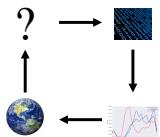


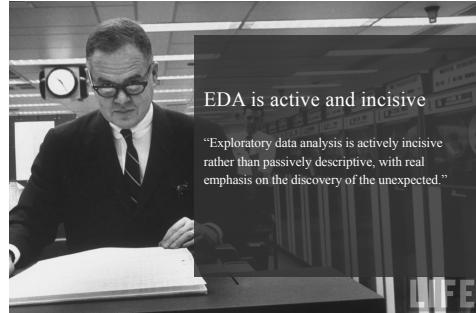
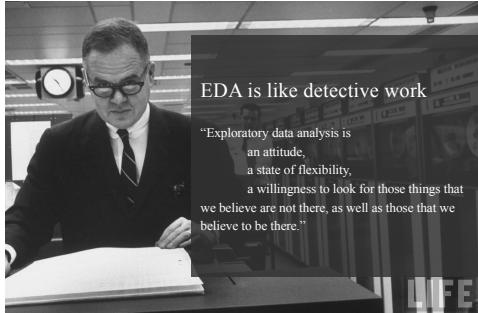
Data Science 100

Lecture 5: Exploratory Data Analysis

Slides by:
Deb Nolan
deborah_nolan@berkeley.edu



Data Analysis & Statistics, Tukey 1965



Philosophy of EDA

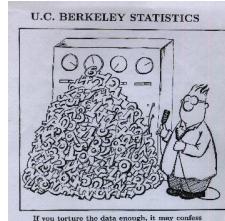
- Confirm understanding of the data
- Keep an open mind and be willing to find something surprising
- Iterate
 - Uncover new aspects of our data
 - Re-examine our understanding of the data
 - Continue exploration

Where does EDA help?

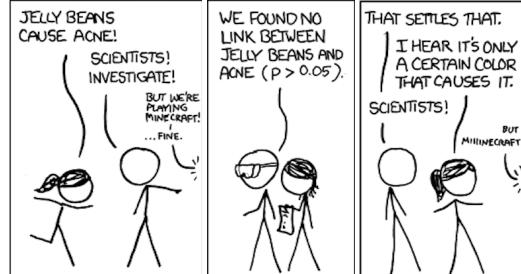
- Clean data
- Transform variables and derive new variables – put data in format suitable for analysis
 - Better understand data/situation (no formal analysis pursued)
 - Inform formal analysis – uncover important features that impact the analysis
- Examine formal analysis – explore output from an analysis, e.g., predictions, parameter estimates, model fit

Caution about EDA

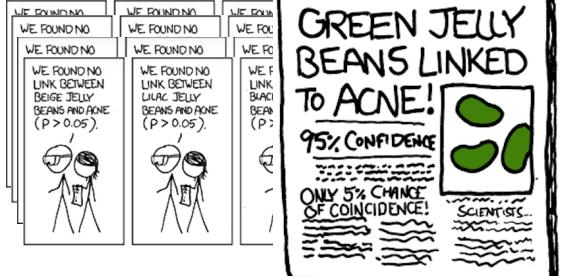
- Considered data snooping
- With enough data, if you look hard enough you will find something "interesting"
- Related to p-hacking



Caution about EDA – ala xkcd



Caution about EDA -



On The Other Hand

EDA can provide valuable insights about a model and its assumptions

What to do?

- Additional Data
 - Pilot study
 - Perform a second experiment
 - Collect more data
 - Short of that – divide data in two at random and explore half
- Reporting
 - Describe EDA that was performed
 - Understand and describe the limitations of your analysis
- Differentiate between descriptive analysis and formal inference

How to Carry out EDA?

- Plot data in multiple ways to get different insights
- Transform variables to symmetrize distributions
- Transform to straighten relationships
- Derive new variables
- Consider effect of other variables on distributions & relationships

Connect what you find to the question and context

A Case Study



Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Question



Taxi driver claim:

When traffic on a freeway begins to break down, the fasts lane breaks down first so you should move immediately to the right.

Can we confirm this phenomena?

