NEW YORK CITY RESTAURANT INSPECTIONS

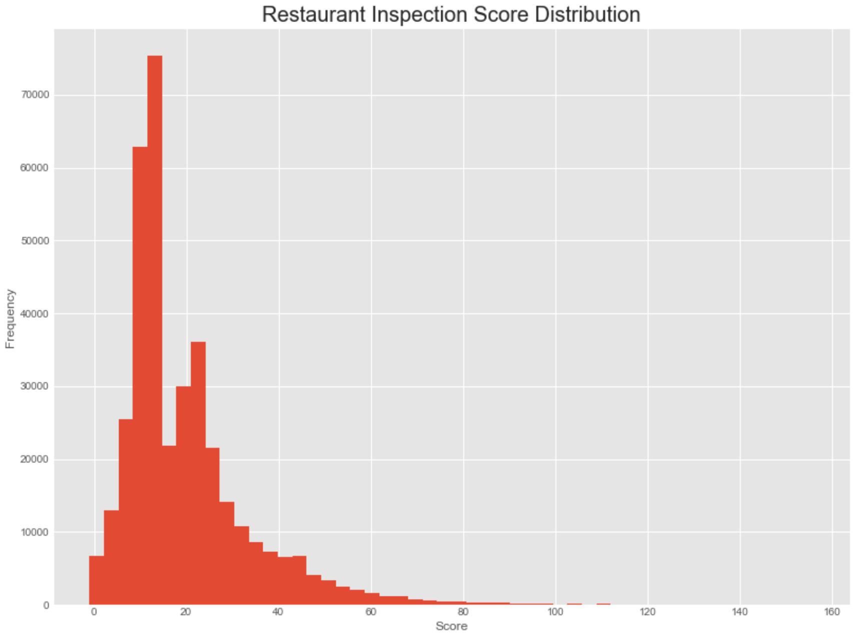
Joyce Fang

AGENDA

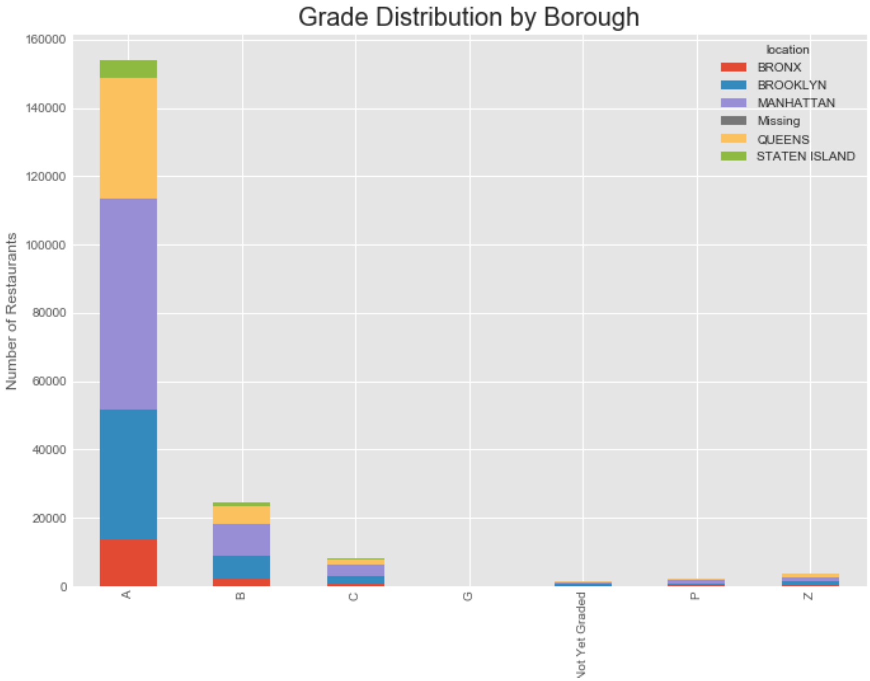
- About the Data
- Clean the Data
- Reorganize the Data
- Model the Data
- Takeaways & Afterthoughts

