EDA, again

Air Pollution

US Air Pollution Data, 2008-10

- The Environmental Protection Agency (EPA) regulates national air quality standards
- One thing it monitors is the level of fine particle pollution (cannot be seen with the naked eye)
- Rule: fine particle pollution averaged over a 3 year time span cannot exceed 12 micrograms per cubic meter

Fine Particle pollution: PM2.5

Source

- Particulate matter, or PM, is the term for particles found in the air, including dust, dirt, soot, smoke, and liquid droplets. Particles can be suspended in the air for long periods of time. Some particles are large or dark enough to be seen as soot or smoke. Others are so small that individually they can only be detected with an electron microscope.
- Many manmade and natural sources emit PM directly or emit other pollutants that react in the atmosphere to form PM.
- PM come in a wide range of sizes. Particles fewer than 10 micrometers in diameter (PM10) pose a health concern; they can be inhaled into and accumulate in the respiratory system.
- Particles less than 2.5 micrometers in diameter (PM2.5) are referred to as "fine" particles and are believed to pose the greatest health risks.
- Because of their small size (approximately 1/30th the average width of a human hair), fine particles can lodge deeply in the lungs.

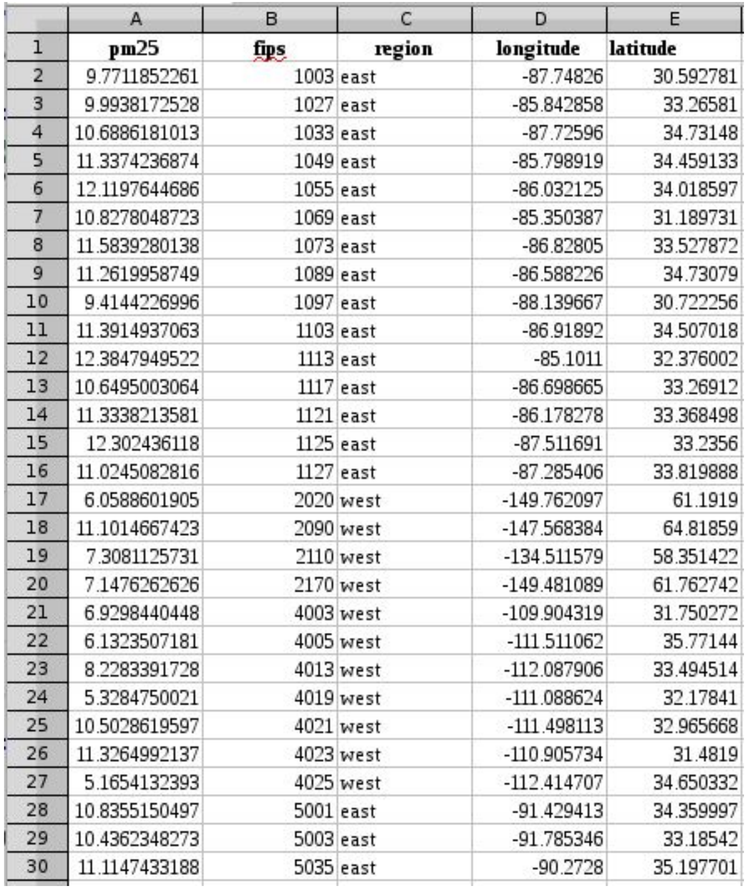
Question: Are there counties that are in violation of the EPA’s set standard for fine particle pollution?

If yes, counties face legal consequences under the Clean Air Act

- States would have to create a SIP and submit it to the EPA
- SIP must consist techniques for reducing air pollution
- SIP must include a reasonable timeline to achieve compliance