Can we turn this claim into a statistics question that is answerable with data?

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Question

When traffic on a freeway begins to break down, the fast lane breaks down first so they move immediately to the right.

What information do we need to study this question? –

- Observe traffic on a freeway
- Include time to see when traffic is heavy and break downs
- Record traffic in lanes to differentiate traffic levels between in lanes

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Simpler Related Questions

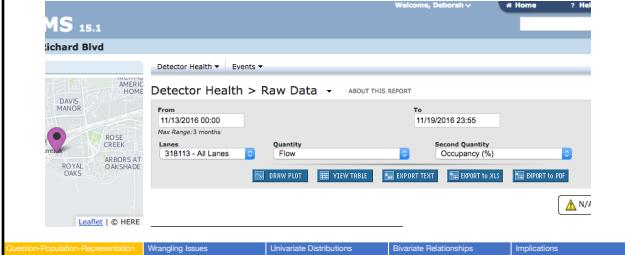
- Do the 3 lanes serve the same amount of traffic?
- When 1 lane is congested, are the other lanes also congested?
- What's the relationship between amount of traffic served and congestion?
- How are answers to these questions impacted by time of day and day of week?

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Let's Get Some Data

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

PEMS Data – Available in online form



Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

PEMS Data – Available in online form

Loop Detector Data:

Number of vehicles that pass over a loop detector in a 5-minute period

Sample Time	318113 Lane 1 Flow	318113 Lane 2 Flow	318113 Lane
11/13/2016 00:00	45	52	2,288 2,924 1,853
11/13/2016 00:15	35	52	1,844 2,444 1,751
11/13/2016 00:30	35	59	1,865 3,255 1,771
11/13/2016 00:45	45	52	1,844 2,444 1,751
11/13/2016 01:00	30	40	1,601 2,422 1,568
11/13/2016 01:15	35	45	1,844 2,444 1,751
11/13/2016 01:30	26	41	1,333 2,279 1,123
11/13/2016 01:45	35	45	1,844 2,444 1,751
11/13/2016 02:00	21	35	1,112 1,845 1,478
11/13/2016 02:15	28	35	1,112 1,845 1,478
11/13/2016 02:30	19	26	1,023 1,378 1,268
11/13/2016 02:45	24	29	1,177 1,488 933
11/13/2016 03:00	27	31	1,333 1,845 1,478
11/13/2016 03:15	18	34	1,023 1,378 1,154
11/13/2016 03:30	28	30	1,078 1,413 1,023
11/13/2016 03:45	21	30	1,023 1,378 1,023
11/13/2016 04:00	17	24	892 1,243 922
11/13/2016 04:15	35	35	1,333 1,845 1,478
11/13/2016 04:30	17	32	879 1,747 1,244
11/13/2016 04:45	17	32	879 1,747 1,244
11/13/2016 05:00	19	17	943 1,413 1,289
11/13/2016 05:15	17	18	892 1,243 922
11/13/2016 05:30	14	21	712 1,145 1,032
11/13/2016 05:45	11	20	9 1,044 1,044

Percentage of time in the 5-minute period that a vehicle was over the loop detector

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Statistical Issues:

What is the context Question?
What is the Population?
Data Represent Population?

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Population

5-minute intervals at any time and place on a freeway

Data:

Time: Nov 13 to Nov 19, 2016 (one week)

Place: Loop 318113 on I 80 Eastbound



Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Representative

Traffic may behave differently at:

- Other locations
- Other times of year

Such as?

Keep this in mind as we analyze the data

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Data Wrangling Issues:

Structure
Granularity
Faithfulness
Time
Scope

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Structure

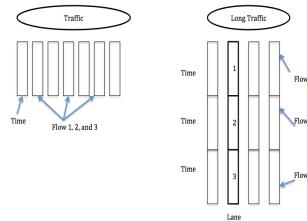
- Rectangular?
 - Record delimiter?
 - Variable value delimiter?
 - Primitive Data Types?

