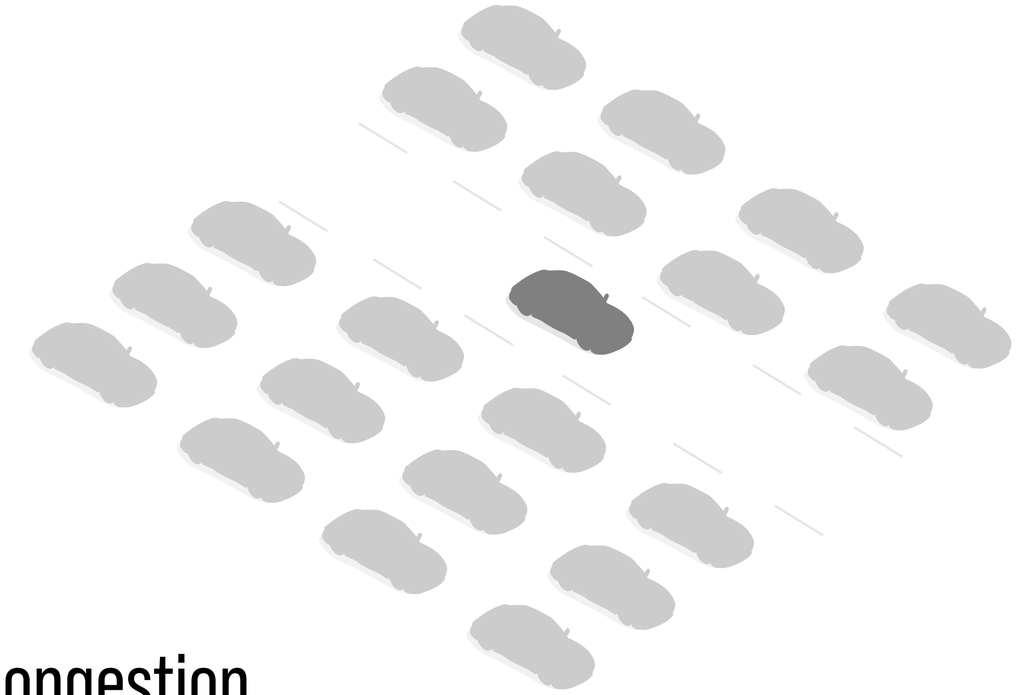


traffic congestion

Duration Prediction • Anomaly Detection

by: Da Guo

Da Guo

da.guo@wsu.edu

linkedin.com/in/da-guo

dagoku.com

medium.com/@opophehu

github.com/opophehu

architecture

retail tenant improvement

web/software development

computer vision, video game, cyber security

data science

statistics, predictive modeling, analytics

the **why**.

the **what**.

the **product**.