SIP = State Implementation Plan

Average PM2.5 by geographic location



	A	B	C	D	E
1	pm25	fips	region	longitude	latitude
2	9.7711852261	1003	east	-87.74826	30.592781
3	9.9938172528	1027	east	-85.842858	33.26581
4	10.6886181013	1033	east	-87.72596	34.73148
5	11.3374236874	1049	east	-85.798919	34.459133
6	12.1197644686	1055	east	-86.032125	34.018597
7	10.8278048723	1069	east	-85.350387	31.189731
8	11.5839280138	1073	east	-86.82805	33.527872
9	11.2619958749	1089	east	-86.588226	34.73079
10	9.4144226996	1097	east	-88.139667	30.722256
11	11.3914937063	1103	east	-86.91892	34.507018
12	12.3847949522	1113	east	-85.1011	32.376002
13	10.6495003064	1117	east	-86.698665	33.26912
14	11.3338213581	1121	east	-86.178278	33.368498
15	12.302436118	1125	east	-87.511691	33.2356
16	11.0245082816	1127	east	-87.285406	33.819888
17	6.0588601905	2020	west	-149.762097	61.1919
18	11.1014667423	2090	west	-147.568384	64.81859
19	7.3081125731	2110	west	-134.511579	58.351422
20	7.1476262626	2170	west	-149.481089	61.762742
21	6.9298440448	4003	west	-109.904319	31.750272
22	6.1323507181	4005	west	-111.511062	35.77144
23	8.2283391728	4013	west	-112.087906	33.494514
24	5.3284750021	4019	west	-111.088624	32.17841
25	10.5028619597	4021	west	-111.498113	32.965668
26	11.3264992137	4023	west	-110.905734	31.4819
27	5.1654132393	4025	west	-112.414707	34.650332
28	10.8355150497	5001	east	-91.429413	34.359997
29	10.4362348273	5003	east	-91.785346	33.18542
30	11.1147433188	5035	east	-90.2728	35.197701

Variables:

- PM2.5 in micrograms per cubic meter
- FIPS: Federal Information Processing Standards
- Region: East or West
- Longitude
- Latitude

Large, unintuitive data set

- Takes time to look through the 577 rows
- Hard to do draw conclusions by simply eyeballing raw data
- A visualization of where the majority of the data lie, in comparison to any potential outliers, might help
- How might we visualize PM2.5 (a quantitative variable)?

Refresher: Visualizations

- Histograms
  - A bar graph in which the area of each bar is proportional to the frequency, so the total area under all bars is 1
- Box (and whisker) plots
  - Contains 5 important variables: min (or lower fence), max (or upper fence), median, and the first and third quartiles (the latter of which encompass half the data)

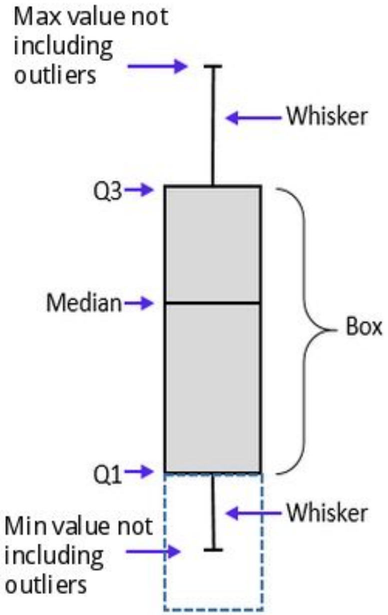

Max value not including outliers
Whisker
Q3
Median
Box
Q1
Whisker
Min value not including outliers

Image Source

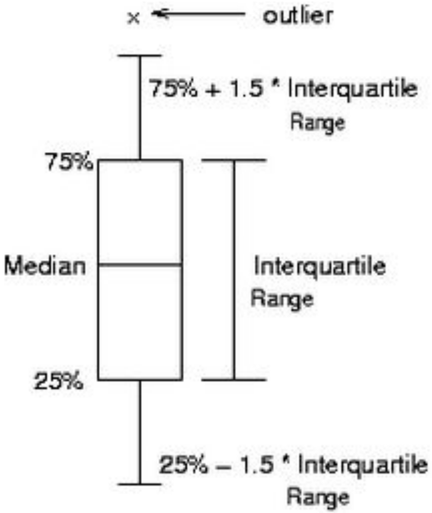

×
outlier
75% + 1.5 * Interquartile Range
75%
Median
Interquartile Range
25%
25% – 1.5 * Interquartile Range

