Chicago Public Taxi Data Analysis (over 200 million rows!)

Sumeet Badgujar Follow 5 min read · Jul 28, 2022

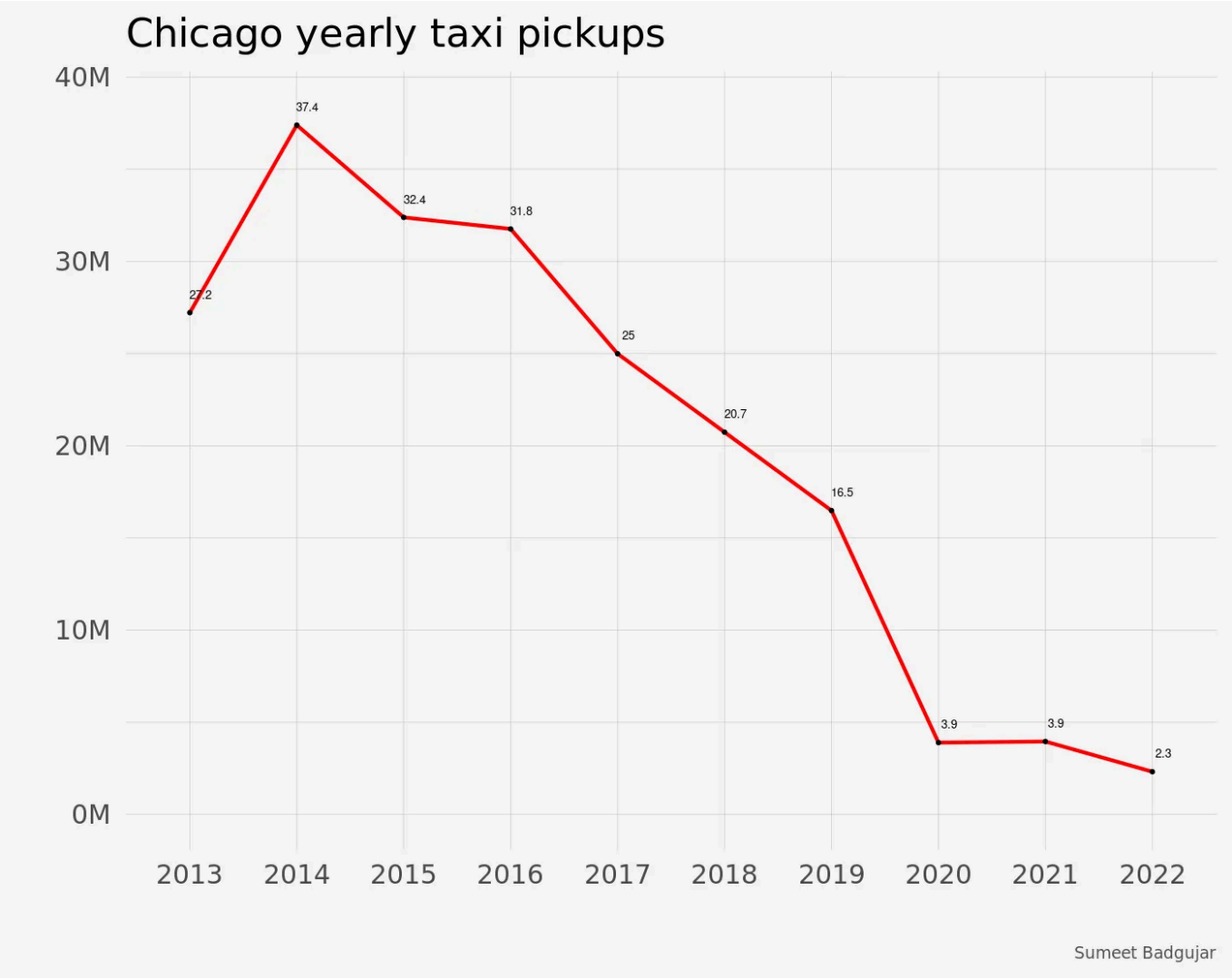
 4    

The city of Chicago has released a public dataset containing over 200 million taxi rides since 2013. The dataset is too big to download ~ around 75 Gigs and run Exploratory Data Analysis on my not-so-muscular laptop. But I know somebody with big muscles, a fast database, and who can run SQL commands amazingly fast! Behold the GOOGLE!

The Chicago taxi trips dataset is hosted on Google public datasets which can be used on Google Cloud Console. The dataset does not include data from ridesharing companies like Uber and Lyft, just public taxis. It is a huge dataset, and a lot of EDA can be done on it. So I said -