ABOUT THE DATA

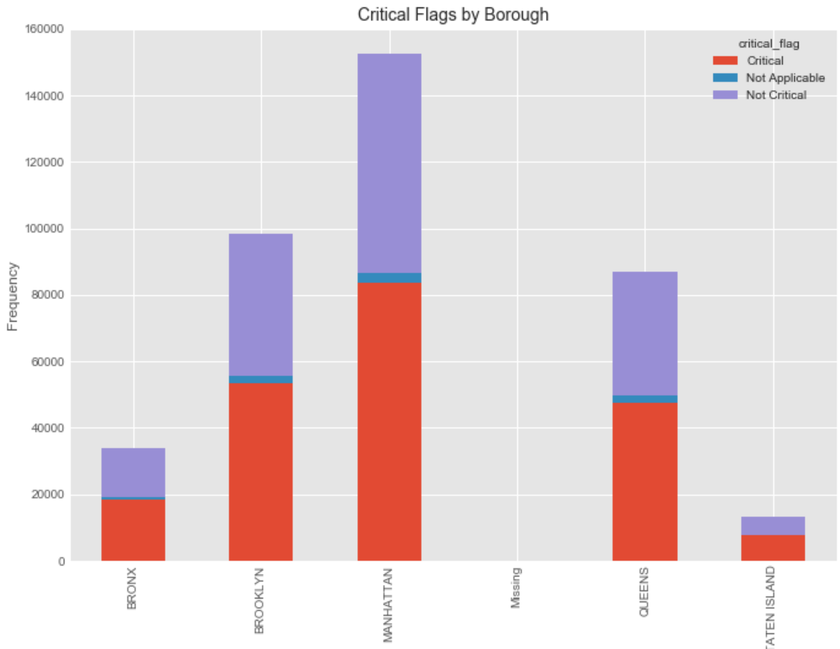
- Last update: March 18, 2019
- Provided by the NYC Department of Health and Mental Hygiene
- 18 Features: id, name, boro, bldg_num, street, zipcode, phone, cuisine, inspect_date, action, violation_code, violation_desc, critical_flag, score, grade, grade_date, record_date, inspect_type
- 385k rows



Restaurant Inspection Score Distribution



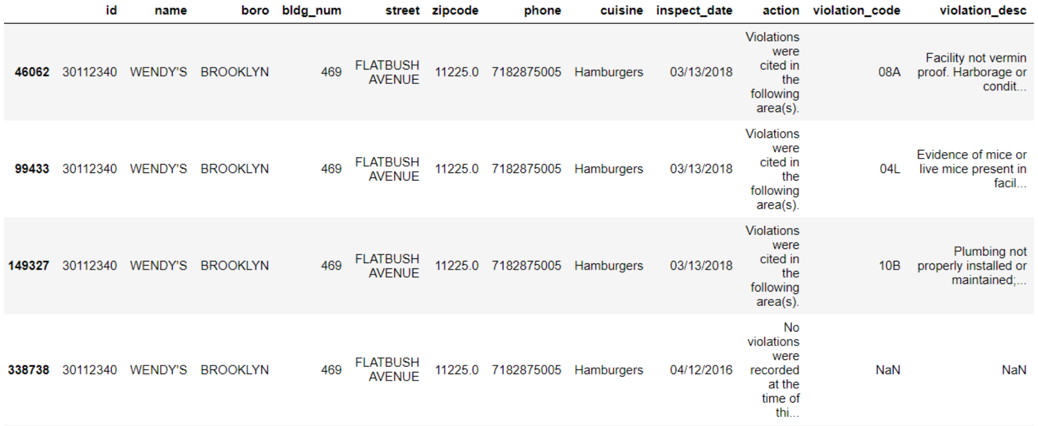
Grade Distribution by Borough



Critical Flags by Borough

CLEAN THE DATA

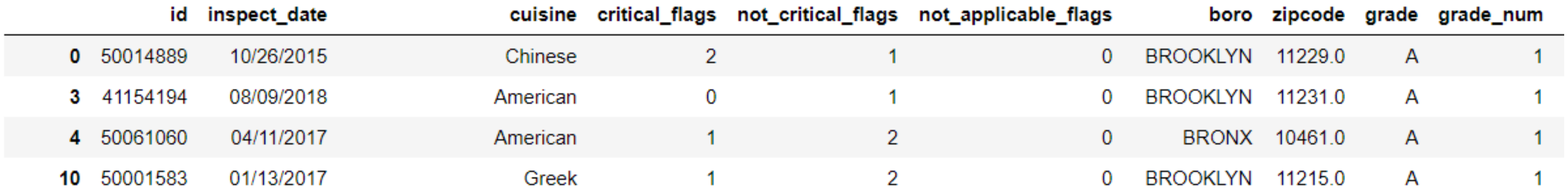
- Each row represents a violation/citation
- Each restaurant can have multiple citations and inspections in one day
- One grade for one inspection type in one day
- Some inspections do not produce grades
  - Dropped citations with no grades or pending grades



	id	name	boro	bldg_num	street	zipcode	phone	cuisine	inspect_date	action	violation_code	violation_desc
46062	30112340	WENDY'S	BROOKLYN	469	FLATBUSH AVENUE	11225.0	7182875005	Hamburgers	03/13/2018	Violations were cited in the following area(s).	08A	Facility not vermin proof. Harborage or condit...
99433	30112340	WENDY'S	BROOKLYN	469	FLATBUSH AVENUE	11225.0	7182875005	Hamburgers	03/13/2018	Violations were cited in the following area(s).	04L	Evidence of mice or live mice present in facil...
149327	30112340	WENDY'S	BROOKLYN	469	FLATBUSH AVENUE	11225.0	7182875005	Hamburgers	03/13/2018	Violations were cited in the following area(s).	10B	Plumbing not properly installed or maintained;...
338738	30112340	WENDY'S	BROOKLYN	469	FLATBUSH AVENUE	11225.0	7182875005	Hamburgers	04/12/2016	No violations were recorded at the time of thi...	NaN	NaN

