LOS ANGELES RESTAURANT DATA ANALYSIS

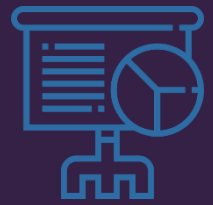
TEAM LINEAR DIGRESSORS

PROJECT OUTLINE

THREE STEPS WE TOOK TO SOLVE THE PROBLEM AND MAKE AN AWESOME MODEL

DATA CLEANING

DATA EXPLORATION; OUTLIER REMOVAL; DATA GROUPING; OVERLAPPING CORRECTION



DATA VISUALIZATIONS

MACHINE LEARNING

RANDOM FOREST, SUPPORT VECTOR CLASSIFIER, NAÏVE BAYES CLASSIFIER

DATA SOURCE

IN ADDITION TO THE GIVEN DATASETS WE HAVE ADDED THE FOLLOWING DATA USING PUBLIC DATASETS.

CLEARGOV

WORLD TRADE CENTER LOS ANGELES (WTCLA)

METADATA

- Population
- Per Capita Income
- Median Household Income
- % Of Population In Workforce
- % With High School Degree
- % With Bachelor Degree
- % With Graduate Degree
- # Of Businesses
- Demographics
- City Division Breakup

WHAT ARE THE KEY FACTORS IN PREDICTING HEALTH SCORES OF THE RESTAURANTS IN LOS ANGELES COUNTY?

TEAM
LINEAR DIGRESSORS

NAÏVE BAYES CLASSIFICATION MODEL

	precision	recall	f1-score	support
1	0.73	0.74	0.74	1996
2	0.54	0.61	0.57	1220
3	0.36	0.04	0.08	228
avg / total	0.64	0.65	0.63	3444

NAIVE BAYES CLASSIFIER FOR MULTINOMIAL MODELS WITH ZIP CODE, FACILITY NAME, OWNER NAME, AND ADDRESS OF THE FACILITY VARIABLES