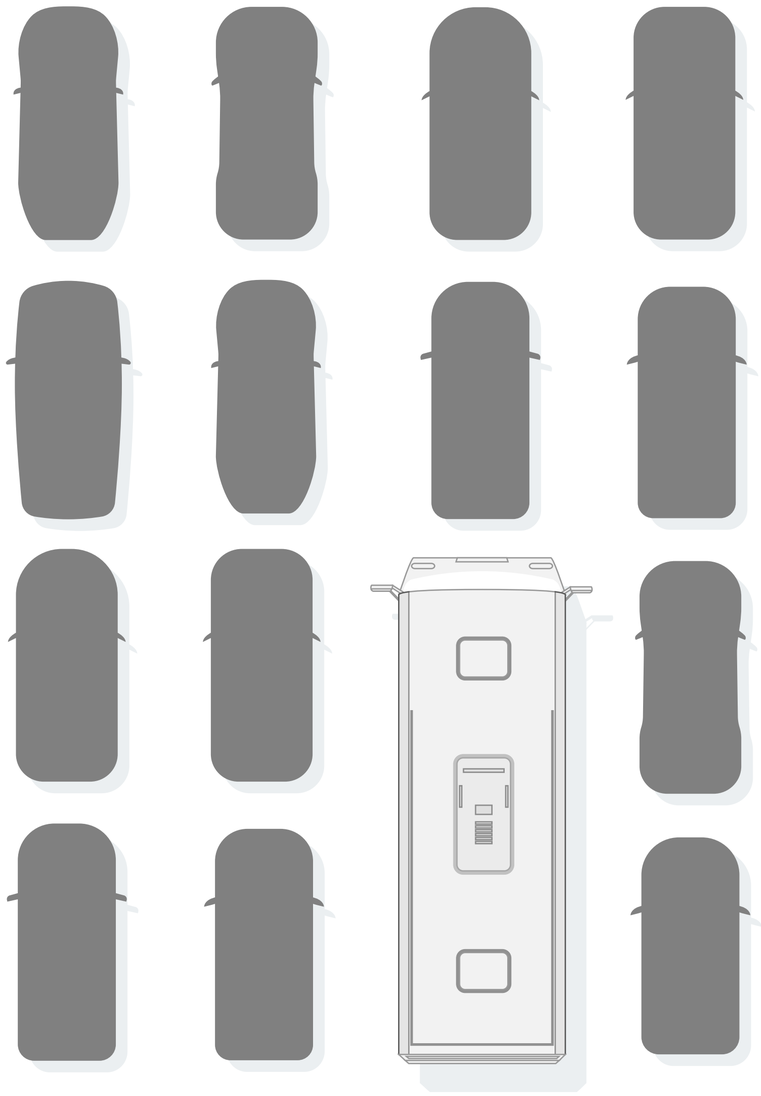
the **how**.

the **findings**.

~ 7 mins

the **why**.

off work. finally.



~~20 mph~~

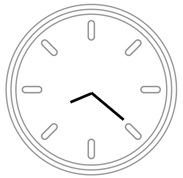
8 mph

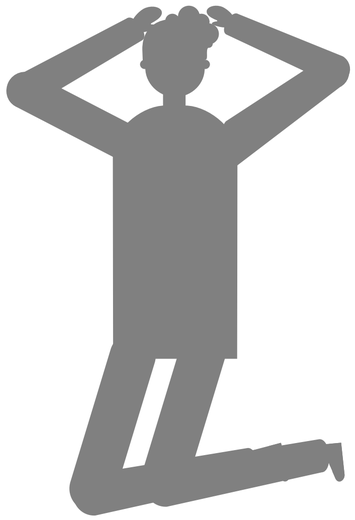
~~50 mins~~

2 Hrs

stuck.

the story told





only if.

the **what**.

what.