Image Source

Let’s start by making a histogram

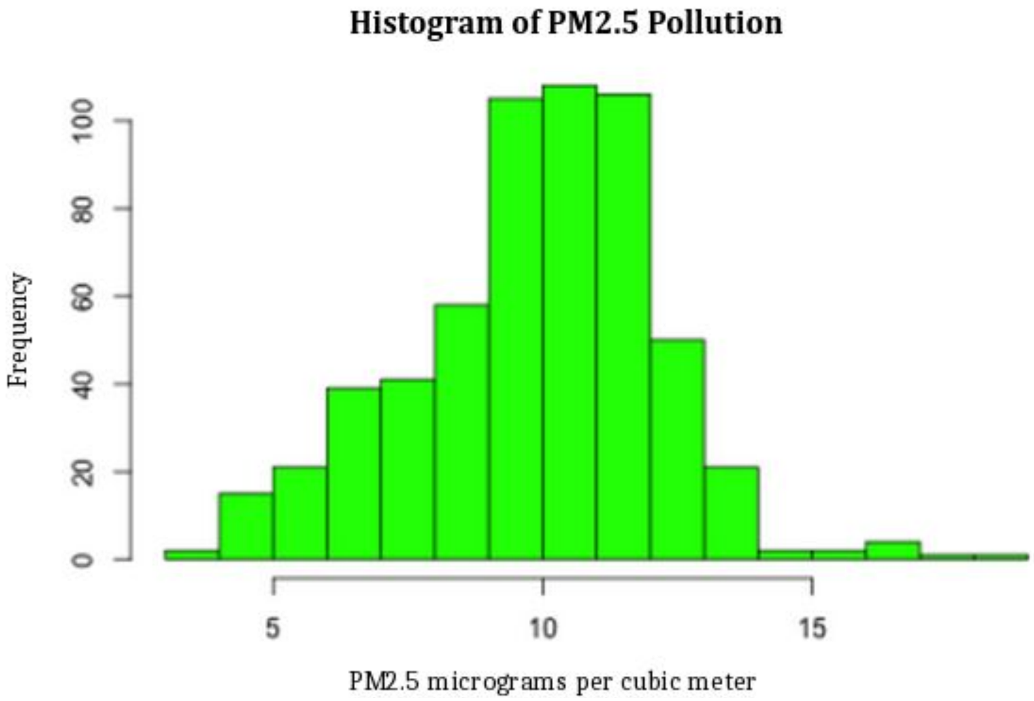
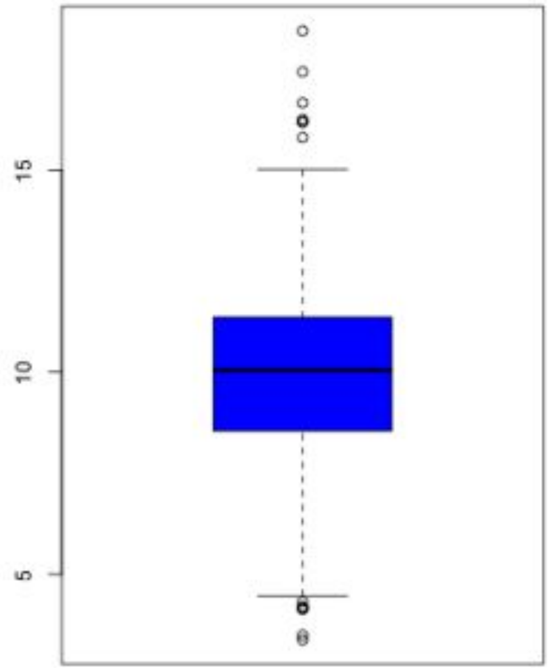


Image Source

- Many counties fall in the range of 9-12 micrograms per cubic meter
- Could be because the cap is 12; most counties barely adhere to it
- Looks like there is a long tail on the right: potential outliers