	Report Date	Site ID	Lane	Flow	Report Date	Site ID	Lane	Flow
1	1/13/2018	31813	lane 1	Flow	1	31813	lane 1	Flow
2	1/13/2018	00:00	5	27	2	2,268	0,04	1,033
3	1/13/2018	00:01	25	16	3	2,268	0,04	1,033
4	1/13/2018	00:02	55	29	4	1,865	0,04	1,033
5	1/13/2018	00:03	25	16	5	1,865	0,04	1,033
6	1/13/2018	00:04	25	16	6	1,865	0,04	1,033
7	1/13/2018	00:05	25	16	7	1,865	0,04	1,033
8	1/13/2018	00:06	25	16	8	1,865	0,04	1,033
9	1/13/2018	00:07	25	16	9	1,865	0,04	1,033
10	1/13/2018	00:08	25	16	10	1,865	0,04	1,033
11	1/13/2018	00:09	25	16	11	1,865	0,04	1,033
12	1/13/2018	00:10	25	16	12	1,865	0,04	1,033
13	1/13/2018	00:11	25	16	13	1,865	0,04	1,033
14	1/13/2018	00:12	25	16	14	1,865	0,04	1,033
15	1/13/2018	00:13	25	16	15	1,865	0,04	1,033
16	1/13/2018	00:14	25	16	16	1,865	0,04	1,033
17	1/13/2018	00:15	25	16	17	1,865	0,04	1,033
18	1/13/2018	00:16	25	16	18	1,865	0,04	1,033
19	1/13/2018	00:17	25	16	19	1,865	0,04	1,033
20	1/13/2018	00:18	25	16	20	1,865	0,04	1,033
21	1/13/2018	00:19	25	16	21	1,865	0,04	1,033
22	1/13/2018	00:20	25	16	22	1,865	0,04	1,033
23	1/13/2018	00:21	25	16	23	1,865	0,04	1,033
24	1/13/2018	00:22	25	16	24	1,865	0,04	1,033
25	1/13/2018	00:23	25	16	25	1,865	0,04	1,033
26	1/13/2018	00:24	25	16	26	1,865	0,04	1,033
27	1/13/2018	00:25	25	16	27	1,865	0,04	1,033
28	1/13/2018	00:26	25	16	28	1,865	0,04	1,033
29	1/13/2018	00:27	25	16	29	1,865	0,04	1,033
30	1/13/2018	00:28	25	16	30	1,865	0,04	1,033
31	1/13/2018	00:29	25	16	31	1,865	0,04	1,033
32	1/13/2018	00:30	25	16	32	1,865	0,04	1,033
33	1/13/2018	00:31	25	16	33	1,865	0,04	1,033
34	1/13/2018	00:32	25	16	34	1,865	0,04	1,033
35	1/13/2018	00:33	25	16	35	1,865	0,04	1,033
36	1/13/2018	00:34	25	16	36	1,865	0,04	1,033
37	1/13/2018	00:35	25	16	37	1,865	0,04	1,033
38	1/13/2018	00:36	25	16	38	1,865	0,04	1,033
39	1/13/2018	00:37	25	16	39	1,865	0,04	1,033
40	1/13/2018	00:38	25	16	40	1,865	0,04	1,033
41	1/13/2018	00:39	25	16	41	1,865	0,04	1,033
42	1/13/2018	00:40	25	16	42	1,865	0,04	1,033
43	1/13/2018	00:41	25	16	43	1,865	0,04	1,033
44	1/13/2018	00:42	25	16	44	1,865	0,04	1,033
45	1/13/2018	00:43	25	16	45	1,865	0,04	1,033
46	1/13/2018	00:44	25	16	46	1,865	0,04	1,033
47	1/13/2018	00:45	25	16	47	1,865	0,04	1,033
