

Group Case Study

Apache Spark

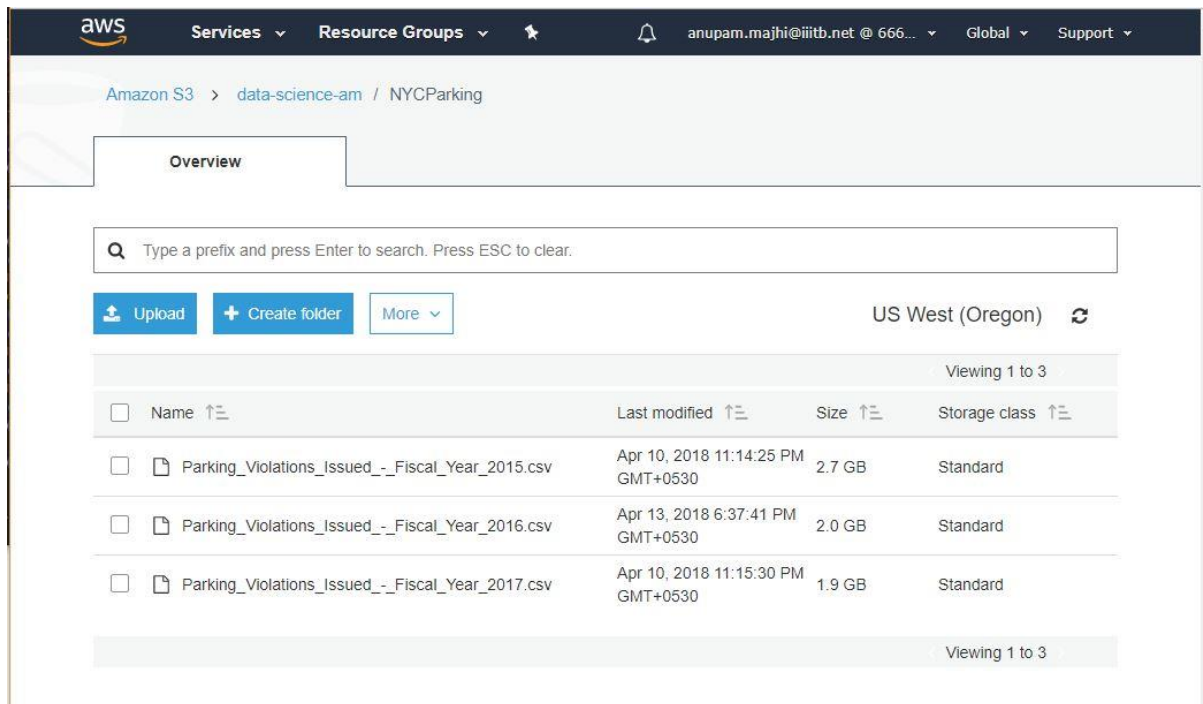
# NYC Parking Tickets: An Exploratory Analysis

Anupam Majhi  
Lijo Thomas  
Hari Nyshadam  
A Kiran Venkat

---

# S3 Screenshot of Data

## 1. Anupam Majhi



aws Services Resource Groups anupam.majhi@iiitb.net @ 666... Global Support

Amazon S3 > data-science-am / NYCParking

Overview

Search: Type a prefix and press Enter to search. Press ESC to clear.

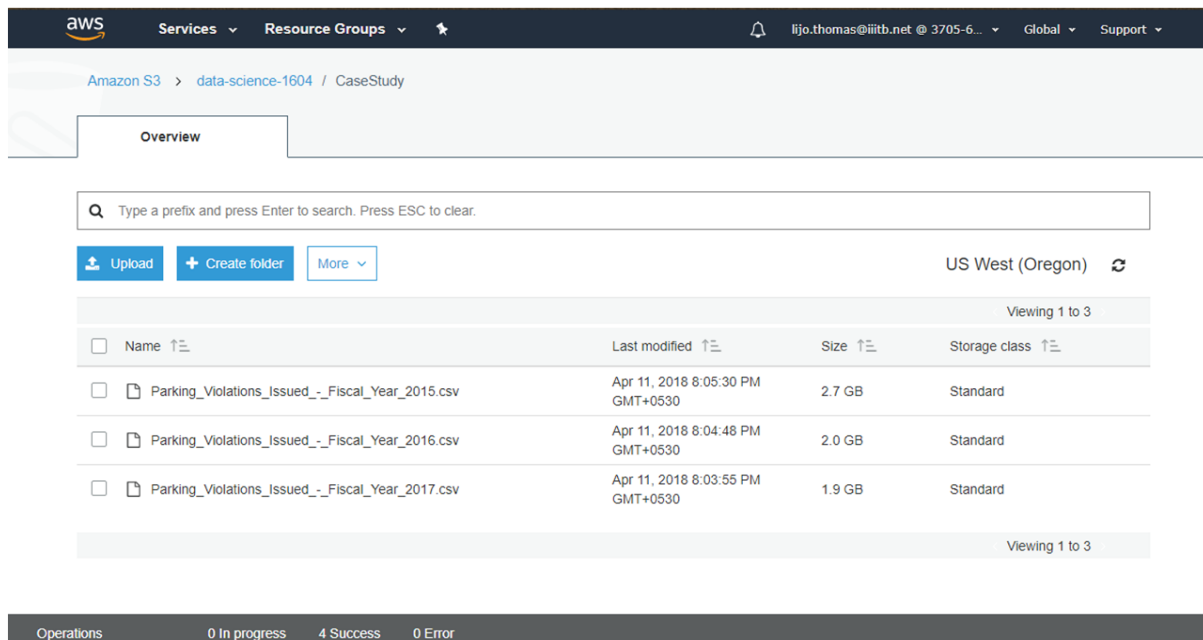
Upload Create folder More

US West (Oregon)

Viewing 1 to 3				
<input type="checkbox"/>	Name ↑	Last modified ↑	Size ↑	Storage class ↑
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2015.csv	Apr 10, 2018 11:14:25 PM GMT+0530	2.7 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2016.csv	Apr 13, 2018 6:37:41 PM GMT+0530	2.0 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2017.csv	Apr 10, 2018 11:15:30 PM GMT+0530	1.9 GB	Standard

Viewing 1 to 3

## 2. Lijo Thomas



aws Services Resource Groups lijo.thomas@iiitb.net @ 3705-6... Global Support

Amazon S3 > data-science-1604 / CaseStudy

Overview

Search: Type a prefix and press Enter to search. Press ESC to clear.