Yearly Taxi Pickups

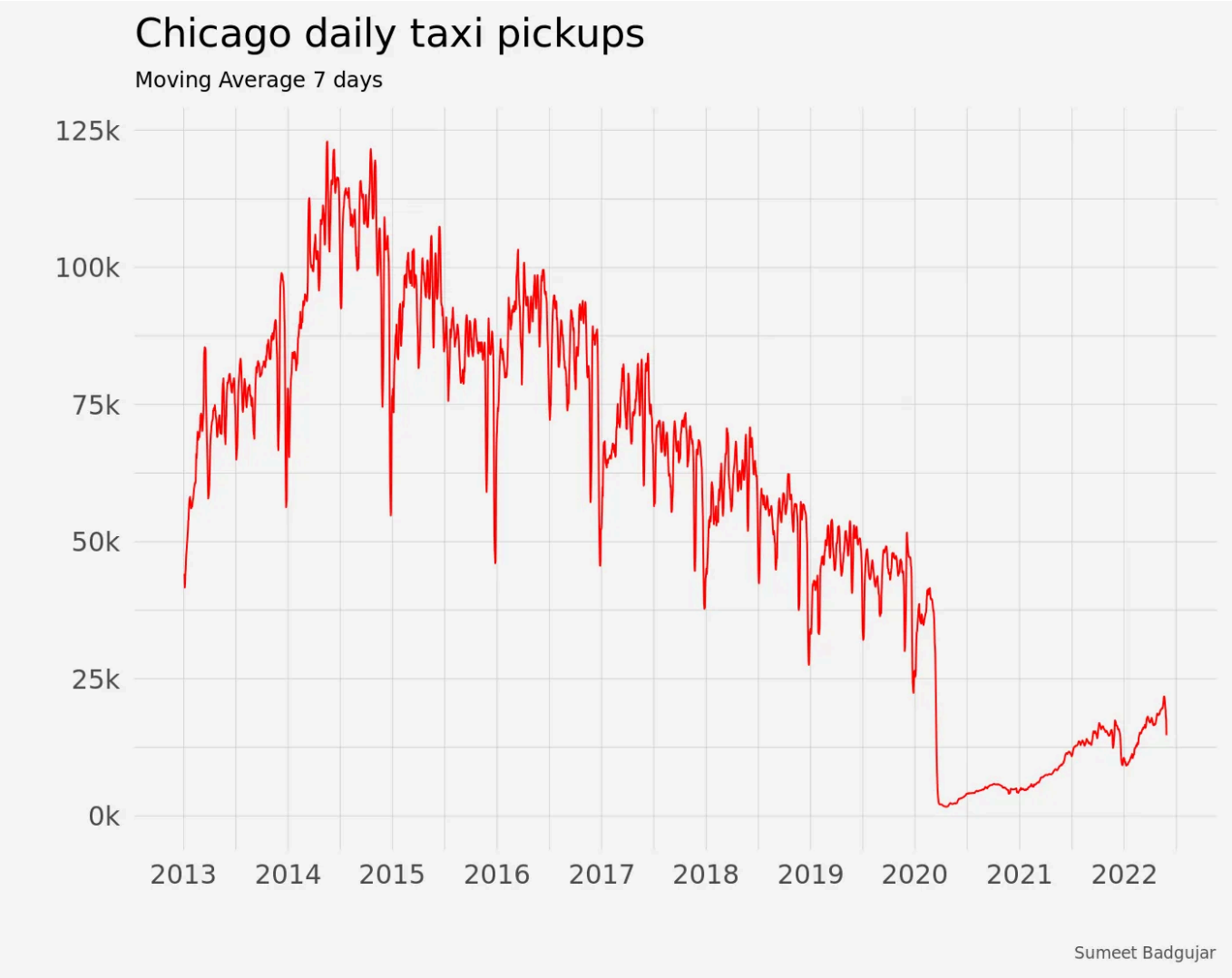
The data does make it clear that taxi ridership has dramatically decreased. Since the peak of 2014, taxi usage has been declining at an average annual rate of 20% till 2019. By 2019 (pre-pandemic), the cumulative drop was 56% since the peak.



If we consider 2020, the drop in ridership was 76.4% from 2019, almost killing the public taxi system. But it can't be compared as all businesses took a hit due to the pandemic. There has been a gradual rise following 2020, but it can't be compared as a comeback.

Daily taxi pickups

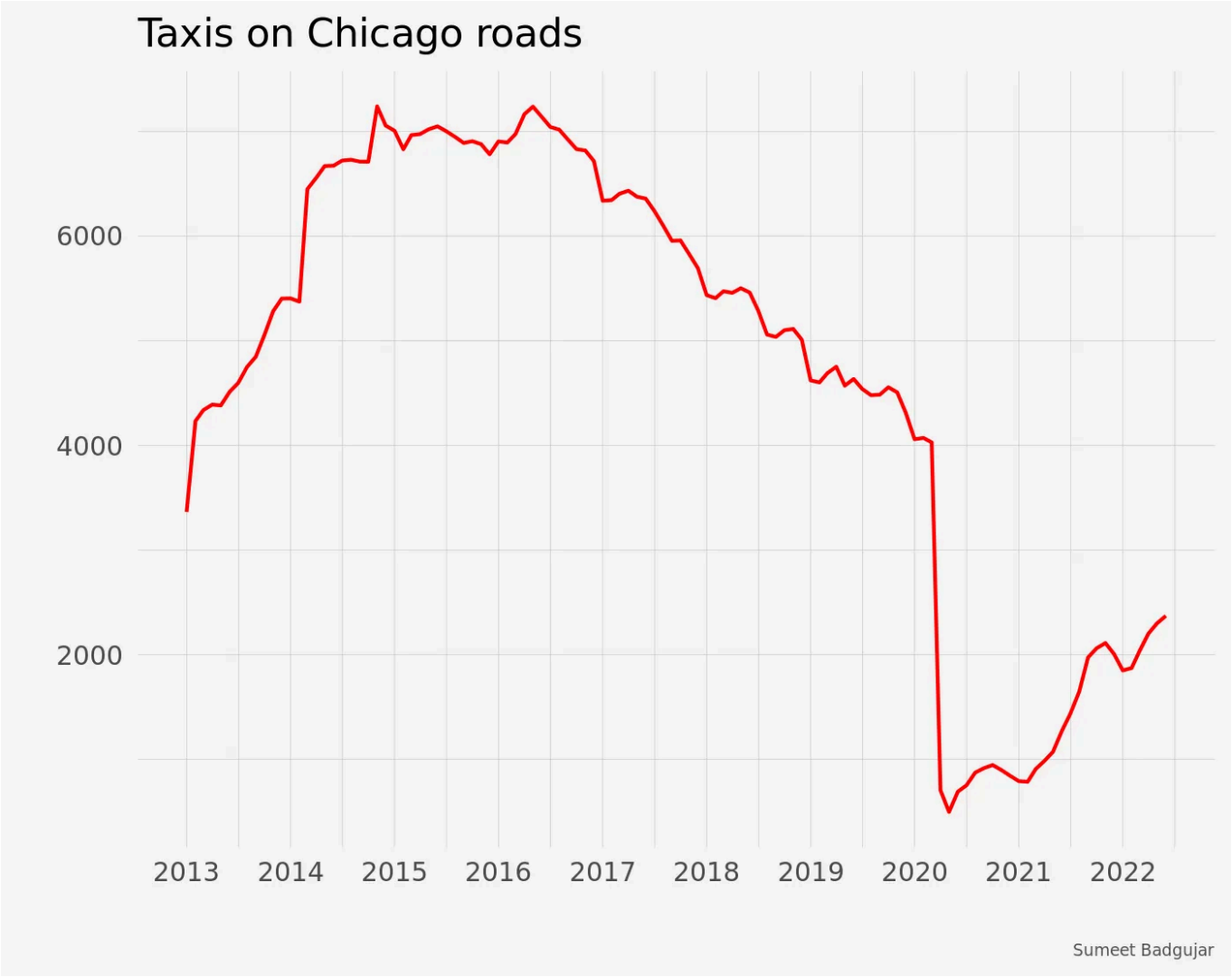
The peak was around the 2014 end of 125k trips per day, a considerable number. Compared to 2019, the peak was around 50k daily trips, a downfall of 60%.



