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
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 less than 10 micrometers in diameter (PM10) pose a health concern because 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 into 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 State Implementation Plan (SIP) and submit it to the EPA
- SIP must consist of a reasonable timeline and techniques for reducing air pollution
### Average PM2.5 by geographic location

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

<table>
<thead>
<tr>
<th></th>
<th>pm25</th>
<th>region</th>
<th>longitude</th>
<th>latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.77116522351</td>
<td>east</td>
<td>-87.74623</td>
<td>30.592781</td>
</tr>
<tr>
<td>2</td>
<td>9.5936172528</td>
<td>east</td>
<td>-85.643858</td>
<td>33.28561</td>
</tr>
<tr>
<td>3</td>
<td>10.6868131013</td>
<td>east</td>
<td>-85.73596</td>
<td>34.73149</td>
</tr>
<tr>
<td>4</td>
<td>11.3978298747</td>
<td>east</td>
<td>-85.708819</td>
<td>34.459338</td>
</tr>
<tr>
<td>5</td>
<td>12.11976414686</td>
<td>east</td>
<td>-86.632125</td>
<td>33.013507</td>
</tr>
<tr>
<td>6</td>
<td>10.8278088273</td>
<td>east</td>
<td>-85.250397</td>
<td>31.189731</td>
</tr>
<tr>
<td>7</td>
<td>11.5936830138</td>
<td>east</td>
<td>-86.828055</td>
<td>33.527872</td>
</tr>
<tr>
<td>8</td>
<td>11.2619587491</td>
<td>east</td>
<td>-86.582263</td>
<td>34.73079</td>
</tr>
<tr>
<td>9</td>
<td>9.4142226996</td>
<td>east</td>
<td>-88.139867</td>
<td>30.722256</td>
</tr>
<tr>
<td>10</td>
<td>11.391497063</td>
<td>east</td>
<td>-88.91892</td>
<td>34.507018</td>
</tr>
<tr>
<td>11</td>
<td>12.3847949522</td>
<td>east</td>
<td>-85.10111</td>
<td>32.379002</td>
</tr>
<tr>
<td>12</td>
<td>10.649003004</td>
<td>east</td>
<td>-86.698605</td>
<td>33.23912</td>
</tr>
<tr>
<td>13</td>
<td>11.3336213581</td>
<td>east</td>
<td>-86.173279</td>
<td>33.368498</td>
</tr>
<tr>
<td>14</td>
<td>12.3024361108</td>
<td>east</td>
<td>-97.511691</td>
<td>33.2356</td>
</tr>
<tr>
<td>15</td>
<td>11.025032816</td>
<td>east</td>
<td>-97.285406</td>
<td>33.819088</td>
</tr>
<tr>
<td>16</td>
<td>7.0588601005</td>
<td>west</td>
<td>-149.762607</td>
<td>61.19109</td>
</tr>
<tr>
<td>17</td>
<td>11.1014667423</td>
<td>west</td>
<td>-147.569394</td>
<td>64.81859</td>
</tr>
<tr>
<td>18</td>
<td>7.3081125731</td>
<td>west</td>
<td>-134.511579</td>
<td>58.354122</td>
</tr>
<tr>
<td>19</td>
<td>7.1476267626</td>
<td>west</td>
<td>-144.810891</td>
<td>61.762742</td>
</tr>
<tr>
<td>20</td>
<td>6.9296440448</td>
<td>west</td>
<td>-106.904319</td>
<td>31.750272</td>
</tr>
<tr>
<td>21</td>
<td>6.1323507181</td>
<td>west</td>
<td>-111.51062</td>
<td>35.77144</td>
</tr>
<tr>
<td>22</td>
<td>8.2283931728</td>
<td>west</td>
<td>-112.087905</td>
<td>33.494514</td>
</tr>
<tr>
<td>23</td>
<td>5.3284750021</td>
<td>west</td>
<td>-111.089624</td>
<td>32.17841</td>
</tr>
<tr>
<td>24</td>
<td>10.5028618597</td>
<td>west</td>
<td>-111.498133</td>
<td>32.903668</td>
</tr>
<tr>
<td>25</td>
<td>11.3256992137</td>
<td>west</td>
<td>-110.905734</td>
<td>31.4819</td>
</tr>
<tr>
<td>26</td>
<td>5.1654132383</td>
<td>west</td>
<td>-112.41707</td>
<td>34.650332</td>
</tr>
<tr>
<td>27</td>
<td>10.835150497</td>
<td>east</td>
<td>-91.429413</td>
<td>34.350997</td>
</tr>
<tr>
<td>28</td>
<td>10.482348273</td>
<td>east</td>
<td>-91.785346</td>
<td>33.18542</td>
</tr>
<tr>
<td>29</td>
<td>11.114743188</td>
<td>east</td>
<td>-90.272891</td>
<td>35.197701</td>
</tr>
</tbody>
</table>
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
- Since PM2.5 is a quantitative variable, we will visualize it using histograms and box plots
Refresher: Methods of visualization

● 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)
Let’s start by making a histogram
Let’s start by making a histogram

- 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

- Yup! There are clear 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 zipcodes 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.
Figure 7. Overlay of 2011 County-level VMT with Transportation Arteries.
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 PM2.5 pollution levels.
Another box plot, differentiating east and west

- Again, there are clear outliers
- Interestingly, all outliers above the allowable level lie in the west, and all below, the east
- Applying the IQR rule of thumb, there are points that fall outside the fences, which can therefore be labeled outliers
What we learned about air pollution (in 2008-10)

● Most counties were complying with EPA’s regulations
● The most severe violations were in California
● The west had more violations than the east
● EDA 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 guide our search for models/reasons
But why visualize?

- Sometimes averages can be deceiving
- In our previous example, we understood that we could get more information from observing regional averages as opposed to national averages after identifying
  - What the outliers look like
  - Where the outliers lie