Upload Create folder More

US West (Oregon)

Viewing 1 to 3				
<input type="checkbox"/>	Name ↑	Last modified ↑	Size ↑	Storage class ↑
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2015.csv	Apr 11, 2018 8:05:30 PM GMT+0530	2.7 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2016.csv	Apr 11, 2018 8:04:48 PM GMT+0530	2.0 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2017.csv	Apr 11, 2018 8:03:55 PM GMT+0530	1.9 GB	Standard

Viewing 1 to 3

Operations 0 In progress 4 Success 0 Error

### 3. Hari Nyshadam

aws

Services

Resource Groups

hari.nyshadam@iitb.net @ 72...

Global

Support

Amazon S3 > nycparkinghari

OverviewPropertiesPermissionsManagement

Q Type a prefix and press Enter to search. Press ESC to clear.

Upload

Create folder

More

US West (Oregon)

Viewing 1 to 3

<input type="checkbox"/>	Name	Last modified	Size	Storage class
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2015.csv	Apr 13, 2018 11:14:45 PM GMT+0530	2.7 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2016.csv	Apr 13, 2018 11:20:28 PM GMT+0530	2.0 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2017.csv	Apr 13, 2018 11:26:41 PM GMT+0530	1.9 GB	Standard

Viewing 1 to 3

### 4. A Kiran Venkat

aws

Services

Resource Groups

kiran.venkat@iitb.net @ 5855-...

Global

Support

Amazon S3 > firsts3bckt

OverviewPropertiesPermissionsManagement

Q Type a prefix and press Enter to search. Press ESC to clear.

Upload

Create folder

More

US West (Oregon)

Viewing 1 to 5

<input type="checkbox"/>	Name	Last modified	Size	Storage class
<input type="checkbox"/>	j-1LPF1CHYU97BE	--	--	--
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2015.csv	Apr 11, 2018 9:35:33 PM GMT+0530	2.7 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2016.csv	Apr 11, 2018 9:35:50 PM GMT+0530	2.0 GB	Standard
<input type="checkbox"/>	Parking_Violations_Issued_-_Fiscal_Year_2017.csv	Apr 11, 2018 9:36:00 PM GMT+0530	1.9 GB	Standard
<input type="checkbox"/>	parking_2014.csv	Apr 11, 2018 10:04:08 PM GMT+0530	1.7 GB	Standard

Viewing 1 to 5

Operations

0 In progress

4 Success

0 Error

# INTRODUCTION

## Problem Statement

New York City is a thriving metropolis. Just like most other metros that size, one of the biggest problems its citizens face, is parking. The classic combination of a huge number of cars, and a cramped geography is the exact recipe that leads to a huge number of parking tickets.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has collected data for parking tickets. Out of these, the data files from 2014 to 2017 are publicly available on Kaggle. We will try and perform some exploratory analysis on this data.

## Data

We have datasets from Kaggle (<https://www.kaggle.com/new-york-city/nyc-parking-tickets/data>)

For the purpose of this study we are using three files each with data roughly organized into Fiscal year (July-June)

The following Analysis is done based on Fiscal Years 2015 to 2017

# DATA PREPARATION AND CLEANING

1. 2015 and 2016 datasets have 51 variables whereas 2017 dataset has 43 variables

Following columns are missing from 2017:

- => Latitude
- => Longitude
- => Community Board
- => Community Council
- => Census Tract
- => BIN
- => BBL
- => NTA

2. Number of rows:
  - a. **2015** data: 11809233 total rows, of which only 10951256 are unique.  
Hence, we filtered the data to get unique rows
  - b. **2016** data: 10626899 unique rows
  - c. **2017** data: 10803028 unique rows.
3. We did some sampling to have an idea about the data. We found:
  - a. Columns from 'No Standing or Stopping Violation' are empty in all three datasets. Hence only column 1 to 40 are valid. So, we removed those extra columns.
  - b. Formatting issues with several date-time columns.
4. We added a column with formatted 'Issue date' and column for 'Fiscal Year' which will be helpful to have a single dataframe for analysis.
5. Invalid issue dates:
  - a. As the data should contain tickets from a particular Fiscal year only, we found that there are several older data, even sometimes ranging decades back.
  - b. Such invalid rows were filtered out.
6. We prepared a single dataframe by joining all the 3 dataframes for easier analysis and to avoid repetitive queries.

## Section I: EXAMINE THE DATA

1. Find total number of tickets for each year.

**Note:** Previously we found some duplicate rows and they were removed from the dataframe.

**Result:**

Fiscal Year	count_SummonsNumber
2015	10598036
2016	10396894
2017	10539563

2. Find out how many unique states the cars which got parking tickets came from.

**Note:** We are assuming that the state code 99 is a valid state code, since we do not have enough information to exclude it from the list.

**Result:**

Fiscal Year	Count_State
2015	69
2016	68
2017	67

3. Some parking tickets don't have addresses on them, which is cause for concern. Find out how many such tickets there are.

**Note:** In the following result we have assumed that 'House Number', 'Street Name' and 'Intersecting Street' makes up a valid address.

**Result:**

Fiscal Year	Frequency_InvalidAddress
2015	3696
2016	2640
2017	2418

**Comments:**

- a. Compared to the size of the data, the invalid address frequency is quite insignificant.
- b. The number of invalid addresses decreased over time

**Note:** In the following result we have assumed that only 'House Number' and 'Street Name' are enough to make up a valid address.