CLASSIFICATION MODEL

REGRESSION MODEL

TOP FEATURES FOR PREDICTION

RANDOM FOREST

- RISK
- PROGRAM STATUS
- PROGRAM ELEMENT
- RACE-(ASIAN)
- RESTAURANT TYPE

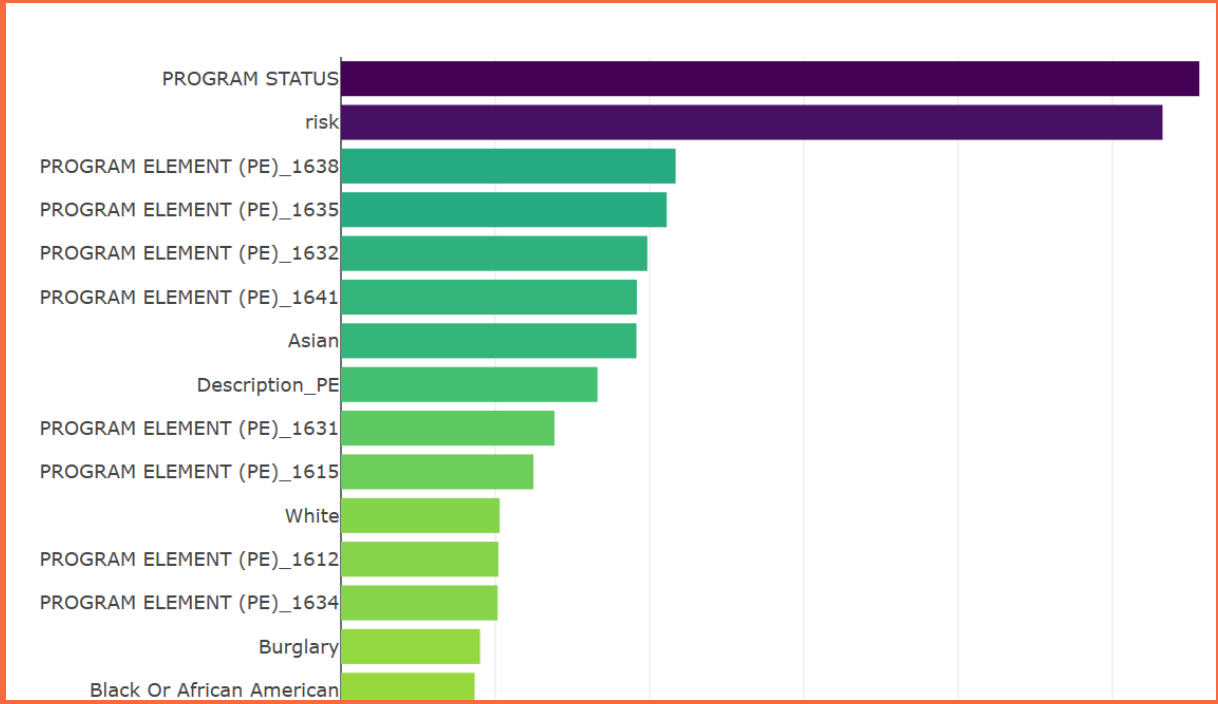
ACCURACY 94.4%

- RODENTS/INSECTS
- RISK TO PUBLIC HEALTH
- HVAC
- HAND WASHING FACILITIES
- FOOD CONTACT SURFACE

ACCURACY 98.1%

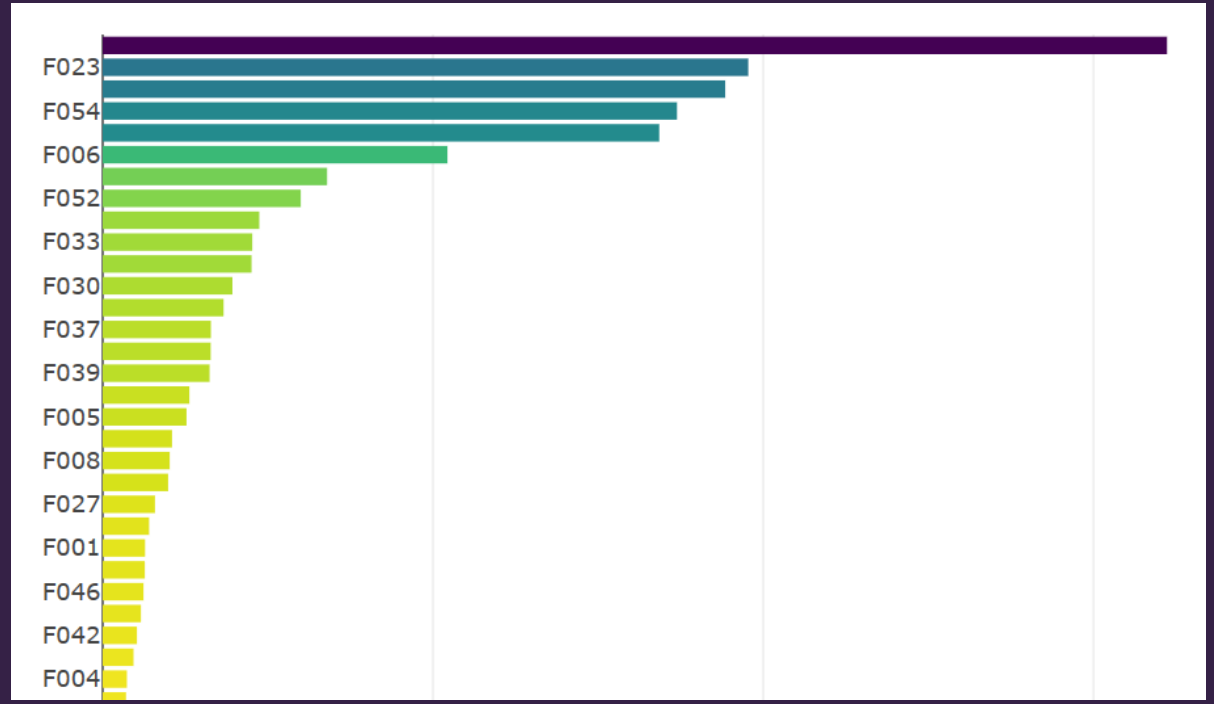
CLASSIFICATION MODEL

TOP FEATURES

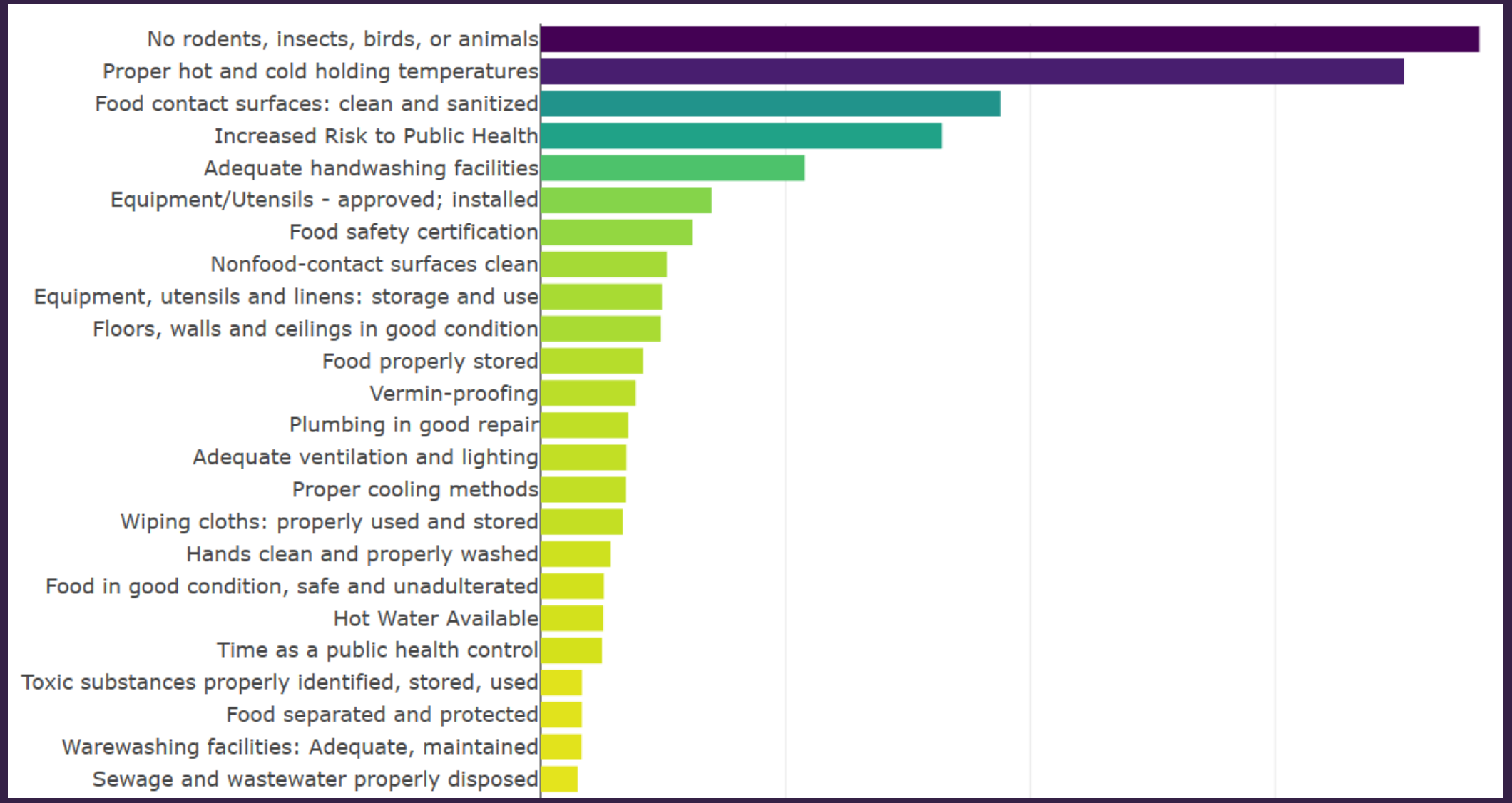


REGRESSION MODEL

TOP FEATURES



MOST CRITICAL VOILATIONS



WHAT ARE THE MOST IMPORTANT FACTORS IN CLASSIFYING RESTAURANTS INTO DIFFERENT "GRADES"?

TEAM
LINEAR DIGRESSORS

EFFECT OF CITY DEMOGRAPHICS ON GRADES

Grade	Population	Per Capita Income	Businesses	Human Service Expense	General Government Expenses	Total Revenue
A	81,982	30,419	2,056	2.0 m	27.1 m	147.4 m
B	76,018	28,044	1,828	1.6 m	22.7 m	108.9 m
C	70,143	27,179	1,756	1.4 m	21.4 m	93.8 m

Highly ranked restaurants can be found in the densely populated areas.

It appears that the healthcare department is taking extra care to routinely check the frequently visited areas.

EFFECT OF POPULATION ON GRADES

Grade	High Population	Low Population
A	94.31%	84.92%
B	5.23%	12.72%
C	0.46%	2.36%

The proportion of A ranked restaurants in top 5 most populated areas are ~10% higher than the proportion of A ranked restaurants in the bottom 5 populated areas.

Health department is alert!

Grade	Division 1	2	3	4	5	24
A	86.71%	85.04%	93.74%	87.07%	92.08%	87.42%
B	11.86%	12.55%	5.62%	11.13%	7.33%	11.11%
C	1.42%	2.41%	0.64%	1.80%	0.59%	1.47%

Division 3 and Division 5 have proportionately higher share of A rated restaurants, indicating a densely populated area.

Grade	Risk LOW	MODERATE	HIGH
A	94.83%	91.89%	86.84%
B	4.78%	7.26%	11.52%
C	0.39%	0.85%	1.64%

The proportion of A ranked restaurants are comparatively lower in when the program is high risk.

Grade	Service Code 1	401
A	94.83%	93.20%
B	4.72%	6.61%
C	0.45%	0.19%

Grade	Incorporated Null	1
A	95.89%	94.53%
B	3.73%	5.01%
C	0.38%	0.46%

Division 3 and Division 5 have proportionately higher share of A rated restaurants, indicating a densely populated area.

Being incorporated/unincorporated does not have an impact on the distribution of Restaurants.

ARE THERE ANY PATTERNS IN TERMS OF HOW HEALTH SCORES OF RESTAURANTS CHANGE OVER TIME?