Let’s also make a box plot

Box and Whisker plot of PM2.5 micrograms per cubic meter

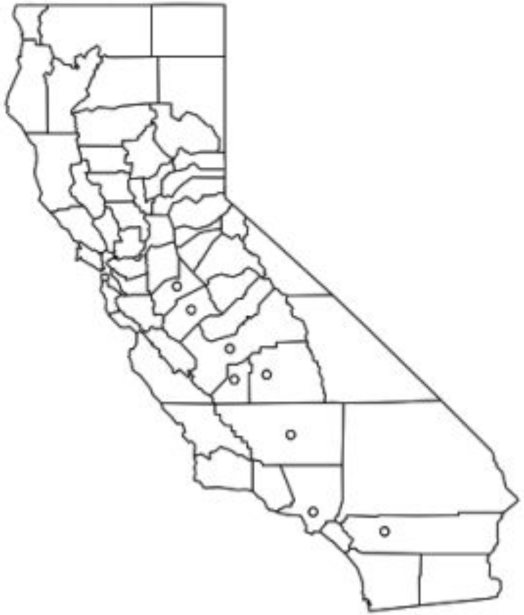


- Yup! There are some outliers: the points above and below the whiskers
- Applying the IQR rule of thumb, there are points that fall outside the fences, which can be labeled outliers

Filter

- Let’s explore our data some more to find out more about the counties in violation
- Specifically, let’s filter our data to find out the locations of counties whose PM2.5 exceeds 15
- This search yields a list of 8 zip codes that all begin with 06
- All the offending counties are in California!

Mapping these 8 counties shows:



- Plotting these data can help us understand what is going wrong
- The next step after the “quick and dirty” graphing is to understand why the graph looks this way

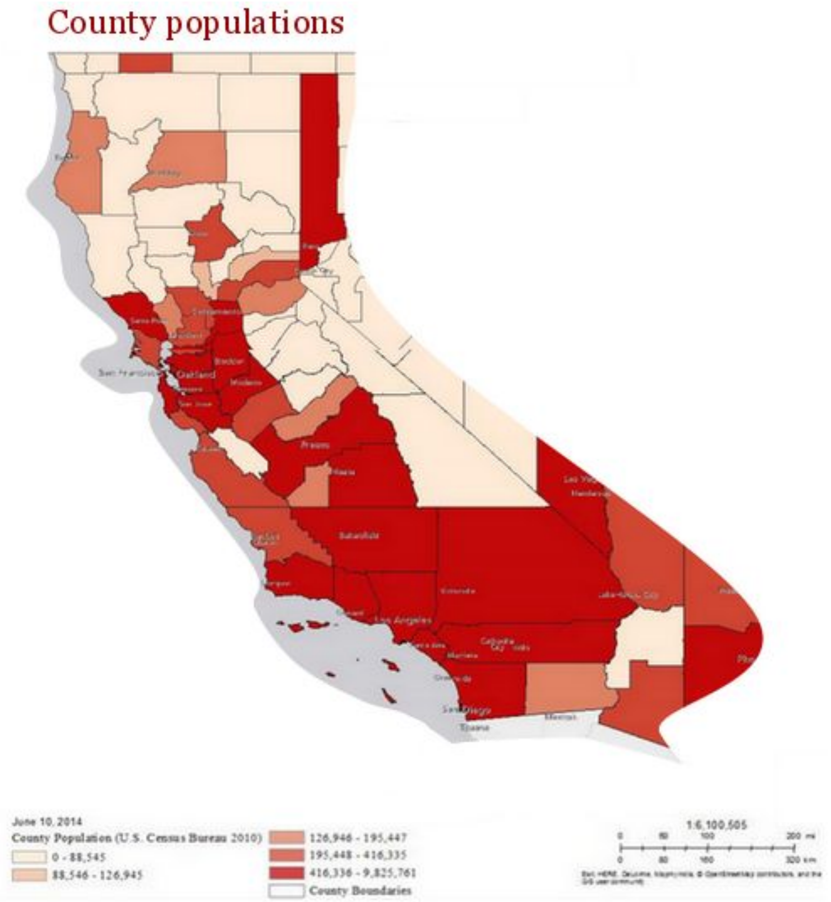
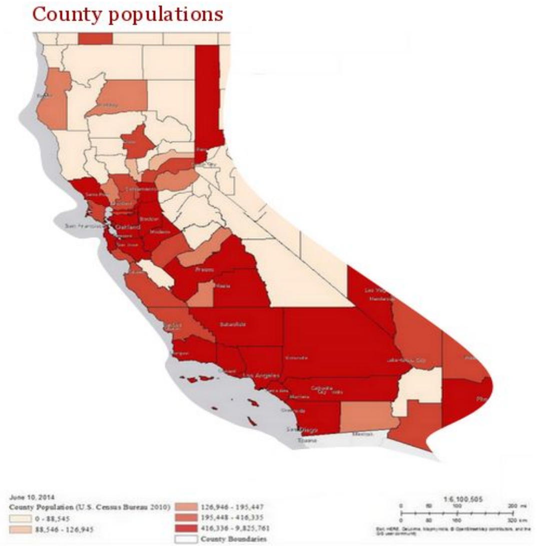
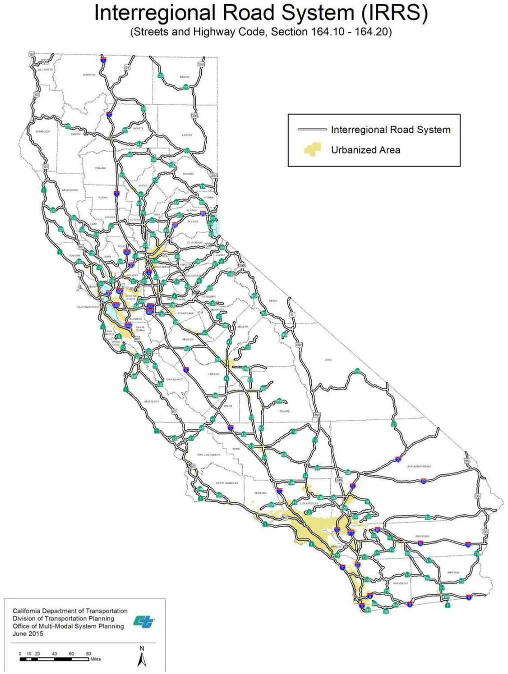