the questions asked

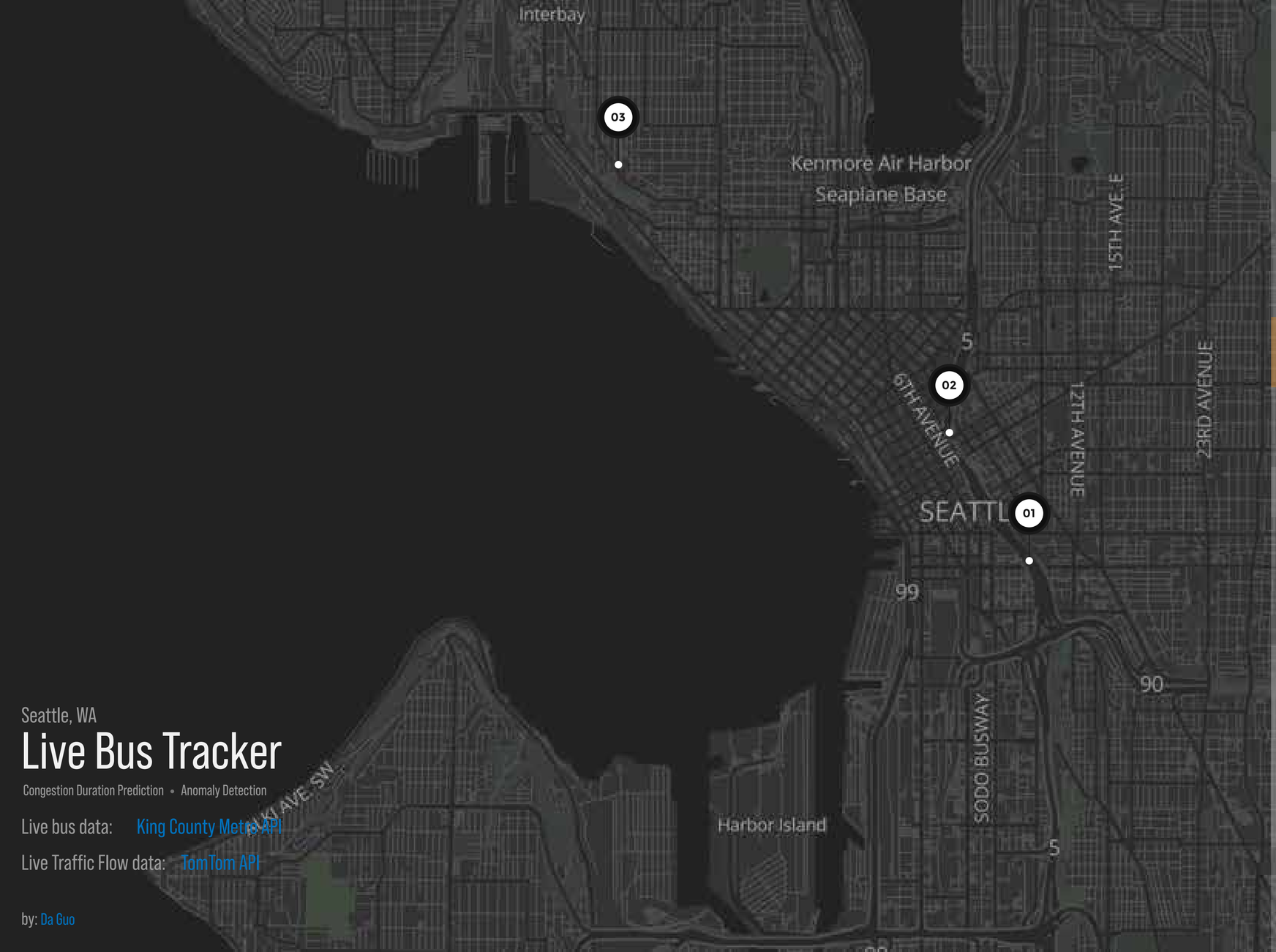
1

Can I predict the **duration of congestion** if that happens?

2

Can I **detect the anomalies** in traffic congestion events?

the product.