Active taxis on roads over the years

Chicago public taxi dataset has anonymized taxi medallion numbers for each trip. This makes it possible to do many things.

1. Count the number of unique taxis per month
2. See area preference by taxi id



One possible reason for fewer rides could be simply attributed to fewer taxis on the road. Though there are multiple factors in play for fewer taxis on the road — foreclosure of medallions, competitive pricing, Ride sharing services, ease of technology, and many more.

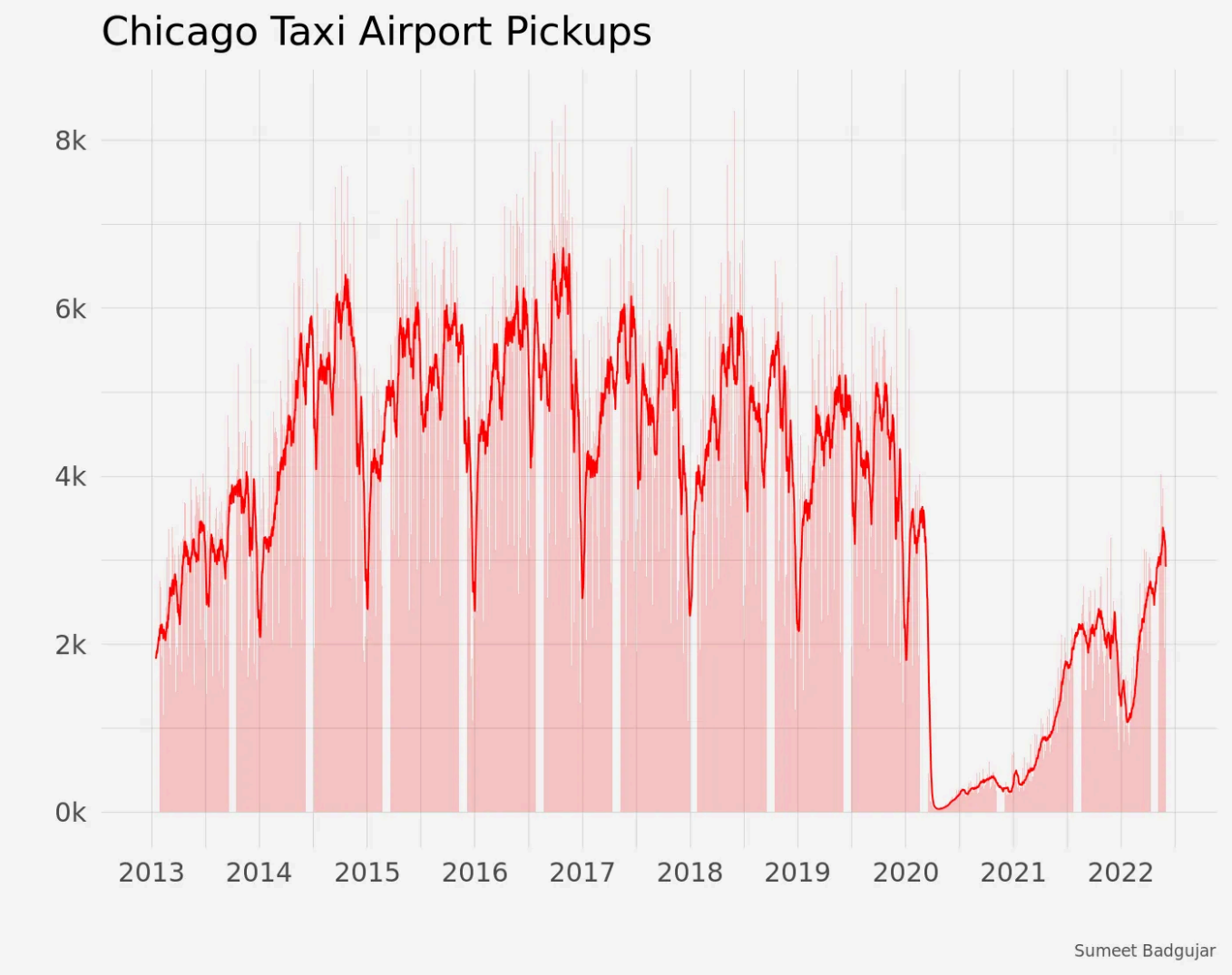
Area wise pickups

The Chicago area is divided into 77 community areas for city planning. The dataset does not pinpoint the location (geocodes); it gives the pickup and dropoff community