Image Source



County populations



Interregional Road System (IRRS)

(Streets and Highway Code, Section 164.10 - 164.20)

California Department of Transportation
Division of Transportation Planning
Office of Multi-Modal System Planning
June 2015

What did we learn?

- We identified problem areas in CA, and mapped them
  - We observed the populations of these areas
  - We also observed their traffic densities
- These visualizations enabled us to formulate hypotheses about the potential causes of the PM2.5 excess
- Are we done? No, a data scientist’s work is never done.
- Next, perhaps we could visualize PM2.5 pollution by region

Two histograms, differentiating east and west

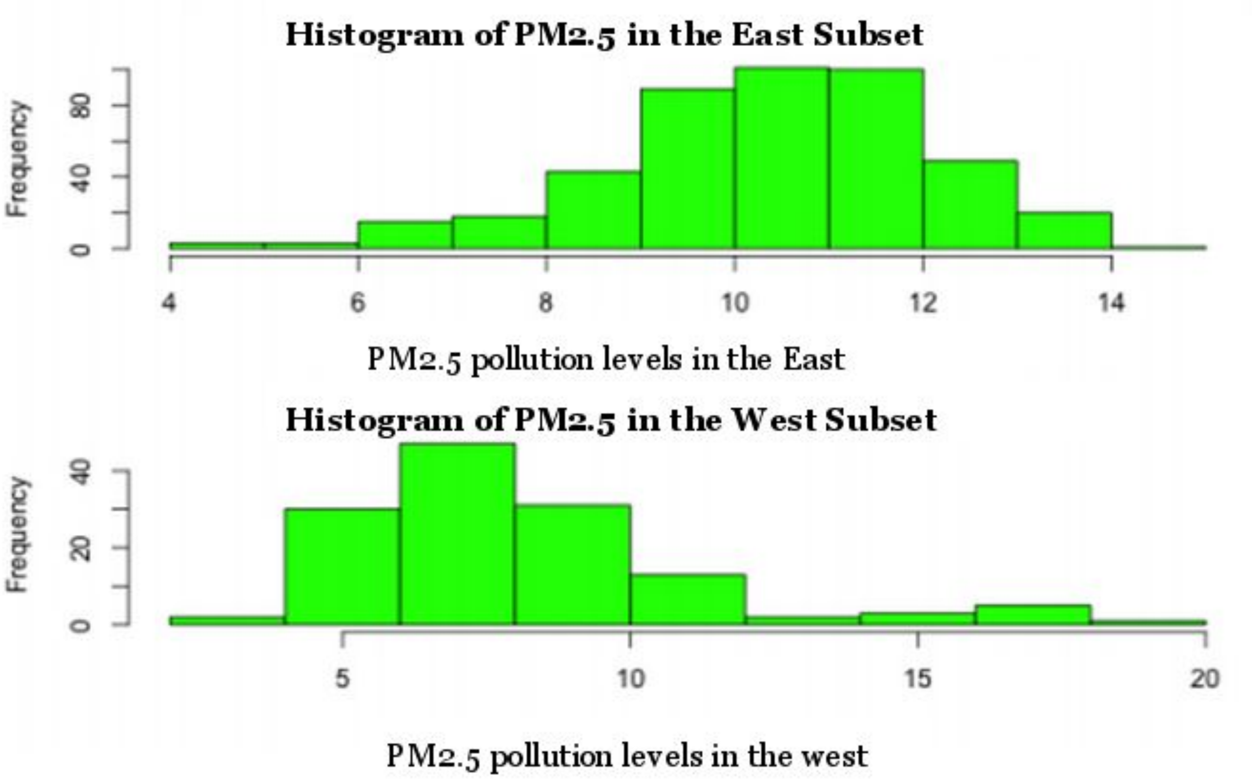


Image Source

Another box plot, differentiating east and west

Box and Whisker plot by East and West

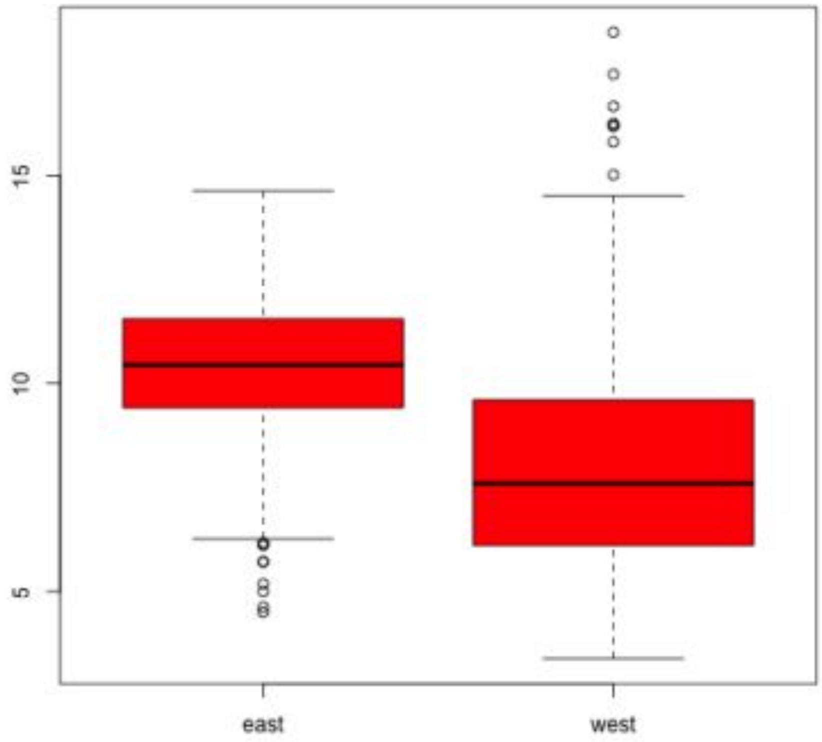


Image Source

- The median for the east is much higher than for the west
- There are outliers in both the east and the west
- But interestingly, all outliers above the allowable level lie in the west, and all below, the east

What we learned about air pollution (in 2008–10)

- Most counties complied with EPA’s regulations
- The most severe violations were in California
- The west had more severe violations than the east

Exploratory Data Analysis

- Allows us to identify suspected problem areas quickly, so we can begin to correct potential problems early on
- It has been called “quick and dirty”
  - It is quick, because, well, it can/should be quick
  - It is dirty, because it does not involve model building of any sort, so it does not necessary uncover the reasons for the associations we find in our data, but EDA can still guide our search for explanations

But why visualize?

- Sometimes averages, and other numerical descriptive statistics based on aggregate data, can be deceiving
- Even aggregate histograms can be deceiving!
  - E.g., the histogram that aggregates east and west data
- In this example, we obtained more information by partitioning the data regionally, into regional as opposed to a national histogram
  - In so doing, we learned where the different sorts of outliers lie