48	1/13/2018	00:46	25	16	48	1,865	0,04	1,033
49	1/13/2018	00:47	25	16	49	1,865	0,04	1,033
50	1/13/2018	00:48	25	16	50	1,865	0,04	1,033
51	1/13/2018	00:49	25	16	51	1,865	0,04	1,033
52	1/13/2018	00:50	25	16	52	1,865	0,04	1,033
53	1/13/2018	00:51	25	16	53	1,865	0,04	1,033
54	1/13/2018	00:52	25	16	54	1,865	0,04	1,033
55	1/13/2018	00:53	25	16	55	1,865	0,04	1,033
56	1/13/2018	00:54	25	16	56	1,865	0,04	1,033
57	1/13/2018	00:55	25	16	57	1,865	0,04	1,033
58	1/13/2018	00:56	25	16	58	1,865	0,04	1,033
59	1/13/2018	00:57	25	16	59	1,865	0,04	1,033
60	1/13/2018	00:58	25	16	60	1,865	0,04	1,033
61	1/13/2018	00:59	25	16	61	1,865	0,04	1,033
62	1/13/2018	00:00	25	16	62	1,865	0,04	1,033
63	1/13/2018	00:01	25	16	63	1,865	0,04	1,033
64	1/13/2018	00:02	25	16	64	1,865	0,04	1,033
65	1/13/2018	00:03	25	16	65	1,865	0,04	1,033
66	1/13/2018	00:04	25	16	66	1,865	0,04	1,033
67	1/13/2018	00:05	25	16	67	1,865	0,04	1,033
68	1/13/2018	00:06	25	16	68	1,865	0,04	1,033
69	1/13/2018	00:07	25	16	69	1,865	0,04	1,033
70	1/13/2018	00:08	25	16	70	1,865	0,04	1,033
71	1/13/2018	00:09	25	16	71	1,865	0,04	1,033
72	1/13/2018	00:10	25	16	72	1,865	0,04	1,033
73	1/13/2018	00:11	25	16	73	1,865	0,04	1,033
74	1/13/2018	00:12	25	16	74	1,865	0,04	1,033
75	1/13/2018	00:13	25	16	75	1,865	0,04	1,033
76	1/13/2018	00:14	25	16	76	1,865	0,04	1,033
77	1/13/2018	00:15	25	16	77	1,865	0,04	1,033
78	1/13/2018	00:16	25	16	78	1,865	0,04	1,033
79	1/13/2018	00:17	25	16	79	1,865	0,04	1,033
80	1/13/2018	00:18	25	16	80	1,865	0,04	1,033
81	1/13/2018	00:19	25	16	81	1,865	0,04	1,033
82	1/13/2018	00:20	25	16	82	1,865	0,04	1,033
83	1/13/2018	00:21	25	16	83	1,865	0,04	1,033
84	1/13/2018	00:22	25	16	84	1,865	0,04	1,033
85	1/13/2018	00:23	25	16	85	1,865	0,04	1,033
86	1/13/2018	00:24	25	16	86	1,865	0,04	1,033
87	1/13/2018	00:25	25	16	87	1,865	0,04	1,033
88	1/13/2018	00:26	25	16	88	1,865	0,04	1,033
89	1/13/2018	00:27	25	16	89	1,865	0,04	1,033
90	1/13/2018	00:28	25	16	90	1,865	0,04	1,033
91	1/13/2018	00:29	25	16	91	1,865	0,04	1,033
92	1/13/2018	00:30	25	16	92	1,865	0,04	1,033
93	1/13/2018	00:31	25	16	93	1,865	0,04	1,033
94	1/13/2018	00:32	25	16	94	1,865	0,04	1,033
95	1/13/2018	00:33	25	16	95	1,865	0,04	1,033
