DSC 465 – Data Visualization

Homework 3

PKeener
2-21-2021

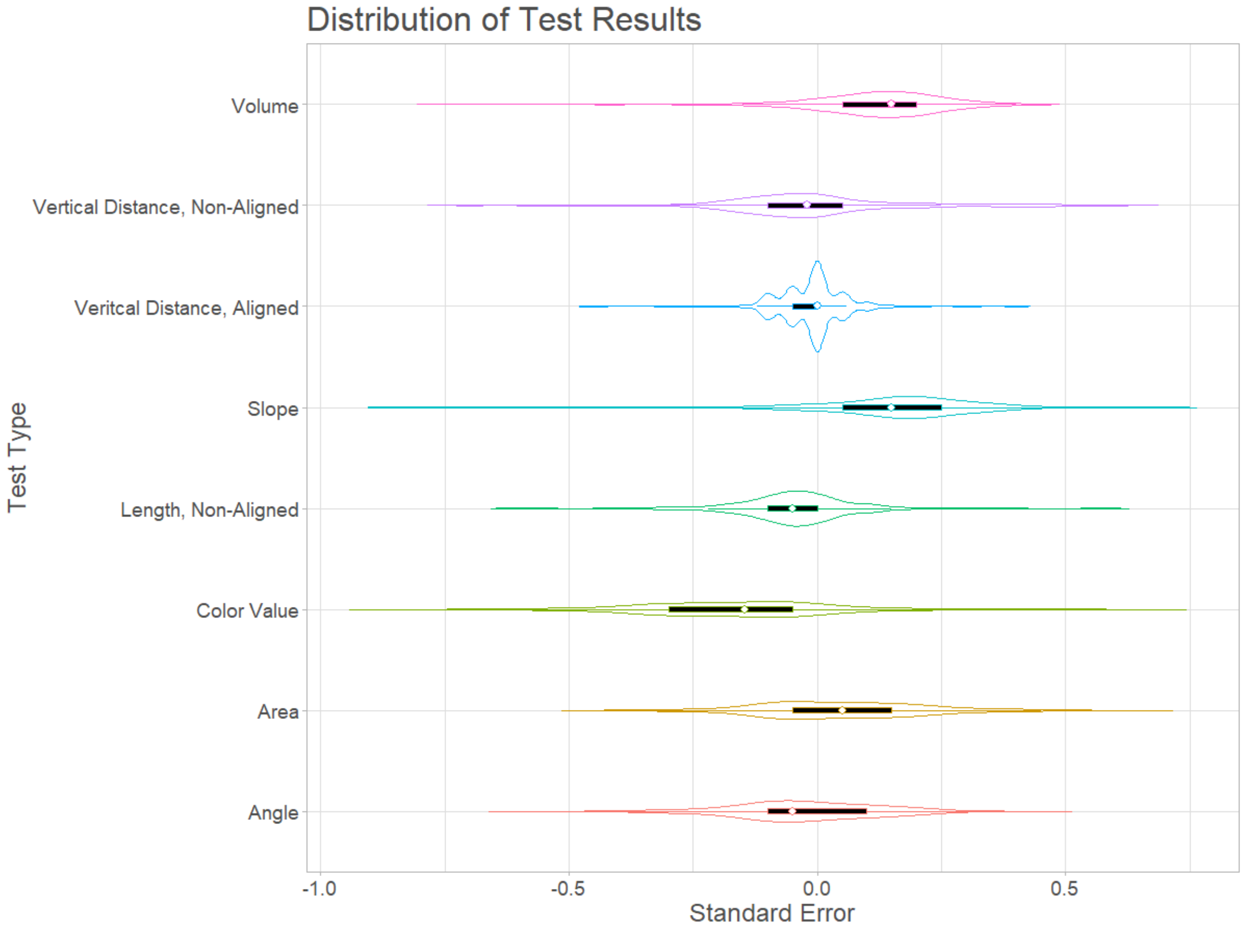
Problem 1

Problem 1 required exploration of the PerceptionExperiment.csv data set.

Corresponding Lines in Code File: 19 to 323

A. Over/Under Estimation of Data

Corresponding Lines in Code File: 39 to 78



The question of whether there were any tests where people generally under- or over-estimated data can be answered using a violin plot of the standard error (distance from response to actual). Typically, violin plots are vertical, but in this case the distribution was the subject of analysis, therefore it seemed more natural to generate them horizontally. Color was used to distinguish tests, and smoothing was reduced slightly as to emphasize distributions that weren't normal, such as the vertical/aligned test which had a multi-modal distribution.

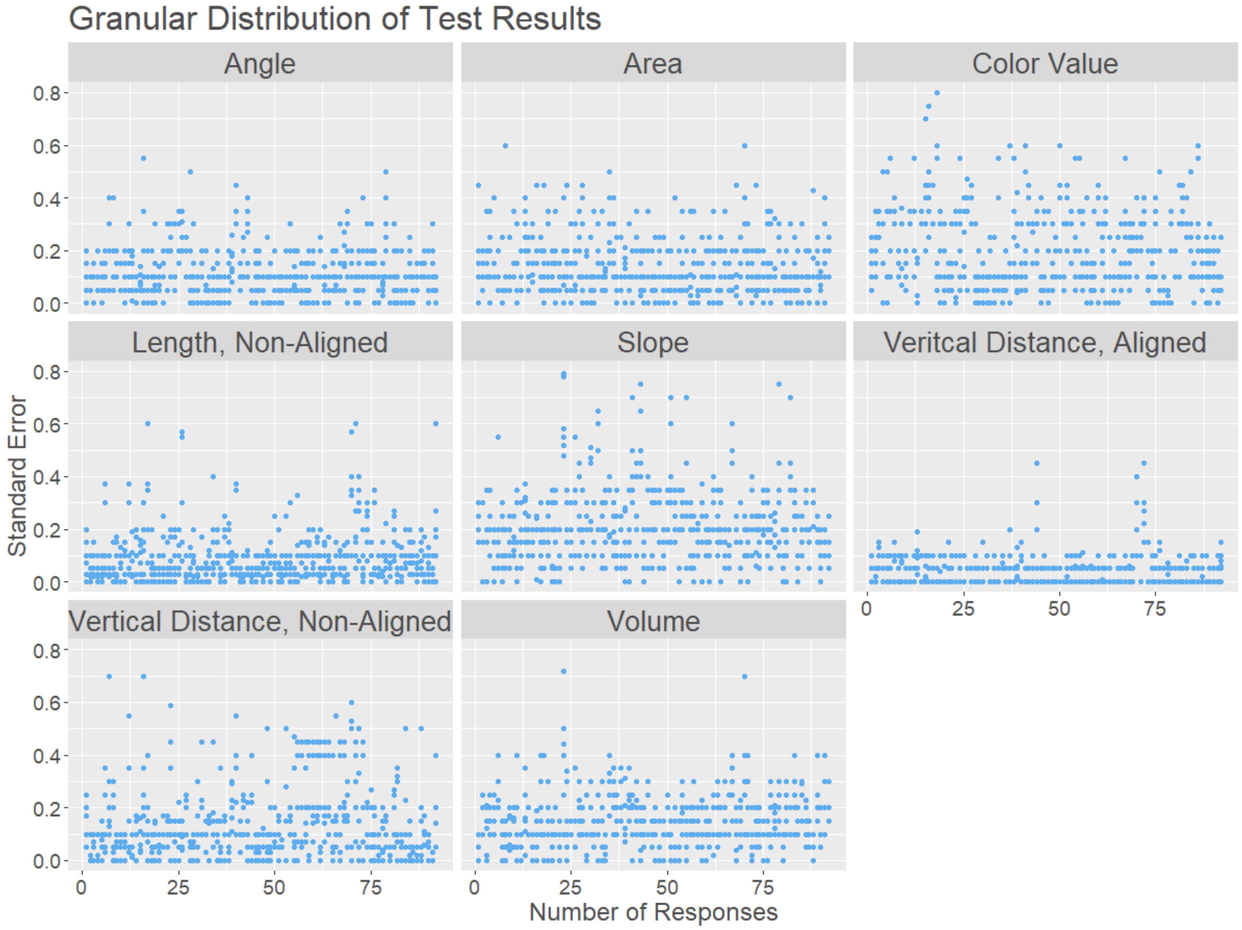
The graph above reveals that people most over-estimated Volume, although it was tightly grouped, so generally it was over-estimated to a common degree. However, Color Value shows under-estimation with a long and thin distribution and an accurate estimate well outside its box area[1]. This indicates that this pathway is not only inaccurate but perceived inconsistently by users. Area and angle have a

[1] The Box area is the black box within the violin plot that represents the 25 to 75 percentile range of values.

similarly wide distribution, but they are clustered much closer to a standard error of 0, so, while the perception is inconsistent, the perceived value is relatively near the actual. Aligned vertical distance has the smallest range and box-area, indicating it is the easiest to decode. Vertical distance unaligned and length are close- while length is farther from the actual results, it is more consistently estimated around a central value. Unaligned vertical distance has a wider range of estimates, but the actual result is near the center of the range.

