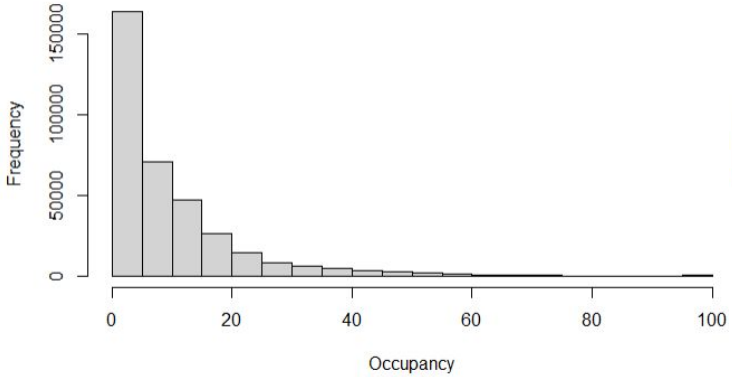
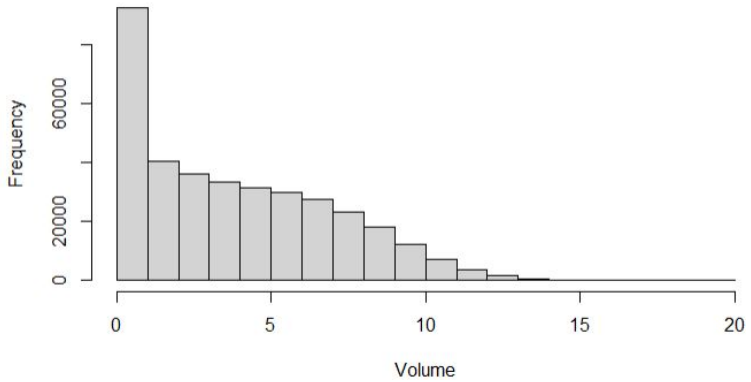
Data Performance Threshold

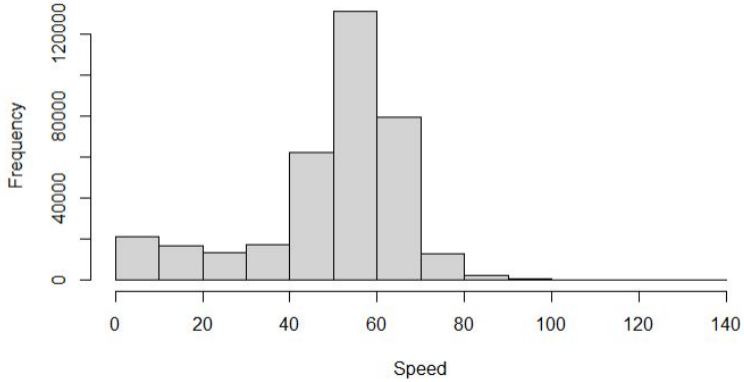
What is the suspicious data and how can we clean it?

Kyuri Kim 2024 Dec.

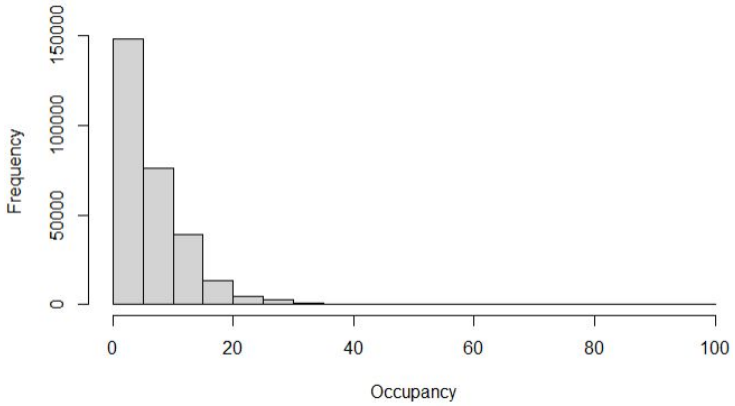
Data Distribution: Histogram (ODOT)

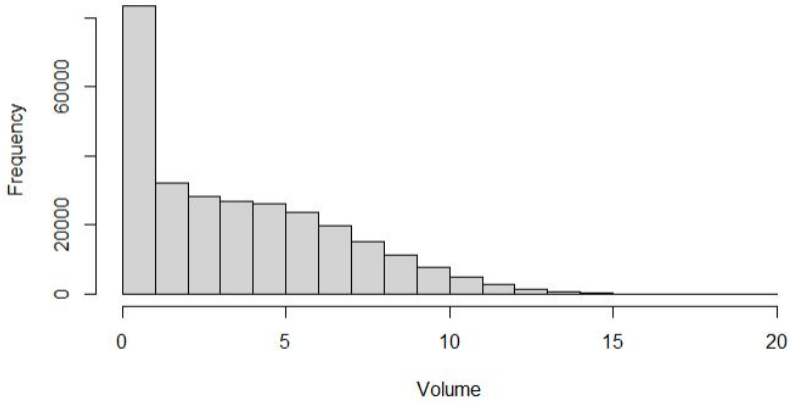


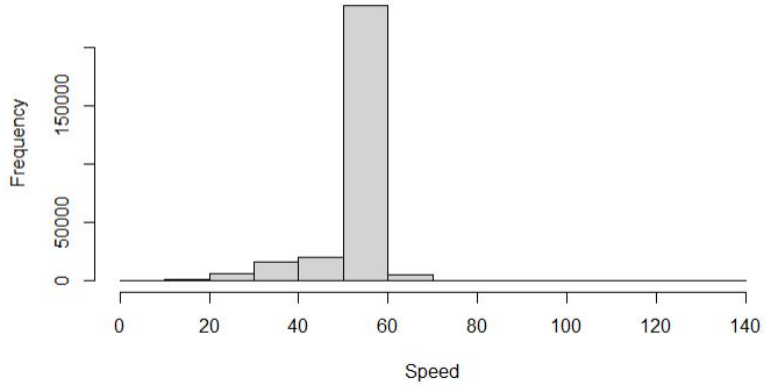




Data Distribution: Histogram (WSDOT)



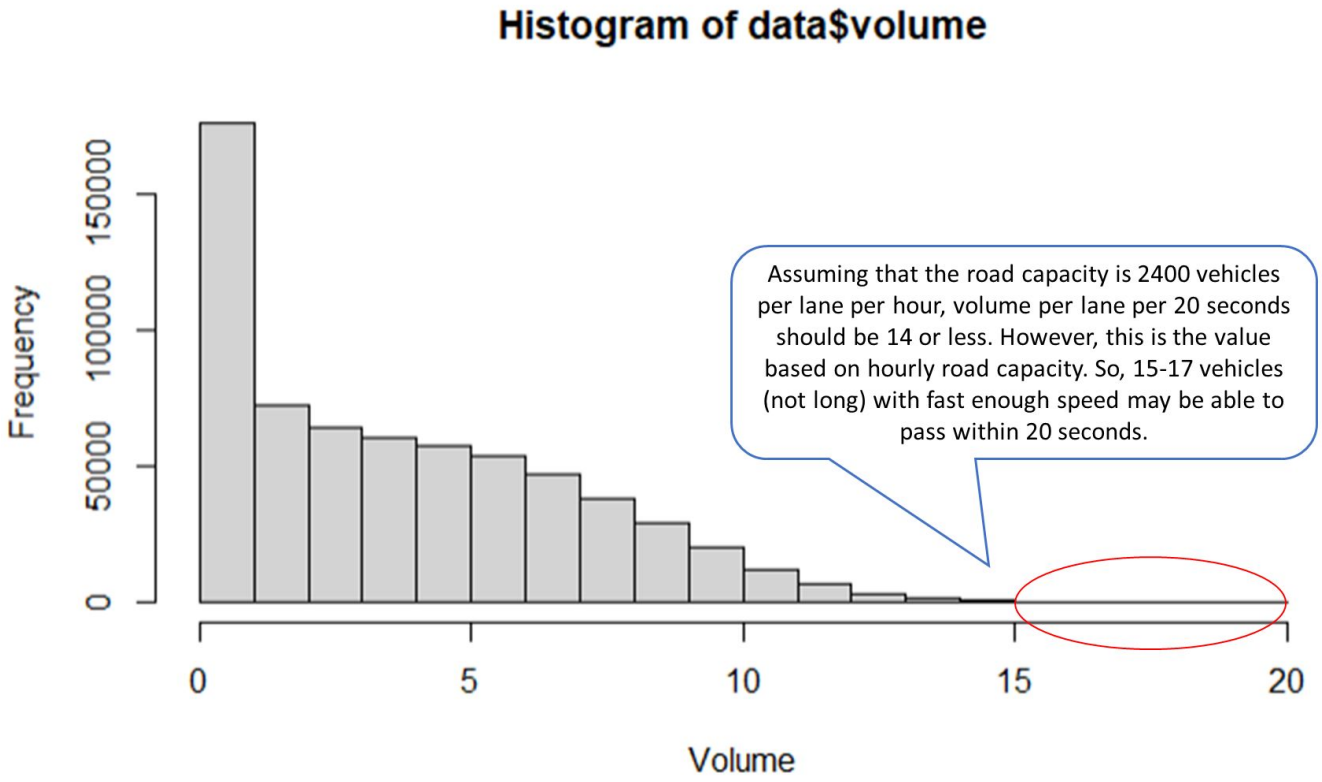




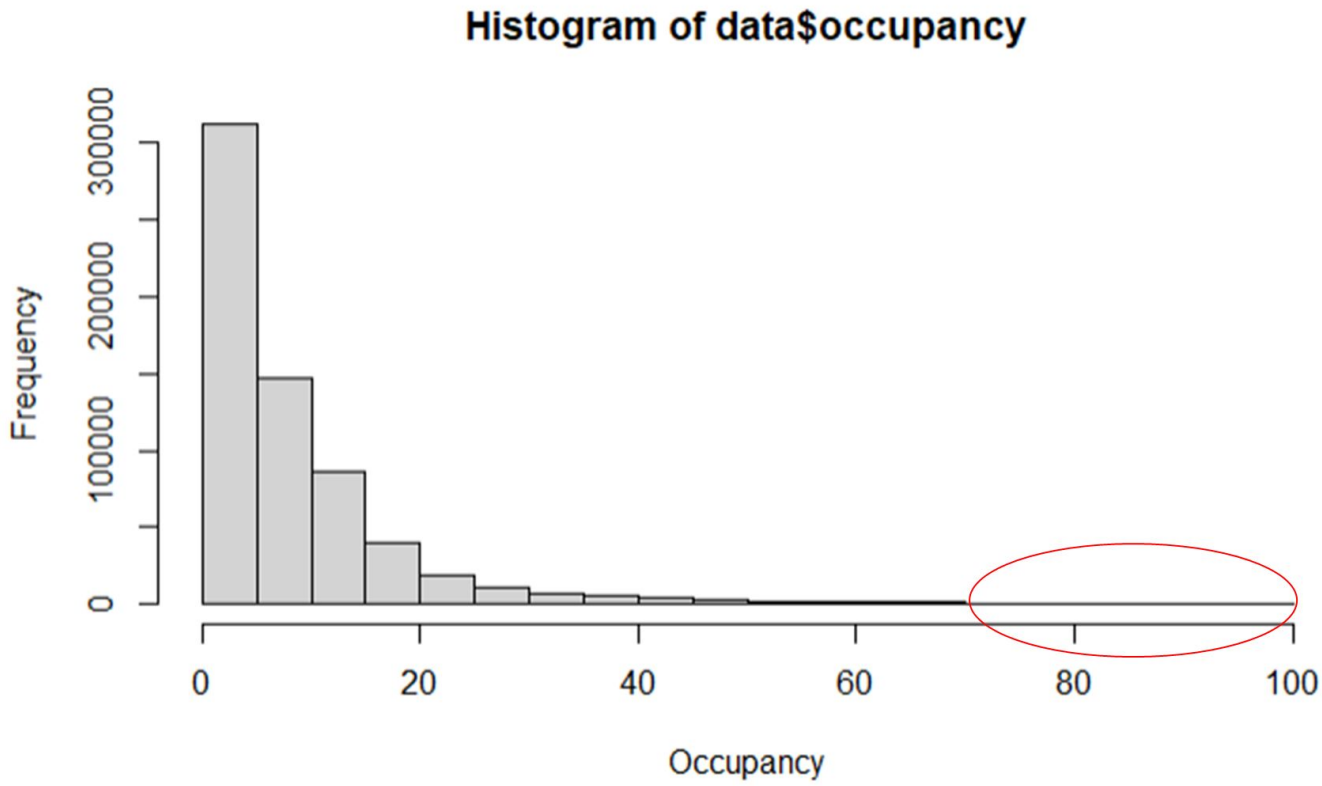
Thresholds for Outliers

- Sample Data
  - Based on ODOT & WSDOT data
  - On November 4-10
  - Detector type: Loop at one place (on three lanes), radar at other places
- Three Criteria
  - Volume: number of vehicle by lane within 20 seconds
  - Occupancy: time period the vehicles are detected within 20 seconds (%)
  - Speed: mile per hour