area. It is done to protect the privacy of the rider and the cab. This makes it challenging to do an in-depth analysis of taxi routes between areas or impossible to do within the area.

But wait!

Interestingly, a community area covers just one single location, its the International airport of Chicago. So EDA can be done for the airport rides.

Compared to the general downtrend of the rides, the expectation was the same for airport pickups. But the reality was different!

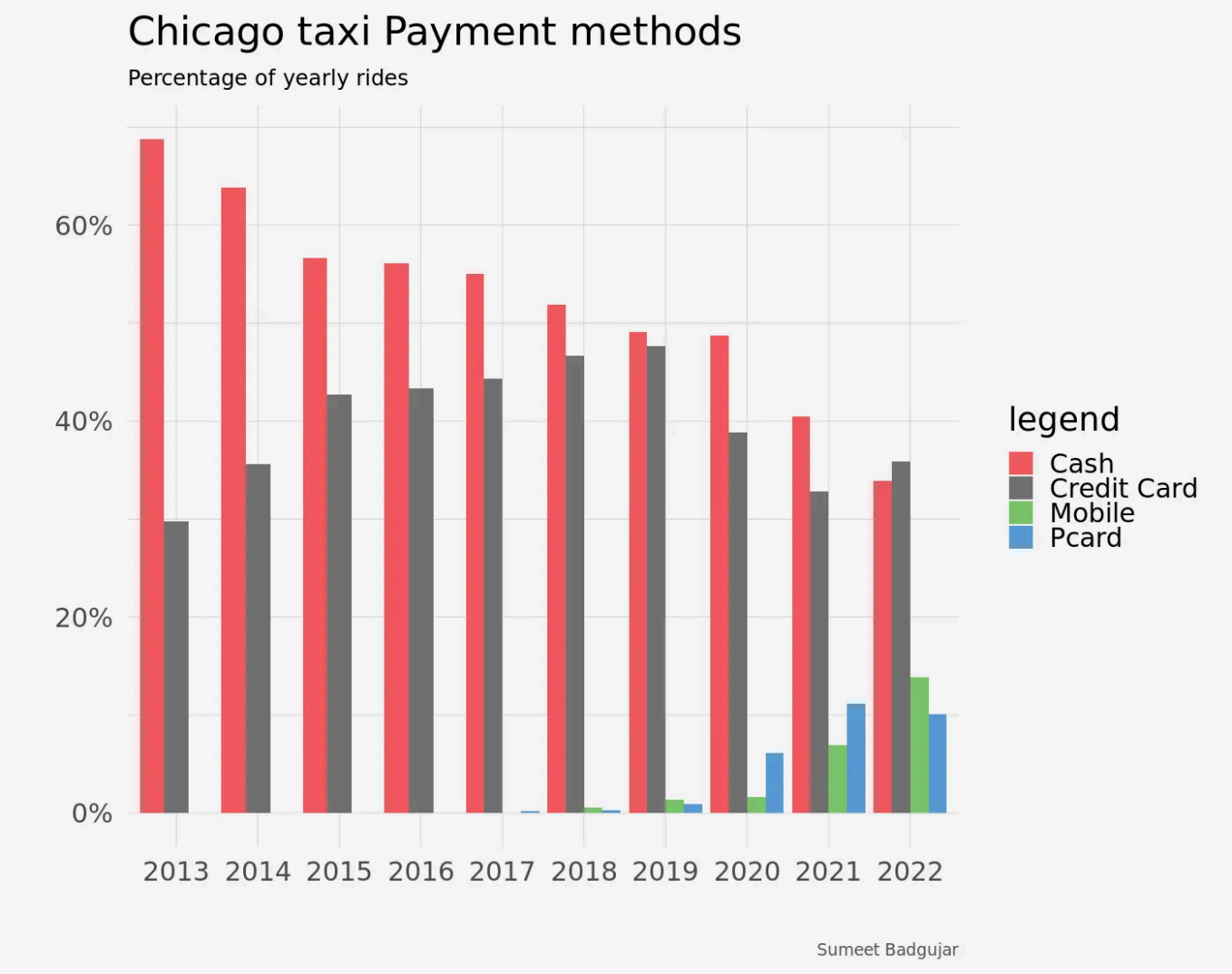


The peak for airport pickups was in October 2016. A simple search of October 2016 Chicago on google informed me of the 2016 World series. The

baseball fans came in numbers, and man, they did spike the airport pickup numbers.