B. Analyze Distributions Using Fine Detail Graphs

Corresponding Lines in Code File: 82 to 115



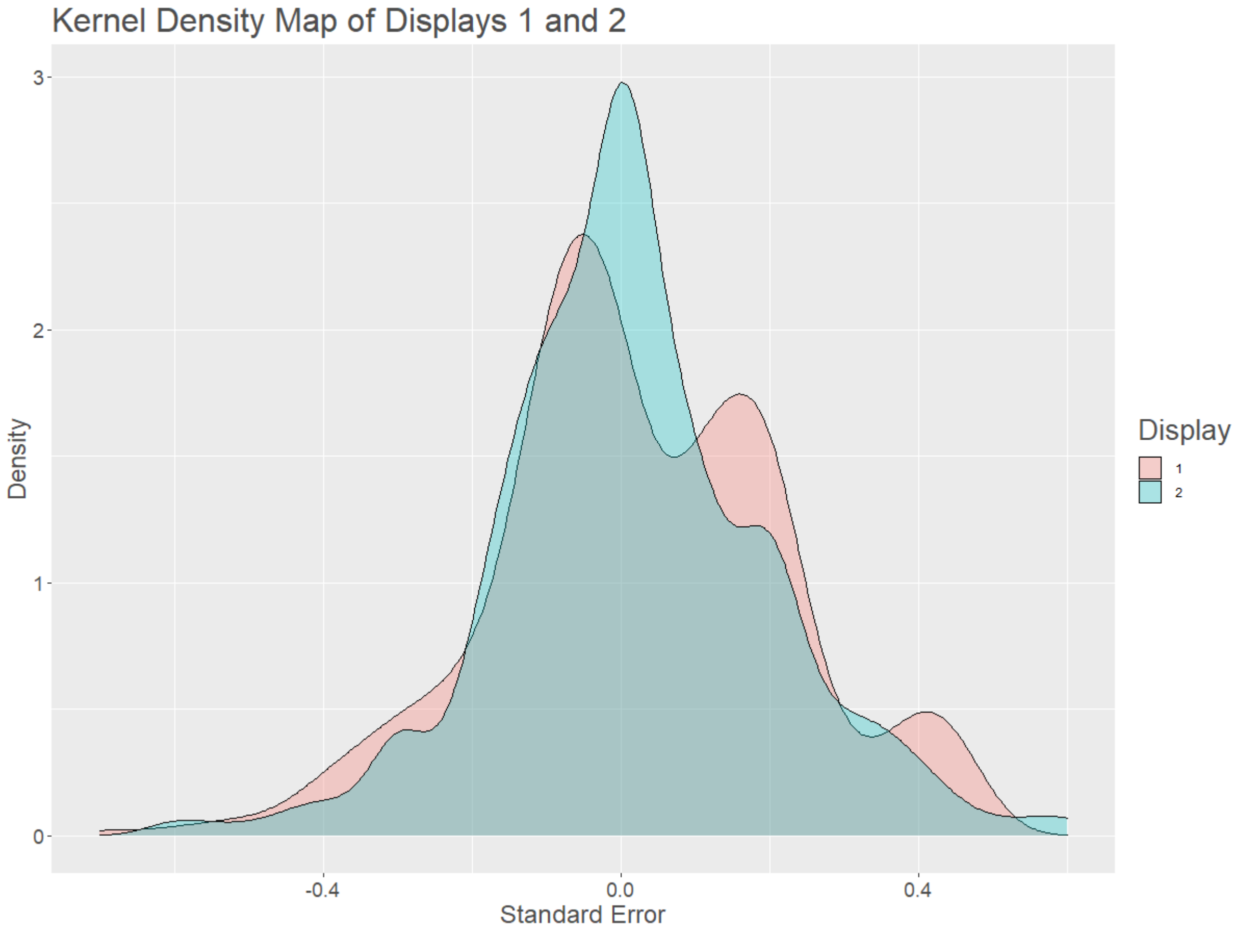
A univariate scatterplot, plotting absolute error for each respondent by test was used to analyze distributions. A chart with perfect estimates would have a single blue line along the x-axis representing errors of 0 for each user, a normally distributed chart will have heavy density near the bottom that thins out as the standard error increases, and a poor chart may have few or no values along the 0 value with many scattered throughout or concentrated near the top, meaning the pathway was interpreted to be significantly different than the actuals.

The highest concentration of values near the bottom is the “Vertical Distance, Aligned” test, followed by the Length, non-aligned”, then Vertical Distance, Non-Aligned. This is largely consistent with expectations set by Cleveland & McGill, whose research found that aligned scales are more accurate than non-aligned scales. There may be a slight deviation in their predictions, as they suggested that distance/unaligned (position on unaligned scale) should be relatively more accurate than length, however, the information around their experiments did not indicate whether it used a common scale when measuring accuracy of length, so this may be totally consistent.

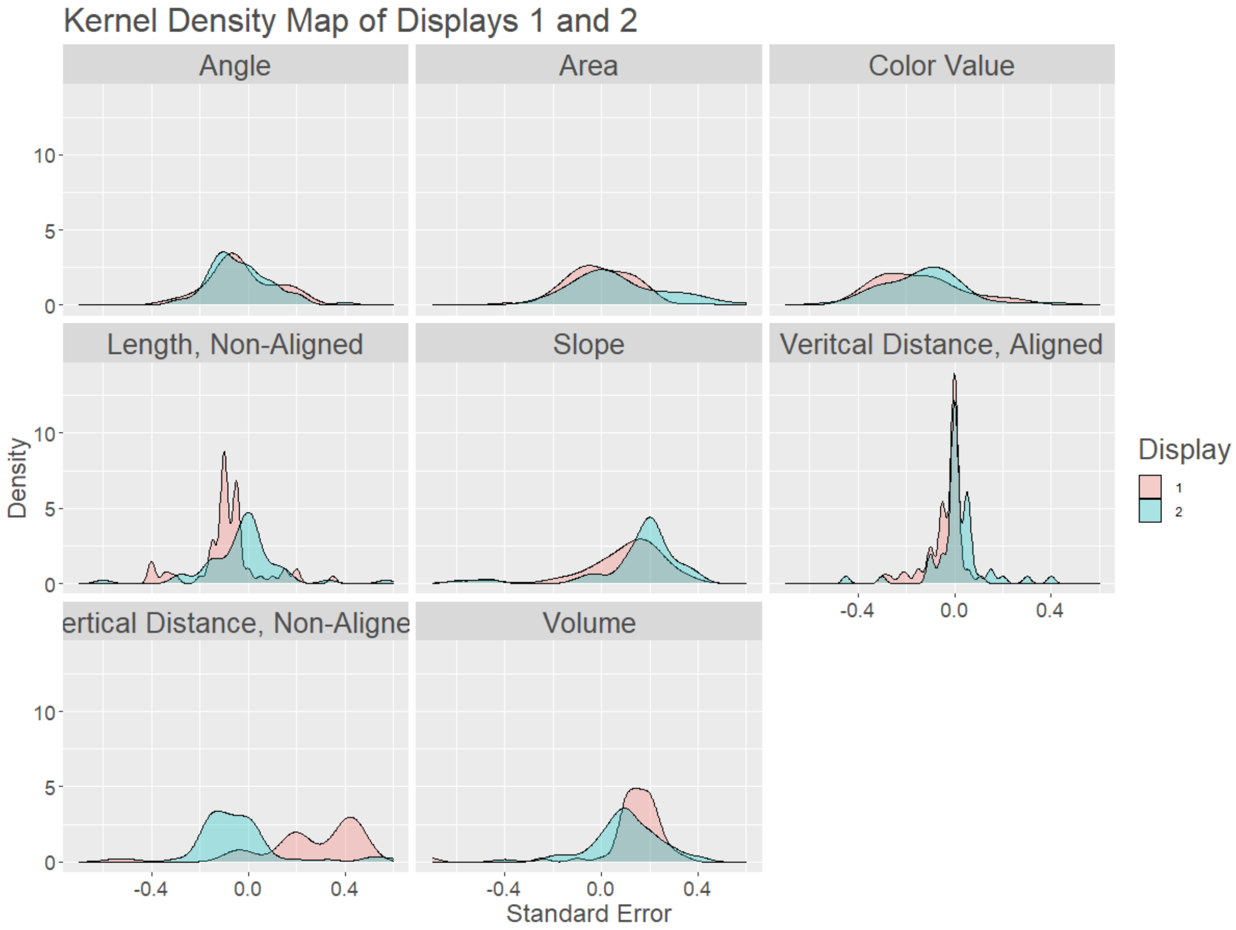
Cleveland & McGill also showed that color would be one of the worst methods of decoding specific values, and these results were consistent with this assertion. Color shows the highest deviations, followed by Slope, Area, and Volume. Area and Volume are too close to determine which has a tighter distribution in the graph.