96	1/13/2018	00:34	25	16	96	1,865	0,04	1,033
97	1/13/2018	00:35	25	16	97	1,865	0,04	1,033
98	1/13/2018	00:36	25	16	98	1,865	0,04	1,033
99	1/13/2018	00:37	25	16	99	1,865	0,04	1,033
100	1/13/2018	00:38	25	16	100	1,865	0,04	1,033
101	1/13/2018	00:39	25	16	101	1,865	0,04	1,033
102	1/13/2018	00:40	25	16	102	1,865	0,04	1,033
103	1/13/2018	00:41	25	16	103	1,865	0,04	1,033
104	1/13/2018	00:42	25	16	104	1,865	0,04	1,033
105	1/13/2018	00:43	25	16	105	1,865	0,04	1,033
106	1/13/2018	00:44	25	16	106	1,865	0,04	1,033
107	1/13/2018	00:45	25	16	107	1,865	0,04	1,033
108	1/13/2018	00:46	25	16	108	1,865	0,04	1,033
109	1/13/2018	00:47	25	16	109	1,865	0,04	1,033
110	1/13/2018	00:48	25	16	110	1,865	0,04	1,033
111	1/13/2018	00:49	25	16	111	1,865	0,04	1,033
112	1/13/2018	00:50	25	16	112	1,865	0,04	1,033
113	1/13/2018	00:51	25	16	113	1,865	0,04	1,033
114	1/13/2018	00:52	25	16	114	1,865	0,04	1,033
115	1/13/2018	00:53	25	16	115	1,865	0,04	1,033
116	1/13/2018	00:54	25	16	116	1,865	0,04	1,033
117	1/13/2018	00:55	25	16	117	1,865	0,04	1,033
118	1/13/2018	00:56	25	16	118	1,865	0,04	1,033
119	1/13/2018	00:57	25	16	119	1,865	0,04	1,033
120	1/13/2018	00:58	25	16	120	1,865	0,04	1,033
121	1/13/2018	00:59	25	16	121	1,865	0,04	1,033
122	1/13/2018	00:00	25	16	122	1,865	0,04	1,033
123	1/13/2018	00:01	25	16	123	1,865	0,04	1,033
124	1/13/2018	00:02	25	16	124	1,865	0,04	1,033
125	1/13/2018	00:03	25	16	125	1,865	0,04	1,033
126	1/13/2018	00:04	25	16	126	1,865	0,04	1,033
127	1/13/2018	00:05	25	16	127	1,865	0,04	1,033
128	1/13/2018	00:06	25	16	128	1,865	0,04	1,033
129	1/13/2018	00:07	25	16	129	1,865	0,04	1,033
130	1/13/2018	00:08	25	16	130	1,865	0,04	1,033
131	1/13/2018	00:09	25	16	131	1,865	0,04	1,033
132	1/13/2018	00:10	25	16	132	1,865	0,04	1,033
133	1/13/2018	00:11	25	16	133	1,865	0,04	1,033
134	1/13/2018	00:12	25	16	134	1,865	0,04	1,033
135	1/13/2018	00:13	25	16	135	1,865	0,04	1,033
136	1/13/2018	00:14	25	16	136	1,865	0,04	1,033
137	1/13/2018	00:15	25	16	137	1,865	0,04	1,033
138	1/13/2018	00:16	25	16	138	1,865	0,04	1,033
139	1/13/2018	00:17	25	16	139	1,865	0,04	1,033
140	1/13/2018	00:18	25	16	140	1,865	0,04	1,033
141	1/13/2018	00:19	25	16	141	1,865	0,04	1,033