Payment Methods over the years

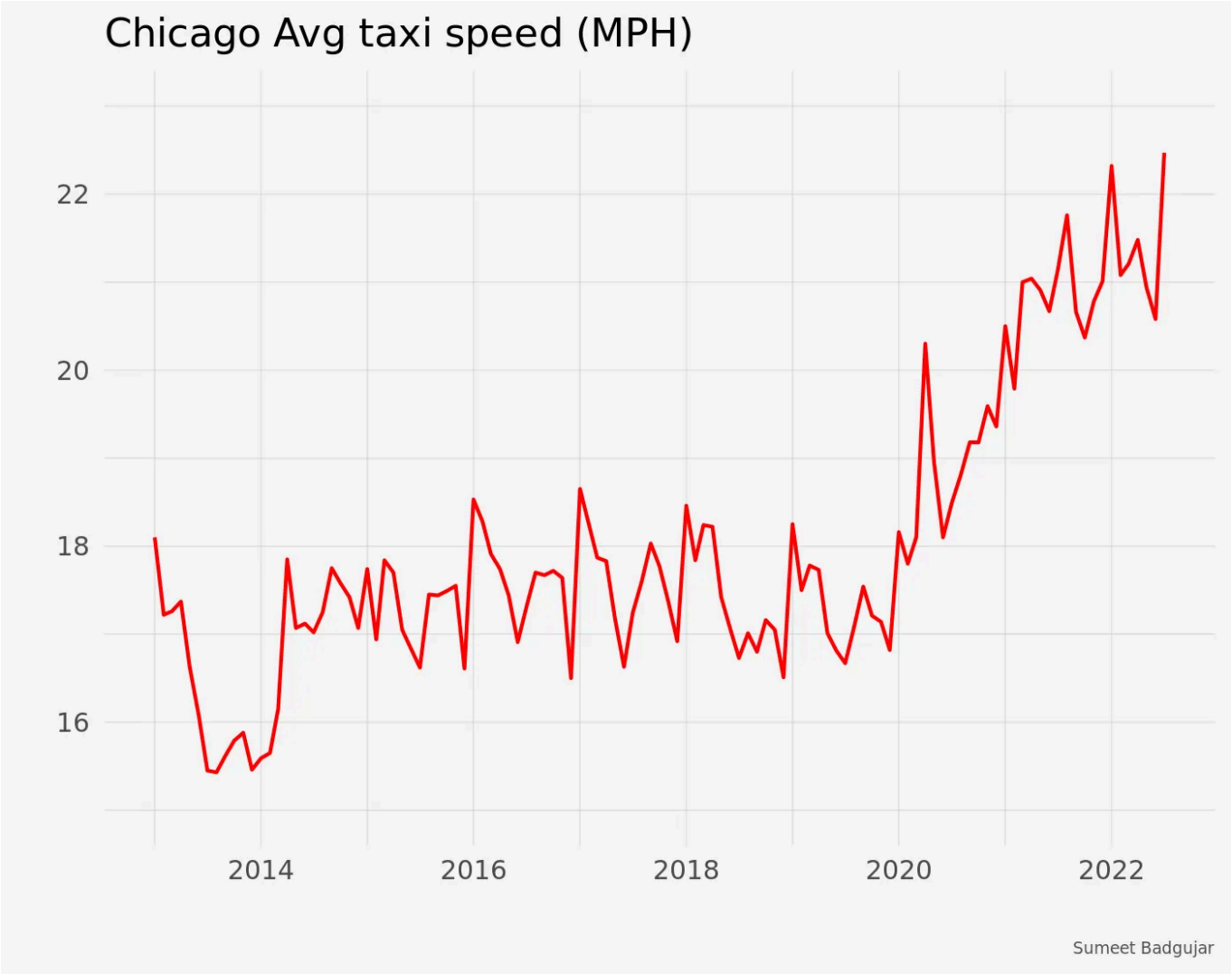
The Chicago database has undefined or unsettled payments too in the payment column i.e. Unknown and Dispute. Excluding those, we can see clearly that Cash payment has been decreasing, and the rise of credit cards compensates for that decrease. The US is late to the party in adopting cardless, cashless Mobile payments. Mobile payments started in 2018 and are slowly increasing; as of 2022, it accounts for 17%.