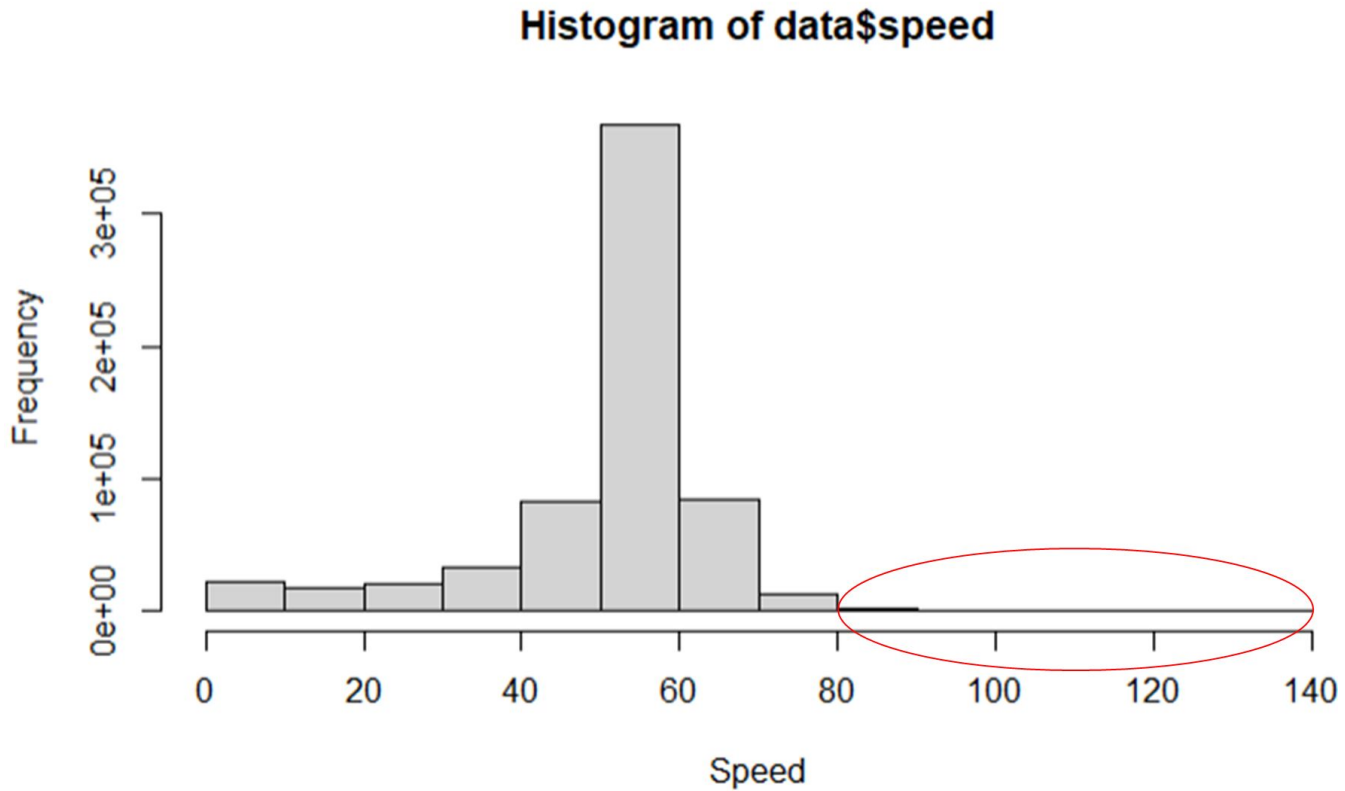
Data Distribution: Histogram (ALL) - Volume



Data Distribution: Histogram (ALL) - Occupancy



Data Distribution: Histogram (ALL) - Speed



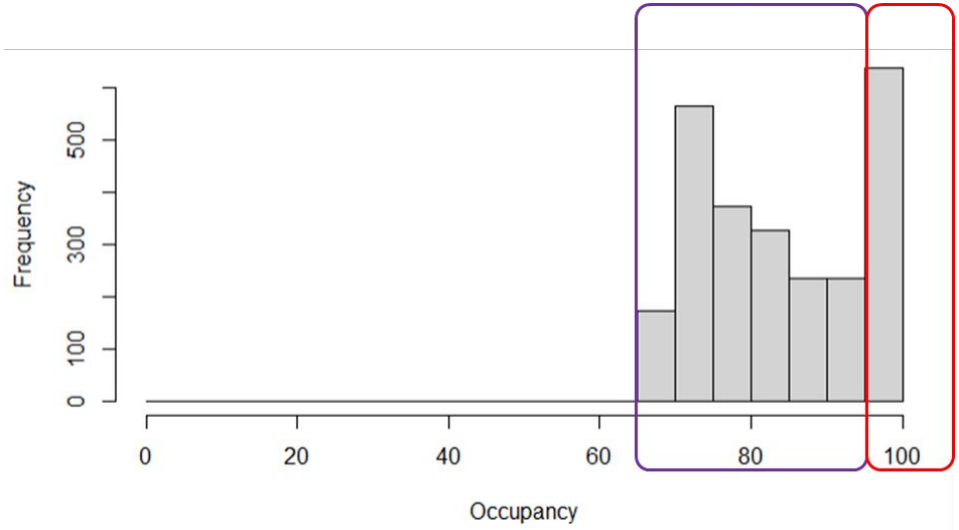
Approximately 6.7k, or 1% can be considered outliers, among the total 643k cases

BUT! Outliers do not necessarily mean measurement errors or machine malfunction, and non-outliers do not necessarily mean that they should be flagged as good.

More Details of Occupancy Outliers

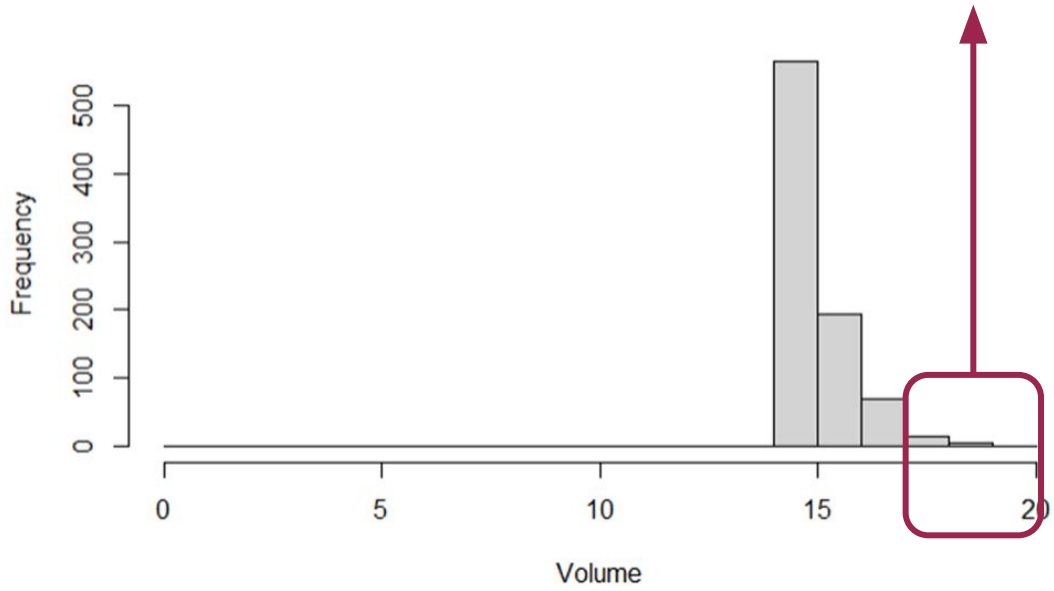
- between 70-100, most cases in 70-75 & 99-100,
- especially 99-100 should be looked in more details (in other words, two groups occ 70-98 and occ 99-100 should be considered as separate situations)

Each group may have different reasons



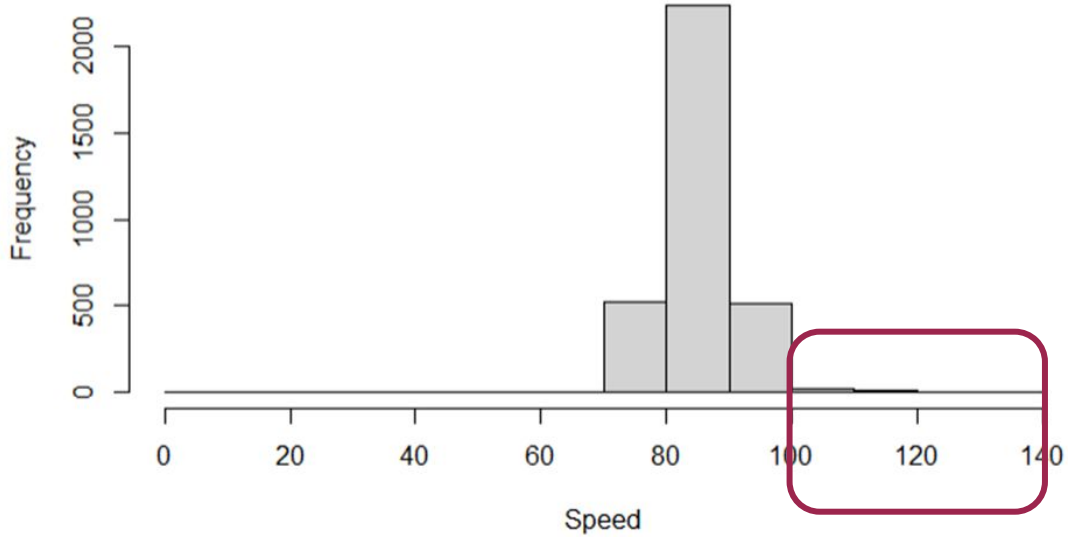
More Details of Volume Outliers

- In outliers, most cases in 15-17,
- Even in outliers, over 17 cases are extremely high-volume cases (*fast enough speed AND high enough occupancy are required or measurement error*)



