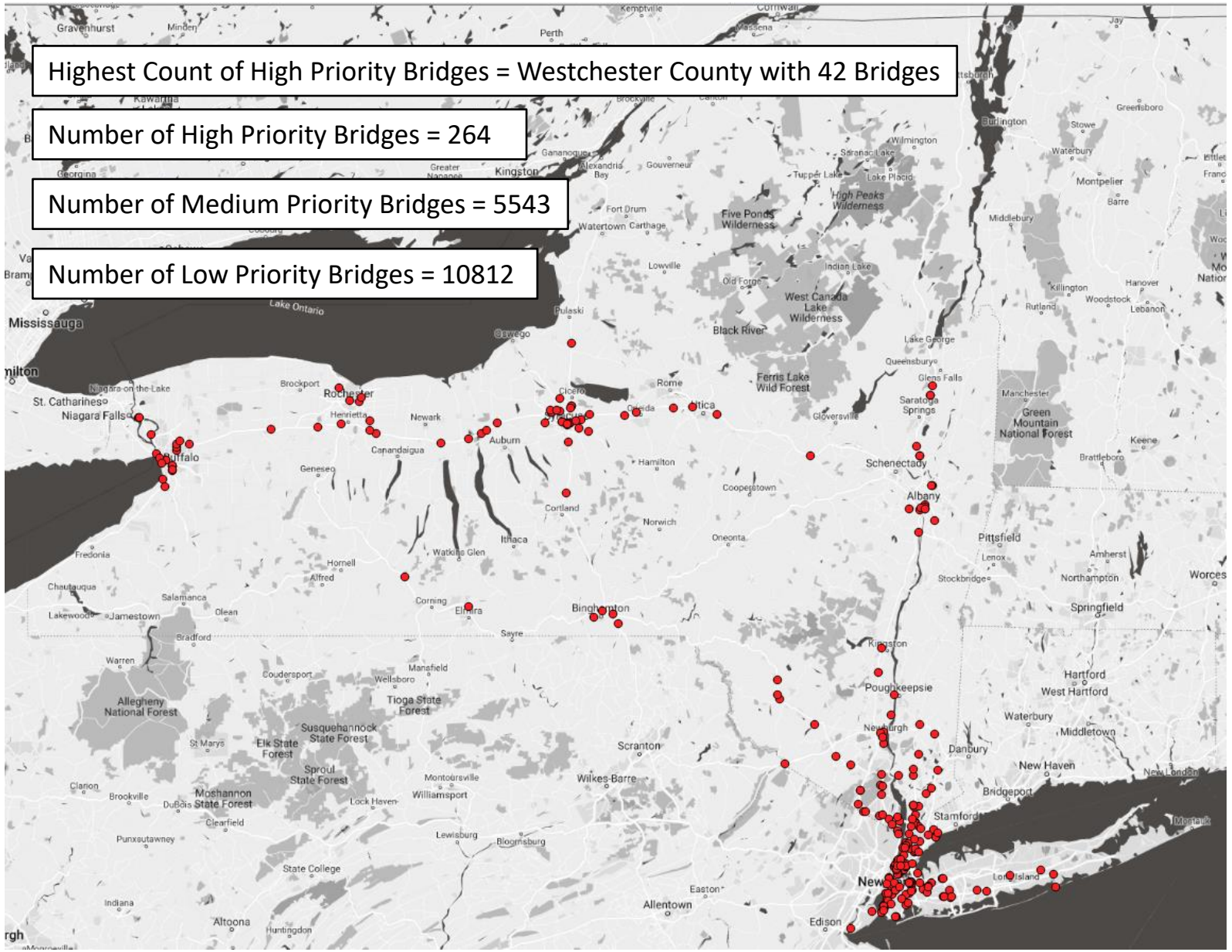


Final Analysis Results

The top 10 worst bridges are in downstate NY, all within the City of New York or nearby

Outside of New York City, there are visual patterns: Many high priority bridges appear to be associated to major highways such as NY-90, NY-390 and NY-87

	bin	county_name	political_unit	total_rank
53	2231439	KINGS	2034 - City of NEW YORK	1.971147
15	1065318	KINGS	2034 - City of NEW YORK	1.969974
70	106531B	KINGS	2034 - City of NEW YORK	1.966604
80	5516340	ROCKLAND	0615 - Town of ORANGETOWN	1.963867
162	2229289	NEW YORK	2034 - City of NEW YORK	1.962122
153	2268650	NEW YORK	2034 - City of NEW YORK	1.961550
25	5521218	RICHMOND	2034 - City of NEW YORK	1.959444
230	2065629	BRONX	2034 - City of NEW YORK	1.959263
22	5521217	RICHMOND	2034 - City of NEW YORK	1.952614
23	552121E	RICHMOND	2034 - City of NEW YORK	1.952494



What Predicts Bridge Condition Rating ?