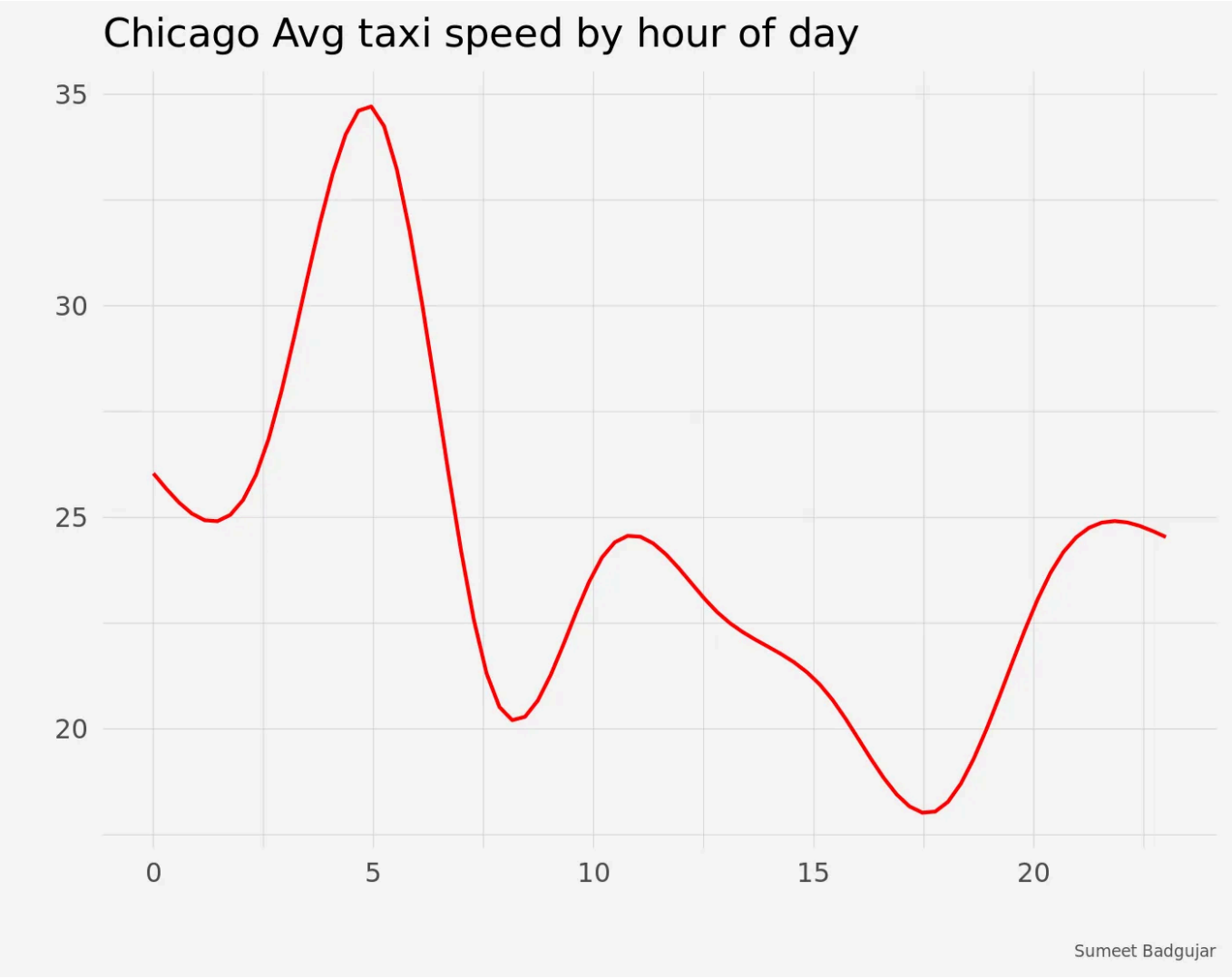
Pcard here is another slowly rising type of card payment. It is a prepaid card, used mainly by business people as a company card. Its sudden rise from 2020 can be explained as employers giving their employees cards for transportation — to bring people back to the office (like an incentive)

Taxi Speed over years

In a city, over the years, traffic can either improve or deteriorate to bumper to bumper. Doing EDA on a dataset and grouping by month, I was surprised to see an overall average speed improvement.



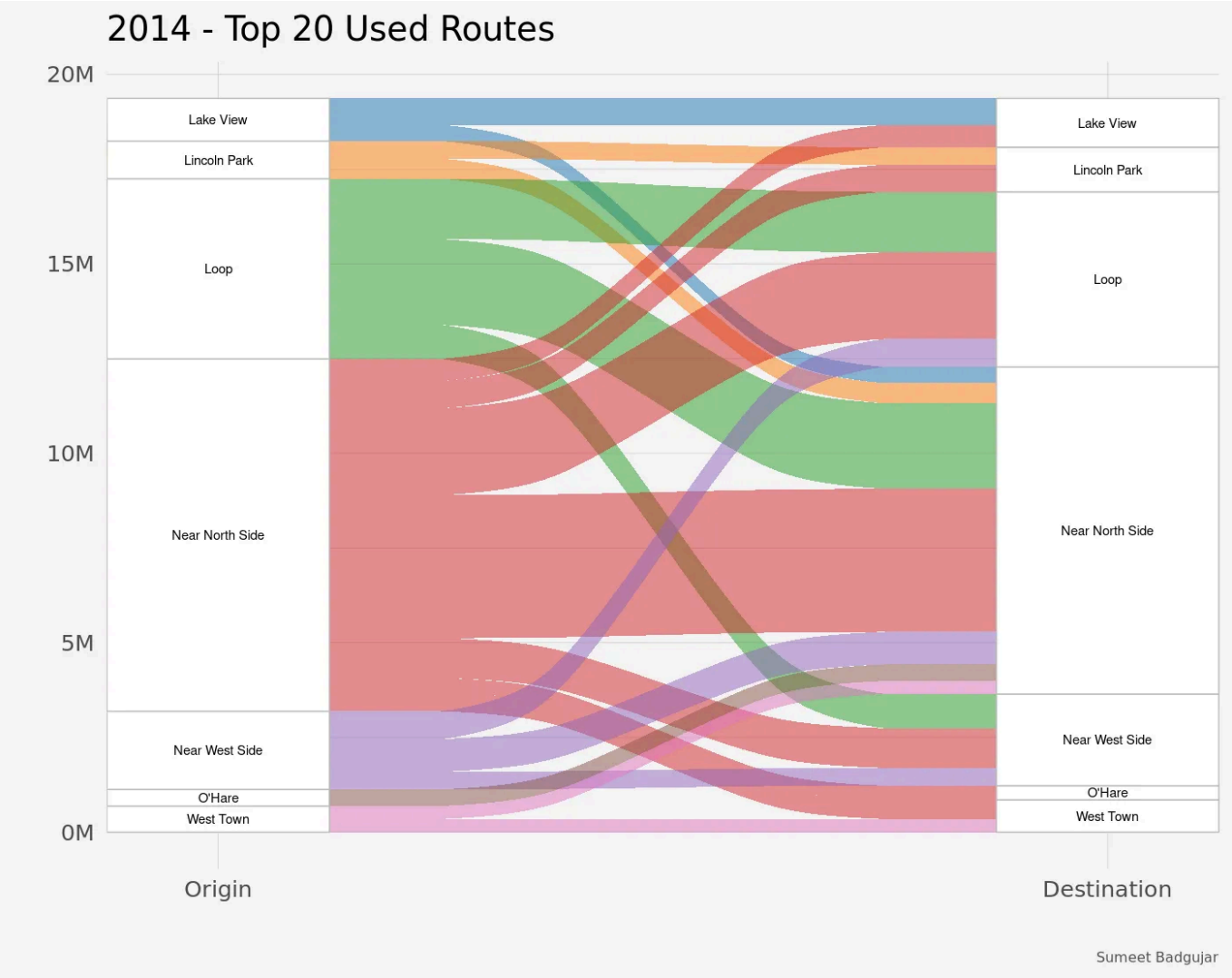
After looking at the plot, I thought this didn't show the complete picture. I increased the granularity of the analysis to an hour of the day.