More Details of Speed Outliers

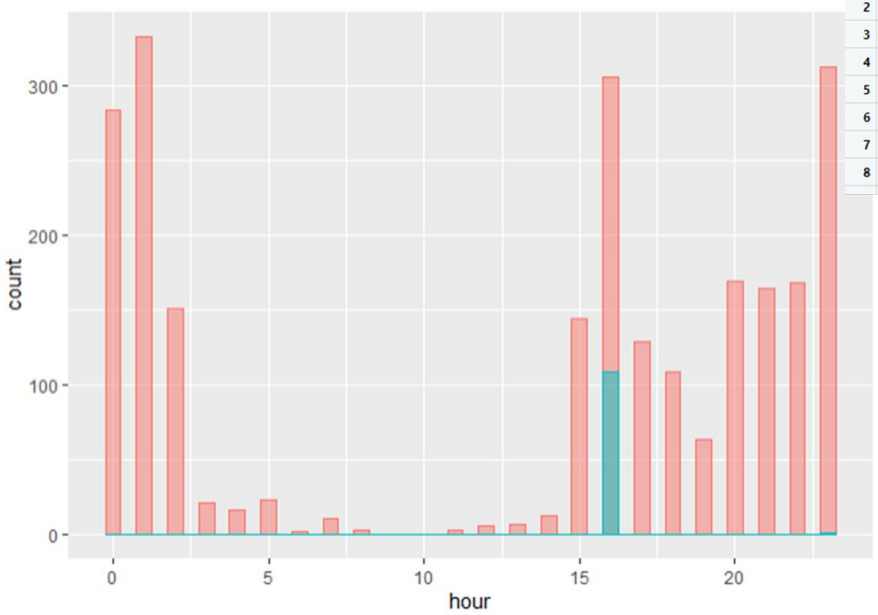
- Speed outliers are primarily between 80 mph and 100 mph
- Some cases have over 100 mph speed but very exceptional



When do outliers occur (outliers by hour)

Outlier (Occupancy >=70, 2544 cases)

- OCC outliers mostly happened in OR, at multiple locations on multiple days and hours
- But Some cases have high occupancy & low speed (congestion), but relatively high volume → **Suspicious**
- Cases in WA at one location multi-lane (I-5, South, radar) on one day a week (Nov. 5th) 4 pm with low speed and low volume → **Good**



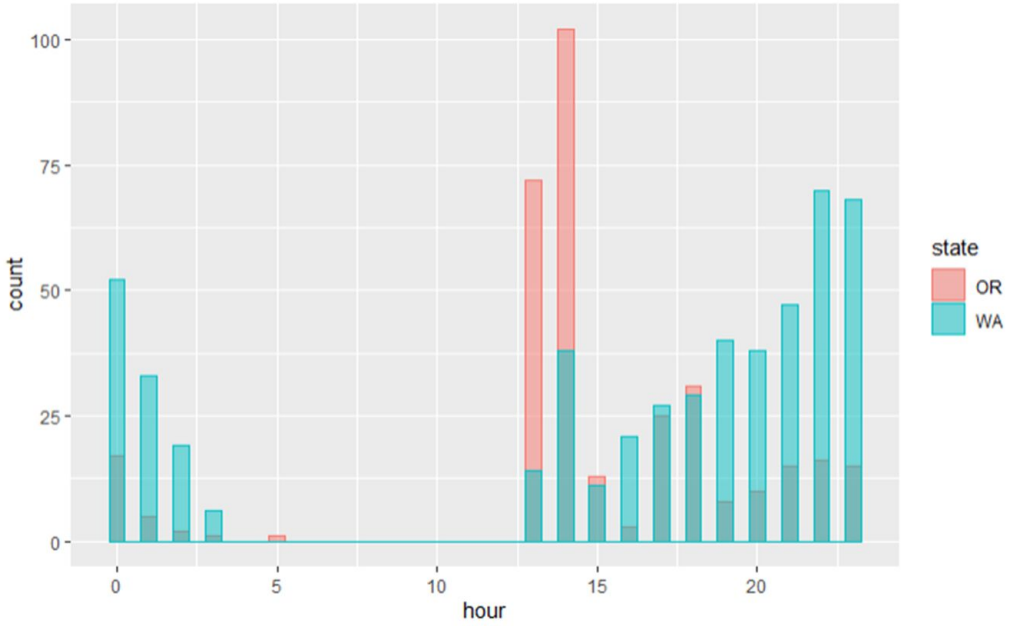
	starttime	detectorid	state	volume	occupancy	speed	date	hour	minute	detectortype	ramp	status
1	2024-11-05 23:26:40	102339	OR	10	74	10	5	23	26	radar	no_ramp	suspicious
2	2024-11-06 23:33:40	102339	OR	11	73	10	6	23	33	radar	no_ramp	suspicious
3	2024-11-08 23:25:00	102339	OR	10	81	10	8	23	25	radar	no_ramp	suspicious
4	2024-11-06 00:36:20	102340	OR	12	74	9	6	0	36	radar	no_ramp	suspicious
5	2024-11-06 00:46:40	102340	OR	10	70	8	6	0	46	radar	no_ramp	suspicious
6	2024-11-04 15:46:20	100894	OR	3	82	10	4	15	46	loop	no_ramp	Good
7	2024-11-04 16:04:20	100894	OR	4	74	9	4	16	4	loop	no_ramp	Good
8	2024-11-04 16:09:20	100894	OR	3	81	7	4	16	9	loop	no_ramp	Good

These cases are possible only when most vehicles are shorter than 15 feet (not bus or truck)

Outlier (Volume >=15, 898 cases)

- Considering the number of sample cases (643k), volume outliers are rare (0.1%)
- In OR, they are highest between 1pm and 3pm at Sandy (2R079) to WB I-84 (loop); 82nd/Halsey (2R080) to WB I-84 (loop) → may be early rush hour
- In WA, volume outliers tend to increase in the late afternoon and decrease through the early morning hours especially between 8 pm and 1 am midnight (all days, multiple places, 270 cases among 513 WA volume outliers): occupancy (16% - 43%) speed (28mph - 71 mph)

These occupancy and speed ranges are acceptable, however, the time zone is questionable (suspicious)



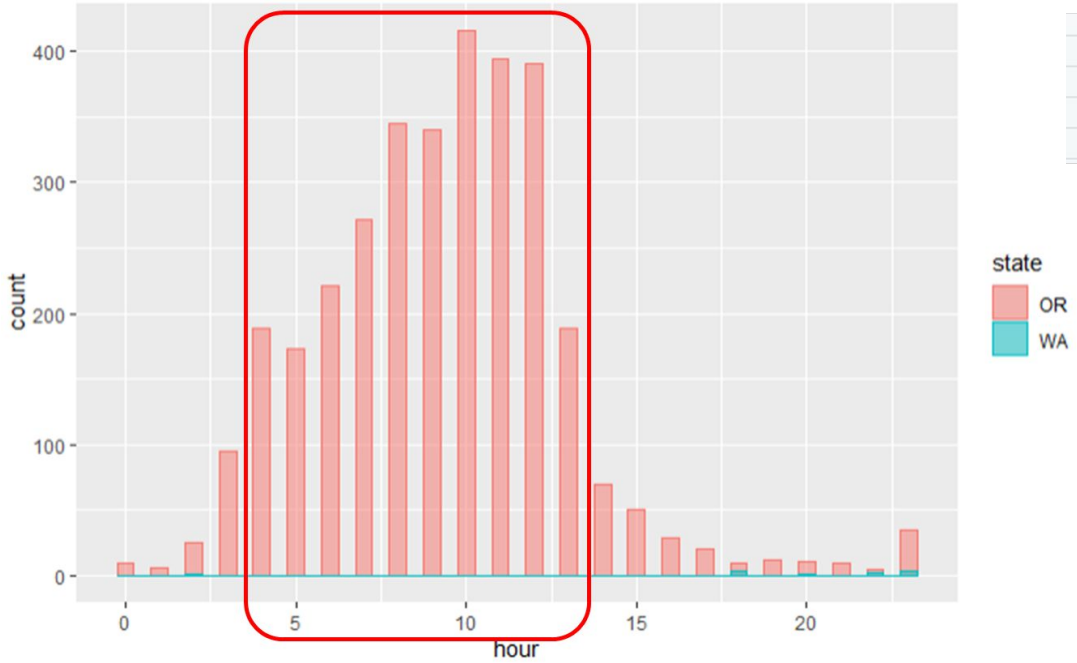
NE Hazel Dell Ave
NE Hazel Dell Ave
NE 2nd Ave
NE 2nd Ave
Burnt Bridge Creek
Burnt Bridge Creek Greenway
Cold Canyon
BLUROCK LANDING
WEST HAZEL DELL
HAZEL DELL
Les Schwab Tire Center
NORTHWEST VANCOUVER
Bill's Tech-Trans
Fourth Plain Rd
LINCOLN
FRUIT VALLEY
W 39th St
W 33rd St
ROSE VILLAGE
CARTER PARK
St Johns Blvd
W Fourth Plain Blvd
HOUGH
ARNADA
Clark College
Marshall Park
E McLoughlin Blvd
Comfort Inn & Suites Vancouver Downtown City...
OREGON
Vancouver
Fort Vancouver National Historic Site
Hayden Island
Smith and Bybee Wetlands Natural Area
Columbia River
HAYDEN ISLAND NEIGHBORHOOD NETWORK
WASHINGTON
OREGON
Smith Lake
N Columbia Blvd
N Lombard
NORTH PORTLAND
Heron Lakes Golf Club

Sandy (2R079) to WB I-84 (**loop**)

82nd/Halsey (2R080) to WB I-84 (**loop**)

Outlier (Volume >=80, 3323 cases)

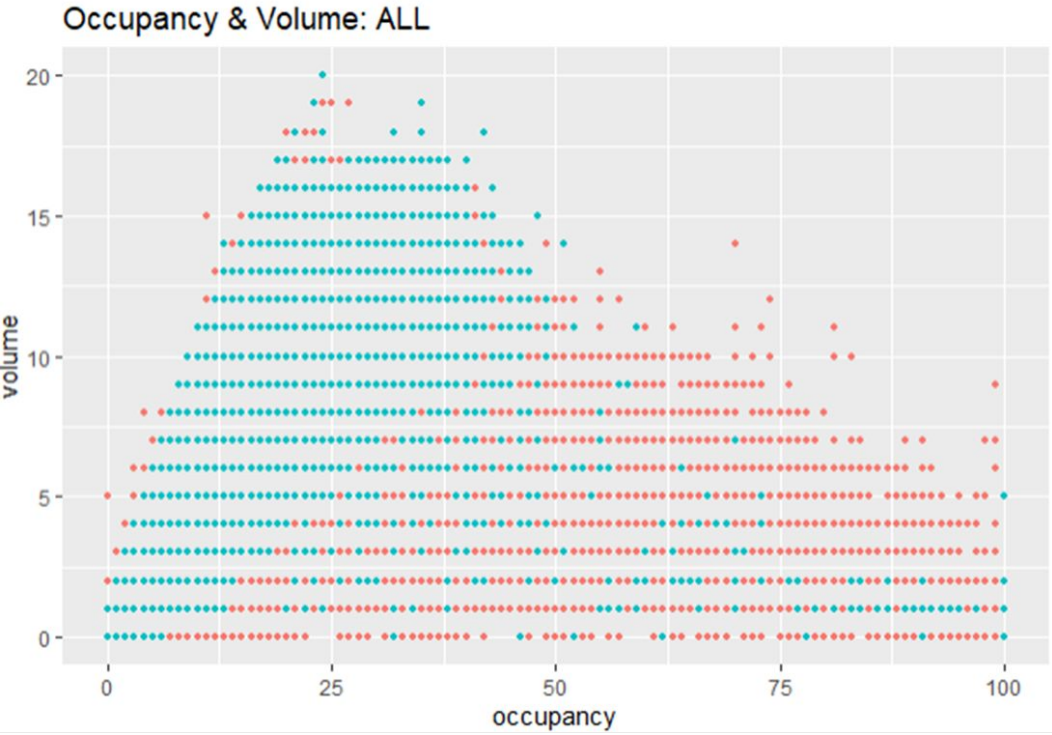
- Speed outliers are mostly measured in OR during a week between 4am – 1pm
- More than 60% speed outliers are measured in OR (Sandy (2R079) to WB I-84, loop)
  - 82nd/Halsey (2R080) to WB I-84 (loop)
  - Russell (2DS150) @ NB&SB I-5 MP302.8 (radar)
  - Russell (2DS150) @ NB I-5 to I-405 MP302.8 (radar)