C. Subjects 56-73 Response Pattern Subpopulation Analysis

Corresponding Lines in Code File: 119 to 190



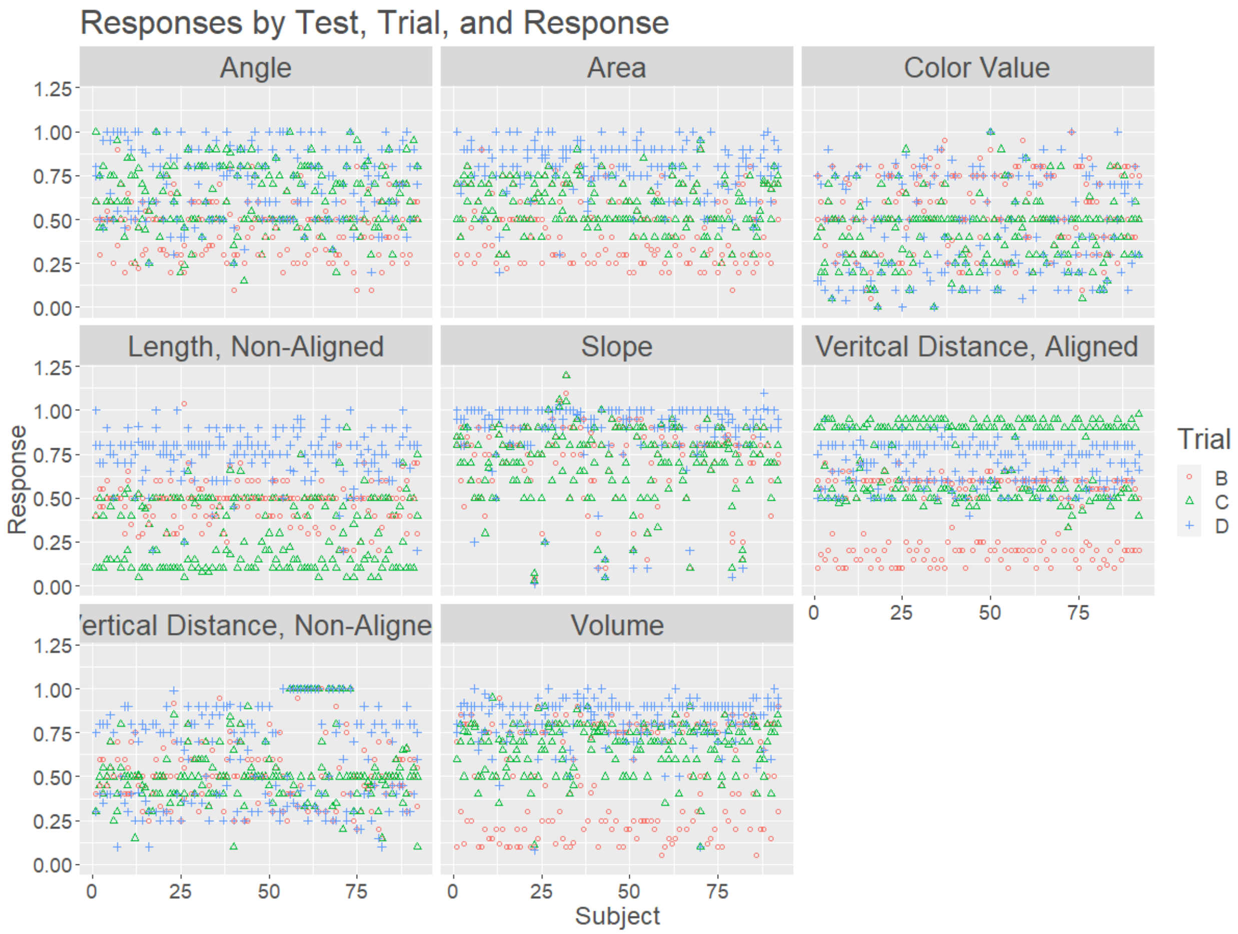
Overall, the participants in this study saw display 1 (red) before display 2 (blue). It is evident in the density map that display 2 is more "peaked" and consists of a single peak, unlike display 1 which has multiple peaks. This indicates that responses in display 2 were more consistent and given that the peak is centered on 0 instead of *around* 0 (like display 1), the responses were more accurate. It appears that multiple exposures to the same tests makes subjects better are judging the values.



Looking at the individual tests, we can see that most of them yield more accurate results in the second pass, with some exceptions. With Slope, Angle, and Area the results become slightly worse. With Slope, the cluster gets tighter, which means interpretation is coalescing, however it is coalescing around a bad value indicating a particularly poor channel choice. It's notable that "Vertical Distance, Aligned" didn't become more accurate, but rather the accuracy was mostly the same with a shift in skew from under-estimation to over-estimation.