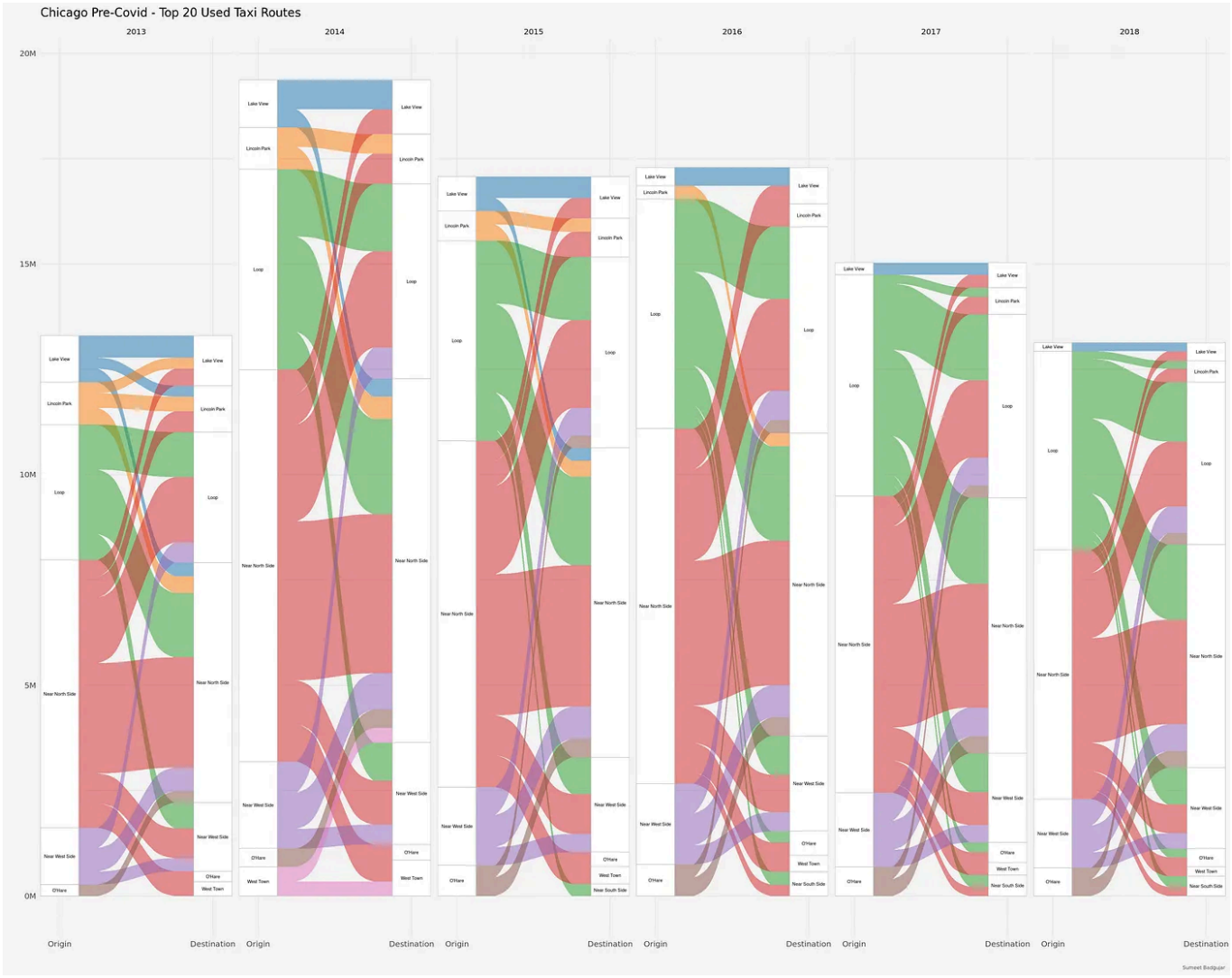
Now here, it somewhat shows the complete picture. In the early morning, the avg speed is the highest, touching 32 mph at its peak. There are two deep valleys in the chart, one around 8 am — office start time and the other around 6 pm — office end time.