**Result:**

Fiscal Year	Frequency_InvalidAddress
2015	3759
2016	2788
2017	2553

**Comments:**

- a. Compared to the size of the data, the invalid address frequency is quite insignificant.
- b. The number of invalid addresses decreased over time

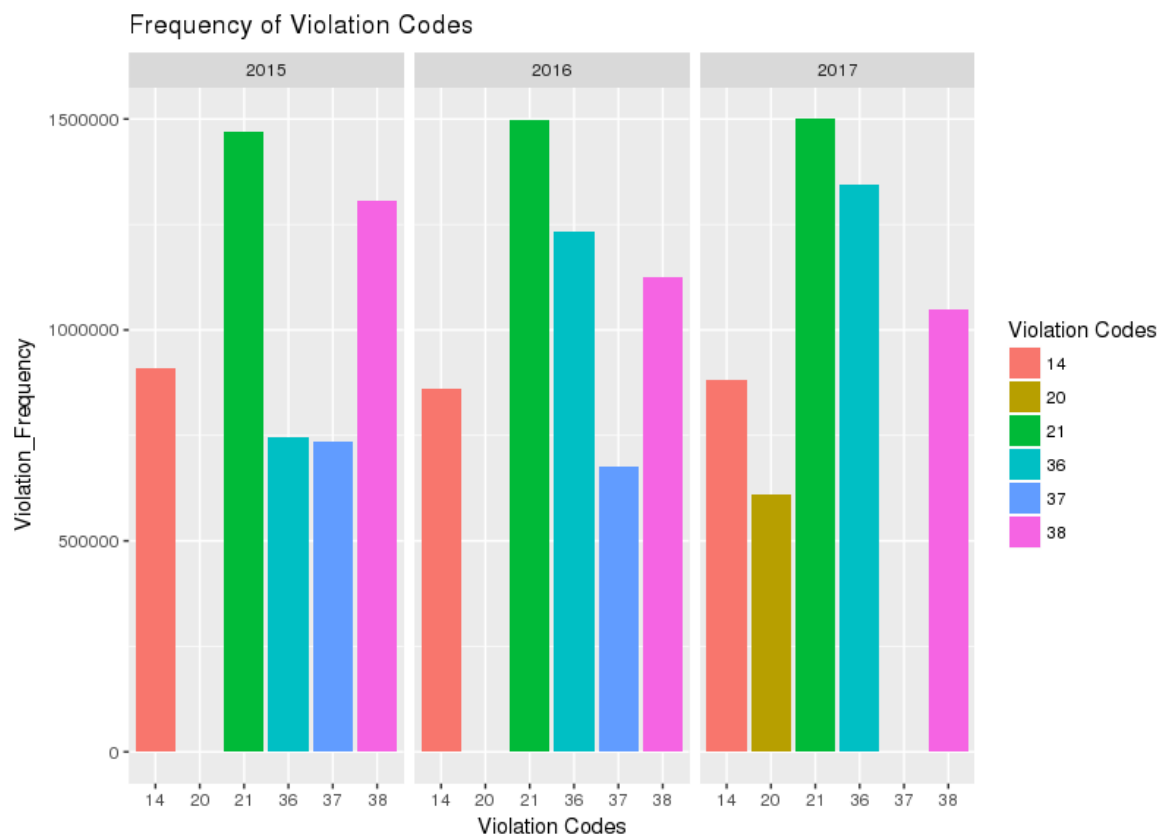
## Section II: AGGREGATION TASKS

1. How often does each violation code occur? (frequency of violation codes - find the top 5)

Result:

Fiscal Year	Violation Code	Violation_Frequency
2016	21	1497269
2016	36	1232952
2016	38	1126835
2016	14	860045
2016	37	677805
2017	21	1500396
2017	36	1345237
2017	38	1050418
2017	14	880152
2017	20	609231
2015	21	1469228
2015	38	1305007
2015	14	908418
2015	36	747098
2015	37	735600

Plot:





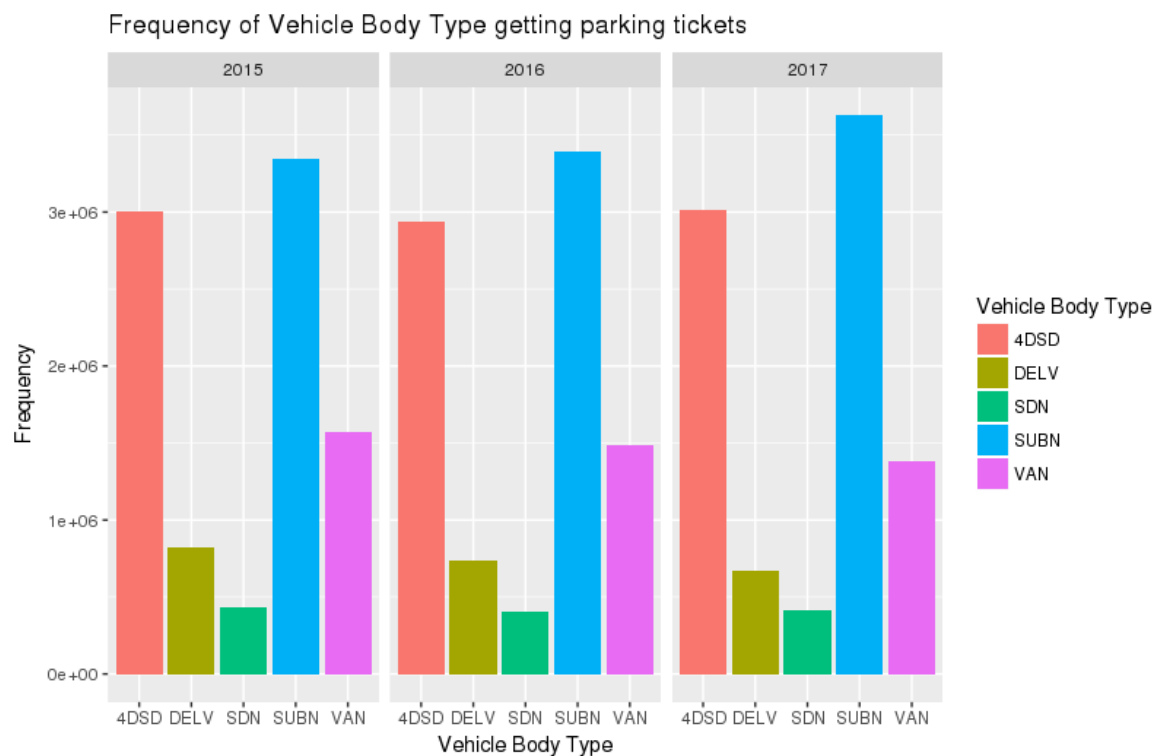
**Comments:**

- a. Violation code 21 is the most frequent.
- b. Violation code 20 comes to top 5 in 2017 which was not the case previously.