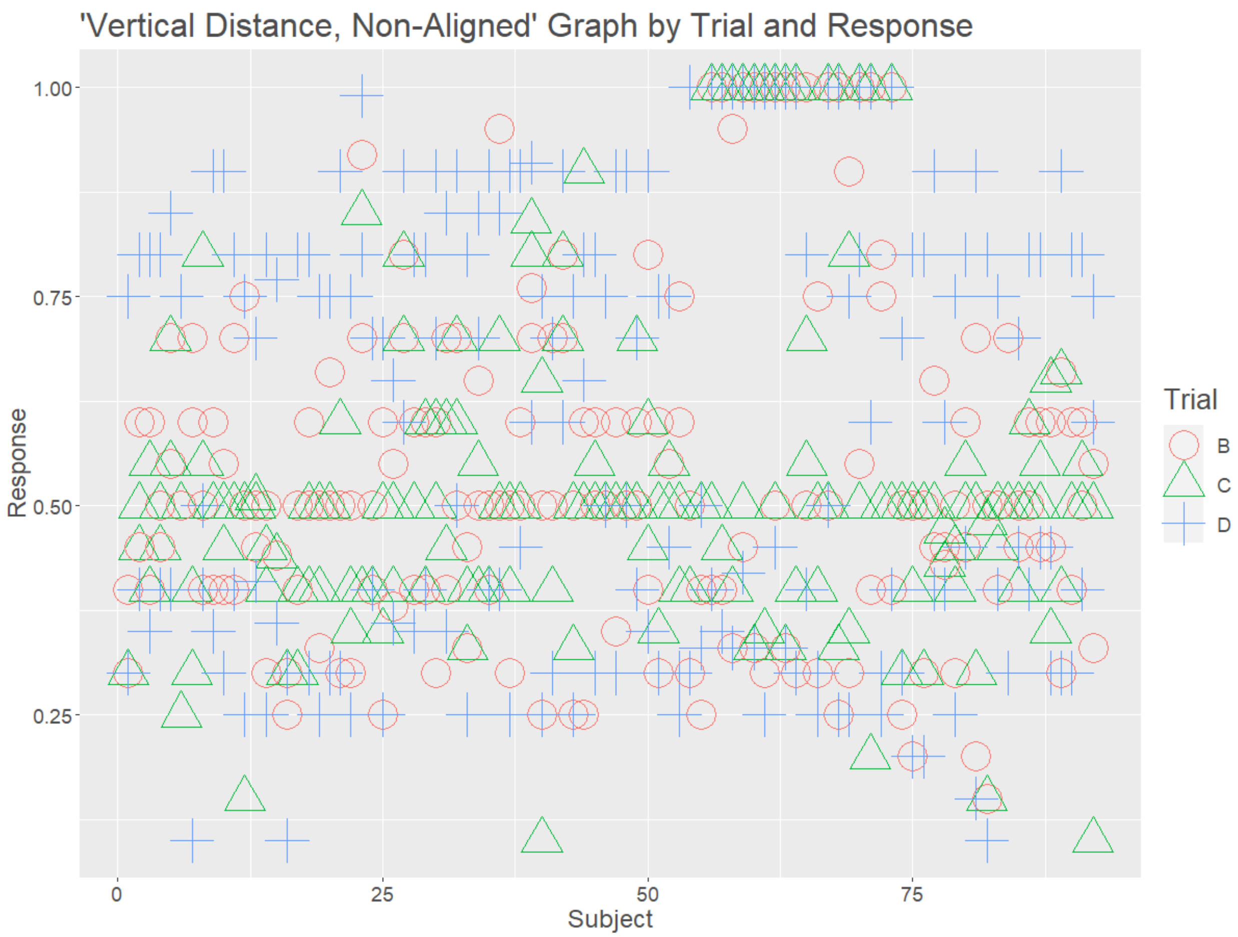
D. Anomalous Data

Corresponding Lines in Code File: 196 to 323

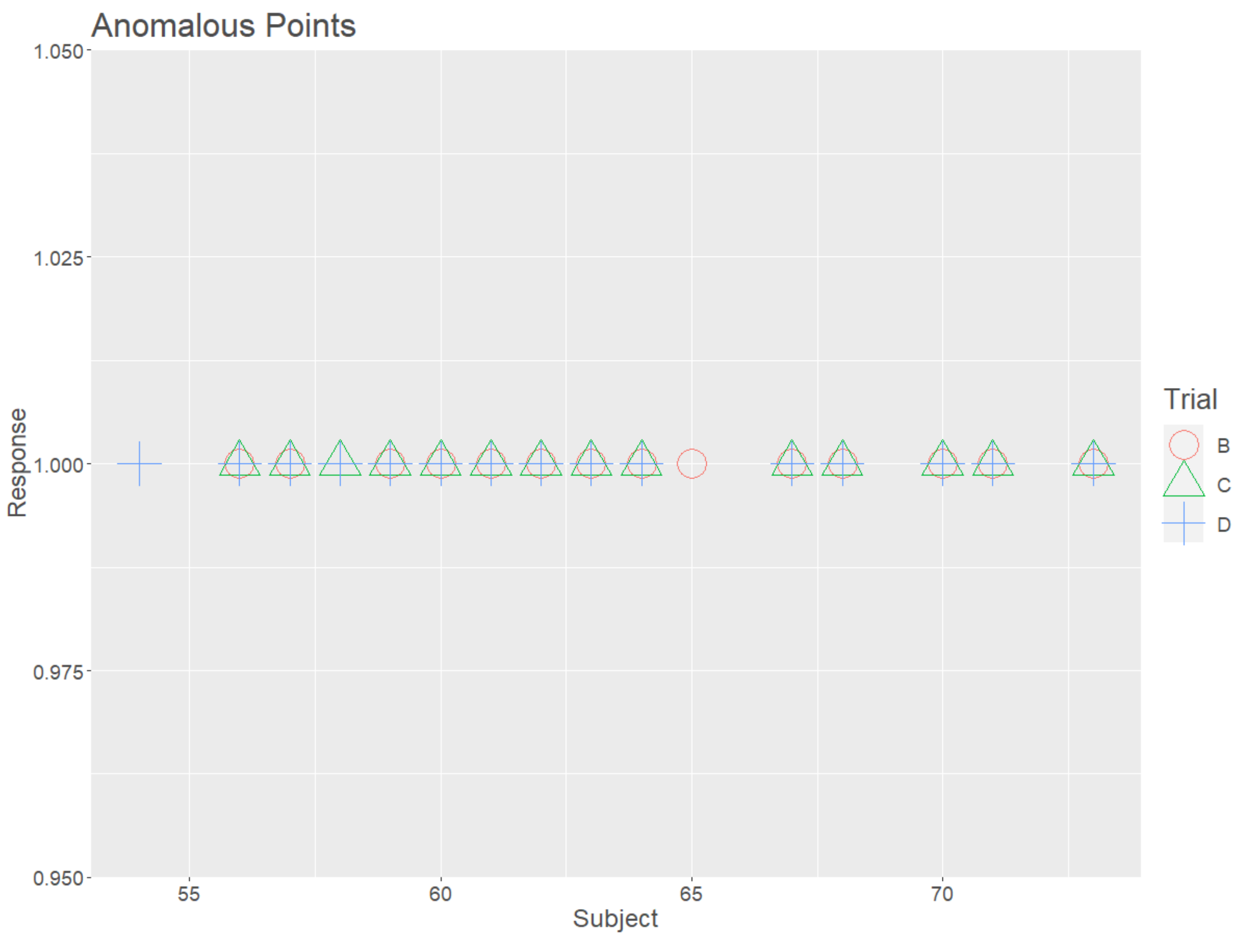


Identify an anomaly in the data can be done using visualization techniques. Detecting problems in the data will help create more meaningful visualizations and more powerful models.

To find the anomaly a visualization was created that included the important axis: The individual trial, the respondent, and their response, segmented by the Test itself which allowed the visualization of four key variables. From the chart above we can quickly identify a questionable pattern in "Vertical Distance, Non-Aligned" where the data seemingly stops following the rest of the pattern. The next step is to analyze this graph itself.



Closer inspection shows multiple subjects where all trials were the same value, and the value was far from the center of the distribution. Not only does it appear as a cluster of points, it also appears as a *lack* of data points in the space below it. A third and final graph is generated to confirm the observations.



The plot shows that all trials are present for most of the subjects. It appears to start with Trial D for number 54 and continue through number 73. Note: This graph only showed responses == 1.

Problem 2

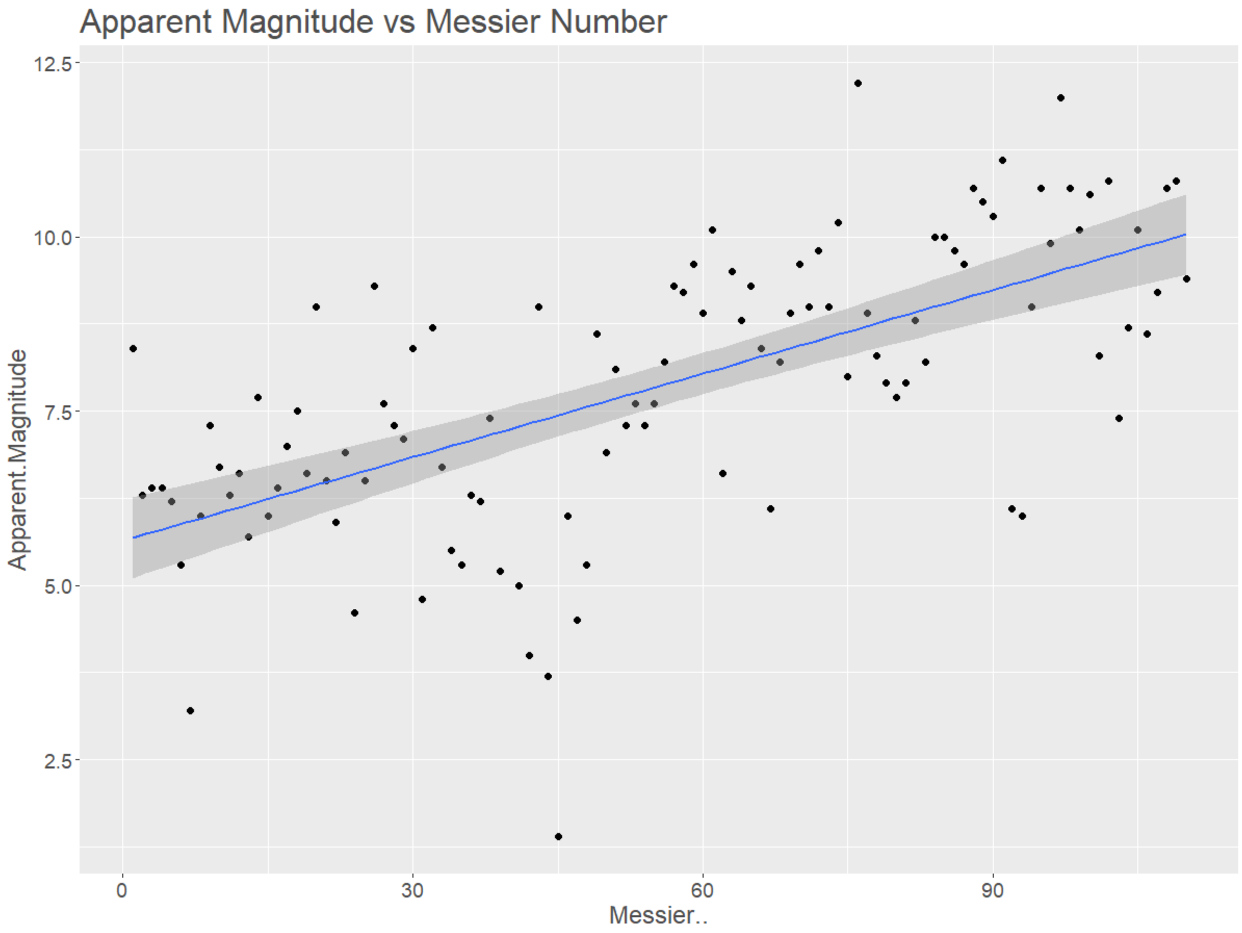
Problem 2 explores the Messier objects data set.

Corresponding Lines in Code File: 330 to 528

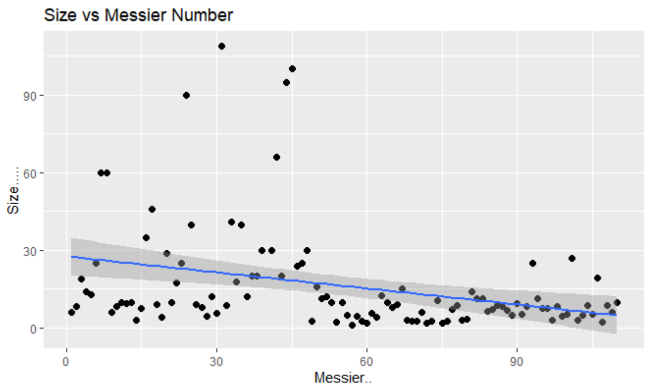
Note: One of the major challenges of this problem was the lack of context. Understanding the meaning behind the variables and how they work together help create more meaningful analysis, especially for counterintuitive variables (like magnitude where smaller is brighter).