Question-Population-Representation

Granularity

- Key – date/time (5 minute interval)
 - Alternative Key – more useful in our analysis:
 - Date + Lane
 - Need to stack our data table
 - Need to create a new variable



Question-Population-Representation Wrangling issues Univariate Distributions Bivariate Relationships Implications

Faithfulness

- Census of all 5-minute intervals in the time period for all 3 lanes
 - Check detector health to see if all working in that time period, missing and faulty values are imputed
 - Dependency –
 - percentages between 0 & 100
 - Number of records 12 (intervals/hn) * 24 (hr/day) * 7 day/wk = 2016

Question-Population-Representation	Wrangling Issues	Univariate Distributions	Bivariate Relationships	Implications
------------------------------------	------------------	--------------------------	-------------------------	--------------

Temporality & Scope

- Temporality - One week in November 2016
 - Scope – Recordings for some (which?) 5-minute intervals in the time period may be imputed from historical data

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications

Ready for 2nd EDA

- Do the 3 lanes serve the same amount of traffic?
- When 1 lane is congested, are the other lanes also congested?
- What's the relationship between amount of traffic served and congestion?

How are answers to these questions impacted by time of day and day of week?

Question-Population-Representation

Distribution of Variable Values

Always a good place to start
Get your head in the game

Distribution-Distribution-Distribution
Marketing-Kommunikation
Institutionelle Distribution
Bürokratische Distribution
Innovationslinie

Distribution of Values

- Rug plots show exact location of values



- This tends not to be useful when we have many values
- Typically, we want a summary of this distribution

Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Histograms & Density Curves

- Histogram bar:
 $\text{Height} * \text{Width} = \text{Area} = \text{Proportion (or count)}$
- Similar property for area beneath a density curve
- No longer see individual values
- Smooth data values over a bin/region (histogram/density curve)
- Focus on the main features of the distribution
- Caution: Smoothing can reveal or disguise features

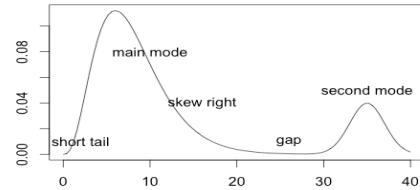
Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Features Seen in a Visual Summary

- Mode(s) - values concentrate around particular points
- Symmetry – skew left, symmetric, skew right distribution of values about center
- Tails - long, short, normal (what expect for normal distribution)
- Gaps - regions where no values observed
- Outliers - unusually large/small values

Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Distribution of Values



Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

What to Expect for Traffic Flow Distribution?

- Skew or Symmetric? Why?
- Long or short tails?
- Number of modes?

Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Flow Density



Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Do the 3 lanes serve the same amount of traffic?

Flow = count of vehicles in 5-minute intervals
Sum flow - left: 11k, center: 13k, right: 8k lanes

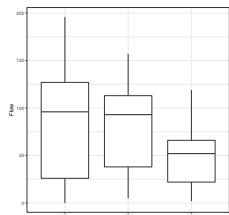
Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Compare Flow Across 3 Lanes



Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Summary Stats for 3 Flow Distributions



- Quartiles
- Skewness
- Longer right tail
- What don't we see?

Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

How Do Lane Flows Vary Together in Time?

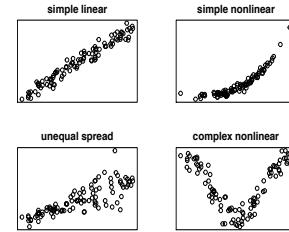
Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | Bivariate Relationships | Implications

Plotting Pairs of Variables

- Scatter plot uncovers form of relationship between 2 variables
- Linear relationships are particularly simple to interpret
- Simple and elegant statistical theory for linear relationships
- Models are typically approximations, choose a simpler model over a complex one

Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | **Bivariate Relationships** | Implications

Plotting Pairs of Variables



Question-Population-Representation | Wrangling Issues | **Univariate Distributions** | **Bivariate Relationships** | Implications

Flow in: Middle & Right Lanes and Left and Right Lanes

Do we expect:

- flow in two lanes increase together?
- flow in two lanes to be linearly associated?
- Slope to be roughly 1?

Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

Flow in Lanes for the Same 5 Minutes

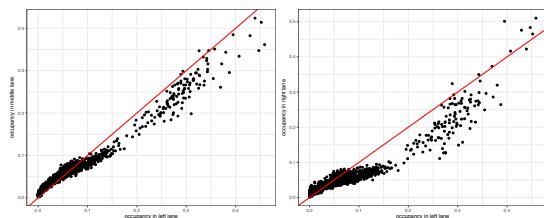


Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

When a lane is congested,
are the other lanes
congested too?

Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

Occupancy in Lanes for 5 Min. Intervals



Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

What's the relationship
between amount of traffic
and congestion?

Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

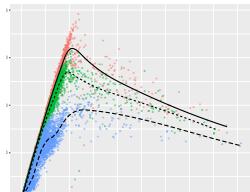
Flow and Occupancy Pairs

Do we expect:

- flow and occupancy to increase together?
- flow and occupancy to be linearly associated?
- 3 lanes to have the same relationship?