2. How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

**VEHICLE BODY TYPE****Result:**

Fiscal Year	Vehicle Body Type	Frequency
2016	SUBN	3393838
2016	4DSD	2936729
2016	VAN	1489924
2016	DELV	738747
2016	SDN	401750
2017	SUBN	3632003
2017	4DSD	3017372
2017	VAN	1384121
2017	DELV	672123
2017	SDN	414984
2015	SUBN	3341110
2015	4DSD	3001810
2015	VAN	1570227
2015	DELV	822041
2015	SDN	428571

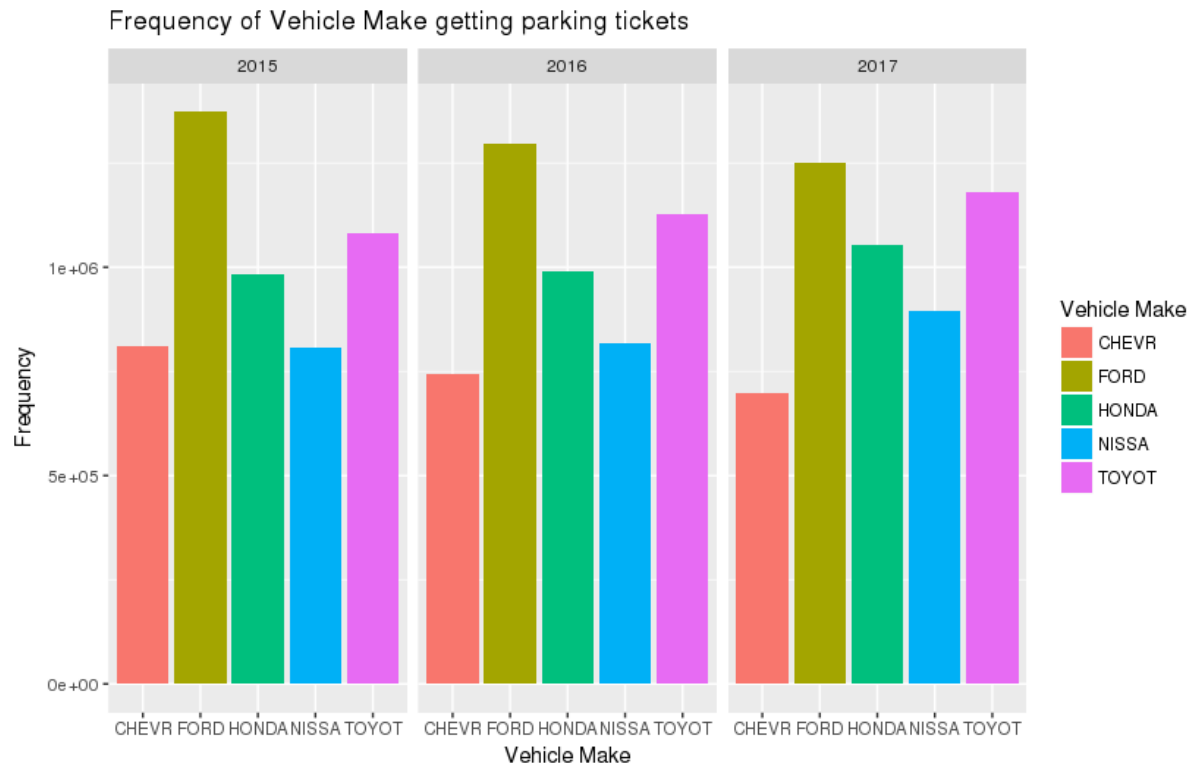
**Plot:****Comments:**

- SUBN is the most frequent occurring body type here.
- Frequency of SUBN increases over time whereas rest of them in top5 show a downward trend.
- However this does not tell us about the cause as this might depend on vehicle popularity too.

**VEHICLE MAKE****Result:**

Fiscal Year	Vehicle Make	Frequency
2016	FORD	1297363
2016	TOYOT	1128909
2016	HONDA	991735
2016	NISSA	815963
2016	CHEVR	743416
2017	FORD	1250777
2017	TOYOT	1179265
2017	HONDA	1052006
2017	NISSA	895225
2017	CHEVR	698024
2015	FORD	1373157
2015	TOYOT	1082206
2015	HONDA	982130
2015	CHEVR	811659
2015	NISSA	805572

**Plot:**



**Comments:**

- FORD is the highest here however it has a decreasing trend, CHEV is also decreasing.
- TOYOTA is increasing along with HONDA and NISSAN.

**3.** A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

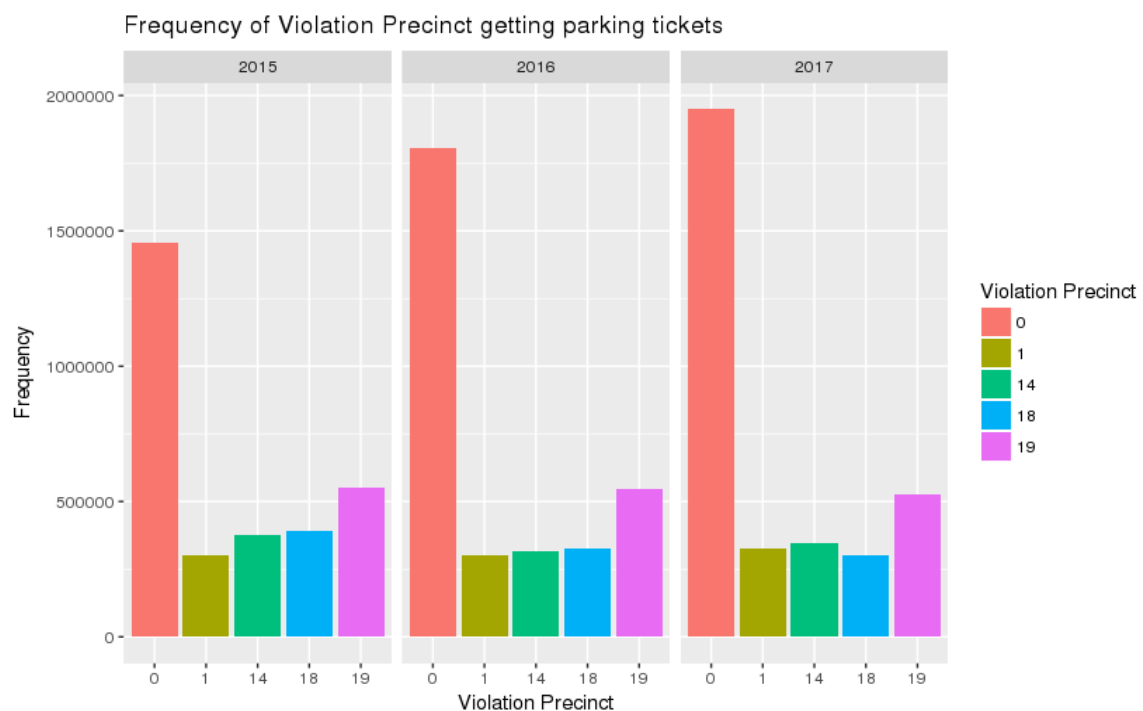
**3.a.** Violating Precincts (this is the precinct of the zone where the violation occurred)