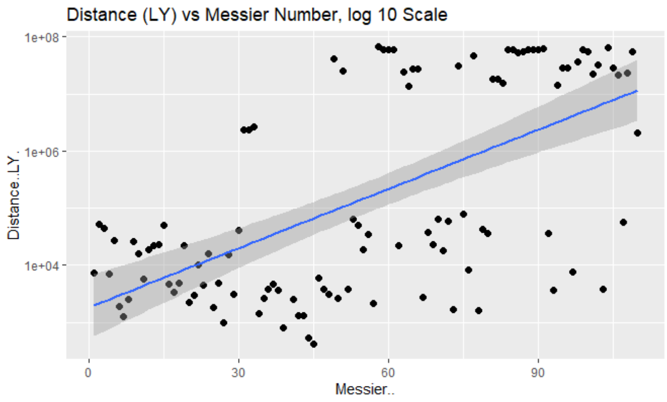
A. Analyze the *Messier Number* in the Context of the Data Set

Corresponding Lines in Code File: 354 to 409



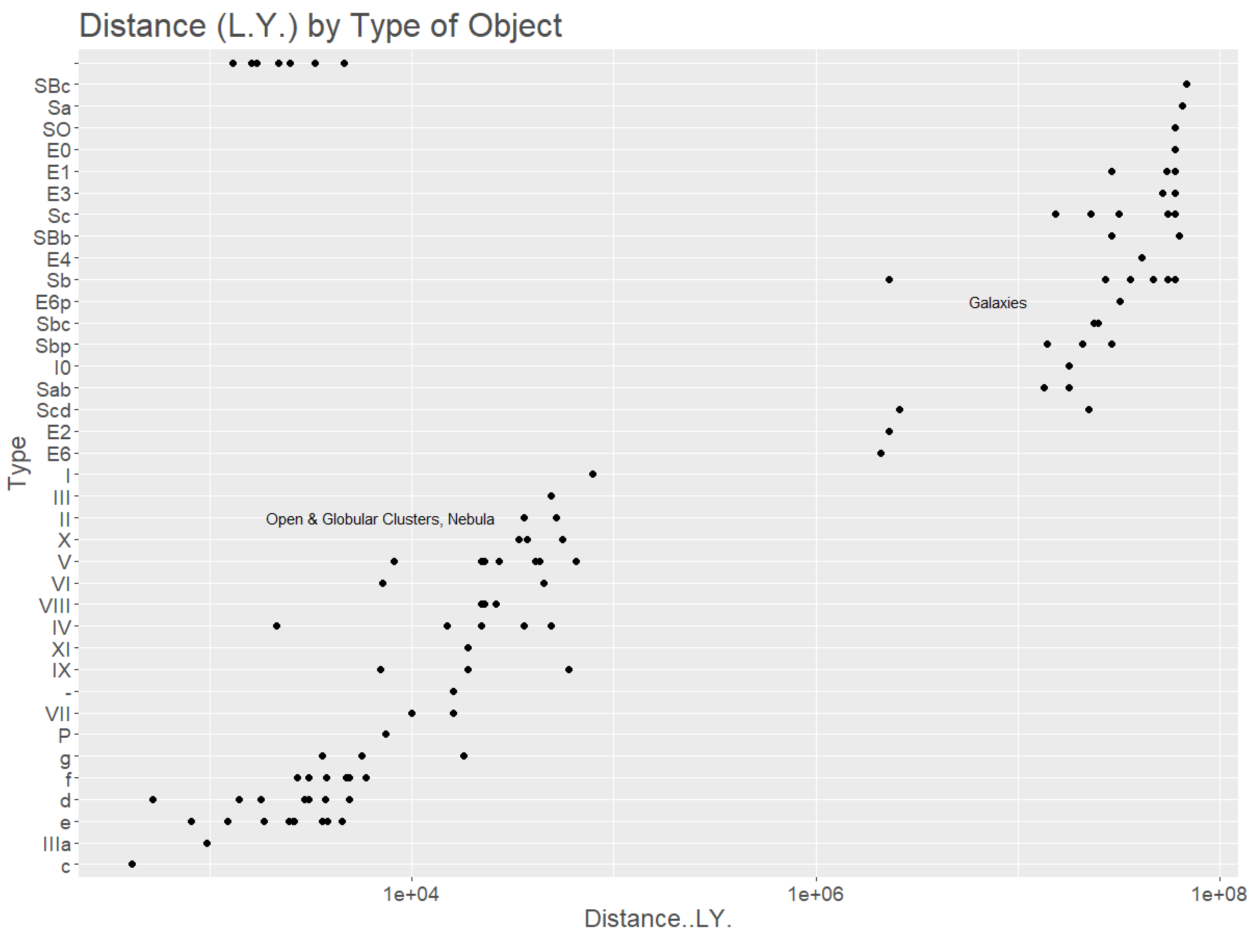
There are several factors that have a positive correlation with their Messier number. Aside from year, which is highly correlated but does not tell us anything, Distance in LY, Size, and Apparent magnitude show some form of correlation with their Messier number. The correlations follow the pattern: more difficult items have a higher Messier number (and discovery year). This is especially true with Apparent Magnitude, where higher magnitudes correspond to less luminous objects. As technology developed, higher magnitude (dimmer) objects were observable. Similarly, higher numbers tend to correspond to smaller and more distance objects.





B. Compare Distributions of Distance to Objects of Each Kind

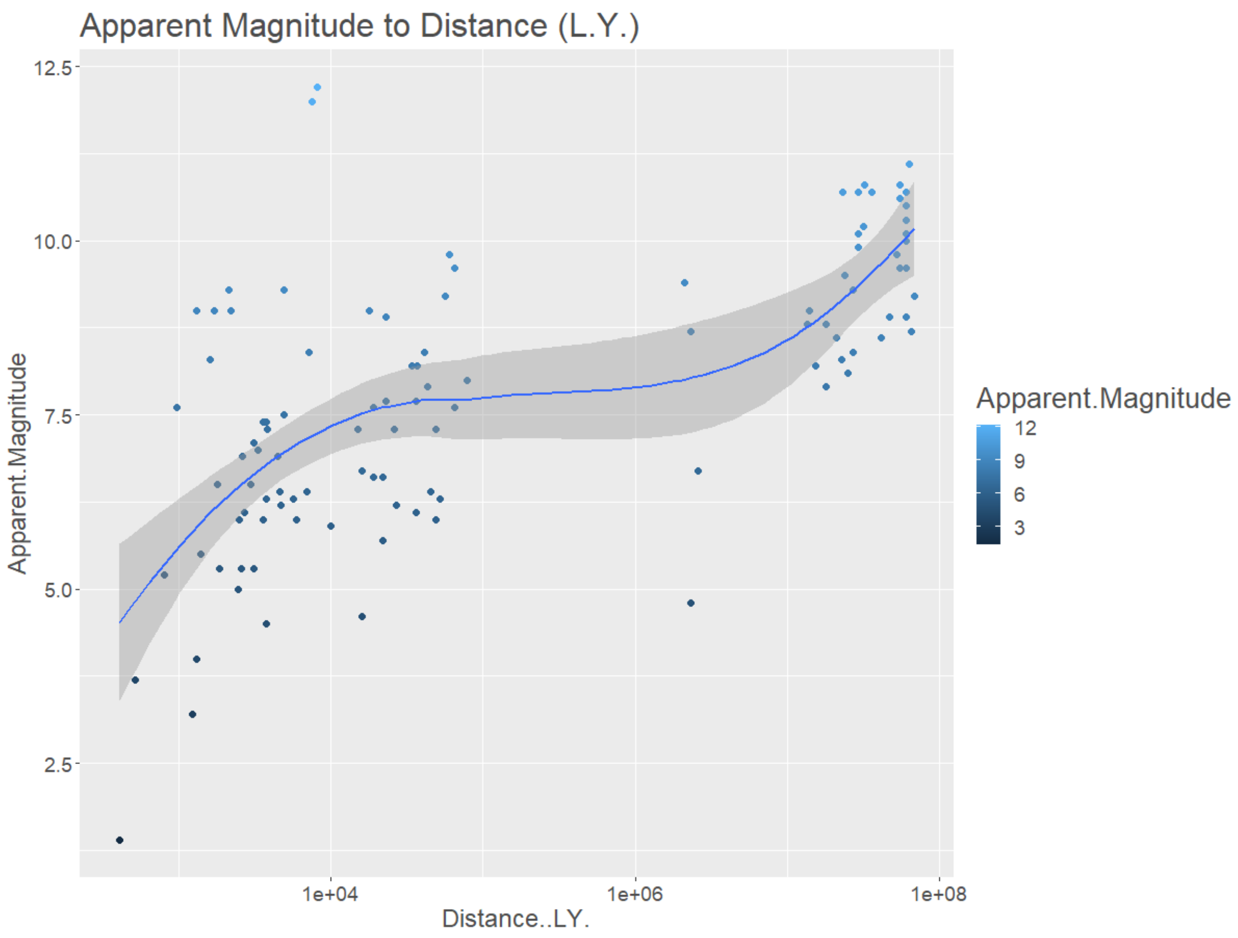
Corresponding Lines in Code File: 415 to 455



This distribution clearly shows multiple groups based on type. Types I to C correspond to open and globular clusters as well as nebulae and types SBc to E6 correspond to galaxies.