REORGANIZE THE DATA

- New unique identifier: restaurant id and inspection date
  - Multiple inspection dates yields multiple rows of data
- Features used: id, inspect_date, cuisine, borough, critical_flags, not_critical_flags, not_applicable flags
- 80889 rows
- Critical flags vs. violations
  - 66 violations
- Model prediction: grade vs. score



	id	inspect_date	cuisine	critical_flags	not_critical_flags	not_applicable_flags	boro	zipcode	grade	grade_num
0	50014889	10/26/2015	Chinese	2	1	0	BROOKLYN	11229.0	A	1
3	41154194	08/09/2018	American	0	1	0	BROOKLYN	11231.0	A	1
4	50061060	04/11/2017	American	1	2	0	BRONX	10461.0	A	1
10	50001583	01/13/2017	Greek	1	2	0	BROOKLYN	11215.0	A	1

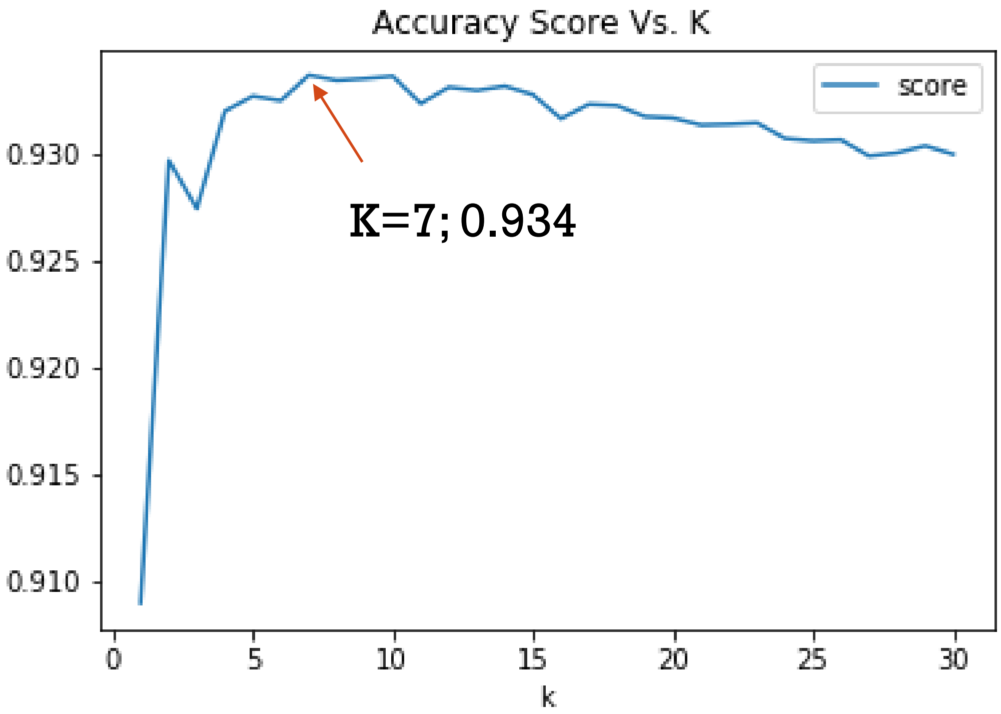
MODEL THE DATA

- Dropped citations that did not yield grades
- Dropped duplicates
- Separated out top ten cuisines and grouped the rest into 'Other'
- Filled in the missing boroughs
- Dummified categorical variables

K NEAREST NEIGHBORS

- Baseline score:0.894
- 5-fold, k=30
- Accuracy score: 0.933

- Exhaustive and time-consuming



K=7; 0.934

LOGISTICAL REGRESSION

- Training data accuracy score: 0.933
- Testing data accuracy score: 0.934
- Baseline score: 0.894
- Confusion Matrix
  - Accuracy: 0.934

	TPR	FPR
A	0.996	0.457
B	0.528	0.018
C	0.023	0.001

N = 20,223		Predicted		
		A	B	C
Actual	A	17,985	61	8
	B	785	889	8
	C	207	269	11

TAKEAWAYS & AFTERTHOUGHTS

- More efficient way to organize data
- Quantitative
  - Predict restaurant location based on score/cuisine
  - Look into why model was bad at predicting B and C grades
  - Try additional models
- Qualitative
  - Restaurant business is really tough