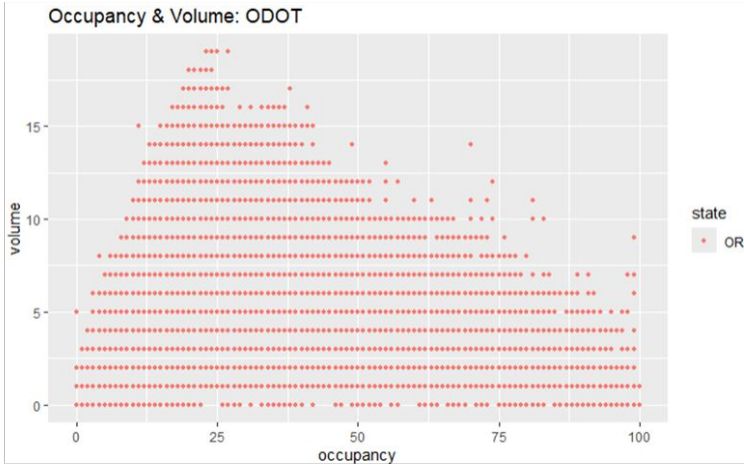


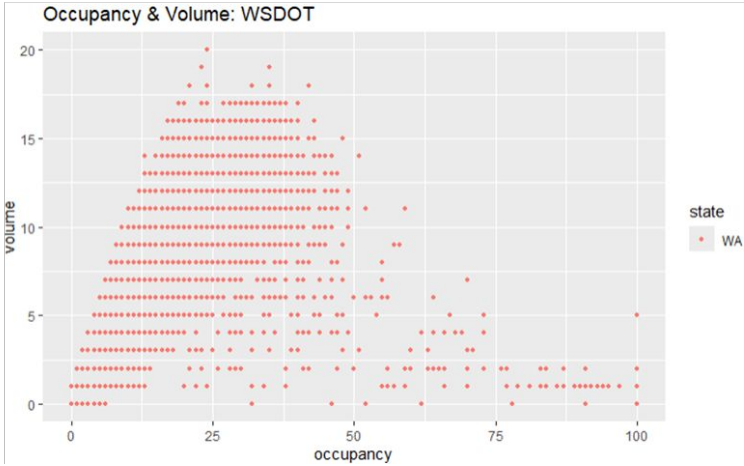
	starttime	detectorid	state	volume	occupancy	speed	date	hour
1	2024-11-06 18:33:20	100894	OR	5	56	81	6	18
2	2024-11-09 04:18:00	100894	OR	8	21	80	9	4
3	2024-11-05 13:50:00	100894	OR	10	20	81	5	13
4	2024-11-05 12:50:40	100894	OR	6	20	83	5	12

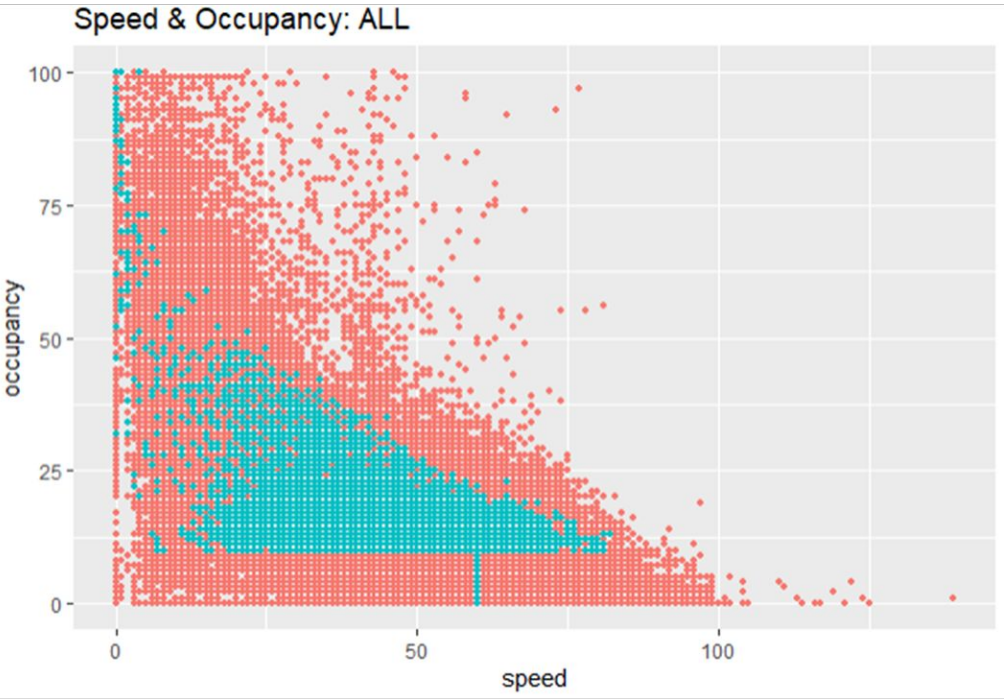
Considering the high speed and occupancy, the volume is not high enough (occ >50 means that vehicles should be detected for more than 10 seconds in 20 seconds). BUT! This is possible when long trucks like road train trucks pass.