**Note:** We are assuming that Precinct number 0 is a valid entry as this could include precincts which do not have a number. E.g. Central Park Precinct.

## Result:

Fiscal Year	Violation Precinct	Frequency
2016	0	1807139
2016	19	545669
2016	18	325559
2016	14	318193
2016	1	299074
2017	0	1950083
2017	19	528317
2017	14	347736
2017	1	326961
2017	18	302008
2015	0	1455166
2015	19	550797
2015	18	393802
2015	14	377750
2015	1	302737

## Plot:



## Comments:

- 0 is the highest occurring Violation Precinct followed by 19.
- Violations in Precinct 0 increases while rest of them tend to have a decreasing trend year on year.

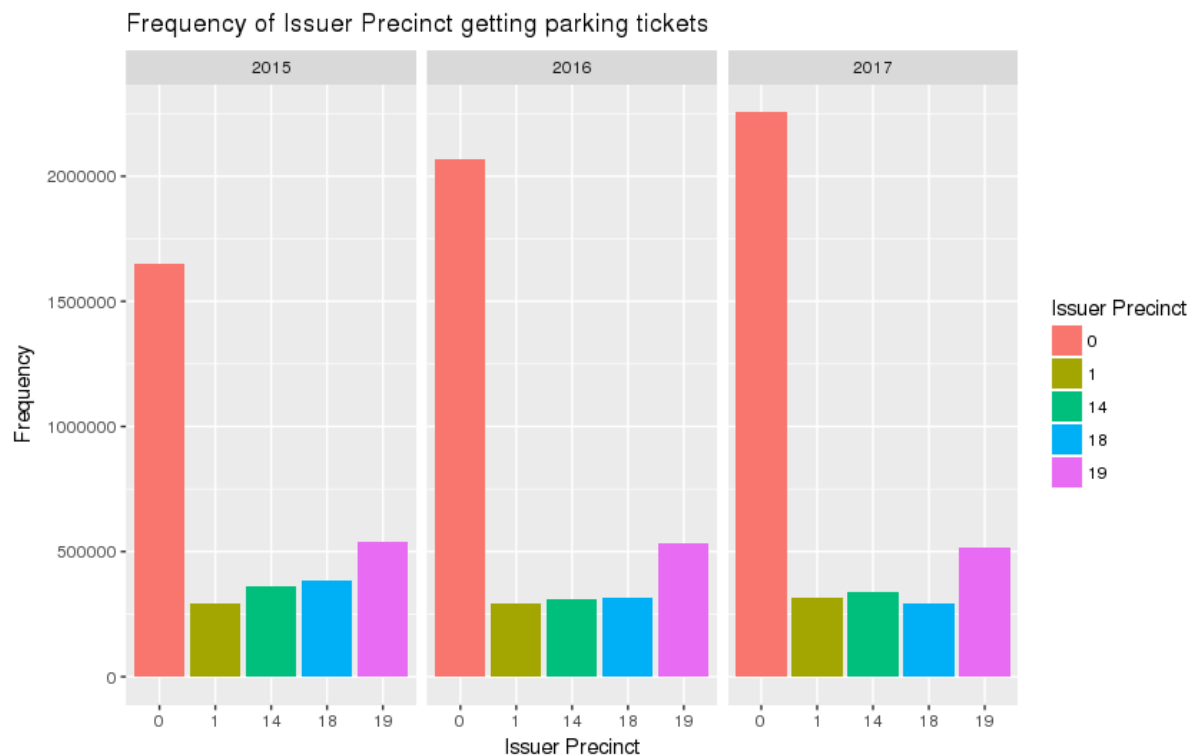
### 3.b. Issuing Precincts (this is the precinct that issued the ticket)

**Note:** We are assuming that Precinct number 0 is a valid entry as this could include precincts which do not have a number. E.g. Central Park Precinct.

**Result:**

Fiscal Year	Issuer Precinct	Frequency
2016	0	2067219
2016	19	532298
2016	18	317451
2016	14	309727
2016	1	290472
2017	0	2255086
2017	19	514786
2017	14	340862
2017	1	316776
2017	18	292237
2015	0	1648671
2015	19	536627
2015	18	384863
2015	14	363734
2015	1	293942