C. Distance to Apparent Magnitude

Corresponding Lines in Code File: 458 to 491

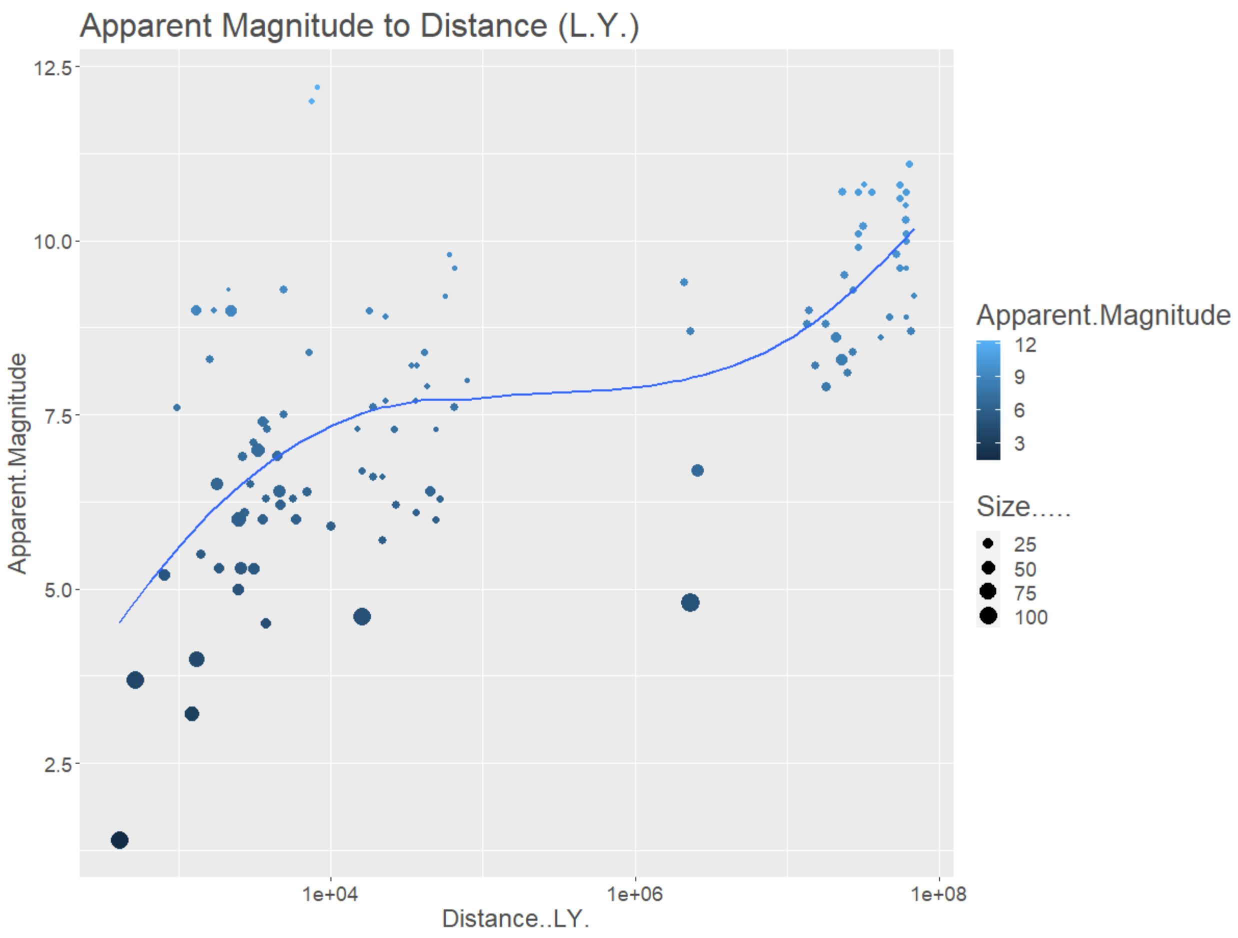


There is a strong correlation between closer objects and brightness. Apparent Magnitude, A.K.A. Magnitude, is the brightness of an astronomical object as it appears on earth. The implication is that farther objects have more interference on their light during its travel to earth.

In this chart, dimmer objects have less contrast with the background while lighter objects have more contrast.

D. Graph Enhancement and Recommendations

Corresponding Lines in Code File: 494 to 528



The addition of size makes some parts easy to miss (the two outliers near (1e04, 12.3)). Additionally, it is difficult to tell the difference between the 75 and 100 circles. There are several enhancements: one option is to add bins to the circles, which would allow for a more granular, although this still would not allow for decoding specific values. If this graph were interactive, it may be helpful to create a 3-d chart where the additional axis has a line indicating its size. It should be interactive so the graph can be 'moved' and data that is obscured can be viewed.

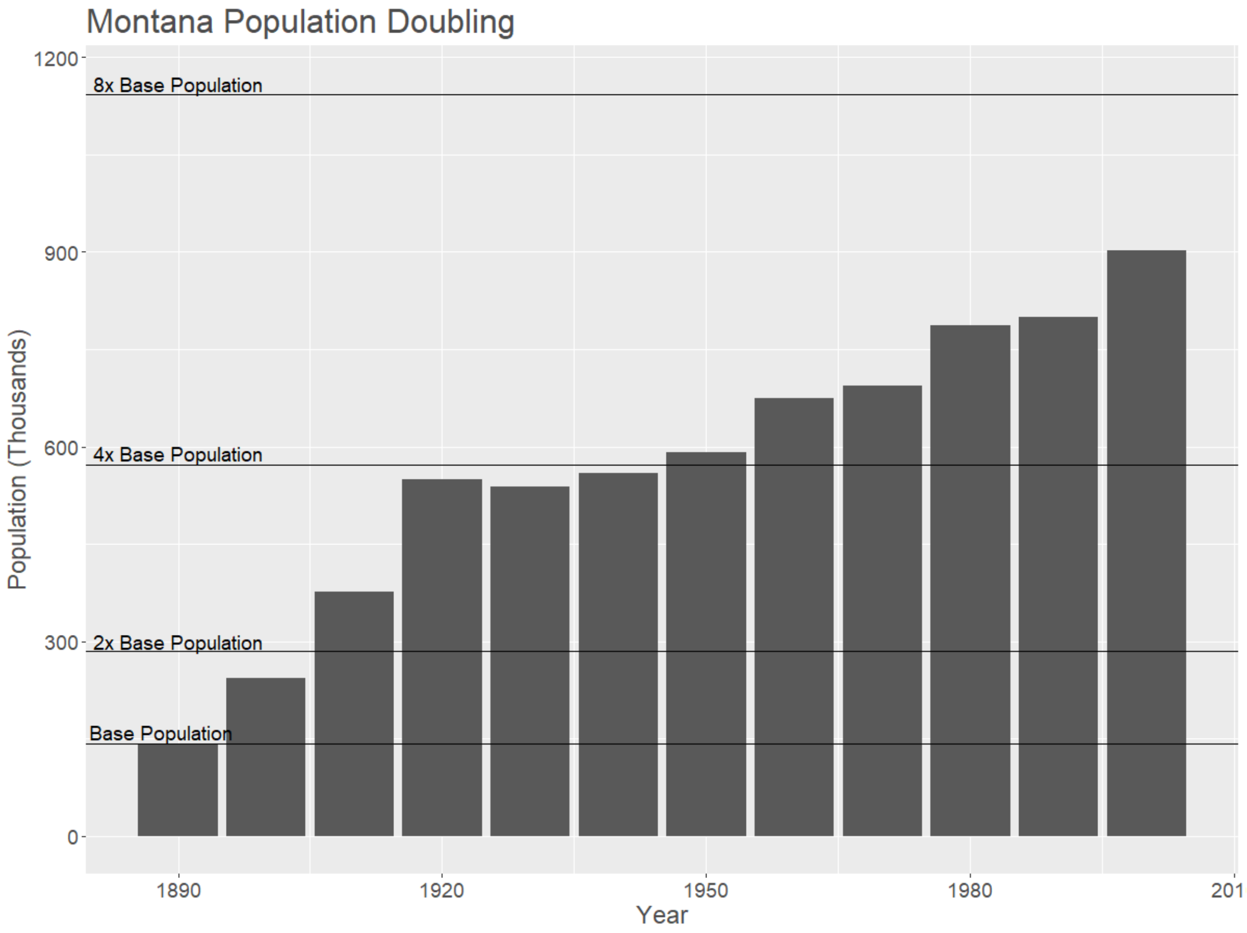
Problem 3

Problem 3 explores the Montana Population data set.