Seattle, WA

Live Bus Tracker

Congestion Duration Prediction • Anomaly Detection

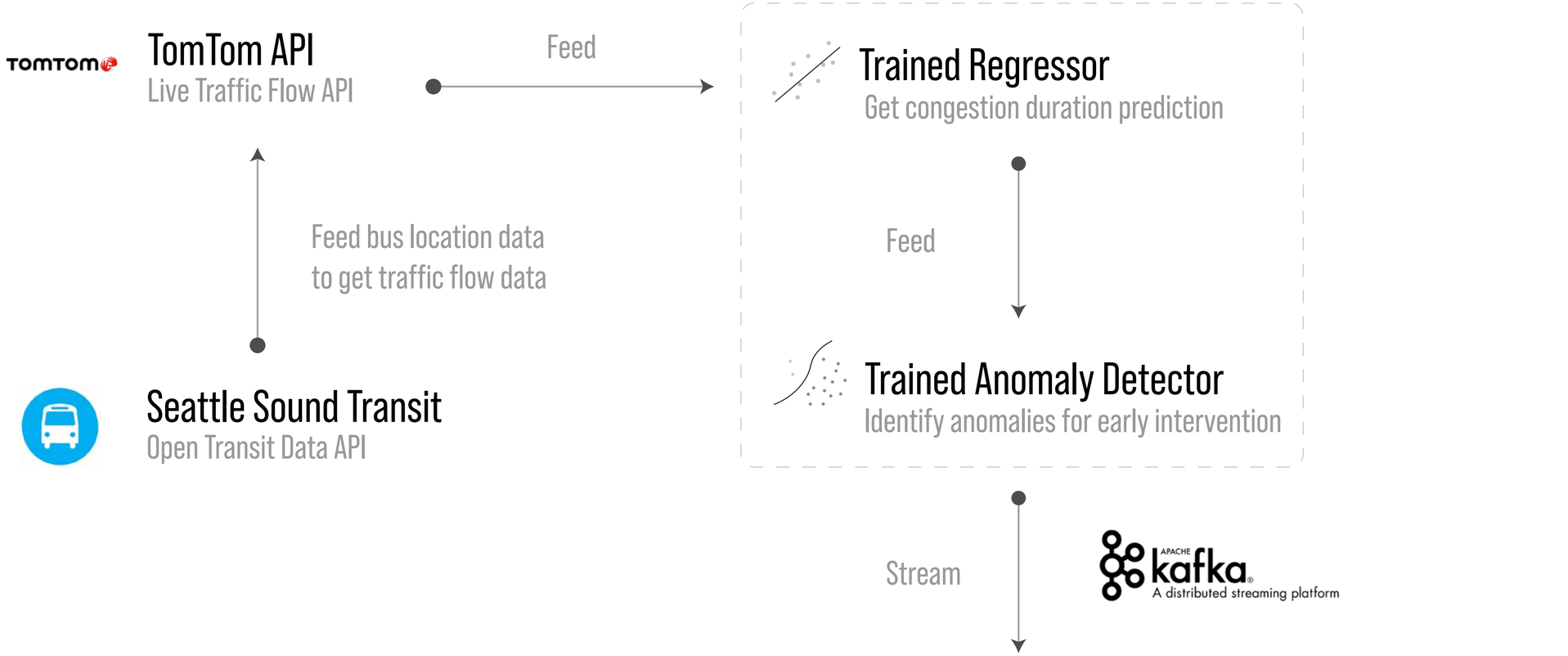
Live bus data: King County Metro API

Live Traffic Flow data: TomTom API

by: Da Guo

Busline	Message	Timestamp	Coordinates
01	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
02	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
01	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
02	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	Congestion detected, estimated duration: 92.21 minutes. !!!This is an anomaly!!! Please contact local traffic authorities.	2020-6-13 9:47:38	47.6398, -122.37031
01	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
02	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
01	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
02	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
01	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
02	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
03	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
01	Congestion detected, estimated duration: 47.68 minutes. This is normal, please be patient.	2020-6-13 9:47:38	47.6398, -122.37031
02	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031
03	No congestion detected.	2020-6-13 9:47:38	47.6398, -122.37031

backend architecture



the **how**.

how.
the process explained

Data

1

Traffic Event Data (25 million)
Congestion, Incidents, etc.
Severity, Type, Location, etc.

+

Weather Event Data (6 million)
Rain, Hail, Fog, etc.
Severity, Type, Location, etc.

+ 2

Population Data
by County, City, State, etc.
Density, Size, Male to Female Ratio, etc.

50 states / partitions

AL
⋮
WA
As focus
⋮
WY

Feature Engineering

Features:

Original (24)

↓

325 — Polynomial Transformation & Feature Interactions

↓

20 — Feature Selection

Sklearn - SelectFromModel
Threshold of .05 - each feature has to impact the model performance by 5%

Classification - Anomaly Detection

With all available information, now we can detect anomalies if there is any.

Unsupervised

Ensemble: IsolationForest

How do we know it's working?

normal

	LocationLat	LocationLng	ZipCode	Density(isqmi)	StartTime(UTC)	Severity	Duration	isNormal
274807	47.29	-122.31	98003	870.90	2019-02-09 07:33:00	1	41	1
271808	47.34	-122.29	98003	870.90	2019-02-09 07:34:00	1	41	1
271828	47.29	-122.30	98003	870.90	2019-02-09 07:43:00	3	47	-1

3 Obs — Same Location — Same Hour — sequential

anomaly

Regression - Predict Duration of Congestion

At any given time, we need to predict possible duration of the congestion if that happens.

Supervised

Ensemble: XGboost

R^2 0.19 → 0.53

RMSE 127.1 → 58.4

1 "Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data."
Moosavi, Sobhan, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath.
In proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.