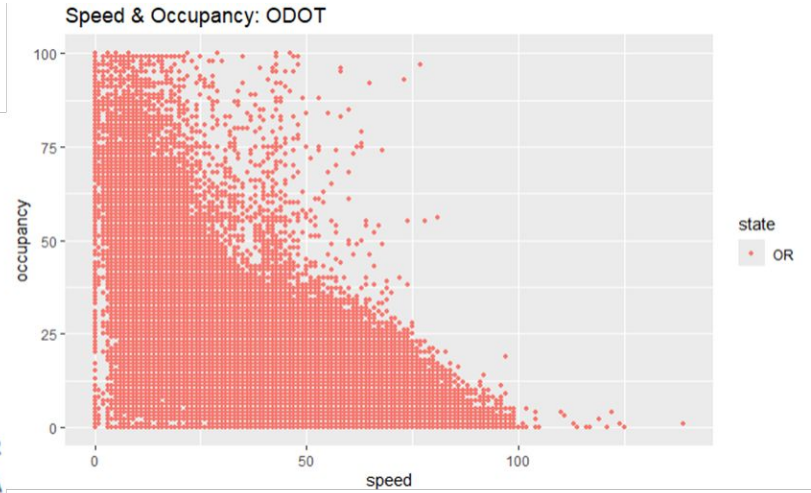
Beyond the outliers

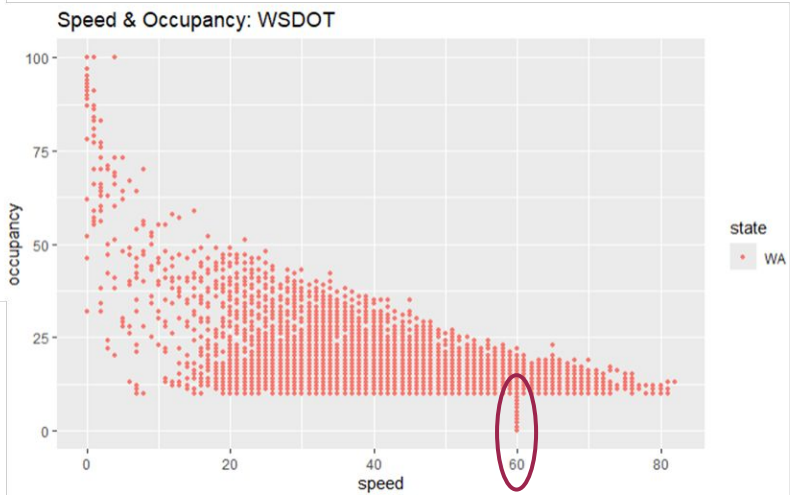

Occupancy & Volume: ALL
volume
occupancy
0
5
10
15
20
0
25
50
75
100
state
OR
WA

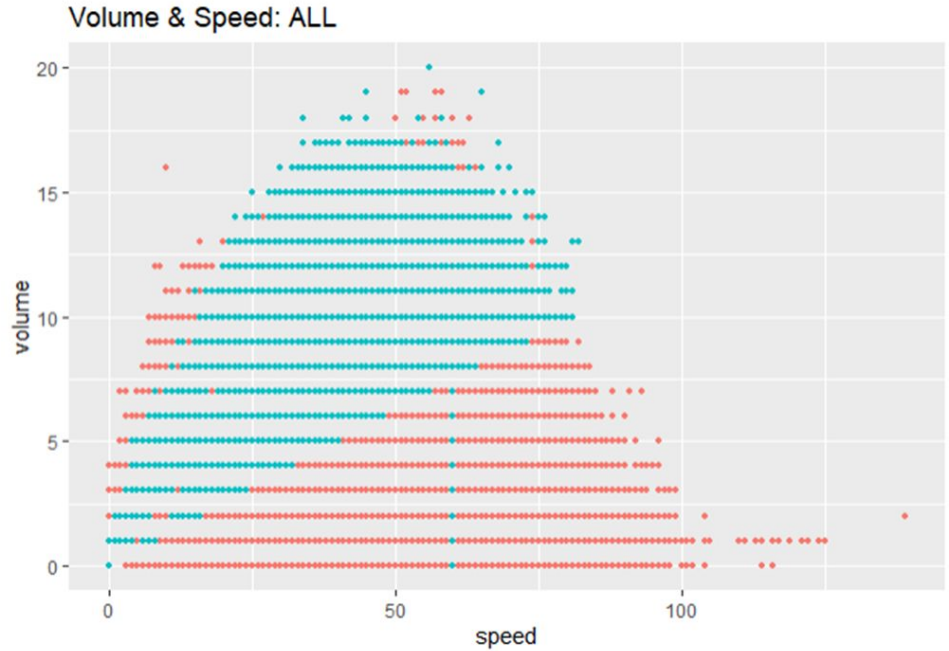

Occupancy & Volume: ODOT
volume
occupancy
0
5
10
15
0
25
50
75
100
state
OR

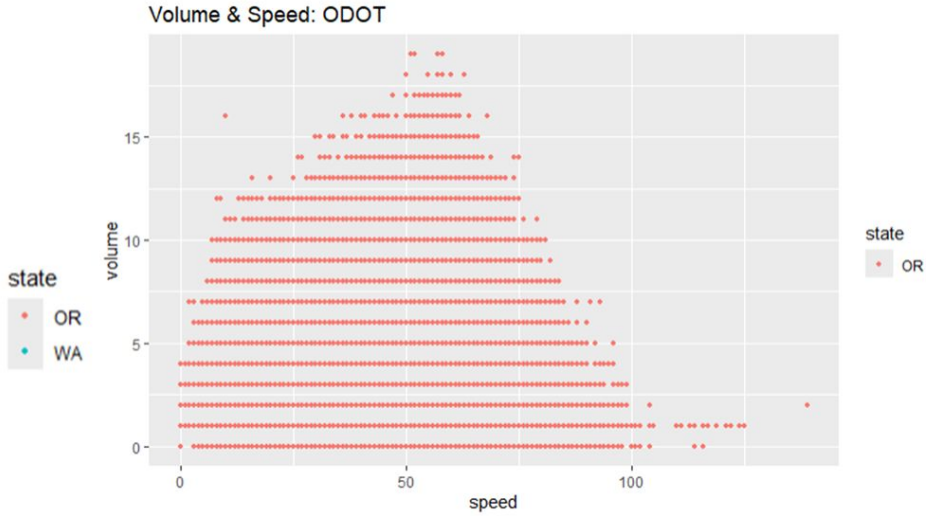

Occupancy & Volume: WSDOT
volume
occupancy
0
5
10
15
20
0
25
50
75
100
state
WA


Speed & Occupancy: ALL
occupancy
speed
0
25
50
75
100
0
50
100
state
OR
WA

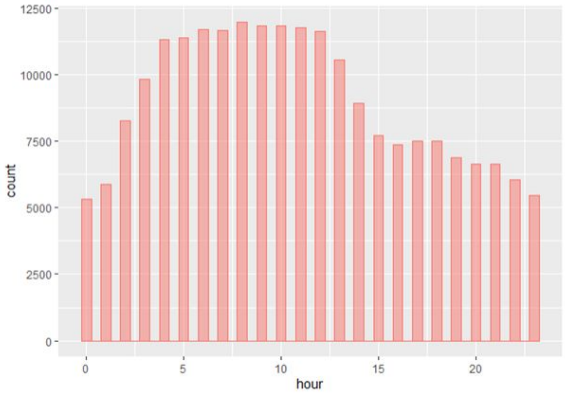

Speed & Occupancy: ODOT
occupancy
speed
0
25
50
75
100
0
50
100
state
OR

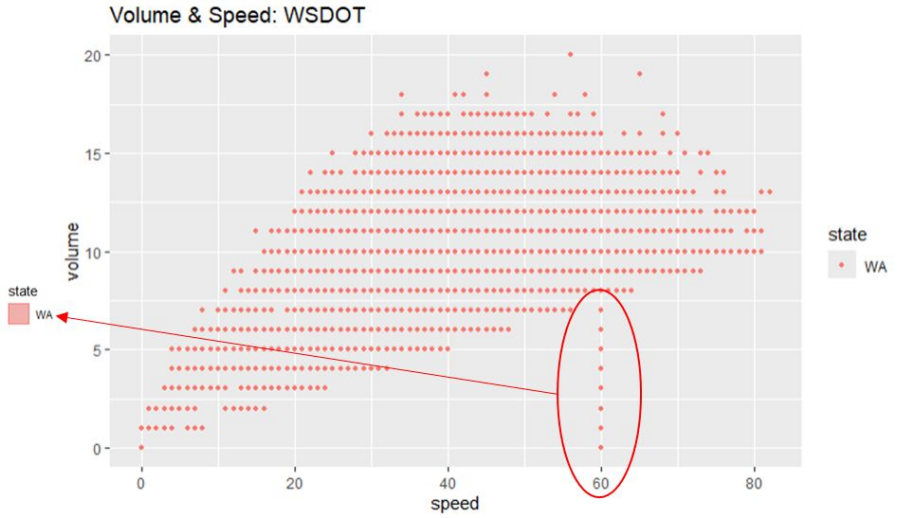

Speed & Occupancy: WSDOT
occupancy
speed
0
25
50
75
100
0
20
40
60
80
state
WA





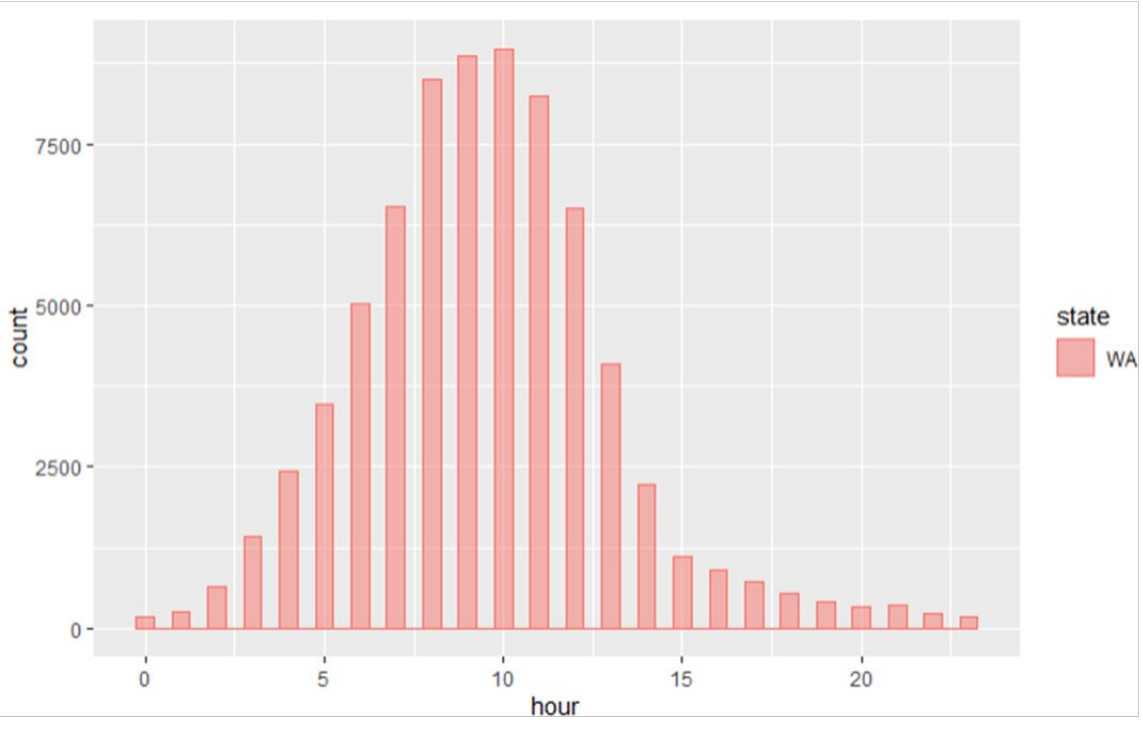
In WA, among 285k cases, the speed of the 220k cases is 60 mph. Many cases of them occupancy and volume are zero or 1.





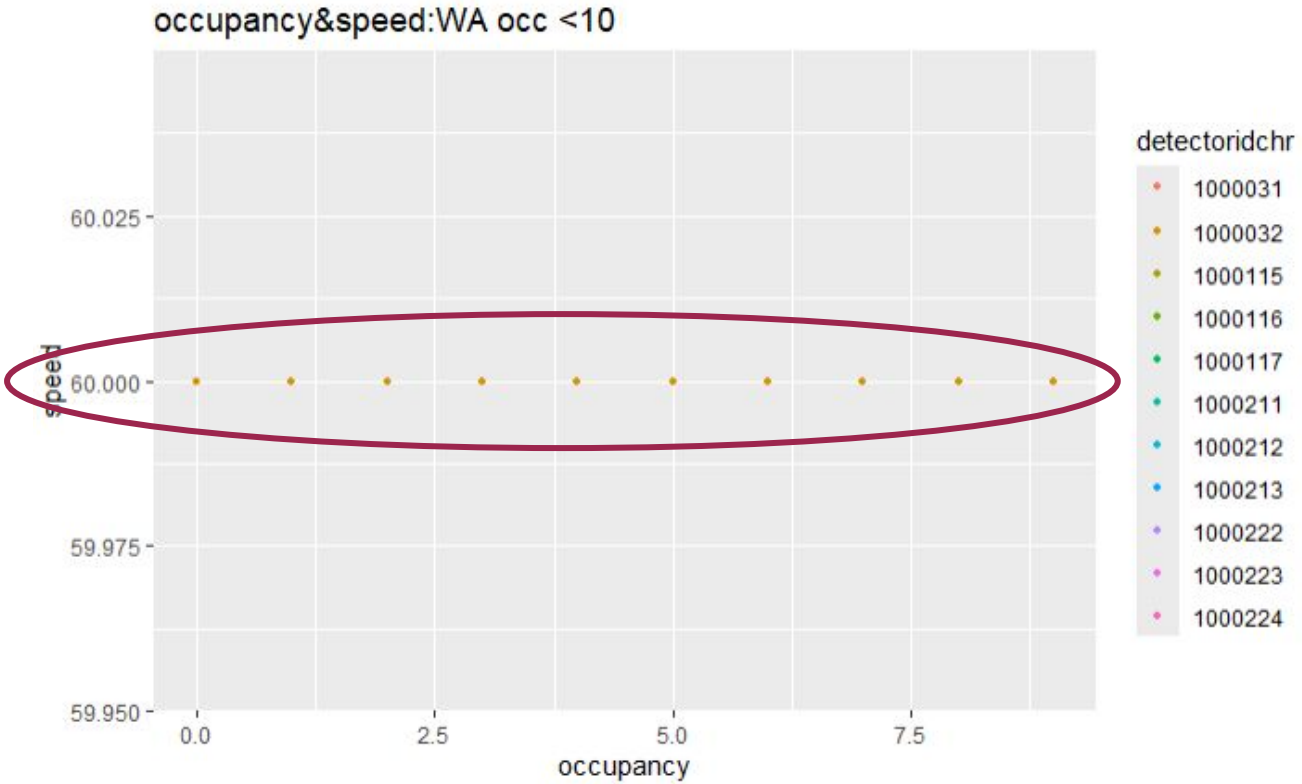
In WA, among 285k cases, the speed of the 220k cases is 60 mph.
43k cases of them have zero occupancy and zero volume.

speed==60 & occupancy <=1 & volume <=1 & state=="WA" by hour



	stationid	highwayname	hour	n
1	10446	WA I-205	10	2087
2	10446	WA I-205	11	1959
3	10446	WA I-205	9	1923
4	10711	WA I-5	10	1922
5	10712	WA I-5	10	1868
6	10712	WA I-5	11	1844
7	10711	WA I-5	9	1835
8	10712	WA I-5	9	1784
9	10711	WA I-5	8	1682
10	10446	WA I-205	8	1656
11	10446	WA I-205	12	1652
12	10712	WA I-5	8	1569
13	10711	WA I-5	11	1492
14	10712	WA I-5	12	1356
15	10711	WA I-5	7	1260
16	10446	WA I-205	7	1014
17	10712	WA I-5	13	1000
18	10446	WA I-205	13	998
19	10429	WA I-205	9	952
20	10712	WA I-5	7	897
21	10711	WA I-5	6	865
22	10429	WA I-205	10	853
23	10429	WA I-205	8	822
24	10711	WA I-5	12	770
25	10446	WA I-205	6	717

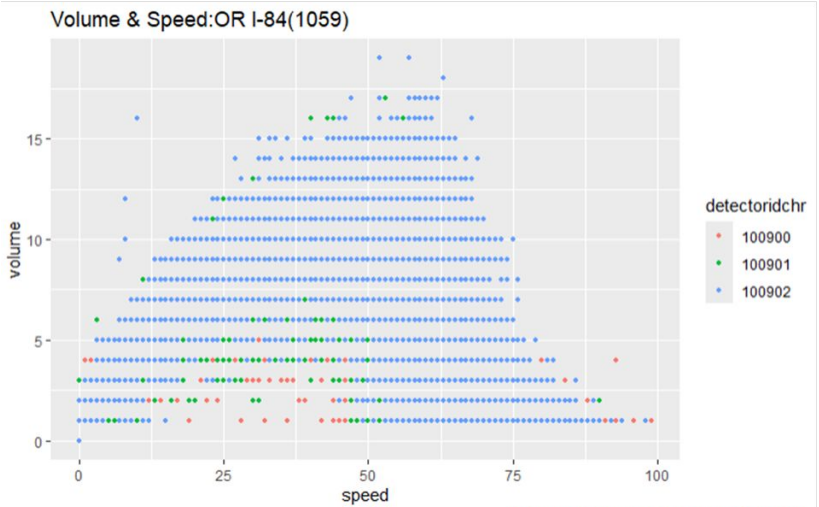
WSDOT’s default speed for occupancy under 10

