Name: Rufus Petrie

HW5 Problem 1: Summaries of age-adjusted mortality

	Sample Size	Sample Mean	Sample SD	Sample Median	Sample IQR
Mortality	60	940.365	62.2048	943.7	88.475

HW5 Problem 2: Summaries of predictors and correlation analyses

	Sample Mean	Sample Median	R	99 % CI for p Value	R_s	Large Samp p Value
% Old	8.79	9	-0.17	(0.1733, 0.2020)	-0.20	0.1144
% Poor	14.37	13.2	0.40	(0.0001, 0.0021)	0.35	0.0058
Density	3876.05	3567.0	0.27	(0.0332, 0.0477)	0.27	0.0365
Hydrocarbons	37.85	14.5	-0.17	(0.1546, 0.1820)	0.30	0.0191
Sulfur Dioxide	53.76	30	0.42	(0.0003, 0.0031)	0.49	0.0001
Rain (in.)	37.35	38	0.51	(0.00002, 0.0018)	0.40	0.0017
January Temp	33.98	31.5	-0.03	(0.8091, 0.8371)	0.02	0.8307
July Temp	74.58	74	0.27	(0.0273, 0.0407)	0.31	0.0160
% Rel Humidity	56.6	57	-0.08	(0.4884, 0.5251)	-0.08	0.5168

HW5 Problem 3: Box plots and comparisons of measures of center

- *Old:* There is a greater variation between the median and lower quartile than there is between the median and upper quartile. There are no outliers, which is consistent with what we might expect because the sample mean and median are relatively close (8.79 and 9).
- *Poor:* There is greater variation between the median and upper quartile than between the median and lower quartile. There are a significant amount of outliers above the upper extreme, which helps to explain why the sample mean is greater than the sample median (14.37 compared to 13.2).
- *Density:* There is a greater variation between the median and upper quartile than between the median and lower quartile. There are also two outliers above the upper extreme, which helps to explain why the sample mean exceeds the sample median in this case (3876.05 as opposed to 3567).
- *Hydrocarbons:* There is greater variation between the upper quartile and median than there is between the lower quartile and median. Additionally, there are a bunch of significant outliers above the upper extreme (two of which appear to be *many* IQRs away from the median). This helps to explain why the sample mean is so much larger than the sample median (37.85 versus 14.5).
- *Sulfur Dioxide:* There is greater variation between the upper quartile and median than there is between

the lower quartile and median. Furthermore, there are four outliers above the upper extreme, but they are not as severe as the outliers for hydrocarbon measurements. Nevertheless, this helps to exlain why the sample mean is much larger than the sample median for sulfur dioxide (53.76 vs. 30).

- *Rain:* The two boxes look about even in terms of deviatiations between the quartiles. There are three outliers below the lower extreme, which helps to explain why the sample mean is below the sample median (37.35 as opposed to 38).

- *January Temperature:* There is greater variation between the upper quartile and median than there is between the lower quartile and median. There is only one outlier above the upper extreme. This helps to explain why the sample mean and sample median are reasonably close (33.98 and 31.5).

- *July Temperature:* There is greater variatiion between the upper quartile and median than there is between the lower quartile and median. There is only one outlier below the lower extreme. This is consistent with how the sample median and sample mean are similar (74.58 and 74).

- *% Relative Humidity:* There is greater variatiion between the upper quartile and median than there is between the lower quartile and median. Also, there are three outliers above the upper extreme and two outliers below the lower extreme. This helps to explain why the sample median and sample mean are similar (56.6 and 57).

HW5 Problem 4: Scatter plots and comparisons of correlation analyses

- *Old:* The scatter plot between *old* and *mortality* looks relatively normal except for around an old age percentage of around 7, at which point the variance in *mortality* increases dramatically compared to other places in the plots. Otherwise, the plot seems consistent with the similar Pearson and Spearman correlations between *old* and *mortality* (-0.20 vs. 0.17).

- *Poor:* The variation in *mortality* looks relatively stable throughout the scatter plot, but the variation in *poor* appears to grow larger as percentage poor reaches 15 percent. This helps to explain why the Pearson correlation is greater than the Spearman correlation (0.40 vs. 0.35) because the values of *poor* start to vary more above a certain threshold.

- *Density:* The scatterplot between *density* and *mortality* looks relatively normal except for one major outlier in the *density* variable. This is consistent with how we observed equivalent Pearson and Spearman correlations between *density* and *mortality*.