2 US Census Bureau
Population Estimate (as of July 1) - 2018 - Both Sexes
https://www.census.gov/data.html

how.
the process explained

Data

1

Traffic Event Data (25 million)
Congestion, Incidents, etc.
Severity, Type, Location, etc.

+

Weather Event Data (6 million)
Rain, Hail, Fog, etc.
Severity, Type, Location, etc.

+ 2

Population Data
by County, City, State, etc.
Density, Size, Male to Female Ratio, etc.

52 states / partitions

AL
⋮
WA
As focus
⋮
WY

Feature Engineering

Features:

24 Original

↓

325 Polynomial Transformation & Feature Interactions

↓

20 Feature Selection

Sklearn - SelectFromModel
Threshold of .05 - each feature has to impact the model performance by 5%

Classification - Anomaly Detection

With all available information, now we can detect anomalies if there is any.

Unsupervised

Ensemble **IsolationForest**

↓ How do we know it's working?

normal

	LocationLat	LocationLng	ZipCode	Density(/sqmi)	StartTime(UTC)	Severity	Duration	isNormal
271807	47.29	-122.31	98003	870.90	2019-02-09 07:33:00	1	41	1
271808	47.34	-122.29	98003	870.90	2019-02-09 07:43:00	1	41	1
271828	47.29	-122.30	98003	870.90	2019-02-09 07:43:00	3	47	-1

3 Obs — Same Location — Same Hour (sequential)

anomaly

Regression - Predict Duration of Congestion

At any given time, we need to predict possible duration of the congestion if that happens.

Supervised

Ensemble **XGboost**

R^2 0.19 ⟶ 0.53

RMSE 127.1 ⟶ 58.4

1. "Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data."
Moosavi, Sobhan, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath.
In proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019.

2. US Census Bureau
Population Estimates (as of July 1) - 2018 - by US States
https://www.census.gov/data.html

the findings.

weather has little effect

In Washington state,

Average Congestion Duration:

	noWthrEvent	Cold	Fog	Hail	Precipitation	Rain	Snow	Storm	wthrAverage	durationDifference
count	502.000	12.000	35.000	3.000	25.000	35.000	33.000	2.000	502.000	502.000
mean	[illegible]	[illegible]	[illegible]	[illegible]	43.066	46.908	[illegible]	[illegible]	[illegible]	0.277
std	10.932	14.189	9.524	7.656	8.716	11.498	6.837	1.591	10.878	1.865
min	17.449	11.000	34.204	29.167	27.750	20.208	11.500	41.500	17.449	-8.141
25%	42.526	[illegible]	[illegible]	[illegible]	38.133	40.094	38.875	[illegible]	[illegible]	0.000
50%	46.912	42.750	46.000	41.000	42.000	45.500	41.000	42.625	46.727	0.000
75%	51.735	51.188	50.371	42.250	46.000	53.261	43.333	43.188	51.427	0.000
max	139.699	57.000	75.000	43.500	69.926	70.703	50.633	43.750	139.699	17.568

with NO weather event

48.6 mins

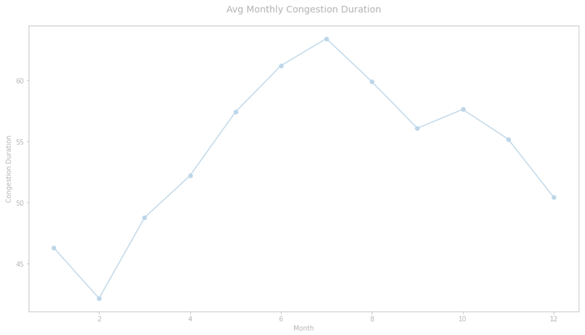
with weather events

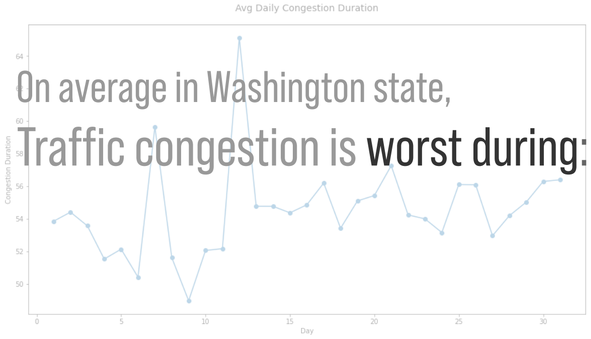
48.4 mins

Average Weather Severity:

1.5 Good-Mild Weather

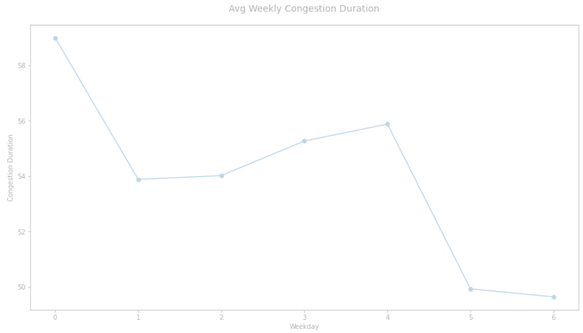
when is congestion worst

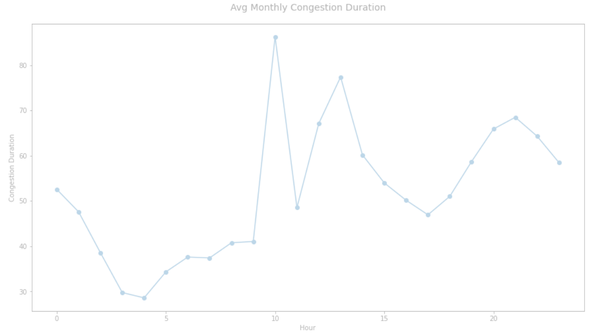




On average in Washington state,

Traffic congestion is worst during:





Most Impactful factor/features:

Summer months

June July Aug

First & Second Weekend of each month

7th & 12th day of each month

Mondays & Fridays

Monday being the worst

10 am, 1 pm, and 9 pm

10 am being the worst

Urban vs. Rural

Severity How severe the congestion is (0 - 3)	x	**Density** Population per SqMi
Severity How severe the congestion is (0 - 3)	x	**Longitude** How west or east you are in the state ex. -122.234

* **Severity** is calculated by speed difference between your current speed vs. speed limit.
i.e. Current speed is 50mph, speed limit is 60mph, 10mph in difference, it's severity 1, 20 is 2, 30+ is 3. 0 means no slow down.

when do anomalies happen

	nor_mean	ano_mean	diff	diffPercent
ZipCode	98253.08	98910.79	657.71	0.67
Duration	51.26	52.12	0.86	1.67
Density(/sqmi)	593.55	122.75	470.80	79.32
AvgAge	37.79	38.16	0.96	2.58
isWeekday	0.35	0.52	0.34	39.52
Mon	0.12	0.13	0.01	12.62
Tue	0.17	0.10	0.07	43.24
Wed	0.18	0.11	0.08	41.19
Thu	0.19	0.10	0.09	48.09
Fri	0.19	0.08	0.11	57.98
Sat	0.11	0.18	0.07	66.96
Sun	0.04	0.31	0.27	668.13
Spring	0.22	0.30	0.08	35.67
Summer	0.23	0.24	0.02	7.77
Autum	0.27	0.23	0.04	15.46
Winter	0.28	0.23	0.05	19.49
Morning(6-12)	0.03	0.35	0.32	1210.72
Noon(12-18)	0.25	0.22	0.03	13.77
Night(18-24)	0.40	0.25	0.15	38.54
midNight(24-6)	0.32	0.19	0.14	42.05
firstHalfHour	0.51	0.51	0.00	0.33
SecondHalfHour	0.49	0.49	0.00	0.34

On average in Washington state,

Traffic congestion anomalies are likely to happen:

Low density counties

Ferry, Lincoln, Columbia, etc.

Spring Months

Summer is least likely to have anomalies.

Mondays & Sundays

Less on weekdays, higher on weekends.

Mornings (6 a.m. to 12 p.m.)

Very unlikely to happen after midnight (12 a.m. to 6 a.m.)

thank you.

Da Guo

da.guo@wsu.edu

linkedin.com/in/da-guo

dagoku.com

medium.com/@opophehu

github.com/opophehu