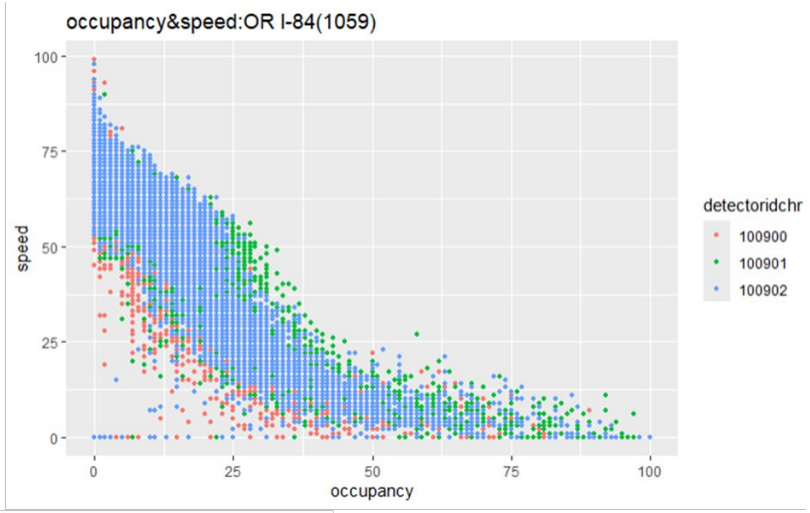


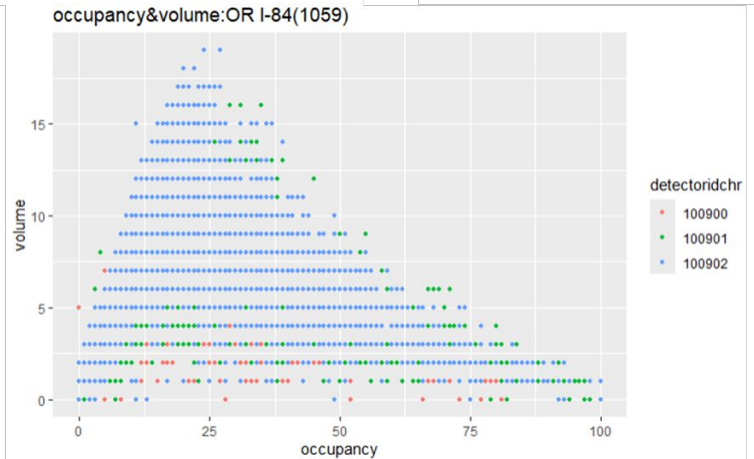
WSDOT
0=< Occupancy <10
Speed = 60 mph

Occupancy >=10
7=< Speed =<81

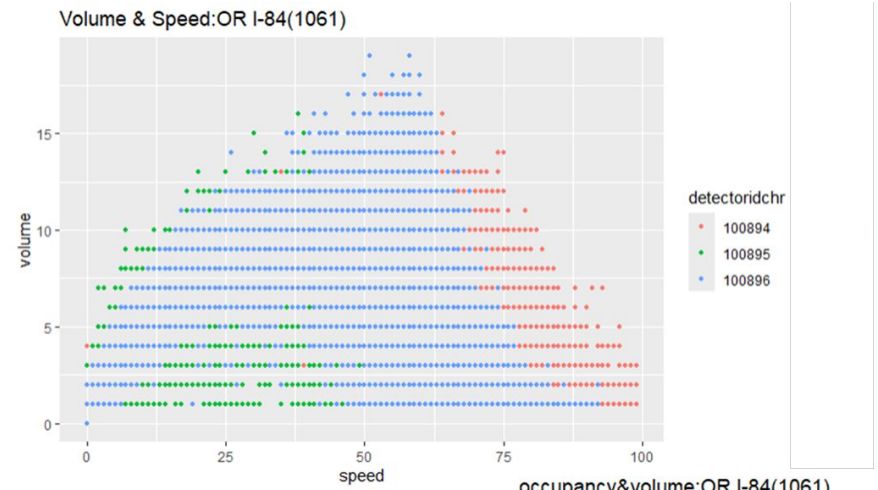
Data Distribution: 2D scatter plot

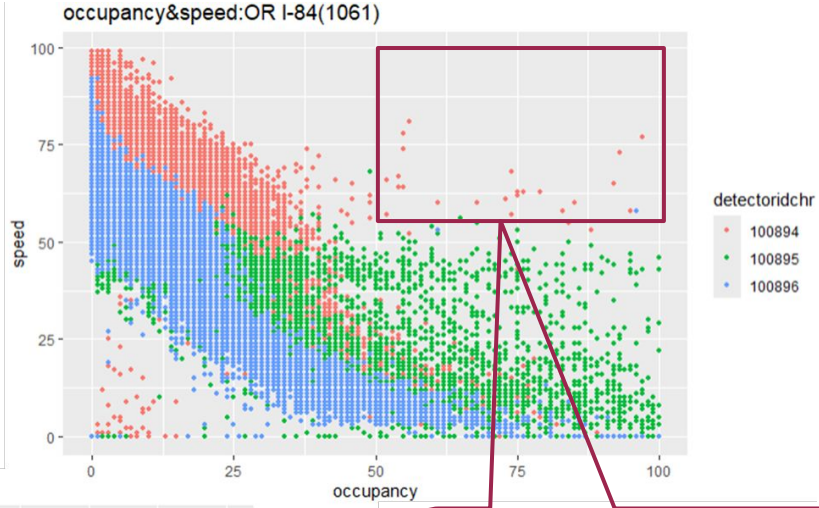


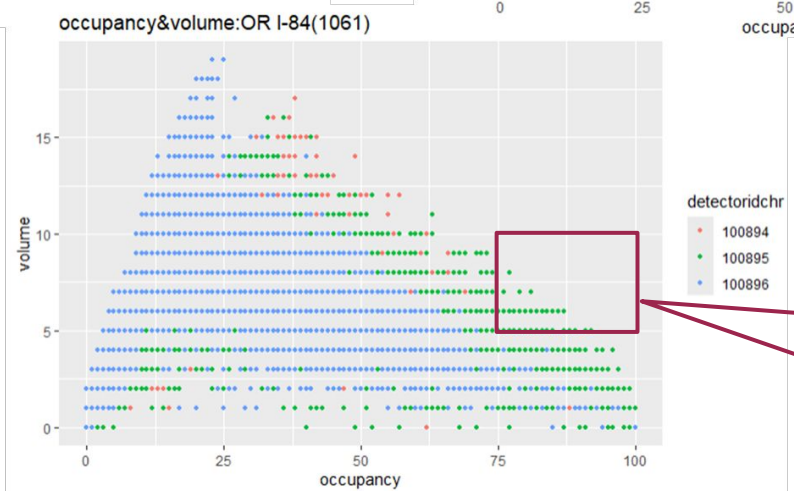




Data Distribution: 2D scatter plot



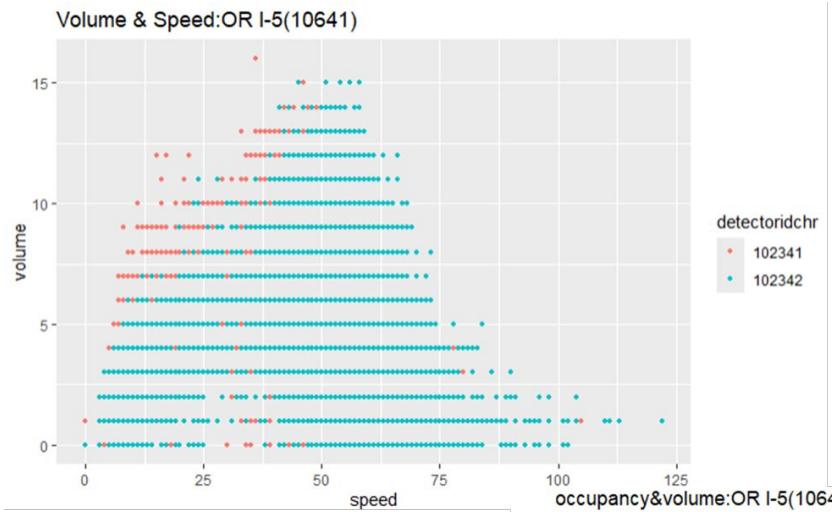


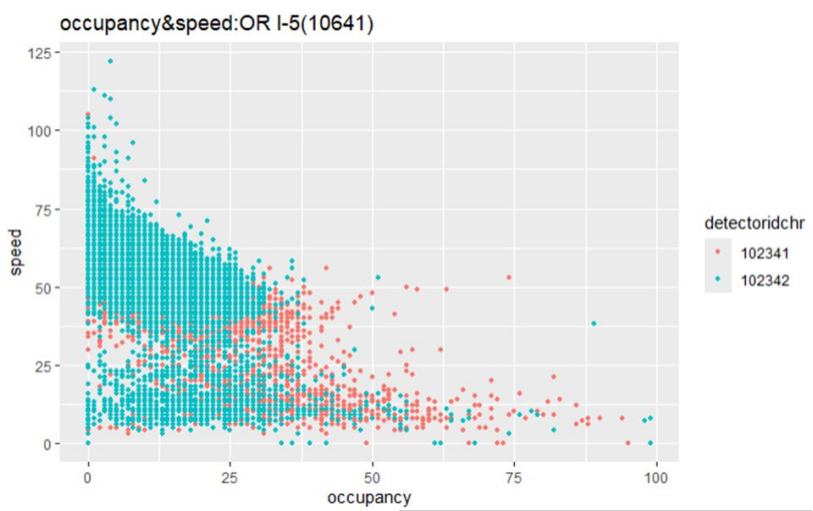


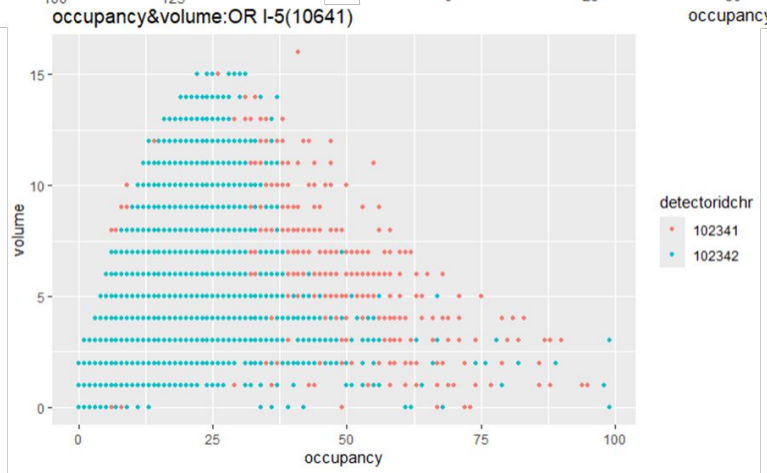
Occupancy more than 50% (congestion) and speed more than 60 mph (suspicious)

Occupancy more than 75% (extremely congestion or stop) but volume more than 5 (suspicious)