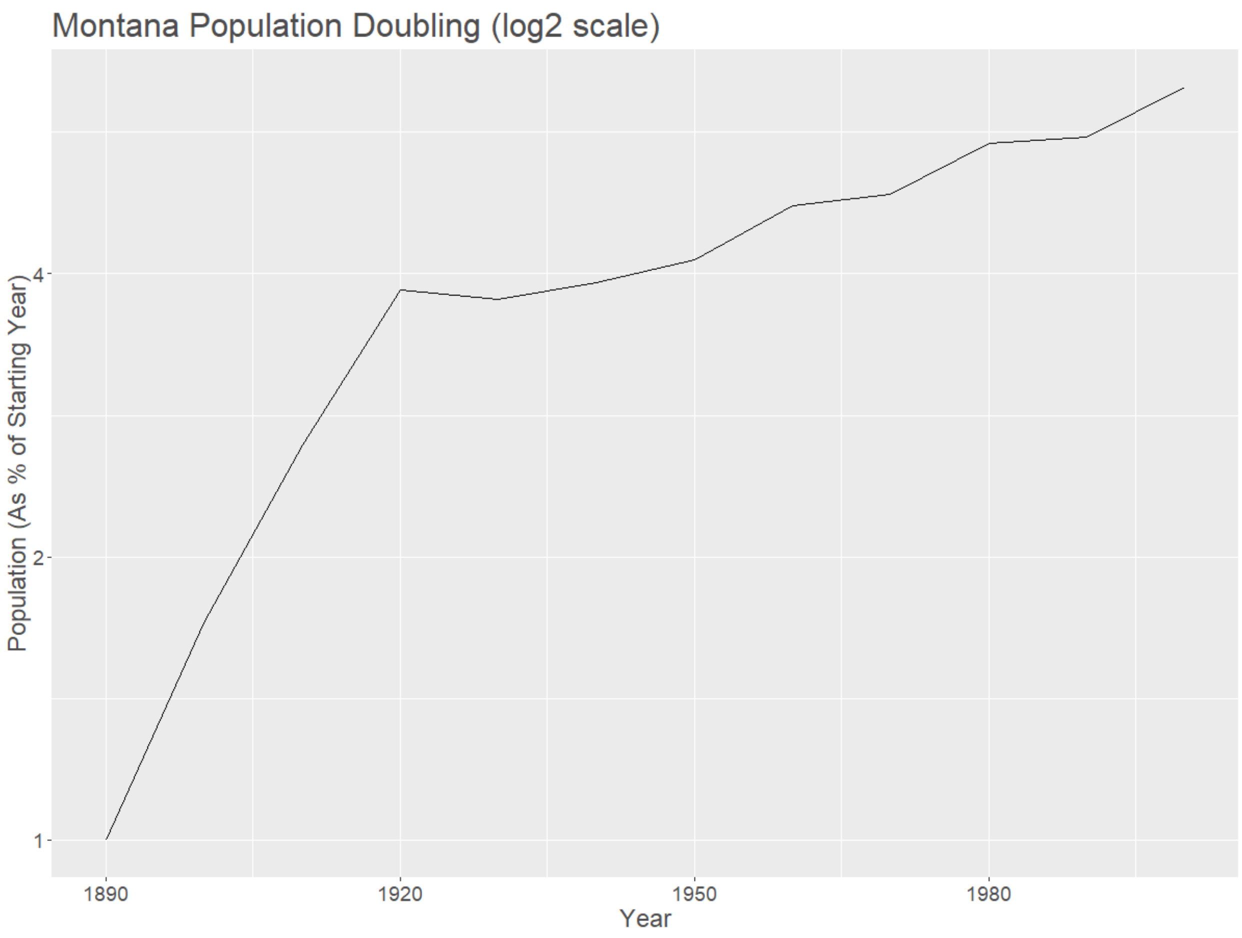
Corresponding Lines in Code File: 535 to 684

A. Times Population Doubled since 1890

Corresponding Lines in Code File: 577to 646

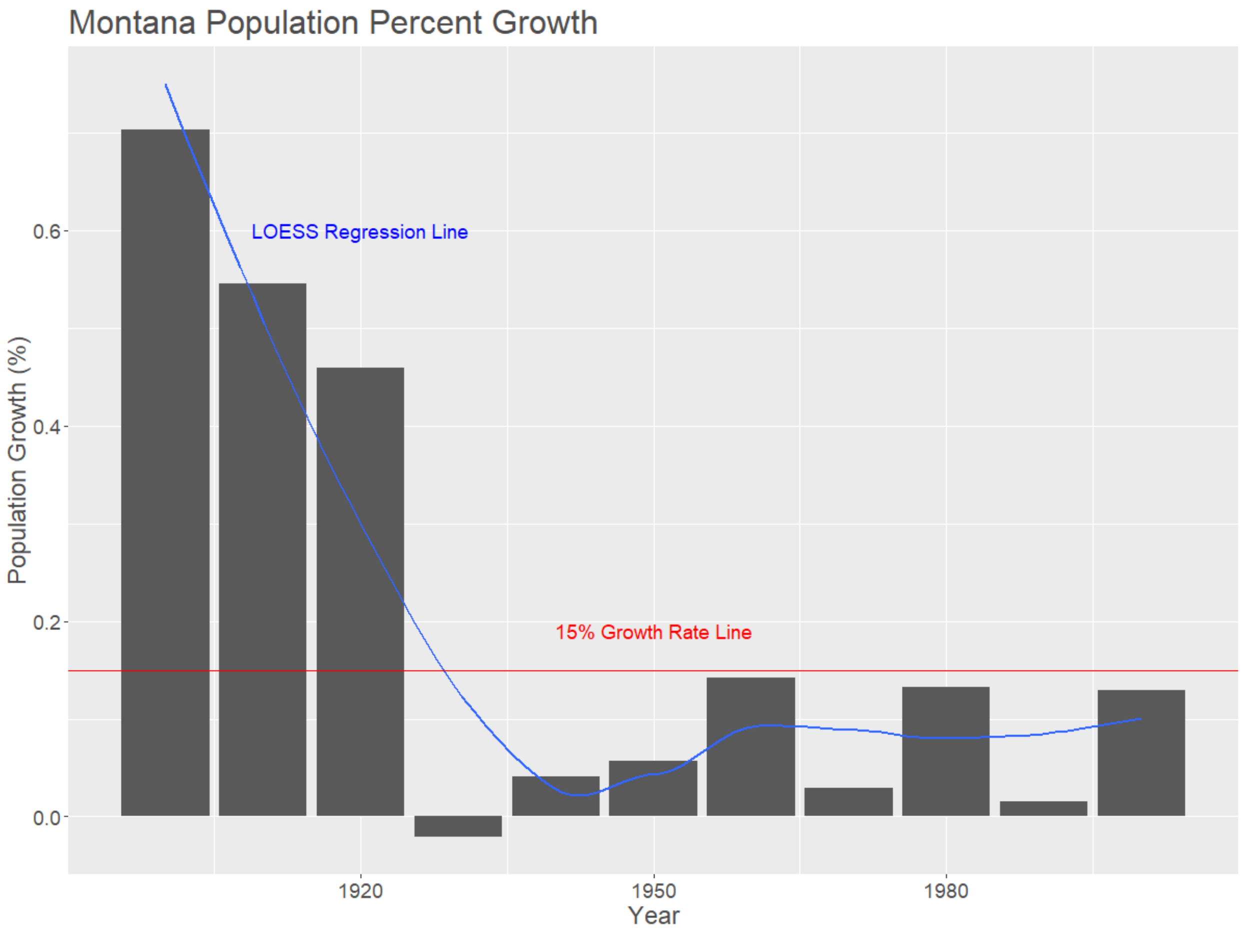


The population has doubled twice. The first visualization does not need log scaling; however a second visualization is provided below in case it's necessary.



B. Change in Percentage Change- Greatest Increase in %

Corresponding Lines in Code File: 648 to 684



As shown by the trendline above (LOESS), the percentage rate of change has decreased over the years. The first 3 decades[2] (after start) showed intense growth (> 40%) which has since decreased significantly, alternating over the decades between growth rates as low as -5% (reduction) up to slightly less than 15% in a clear bimodal distribution.

C. Years w/ growth > 15%

Note: This uses the chart from problem 3B.

The population percentage increase was greater than 15% in 1900, 1910, and 1920.

[2] Since this is a growth rate there is no value for the starting year (1890).