TEAM
LINEAR DIGRESSORS

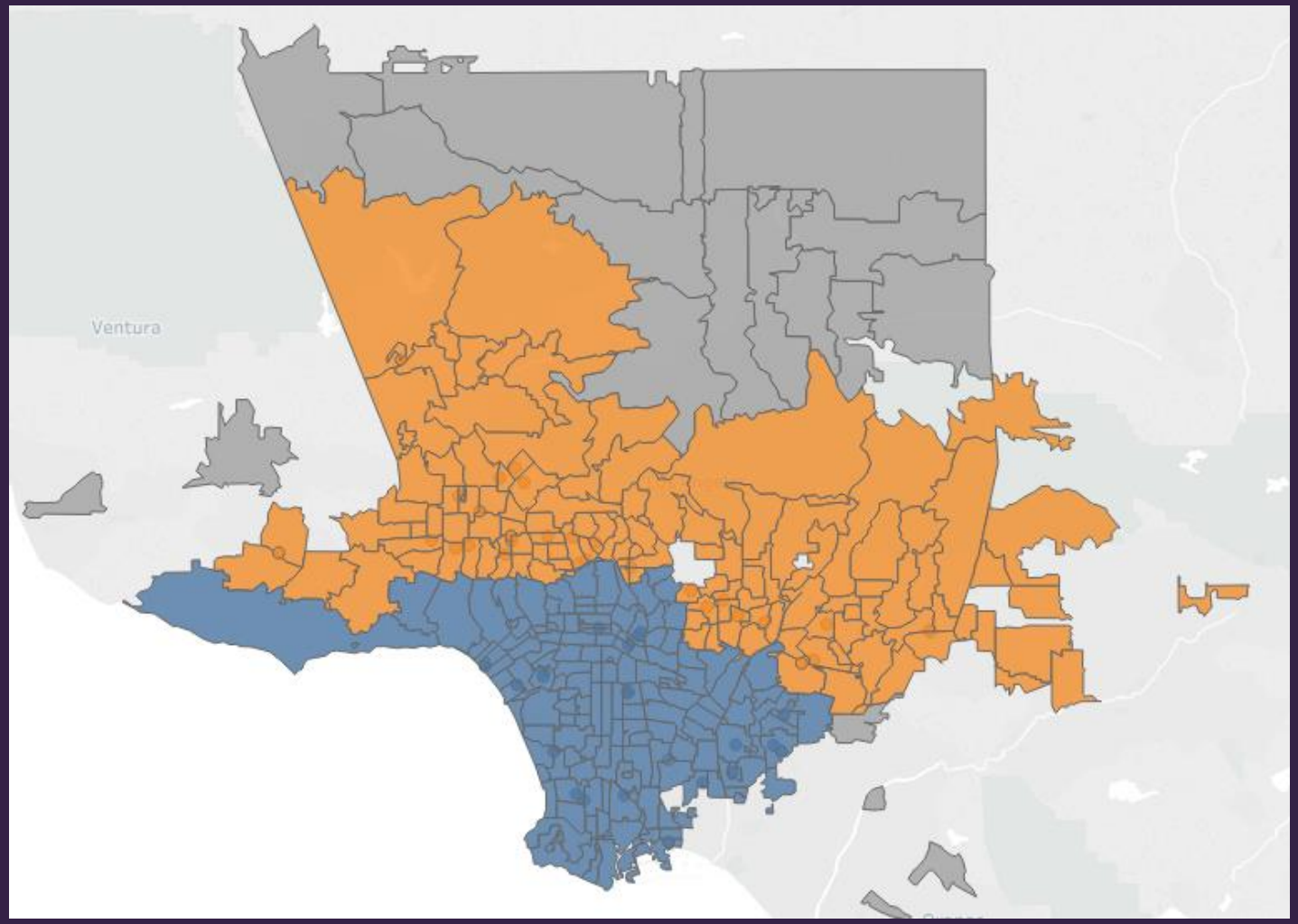
EXPERTS vs NON EXPERTS

		Restaurant Type	
Year of A..	Grade	Chain Restaurants	Other Restaurants
2016	A	99.26%	94.73%
	B	0.74%	5.04%
	C		0.23%
2017	A	98.59%	93.84%
	B	1.37%	5.55%
	C	0.04%	0.61%
2018	A	98.78%	94.27%
	B	1.16%	5.18%
	C	0.06%	0.56%

Grade
A
B
C

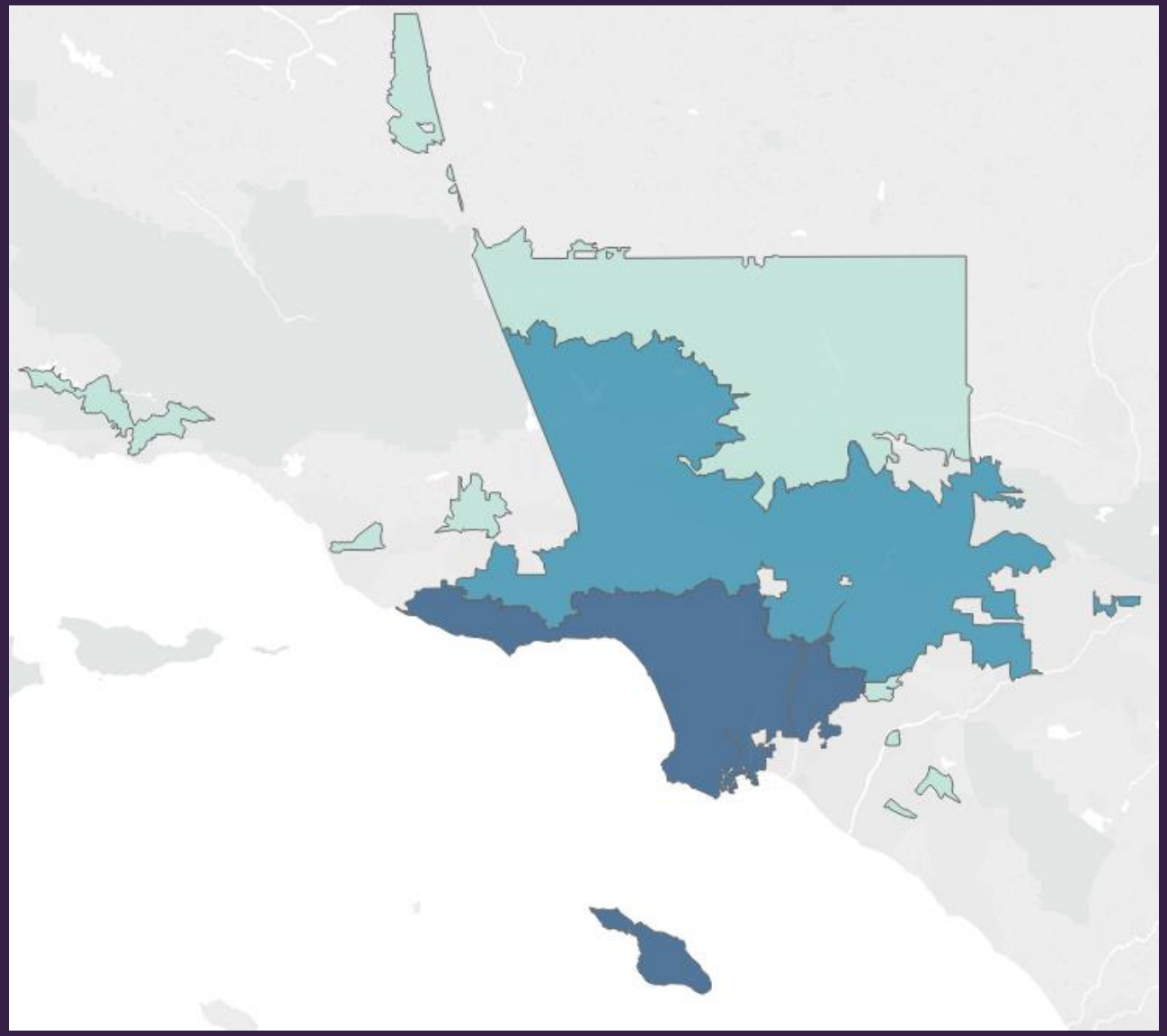
There is a 5% difference in the total number of restaurants which are Graded A between chain restaurants and other restaurants

CLUSTERING LOS ANGELES – K Means



Clustering the data using K-Means Algorithm to segregate 88 incorporated and other unincorporated counties into clusters to analyze.