Most used routes

In Chicago, most rides start from downtown areas — Near North Side and the Loop. Out of the top 20 most used routes, 6–7 start from the Near North Side and 3–4 from Loop. For 2014 (the peak period), rides starting from these two areas accounted for 45% of the 2014 rides.



Here's a plot of all pre-covid years top 20 used routes for side-by-side comparison.



Link for Interactive Analysis —

https://public.tableau.com/app/profile/sumeet.badgujar/viz/ChicagoPublicTaxiAnalysis/ChicagoPublicTaxiAnalysis

Data Science | Sql | Data Visualization | Big Data | Analytics

Written by Sumeet Badgujar

Follow

29 Followers · 12 Following

A guy interested in Data Science and Ex-Machine Learning Engineer, doing data analysis and fun AI projects. “Ore wa Kaizoku Ou ni naru!”

No responses yet

Write a response

What are your thoughts?

More from Sumeet Badgujar

$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1

In Analytics Vidhya by Sumeet Badgujar

Creating MobileNetsV2 with TensorFlow from scratch

MobileNet models are very small and have low latency. The MobileNet models can be...

Sumeet Badgujar

Colorizing Manga using Neural Networks (P1)

Being an avid reader of manga, I see digital artists coloring the black and white pages an...

Jul 17, 2021 36 1

Aug 6, 2021 6 2

Sumeet Badgujar

5 Reasons You should read Manga

Manga: The New Pop Culture

Jul 23, 2019 4

Sumeet Badgujar

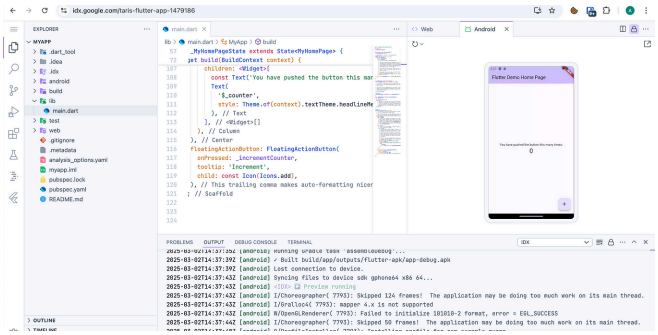
SqueezeNet implementation in TensorFlow

SqueezeNet provides a smart architecture that achieves AlexNet-level accuracy on...

Jul 26, 2021 54

See all from Sumeet Badgujar

Recommended from Medium

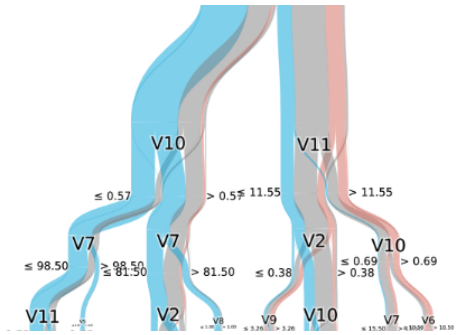


In Coding Beauty by Tari Ibaba

This new IDE from Google is an absolute game changer

This new IDE from Google is seriously revolutionary.

Mar 11 4.9K 282



In Top Python Libraries by ZHEMING XU

How to visualize Decision Trees and Random Forest Trees?

Look at the trees with your eyes

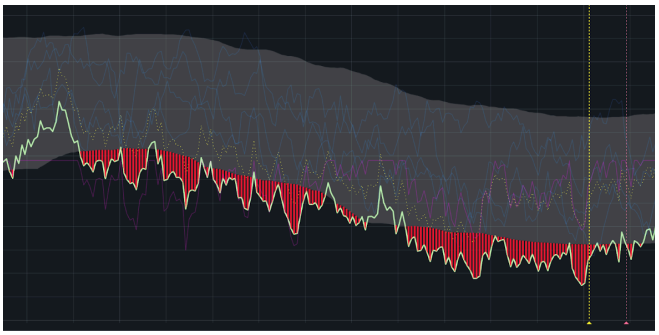
Mar 31 33

In Learn AI for Profit by Nipuna Maduranga

You Can Make Money With AI Without Quitting Your Job

I'm doing it, 2 hours a day

Mar 24 7.7K 361

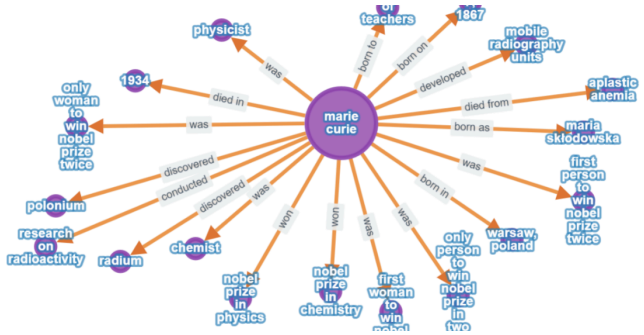


In Booking.com Engineering by Ivan Shubin

Anomaly Detection in Time Series Using Statistical Analysis

Setting up alerts for metrics isn't always straightforward. In some cases, a simple...

Apr 15 234 5



In Level Up Coding by Fareed Khan

Converting Unstructured Data into a Knowledge Graph Using an End...

Step by Step guide

Apr 17 1.5K 25

In Psyc Digest by Alessia Fransisca

The 1-Minute Introduction That Makes People Remember You...

A Behavioral Scientist's Trick to Hack the "Halo Effect"

Apr 12 21K 491

See more recommendations