Question-Population-Representation | Wrangling Issues | Univariate Distributions | **Bivariate Relationships** | Implications

Flow & Occupancy



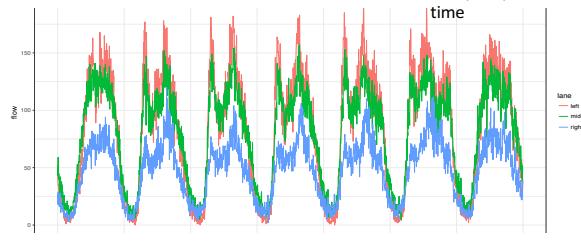
- Flow and occupancy are linearly related for occupancy below 10%
- As occupancy (congestion) increases traffic breaks down and flow decreases
- Breakdown appears to occur at different occupancy levels for the lanes

Question-Population-Representation | Wrangling Issues | Univariate Distributions | Bivariate Relationships | Implications

How are these findings affected by time of day and day of week?

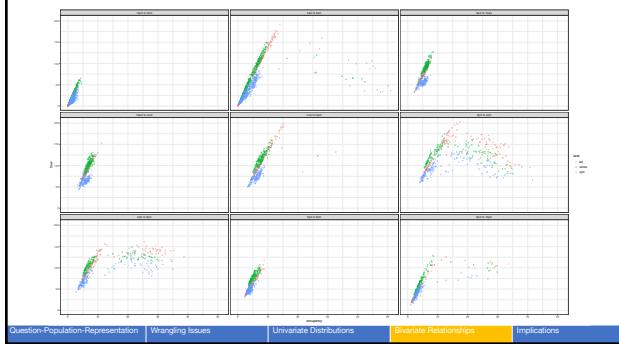
Question-Population-Representation | Wrangling Issues | Univariate Distributions | Bivariate Relationships | Implications

Examine Flow in Time



Question-Population-Representation | Wrangling Issues | Univariate Distributions | Bivariate Relationships | Implications

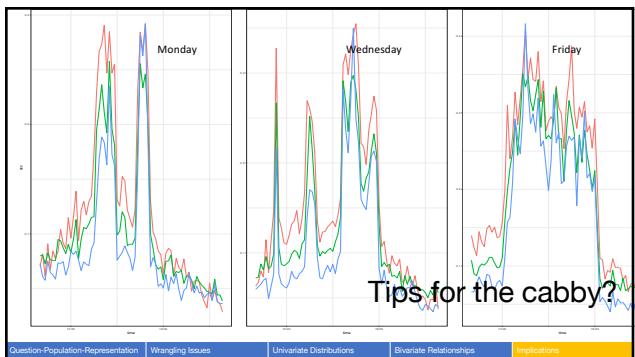
We want to
see flow AND
occupancy in
time



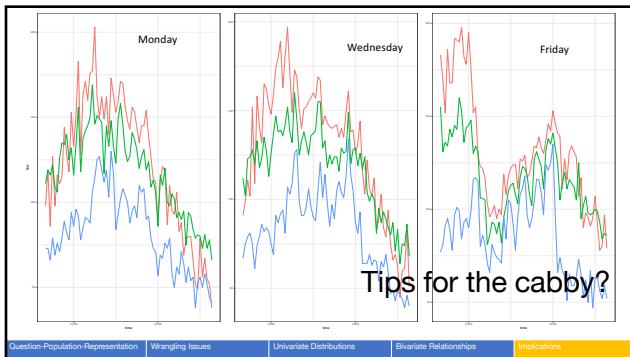
Question-Population-Representation | Wrangling Issues | Univariate Distributions | Bivariate Relationships | Implications

Focus on Certain Times of Day

Question-Population-Representation | Wrangling Issues | Univariate Distributions | Bivariate Relationships | Implications



Tips for the cabby?



Implications for Formal Analysis

- Lane matters – distributions are similar in shape, but locations of peaks and size of spread differ
- Relationship between flow and occupancy is linear until traffic breaks down
- Traffic jams do not have the same relationship, spread increases, negative association between flow and occupancy
- Distinct patterns within a day and for day of week

Question-Population-Representation Wrangling Issues Univariate Distributions Bivariate Relationships Implications