**Plot:**



## Comments:

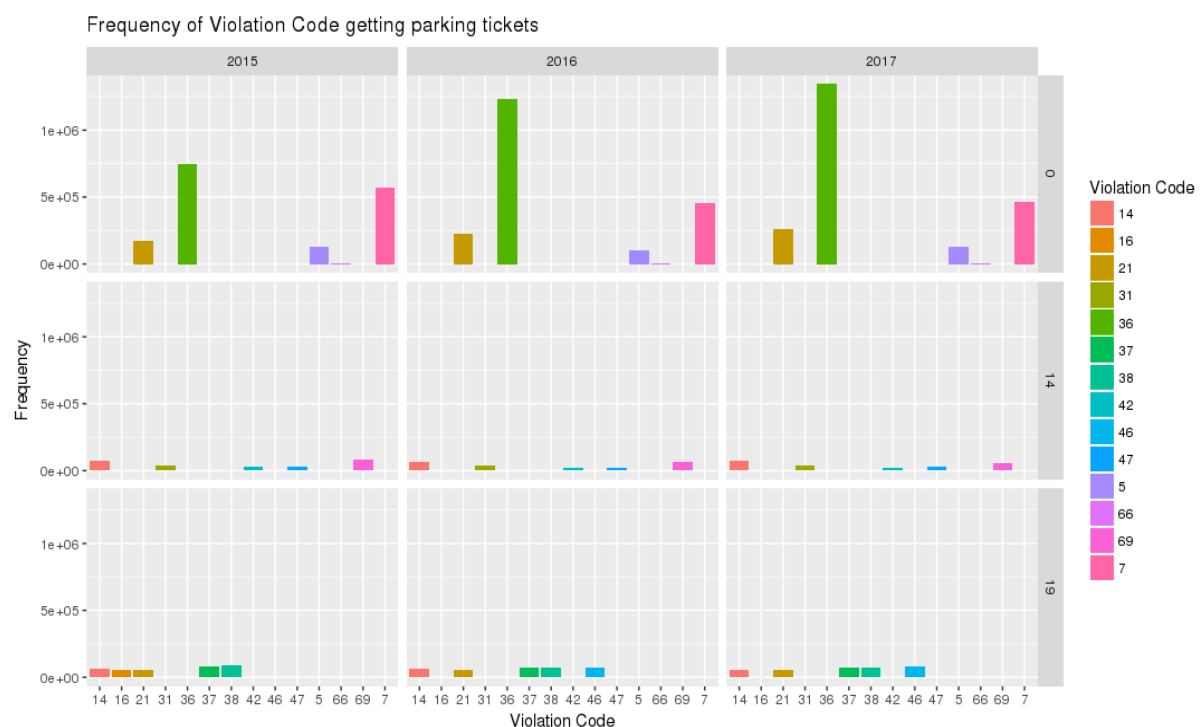
- 0 is the highest occurring Issuer Precinct followed by 19.
- More tickets are issued in Precinct 0 and increases while rest of them tend to have a decreasing trend year on year

- Find the violation code frequency across 3 precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes? Are these codes common across precincts?

## Result:

Year	Precinct	Violation Code (descending order)
2017	0	36,07,21,05,66
	14	14,69,31,47,42
	19	46,38,37,14,21
2016	0	36,07,21,05,66
	14	69,14,31,47,42
	19	38,37,46,14,21
2015	0	36,07,21,05,66
	14	69,14,31,42,47
	19	38,37,14,16,21

## Plot:



**Comments:**

- a. The highest occurring Violation code for a particular precinct remains fairly similar over years.

**5.** You'd want to find out the properties of parking violations across different times of the day:

**5.a.** The Violation Time field is specified in a strange format. Find a way to make this into a time attribute that you can use to divide into groups.

**AND**

**5.b.** Find a way to deal with missing values, if any.

**Note:** We are assuming that the string format is 12 hours format followed by A or P based on AM or PM.

- a. We have converted it into timestamp using methods as in code.
- b. To impute null values we used the column 'From Hours in Effect' as it seems close to the definition of Violation Time.

**Result:**

We have 2684 Missing Values. We have removed these rows

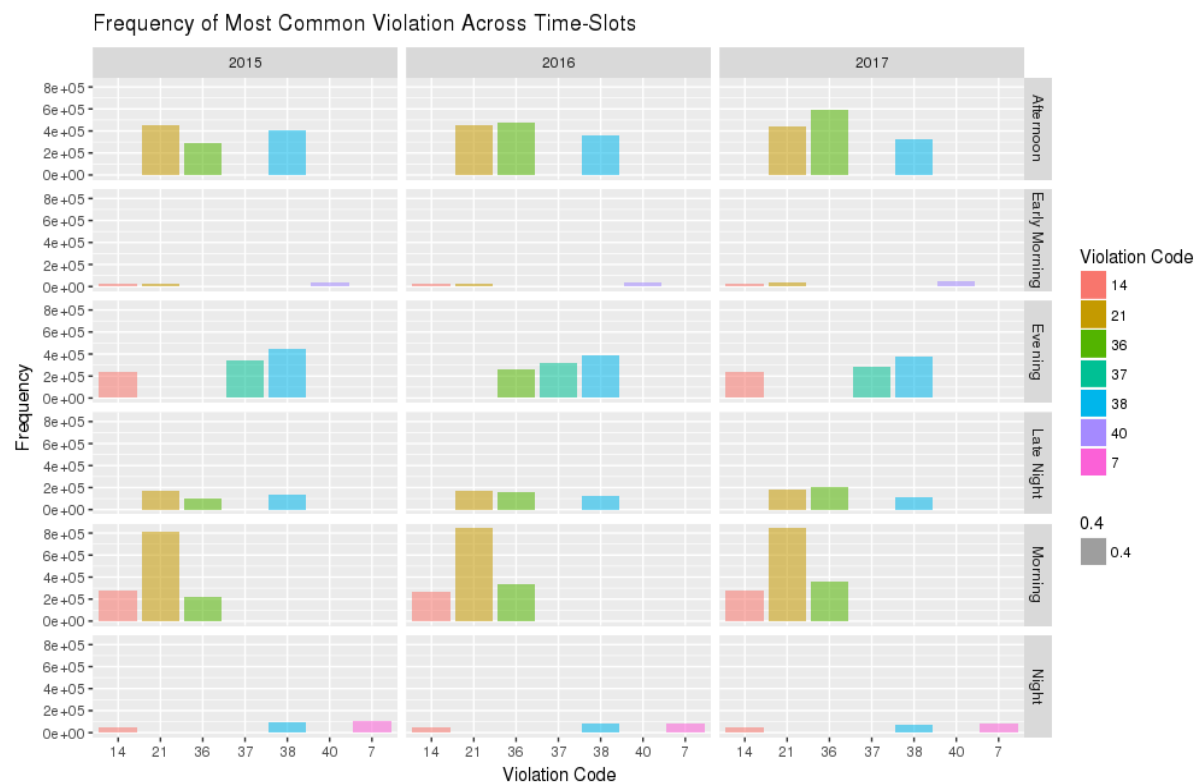
**5.c.** Divide 24 hours into 6 equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the 3 most commonly occurring violations