Variable *county* plays a significant role in model's R2 calc

OLS Regression Results

Dep. Variable:	condition_rate	R-squared:	0.702
Model:	OLS	Adj. R-squared:	0.627
Method:	Least Squares	F-statistic:	9.421
Date:	Fri, 07 Jun 2019	Prob (F-statistic):	2.35e-05
Time:	17:45:38	Log-Likelihood:	-21.783
No. Observations:	31	AIC:	57.57
Df Residuals:	24	BIC:	67.60
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-27.0633	13.526	-2.001	0.057	-54.979	0.852
county	0.0146	0.011	1.319	0.199	-0.008	0.038
curb_to_curb_width_ft	0.0072	0.002	3.089	0.005	0.002	0.012
deck_area_sq_ft	-1.172e-06	3.1e-07	-3.781	0.001	-1.81e-06	-5.32e-07
material	-0.1571	0.068	-2.321	0.029	-0.297	-0.017
structure	-0.0573	0.027	-2.092	0.047	-0.114	-0.001
year	0.0165	0.007	2.393	0.025	0.002	0.031

Omnibus:	3.443	Durbin-Watson:	1.925
Prob(Omnibus):	0.179	Jarque-Bera (JB):	2.280
Skew:	-0.647	Prob(JB):	0.320
Kurtosis:	3.298	Cond. No.	5.24e+07

- **Manual iteration was performed amongst the variables to assess model influence and P <0.05 evaluations as the initial model had a high degree of variability explained at an R^2 of 0.791. Variables didn't change significantly R^2 until the variable "county" was dropped.**

- **The "county" variable has the highest significance even though its "insignificant!**
  - *in terms of a linear model suggesting "location" of a bridge plays a critical role in both usage, value, work performed, and rating. Effort was spent on the graphing of the linear model but had difficulty achieving the visualization in Python.*

- **In our final model, although the "county" variable isn't significant we wouldn't have a prediction without it. "County" captures bridge flow traffic (aadt) and population dynamics impacting bridge condition.**

Predicting existing bridge future conditions becomes a function of its (year built + structure + material & square footage). Bridge inspection dates and rating are by products of the structure itself.

Traffic Tweets

	Word	Freq	Word	Freq
Word:	with	29	nbamlb	24
Word:	my	29	near	22
Word:	s	29	that	21
Word:	traffic	28	your	21
Word:	just	28	me	20
Word:	jasonnym	28	…	20
Word:	lisalgm1	28	&	20
Word:	thruwaytraffic	25	amp	20
Word:	(	24	he	20
Word:	wnyt	24	tractor	20
Word:	trafficmanmatt	244	trailer	20

- Twitter preprocessing turned into be quite a learning situation as removal of characters of detracting value was difficult. These speaks to using techniques, such as regedit & regular expressions, to help address unique situations.
- For example: a prominent “tweeter” is “@Trafficmanmatt.” This individual is very active on traffic states providing good volume and quality texts.
- However, one of his approaches is to include his handle “@Trafficmanmatt” in his tweet to build his brand. Such learnings confirm need to scrub and learn data pull nuisances to result in more significant text and sentiment data mining.

Example of a Good Data Pull (of a Bad day)		Example of What Happens When Expressions Not Addressed	
Word	Count	Word	Count
hudson	9	@	1950
valley	9	:	475
lower	8	https	249
to	7	the	226
nb	7	.	217
i-87	7	trafficmanmatt	214
traffic	7	,	195
sb	7	!	170
blocked	6	rt	159
#	5	suzan916	116
a	5	to	110
accident	5	a	104
slow	5	frecklequeen45	102
service	5	dizzymom64	100

Conclusions:

- No distinct spatial bridge patterns by condition rating
- Good, bad, and ugly top contender?
- Yes you guessed it = New York City
- Overall: 2% high priority(264), 33% medium (5543), & 65% low or 10812 bridges
- Twitter word heat maps would be an ideal means to consolidate tweets
- “Population” near a bridge didn’t contribute meaningfully to its condition