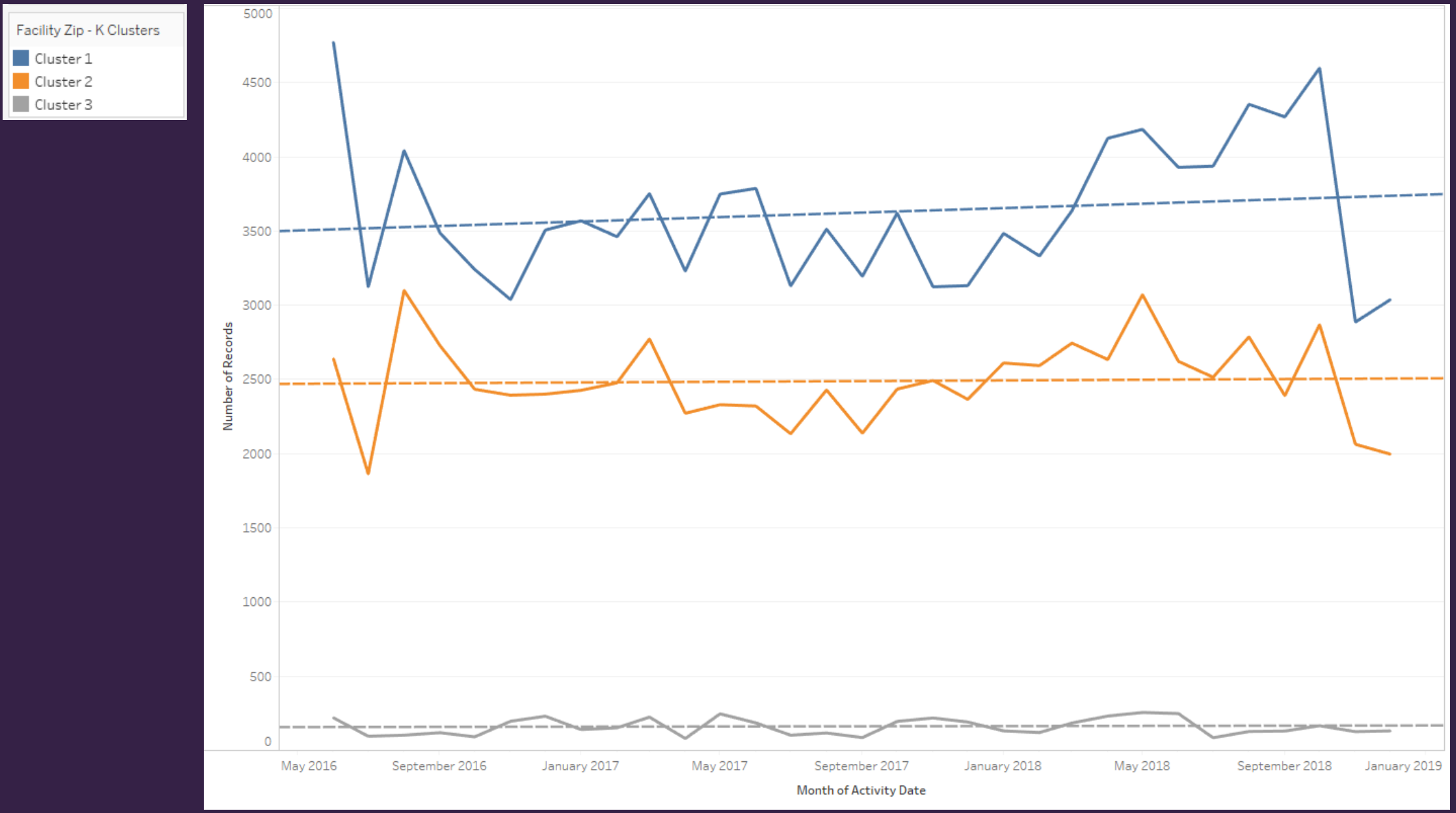
DATA DISTRIBUTION



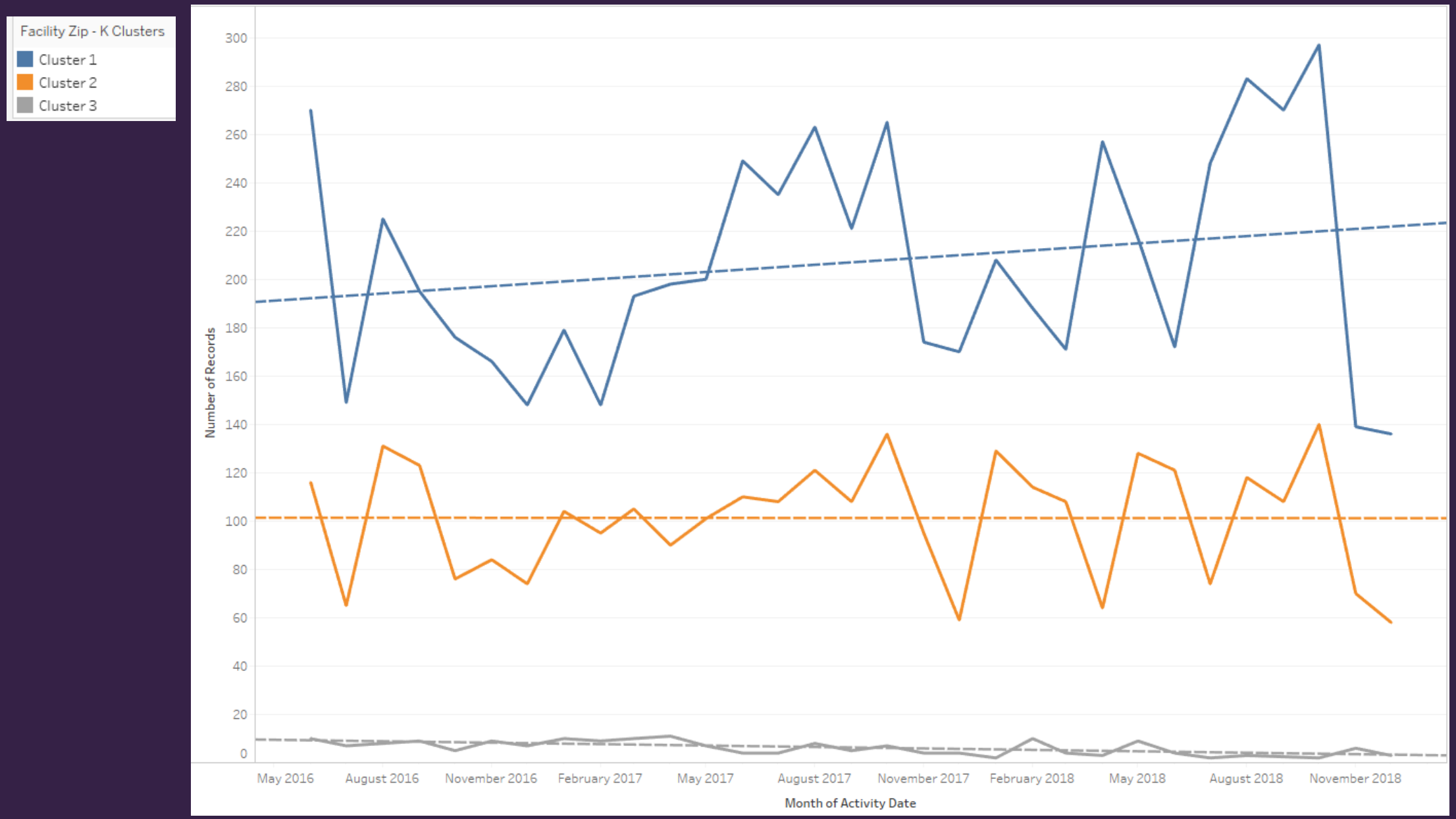
The clusters show a similar patter for:

- Number of violations
- Population
- %Population in workforce
- %People with bachelors and graduate degree
- Per Capita Income
- Revenue & Taxes

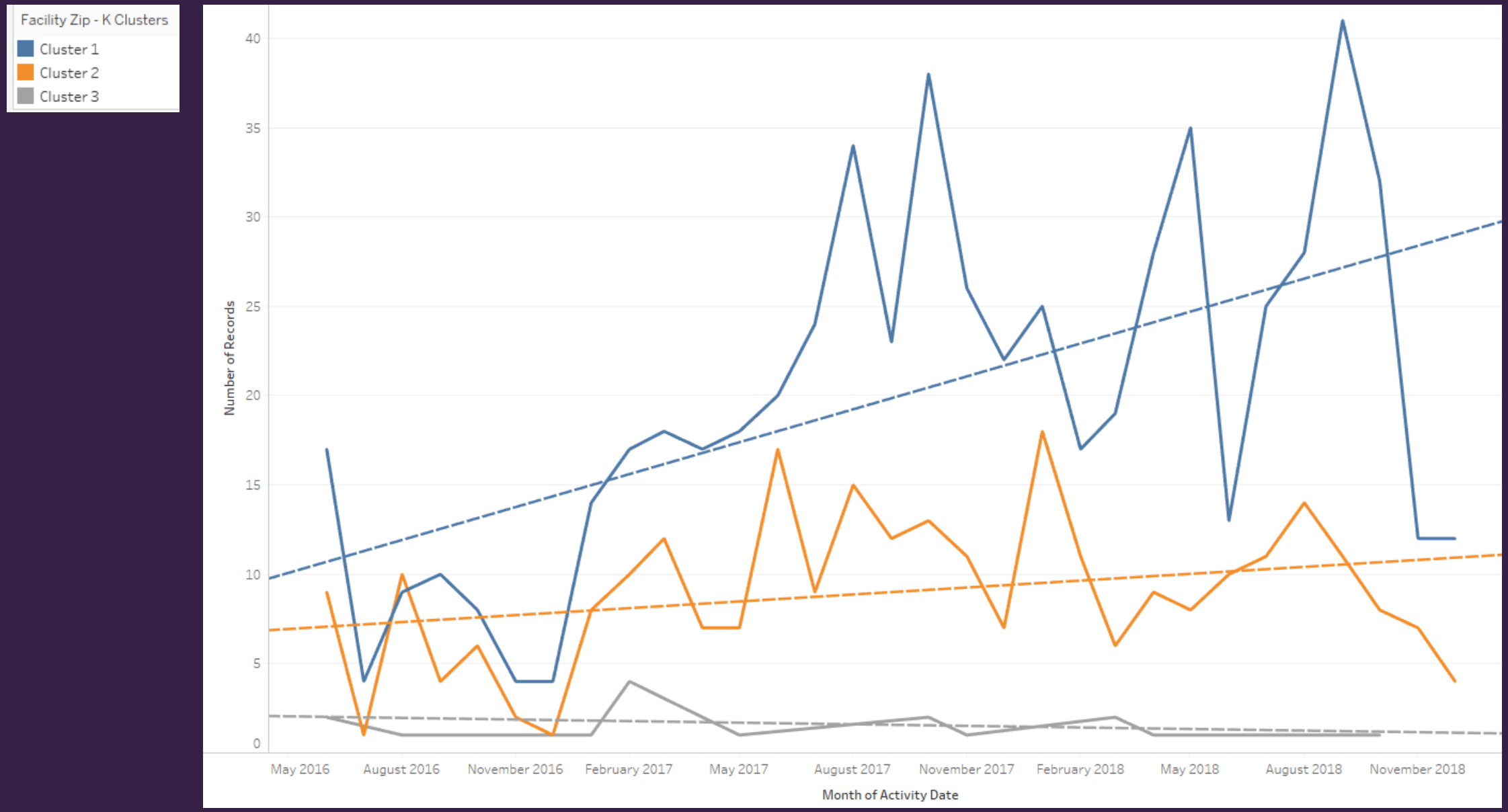