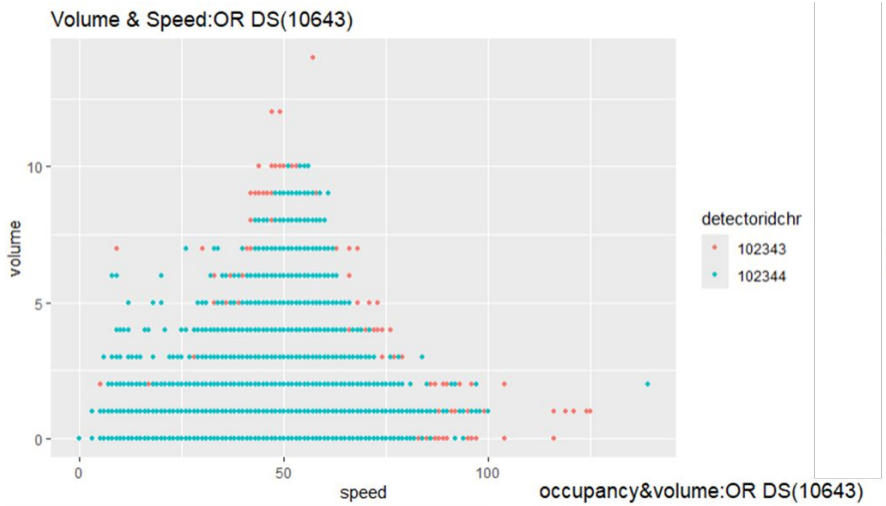
Data Distribution: 2D scatter plot

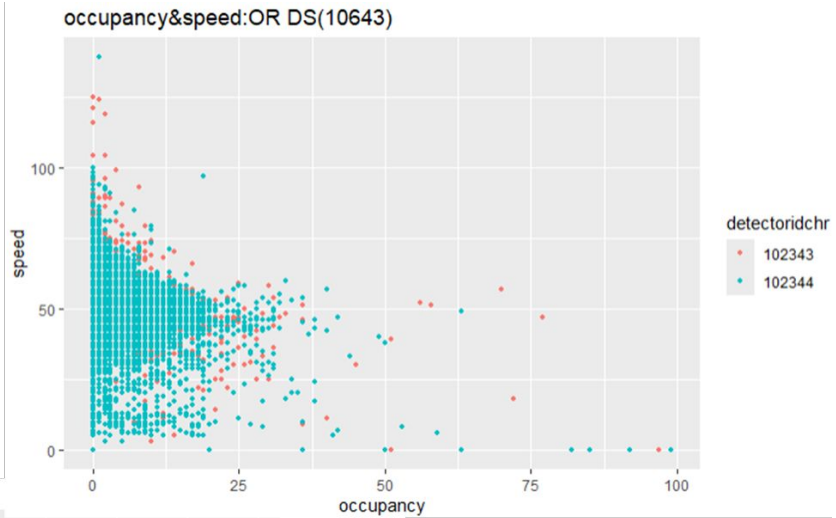


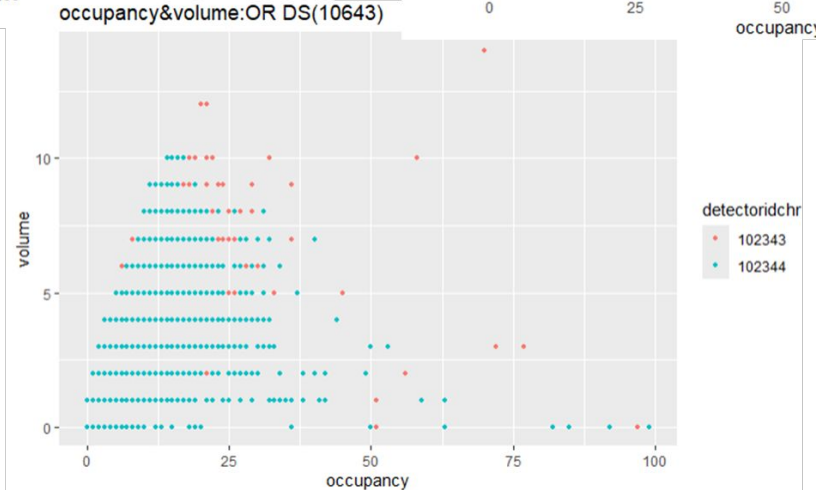




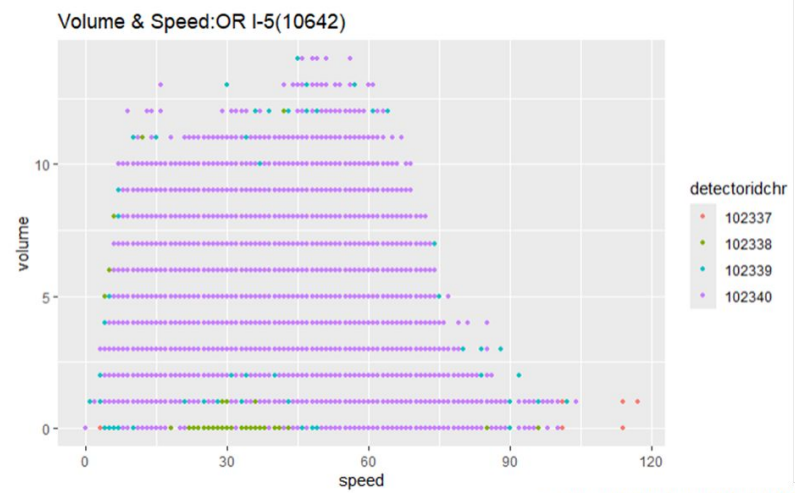
Data Distribution: 2D scatter plot

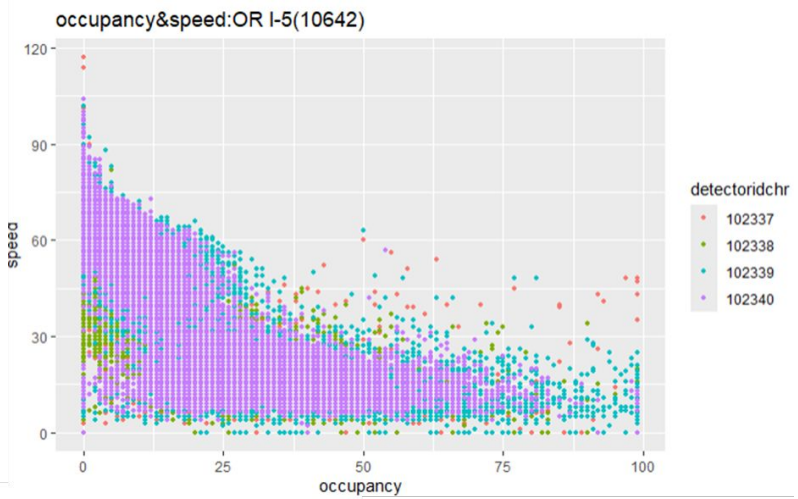


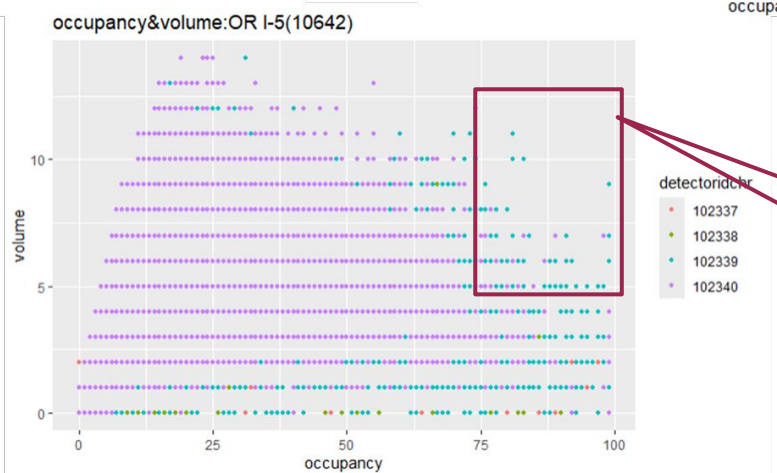




Data Distribution: 2D scatter plot

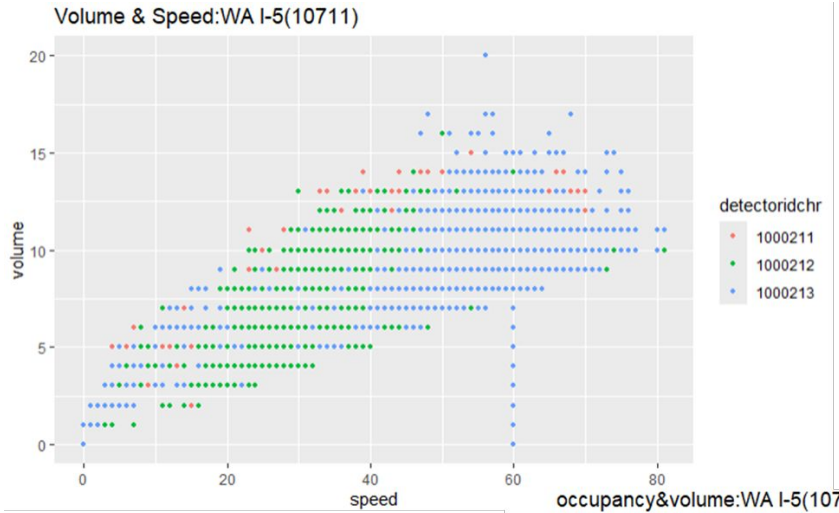


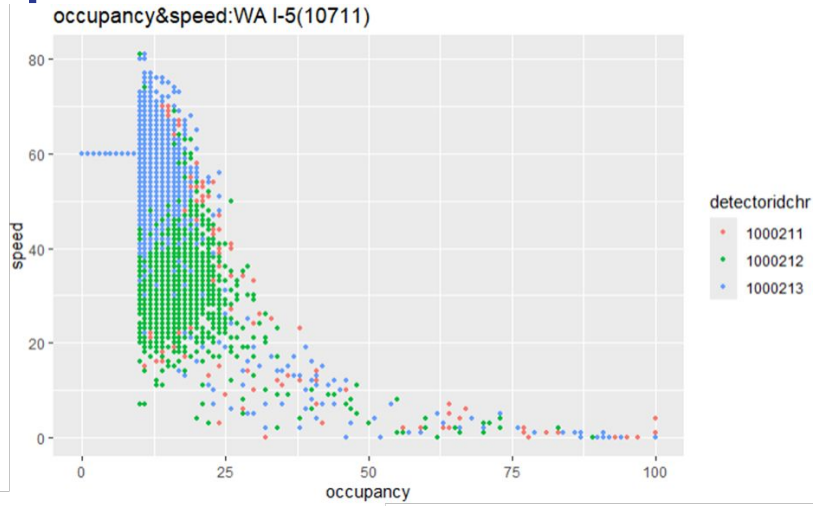


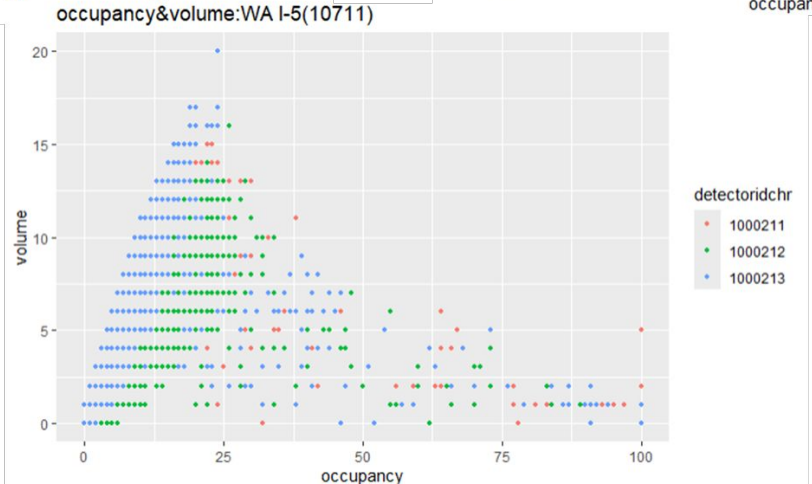


Occupancy more than 75% (extremely congestion or stop) but volume more than 5 (suspicious)

