Satyam Kumar (*satyamk@iisc.ac.in*)
Siva Kranthi Kumar Mallipeddi (*sivam@iisc.ac.in*)
Sreedhar Reddy Vundela (*sreedharv@iisc.ac.in*)
Sudhakar Kulkarni Mukayya (*sudhakark@iisc.ac.in*)


# NYC Parking Violation Data Analysis

Mentor: **Yogesh Simmhan**
simmhan@iisc.ac.in

IISC | M.Tech (Online) | DSBA
DA 231-O Data Engineering at Scale

# TEAM


**Satyam Kumar**
satyamk@iisc.ac.in
9092696415



**Siva Kranthi Kumar Mallipeddi**
sivam@iisc.ac.in
9986778909



**Sreedhar Reddy Vundela**
sreedharv@iisc.ac.in
9886702749



**Sudhakar Kulkarni Mukayya**
sudhakark@iisc.ac.in
9035076656

# PROBLEM

→ Analyzing **NYC Parking violation** data from **2017 – 2021** https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/pvqr-7yc4

→ Infer from the data analysis findings

# Environment

- → Python 3.10
- → Spark 3.1.1
- → Spark RDDs | Spark Data Frames | Spark SQL
- → 2017-2021 datasets from NYC Parking Violation
- → Uses Google Colab for execution

# APPROACH: Data Preprocessing

- → Selecting subset of the data as it is huge
- → Load data into RDD. Use ***Summons Number*** as key & remaining columns as value
- → Replace/discard Invalid data
- → Converting date to proper date/datetime format
- → Find a way to deal with missing values, if any

# APPROACH: Basic Analysis

1) How often does each violation code occur? (frequency of violation codes find the top 5)
2) How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

# APPROACH: Precinct based Analysis

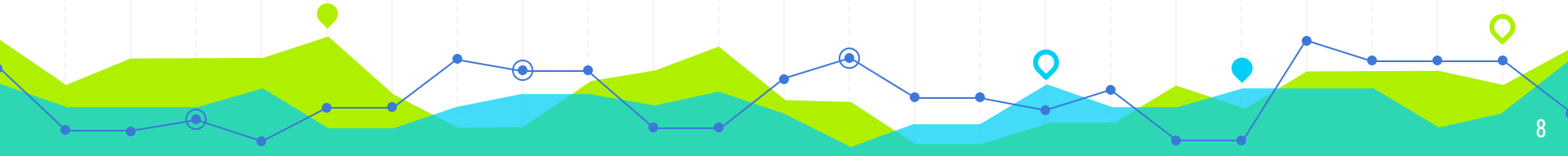
A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

1) Violating Precincts (this is the precinct of the zone where the violation occurred)
2) Issuing Precincts (this is the precinct that issued the ticket)
3) Find the violation code frequency across 3 precincts which have issued the most number of tickets

# APPROACH: Time based Analysis

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

1) Divide 24 hours into 6 equal discrete bins of time. For each of these groups, find the 3 most commonly occurring violations
2) For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part)



## APPROACH: Year / Season based Analysis

1) What is the average reduction in violations for the year 2020 compared to 2019 (due to COVID), and year 2019 compared to 2018
2) Divide the year into 3 number of seasons, and find frequencies of tickets for each season
3) Find the 3 most common violations for each of these season

# APPROACH: Revenue based Analysis

The fines collected from all the parking violation constitute a revenue source for the NYC police department. Gather the fine amounts for each code from NYC site

1) Find the total amount collected year wise
2) Find the top 5 violation codes which collected highest amount

# Evaluation

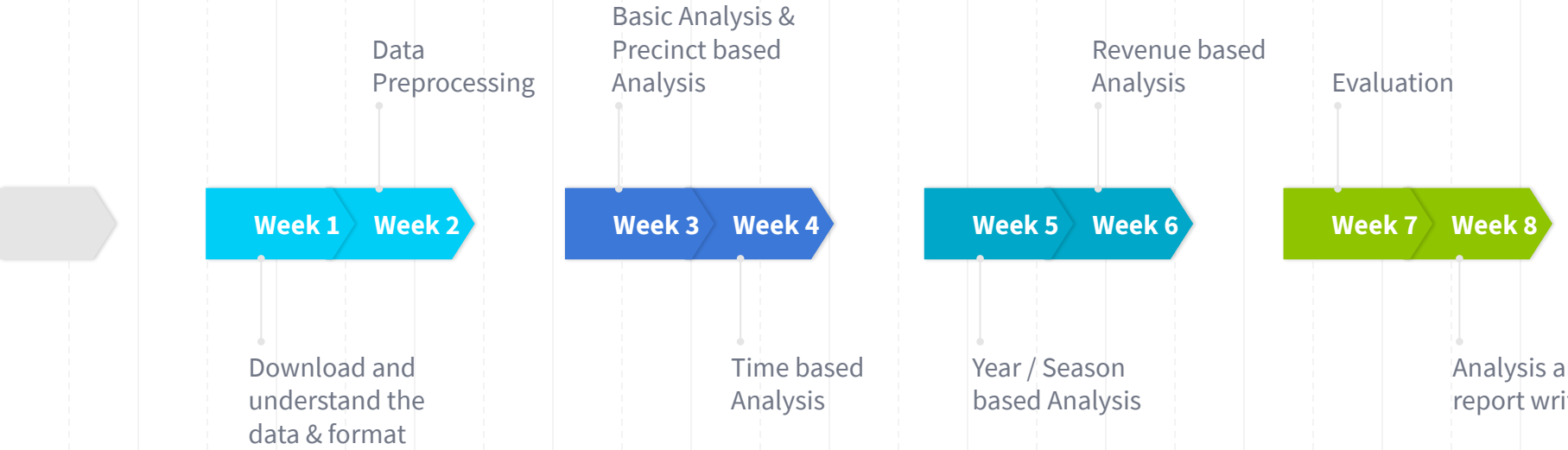
Test on

- → 10k, 100k, 1M, 10M records from 2017-2021
- → Evaluate the time taken by each analysis for different data sizes

Metrics

- → Use Google Colab
- → Ability to complete data analysis for 10M records within 1 hour

# TIMELINE


Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Data Preprocessing
Basic Analysis & Precinct based Analysis
Revenue based Analysis
Evaluation
Download and understand the data & format
Time based Analysis
Year / Season based Analysis
Analysis and report writing

# THANKS!

www.kaggle.com/sarthaksarbahi/nyc-parking-tickets-analysis
www.slidescarnival.com