MONTLY PATTERN – A GRADE



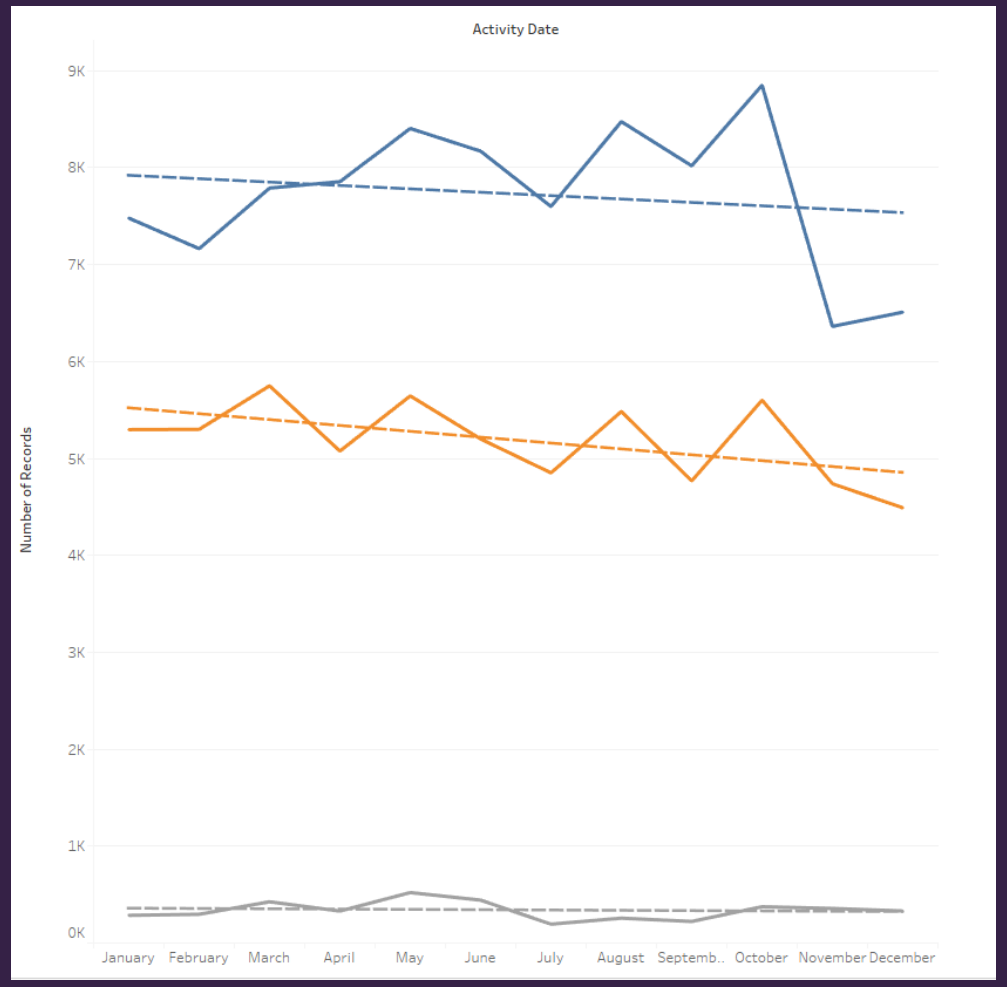
MONTLY PATTERN – B GRADE

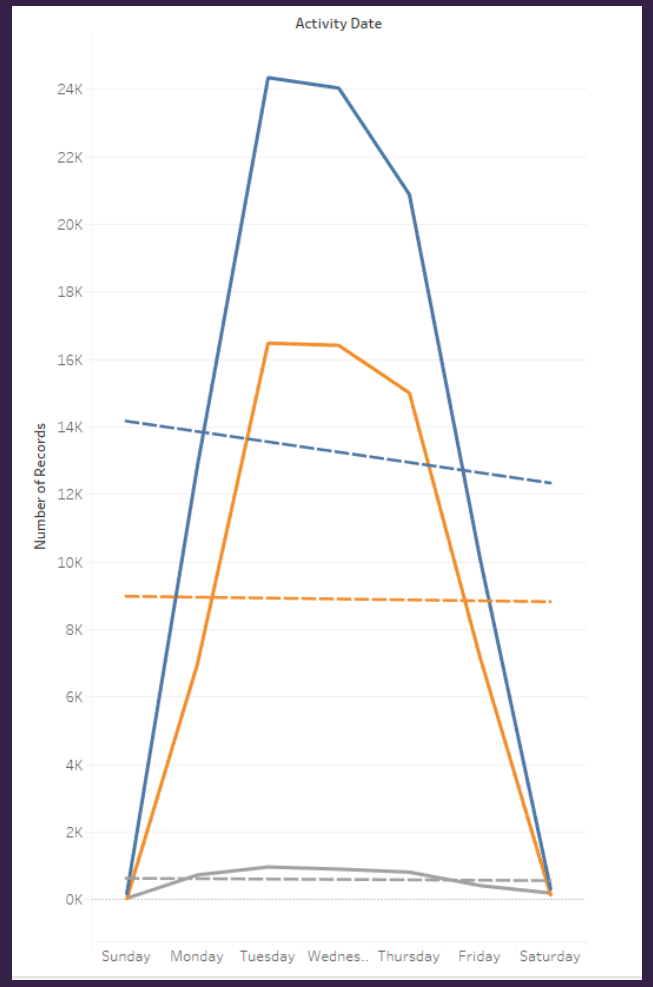


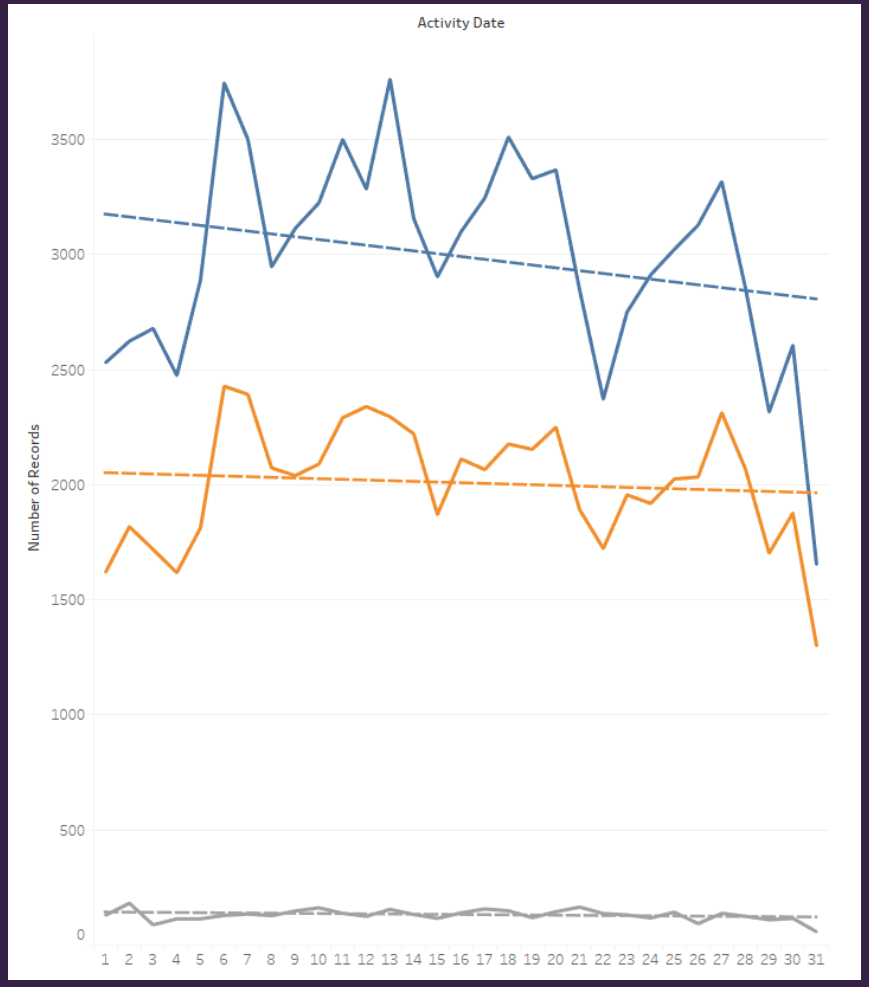
MONTLY PATTERN – C GRADE