- *Hydrocarbons:* The scatterplot between *hydrocarbons* and *mortality* shows that the majority of *hydrocarbon* levels fall near zero with a bunch of sever upper outliers. This helps to explain how we saw differ ent signs between the Pearson and Spearman correlations because the extreme *hydrocarbon* outliers correspond to cities with relatively low *mortality* rates.

- *Sulfur Dioxide:* The scatterplot for *sulfur dioxide* looks relatively normal except for the increase in variation among the *sulfur dioxide* levels as they surpass 100. This probably explains the slight difference between the Pearson and Spearman correlations (0.42 vs. 0.49).

- *Rain:* In the scatterplot for *rain* and *mortality*, there are a couple of outliers associated with both high and low levels of rain, all of which happen in instances where the city has a relatively low mortality rate. This probably helps to explain why the Pearson correlation is slightly larger than the Spearman cor-

relation (0.51 vs. 0.40).

- *January Temperature:* The January temperature vs. mortality scatterplot looks relatively normal except for some increased variation at high and low temperature levels, but there isn't much difference between the Pearson and Spearman correlations.

- *July Temperature:* There doesn't seem to be anything wrong with the mortality rate vs. July temperature scatterplot, and the Pearson and Spearman correlations for those two variables are relatively close.

- *% Relative Humidity:* The scatterplot between *mortality* and *humidity* looks relatively normal except for a couple of positive and negative outliers outside of the main cluster. These don't seem to carry much statistical weight though because we observe equivalent Pearson and Spearman correlations.

HW5 Problem 5: Discussion

From our summary statistics, we see that percentage poor, population density, hydrocarbon levels, sulfur dioxide levels, rain levels and July temperature all had a significant correlation with mortality rates at the 5% level of significance. Additionally, all of the confidence intervals for these estimated p-values (except for relative humidity) are entirely below the 5% threshold, which supports the robustness of our relsults. Percentage old, relative humidity, and January temperatures did not have significant correlation with mortality rates at the 5% level of significant.

Most of our statistically significant correlations seem to make sense. It makes sense that percentage poor has a positive and statistically significant correlation with mortality because having little money implies that one may not be able to access healthcare and other services that help to increase longjevity. It makes sense that population density has a positive and statistically significant correlation with mortality because living in densely packed areas usually indicate that somebody lives in an urban area, which has risks like pollution, increased crime rates, etc. It makes sense that hydrocarbons would have a statistically significant relationship with adjusted mortality rates because whatever their effect, atmospheric chemicals tend to have incredibly consistent effects on people's health. Because of the extreme outliers in hydrocarbon levels that accompanied low mortality rates, it's difficult to discern their effect from this analysis, bit it appears that more investigation may reveal that they play a role in increasing mortality rates. Like hydrocarbons, it also makes sense that sulfur dioxide had a strong statistical correlation with mortality rates because it's an atmospheric gas. However, because of the lack of outliers, we can clearly see the positive correlation between sulfur dioxide and mortality rates, which needs little explanation. Furthermore, I am not certain why rainfall levels have such a strong positive correlation with mortality rates, but I suspect it may have something to do with supporting diseases, causing property damage, flooding, etc. Finally, July temperatures have a significant positive correlation with mortality rates, which may be a result of increases in thing like heatstroke, dehydration, etc.

It makes sense that percentage old doesn't necessary correlate with mortality in this circumstance because we used adjusted mortality rates. This means that we're taking the correlation of mortality rates calculated with national age proportions and comparing them to age proportions idiosyncratic to one area. Thus, it makes sense that there wouldn't be a significant correlation between the two because despite the fact that localized age demographics serve an important role in determining mortality

rates, these demographics are ignored in calculating the age adjusted mortality rates. This means that this correlation captures the effects of increased mortality in a local area due to old living people living there, but not because of their age. Additionally, it makes sense that relative humidity levels may not have a statistically significant impact on mortality rates because of the relatively low amount of varaition that we see in the humidity variable. Finally, I am not quite sure why January temperatures did not have a statistically significant impact on mortality rates despite the fact that July temperatures did. This might be because people show more care when preparing for cold weather.

Overall, I have reached a fair amount of intuitive conclusion from this analysis, but it would be interesting in the future to learn specifically about why rain levels have such a strong positive correlation with mortality rates. In addition, it would be interesting to try and pinpoint an estimate for the relationship between mortality rates and hydrocarbon levels that bypasses the outlier problems that I encountered. Finally, it may be interesting to try and find why January temperature levels have an insignificant correlation with mortality rates and why it's the opposite with July temperature levels.