**Note:** We have split into 6 equal discrete bins of 4 hours starting from 2AM e.g. (2am to 6 am)

## Result:

Year	Time of Day	Most Occuring Violation Code (Descending Order)
2017	Morning	21,36,14
	Afternoon	36,21,38
	Evening	38,37,14
	Night	07,38,14
	Late Night	36,21,38
	Early Morn	40,21,14
2016	Morning	21,36,14
	Afternoon	36,21,38
	Evening	38,37,36
	Night	38,07,14
	Late Night	21,36,38
	Early Morn	40,14,21
2015	Morning	21,14,36
	Afternoon	21,38,36
	Evening	38,37,14
	Night	07,38,14
	Late Night	21,38,36
	Early Morn	40,14,21

## Plot:





## Comments :

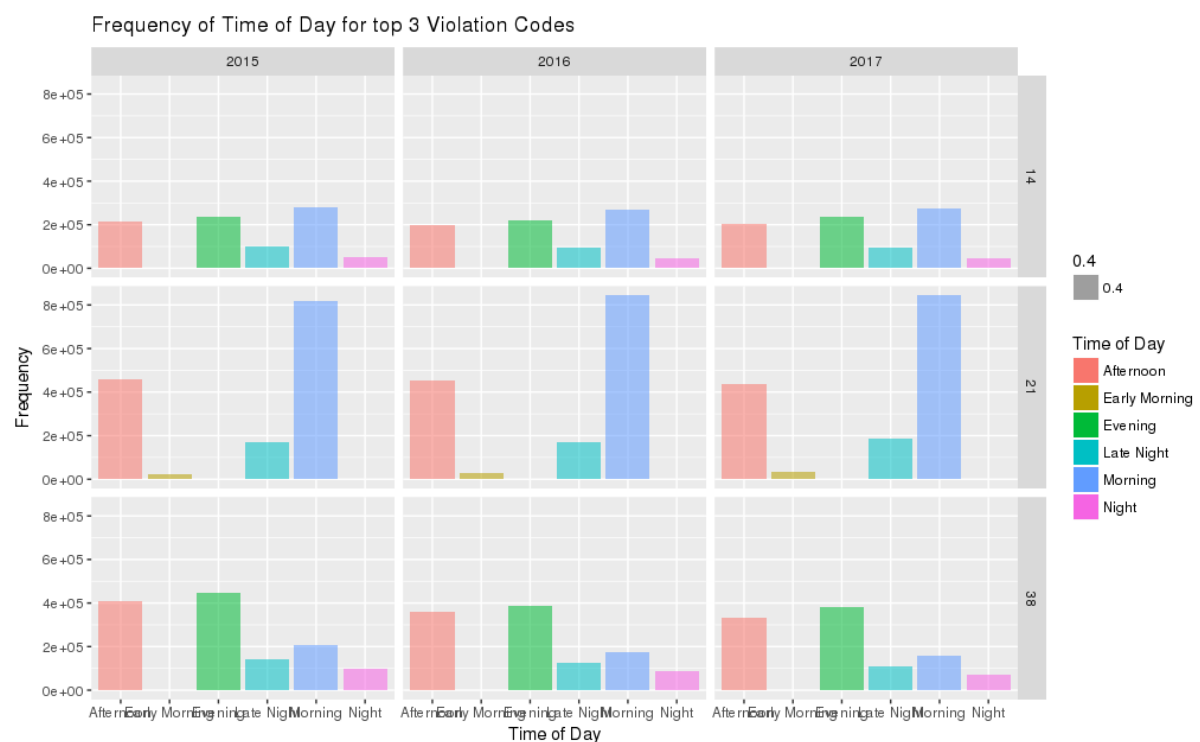
- The violation codes for a particular time of day remains fairly similar across years.
- In Morning Violation code 21 is always high.
- In Evening Violation code 38 is always high

**5.d.** Now, try another direction. For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part)

## Result:

Year	Violation Code	Time of Day most occurring
2017	14	M,E,A,LN,N
	21	M,A,LN,EM,E
	38	E,A,M,LN,N
2016	14	M,E,A,LN,N
	21	M,A,LN,EM,E
	38	E,A,M,LN,N
2015	14	M,E,A,LN,N
	21	M,A,LN,EM,E
	38	E,A,M,LN,N

## Plot:



6. Let's try and find some seasonality in this data

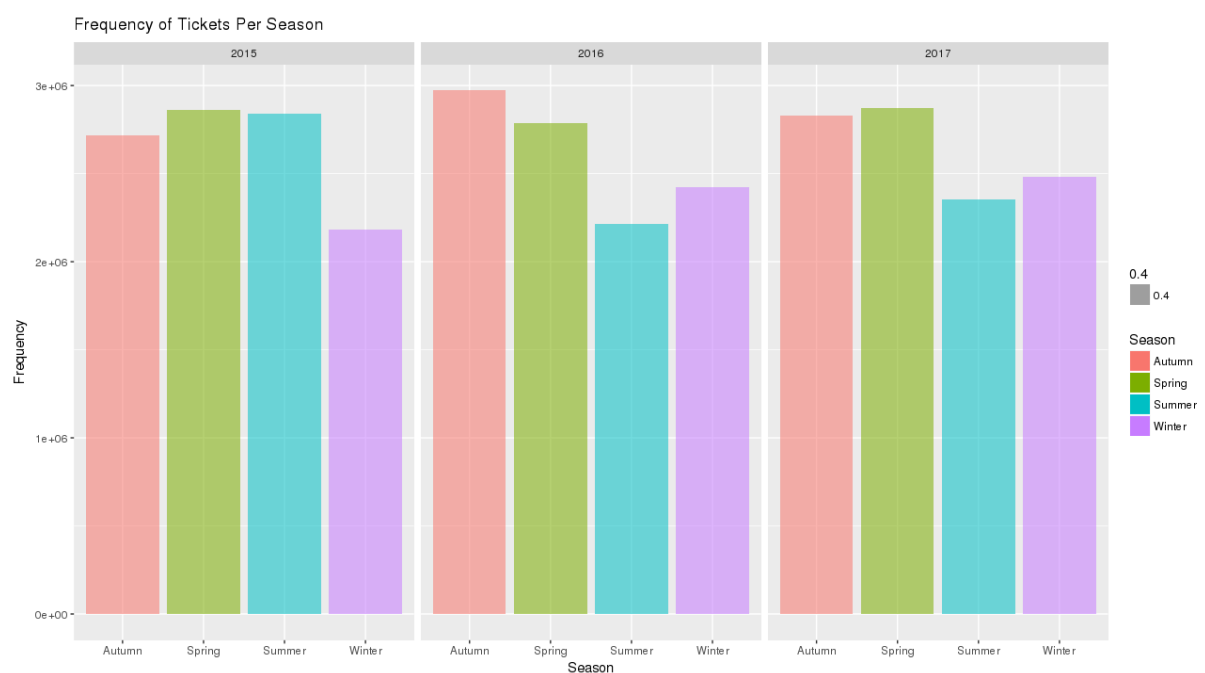
6.a. First, divide the year into some number of seasons, and find frequencies of tickets for each season.