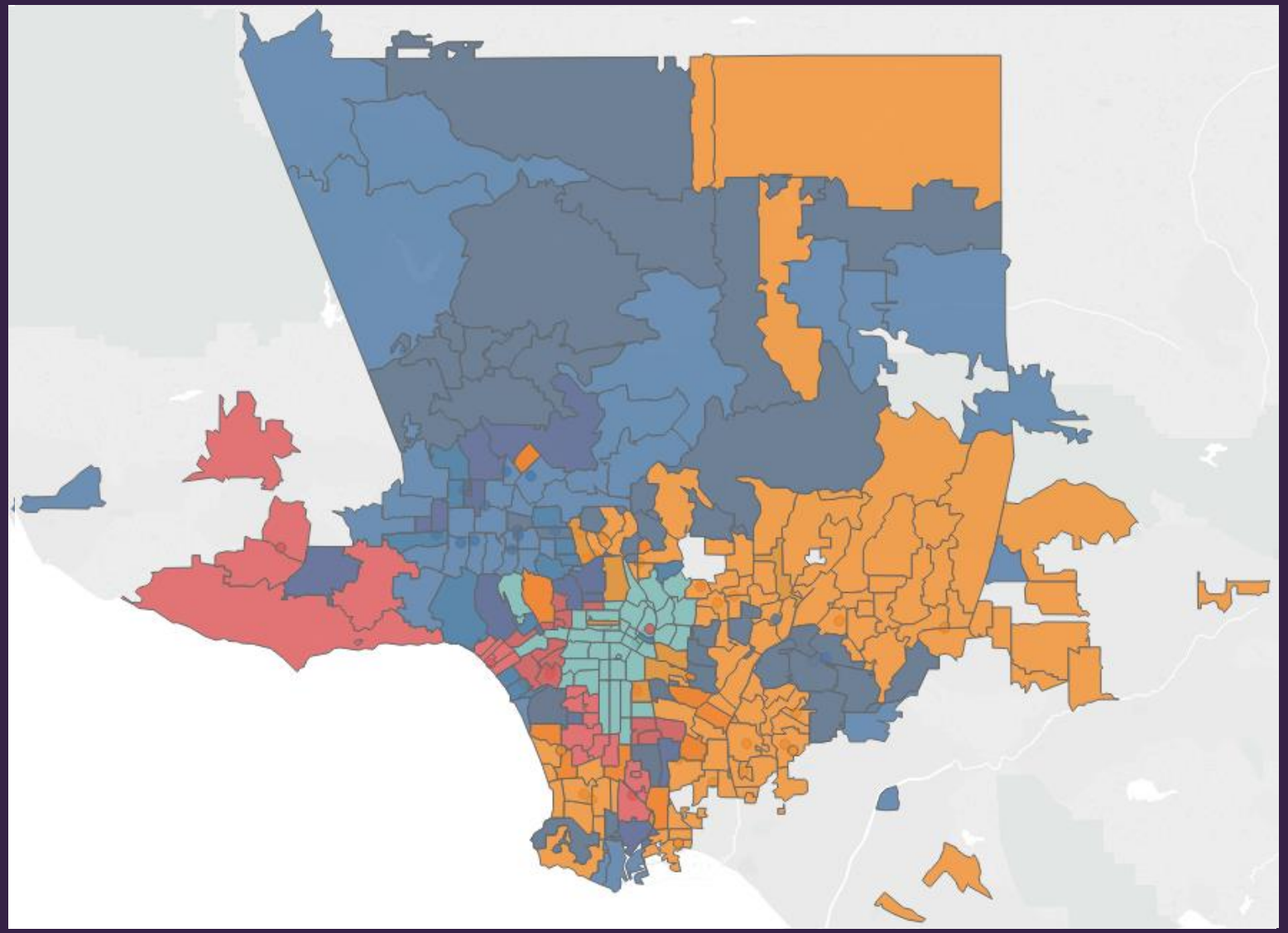
MONTLY-WEEKLY-DAILY PATTERN





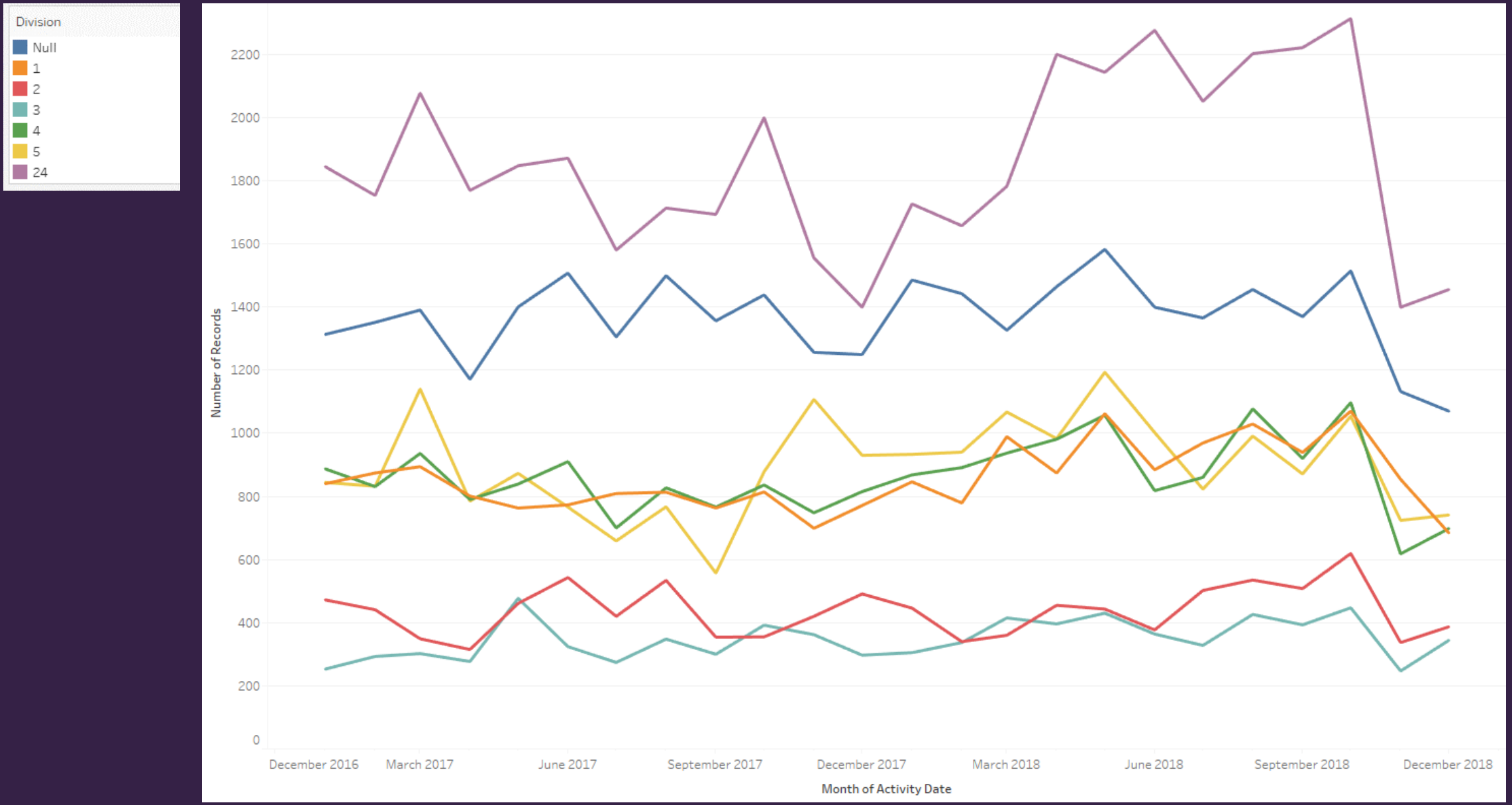


CLUSTERING BY DIVISION

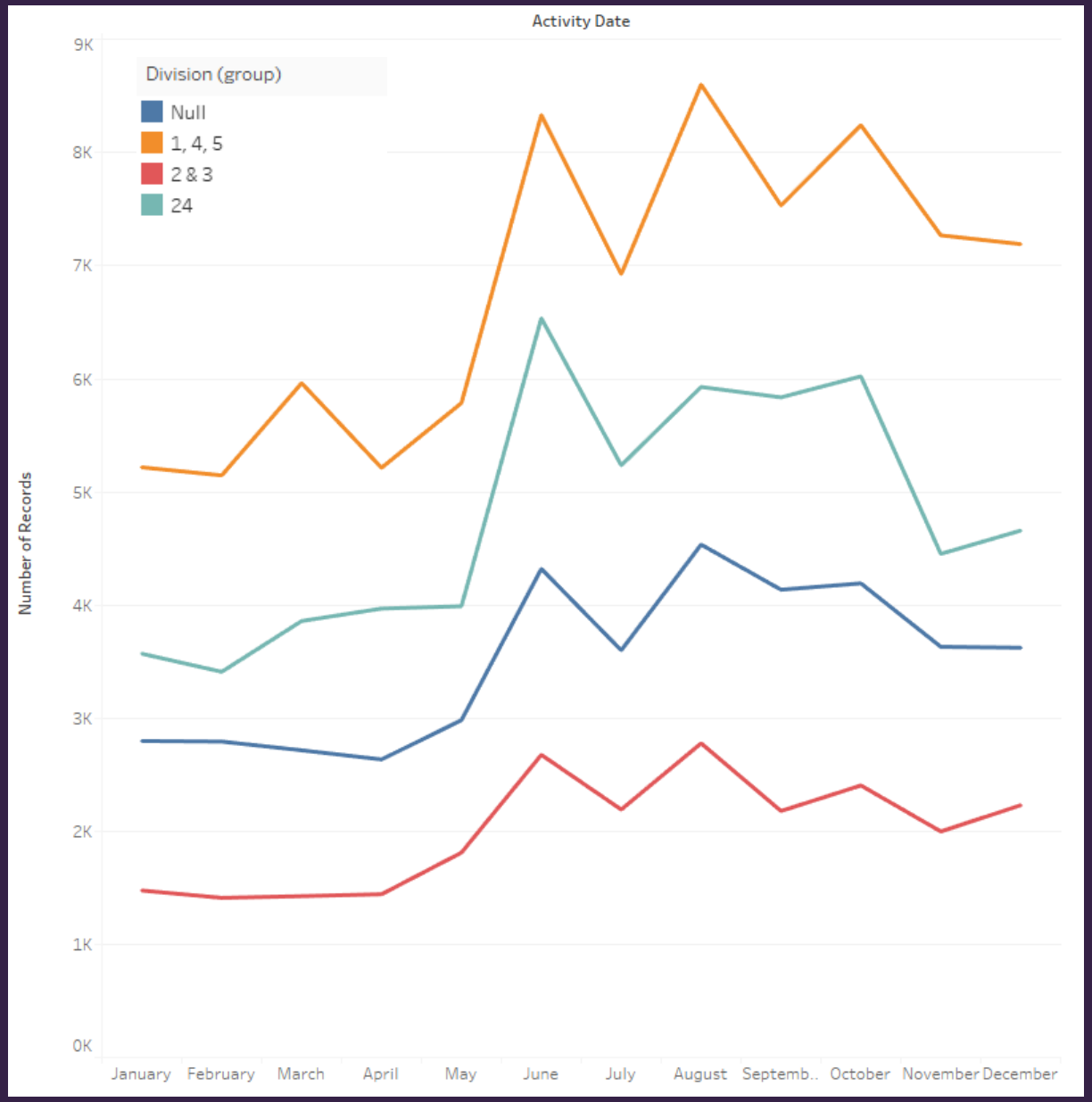