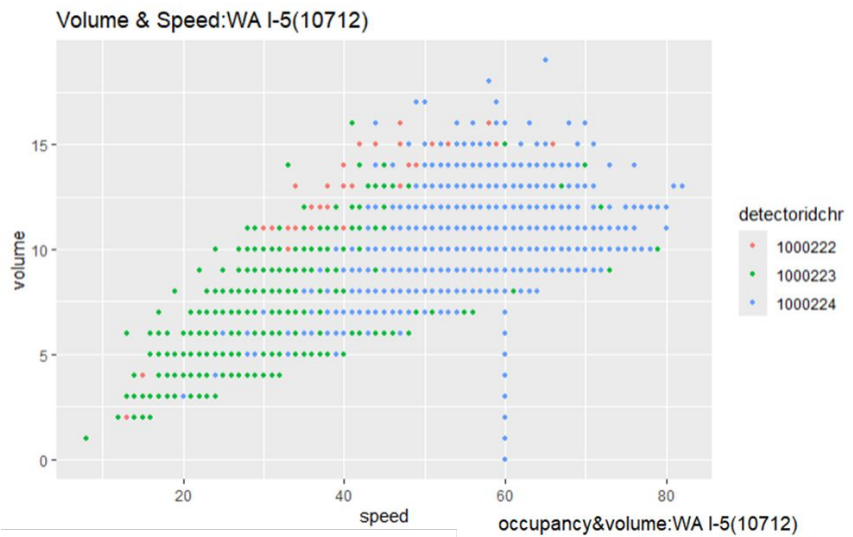
Data Distribution: 2D scatter plot

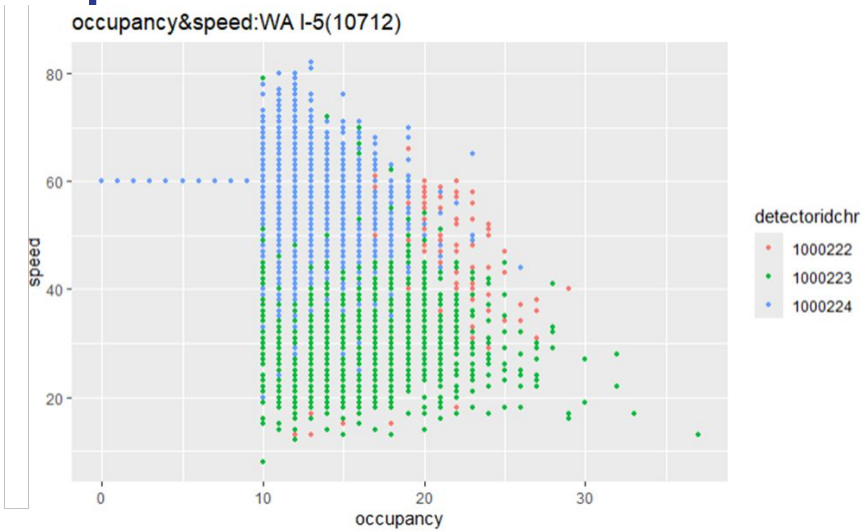


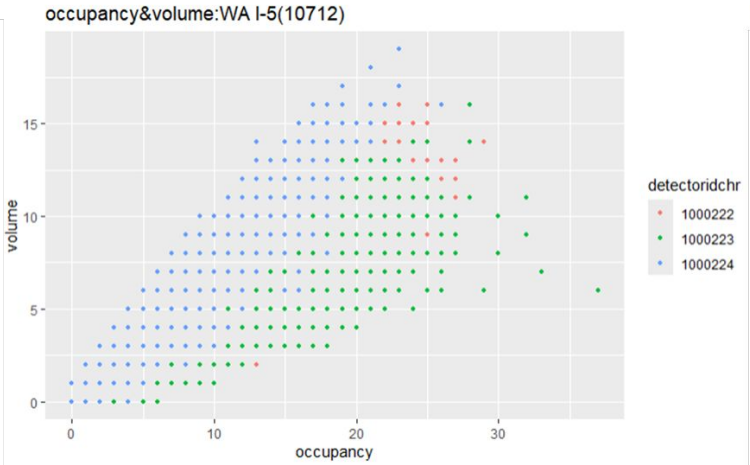




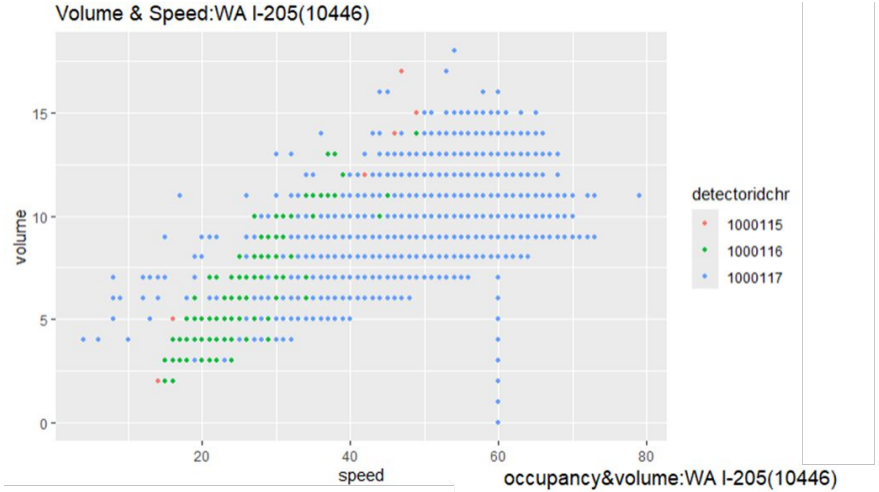
Data Distribution: 2D scatter plot

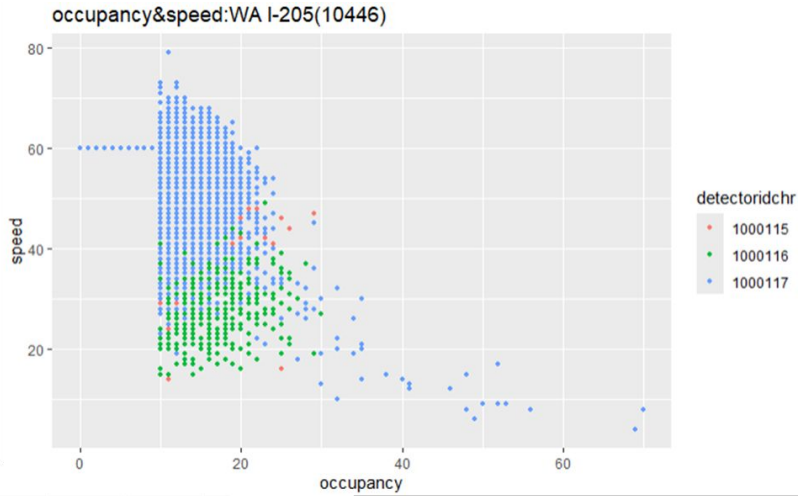


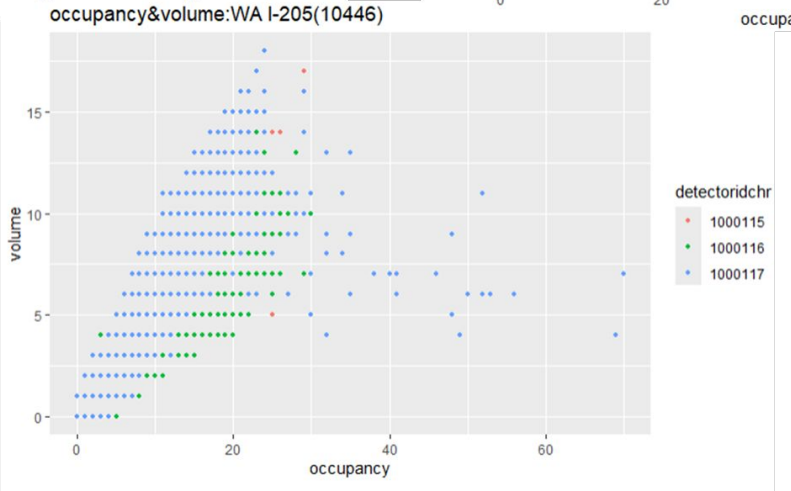




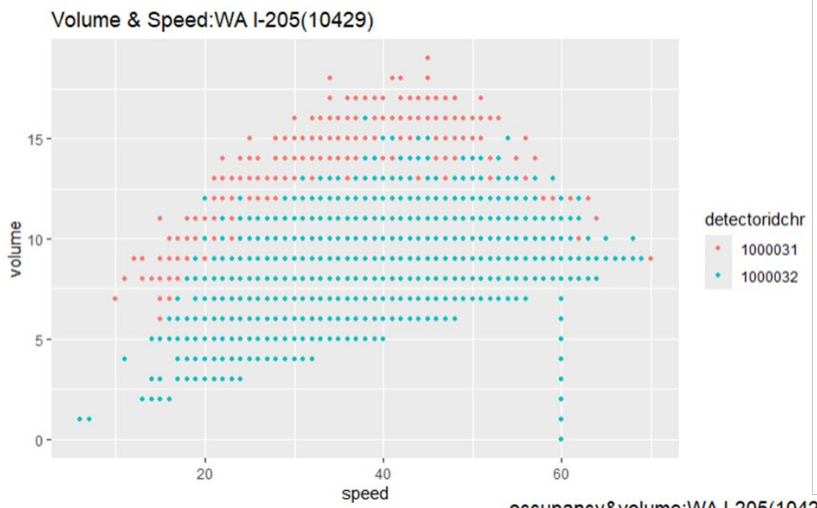
Data Distribution: 2D scatter plot

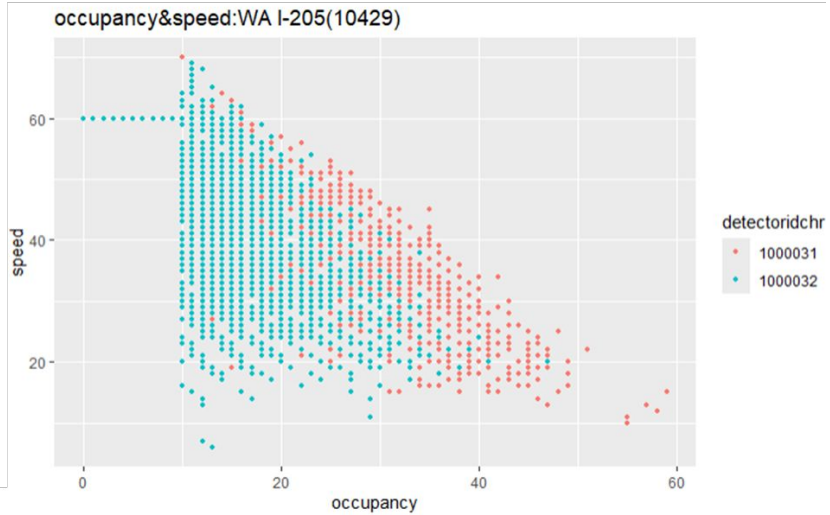


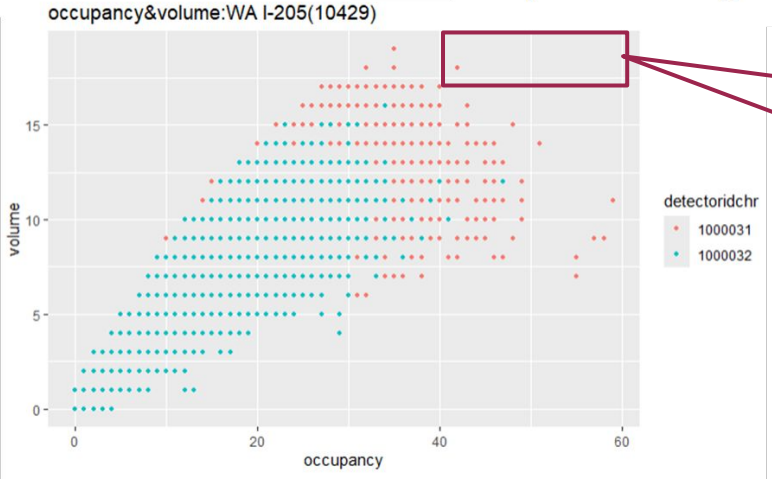




Data Distribution: 2D scatter plot







Occupancy more than 40% and volume more than 17 (suspicious)