Note: NA

Result:

Year	Season	Frequency
2017	Summer	2353920
	Autumn	2829224
	Winter	2483036
	Spring	2873383
2016	Summer	2214536
	Autumn	2971672
	Winter	2421620
	Spring	2789066
2015	Summer	2838306
	Autumn	2718502
	Winter	2180241
	Spring	2860987

Plot:



**6.b.** Then, find the 3 most common violations for each of these season

**Result:**

Year	Season	Violation Code Most Occuring
2017	Summer	21,38,14
	Autumn	36,21,38
	Winter	21,36,38
	Spring	21,36,38
2016	Summer	21,38,14
	Autumn	36,21,38
	Winter	21,36,38
	Spring	21,36,38
2015	Summer	21,38,14
	Autumn	21,38,14
	Winter	38,21,14
	Spring	21,38,14

**Plot:**



**7.** The fines collected from all the parking violation constitute a revenue source for the NYC police department. Let's take an example of estimating that for the 3 most commonly occurring codes.

**7.a.** Find total occurrences of the 3 most common violation codes

**AND**

**7.b.** Then, search the internet for NYC parking violation code fines. You will find a website (on the nyc.gov URL) that lists these fines. They're divided into two categories, one for the highest-density locations of the city, the other for the rest of the city. For simplicity, take an average of the two.

**AND**

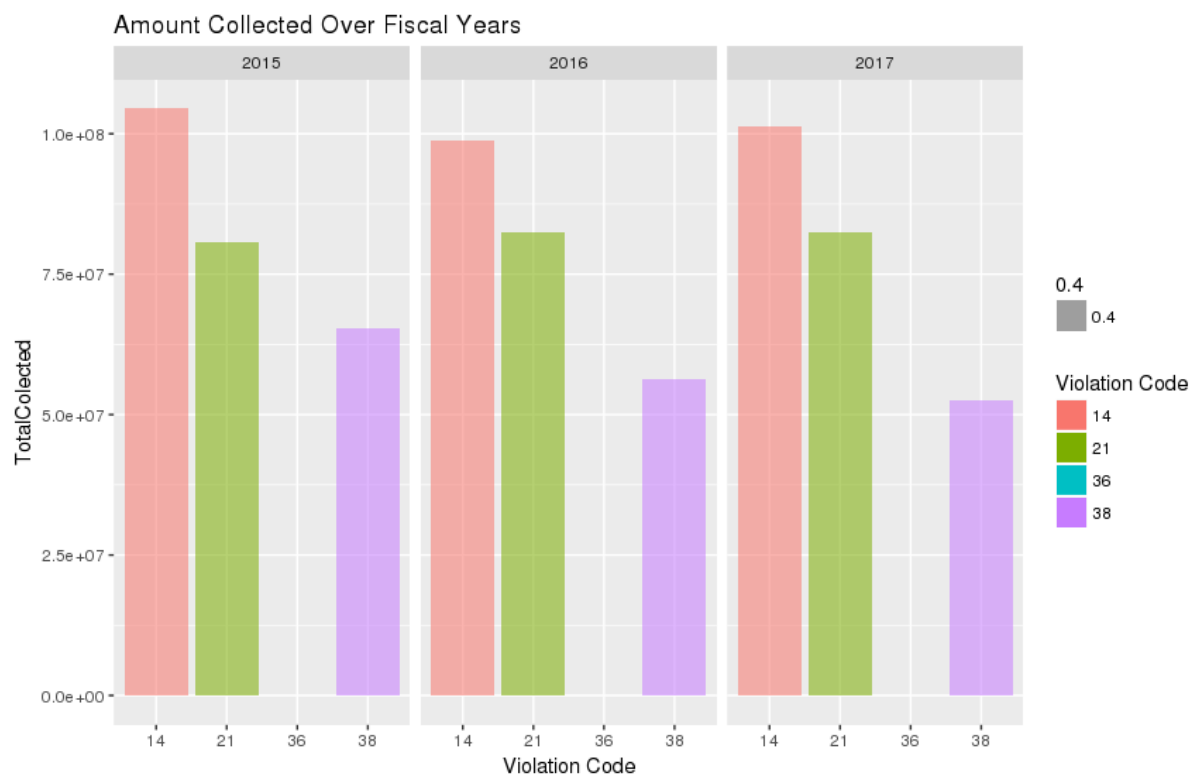
**7.c.** Using this information, find the total amount collected for all of the fines. State the code which has the highest total collection.

**Note:** The Fine for Violation Code 36 has not been mentioned on the government website. Therefore, we code it as zero. As a result we will also find the 4th frequent code

**Result:**

Year	Code	Amount Collected
2017	14	101217480
	21	82521780
	38	52520900
	TOTAL	236260160
2016	14	98905175
	21	82349795
	38	56341750
	TOTAL	237596720
2015	14	104468070
	21	80807540
	38	65250350
	TOTAL	250525960

**Plot:**



**7.d.** What can you intuitively infer from these findings?

**Result:**

- We can see that the revenue collected keeps decreasing every year
- Violation Code 14, brings the maximum Revenue every year

## CONCLUSION

SUBN is the most frequent occurring Body Type across all years. Frequency of SUBN increases over time whereas rest of them in top5 show a downward trend. However, this does not tell us about the cause as this might depend on vehicle popularity too.

FORD is the highest Vehicle Make however it has a decreasing trend, CHEV is also decreasing. TOYOTA is increasing along with HONDA and NISSAN.

The highest occurring Violation code for a particular precinct remains fairly similar over years. The violation codes for a particular time of day remains fairly similar across years.

More tickets are issued in Precinct 0 and increases while rest of them tend to have a decreasing trend year on year.

We can see that the revenue collected keeps decreasing every year.

Violation Code 14, brings the maximum Revenue every year.