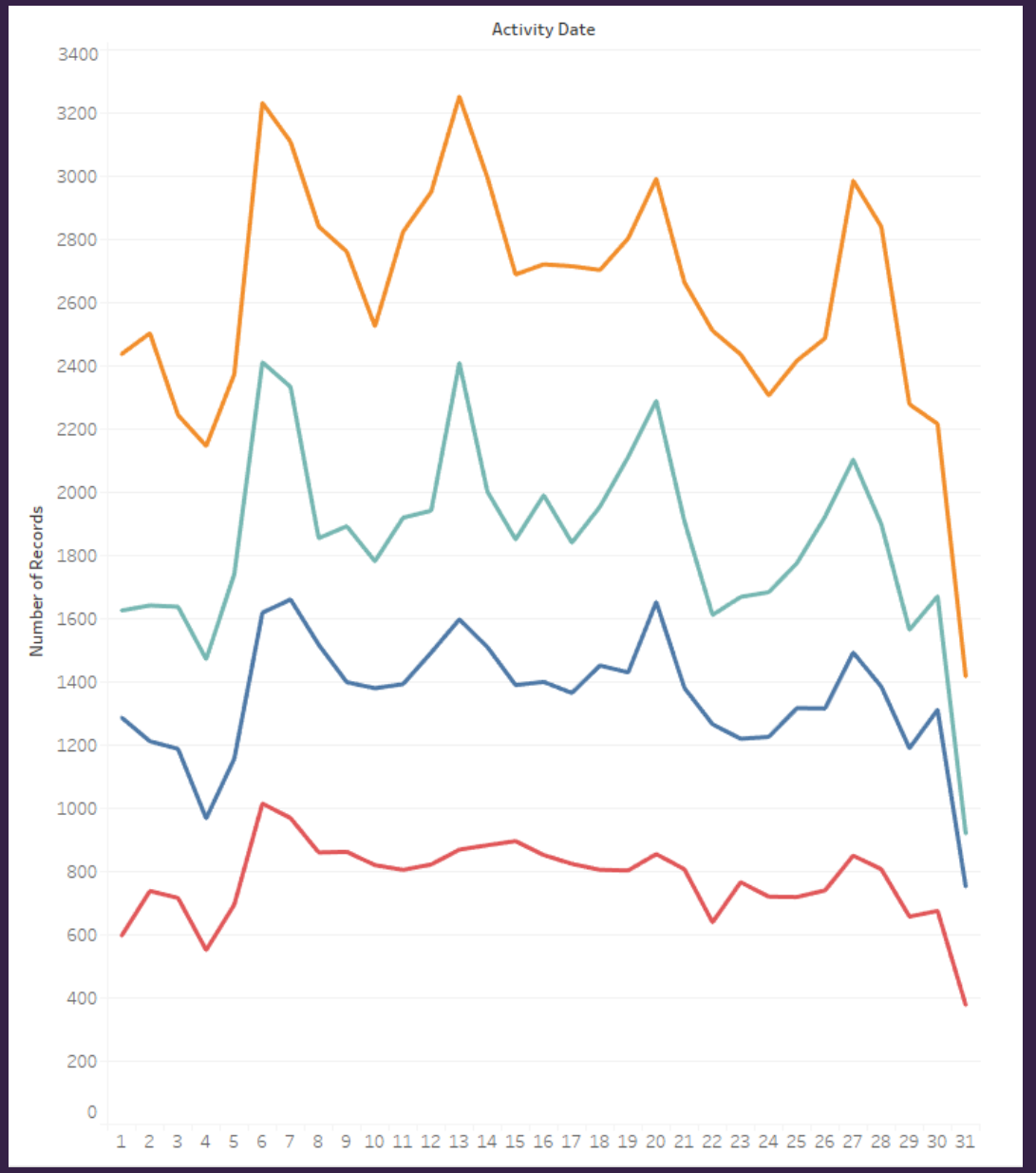
Clustering the data using the administrative divisions in LA County

CLUBBING DATA



MONTLY-DAILY PATTERN





THANK YOU

TEAM
LINEAR DIGRESSORS