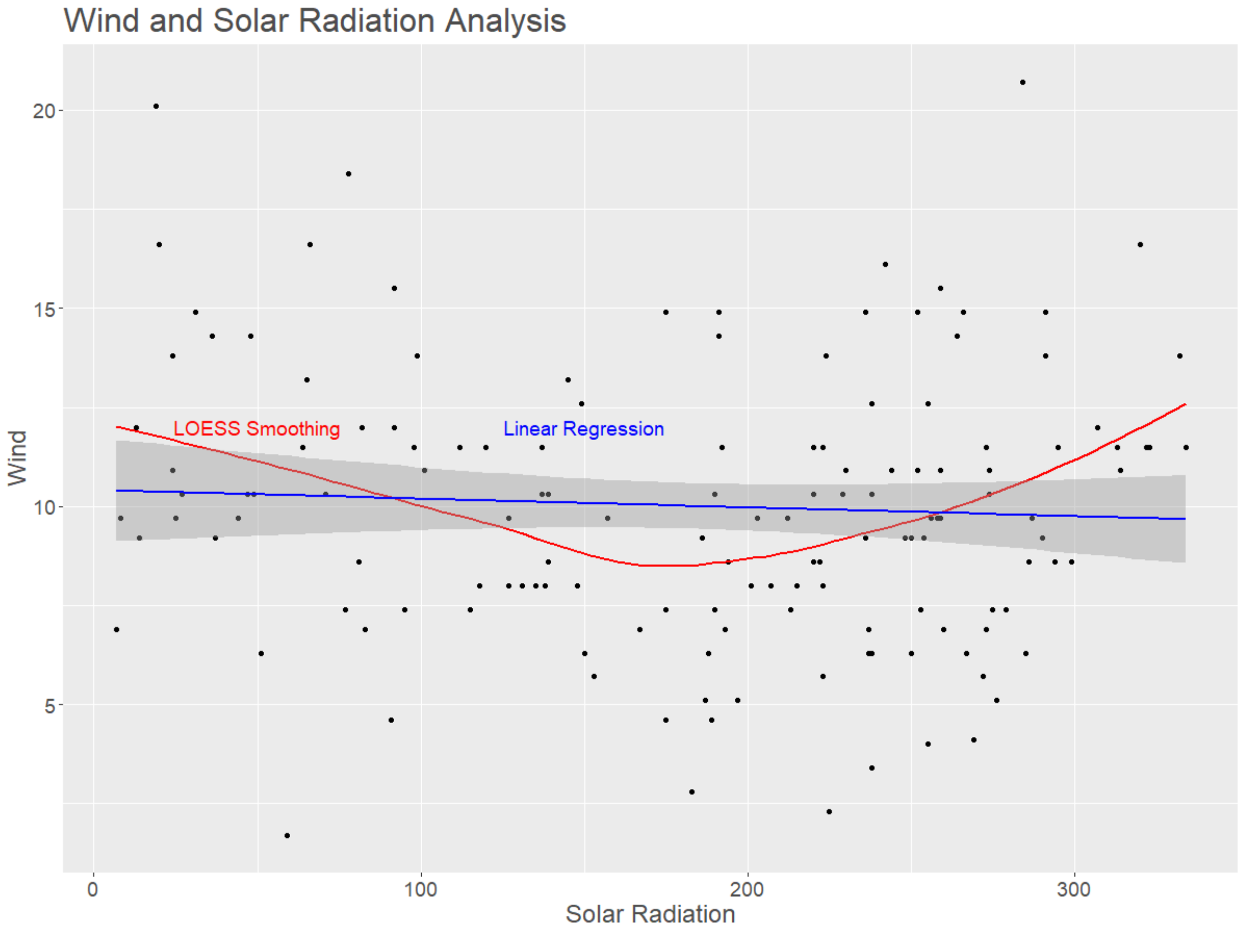
Problem 4

Problem 4 explores the Air Quality data set.

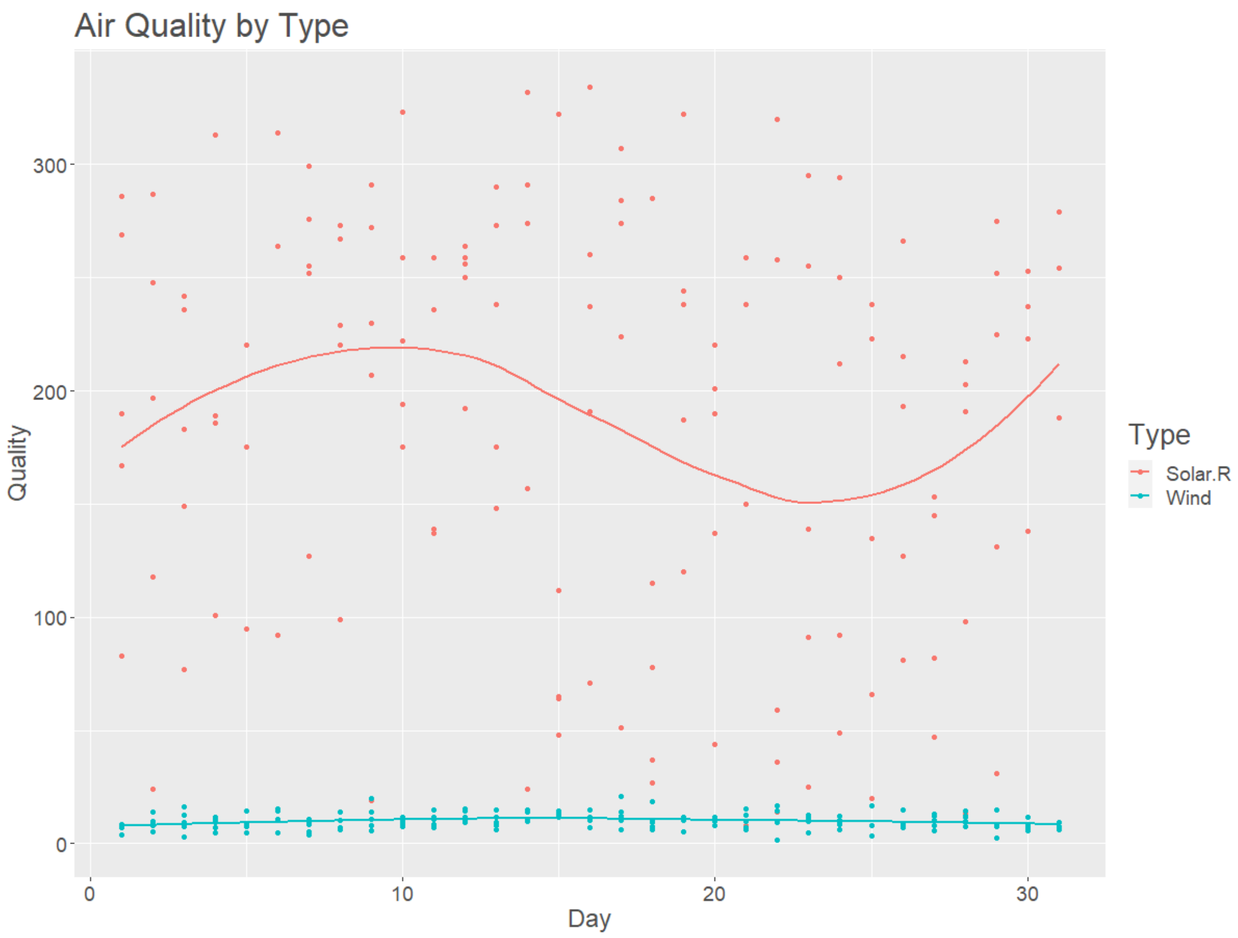
Corresponding Lines in Code File: 691 to 905

A. Relationship between Wind and Solar

Corresponding Lines in Code File: 719 to 749



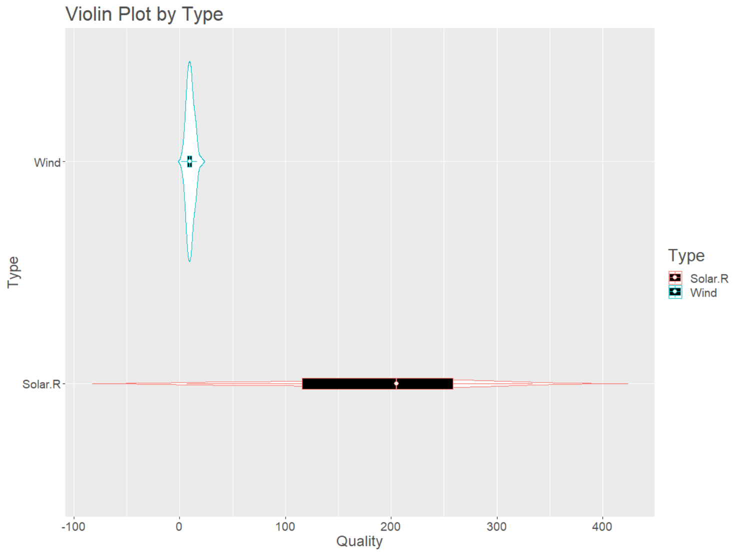
Two trend lines have been added. One can see that there is generally no relationship between the two.

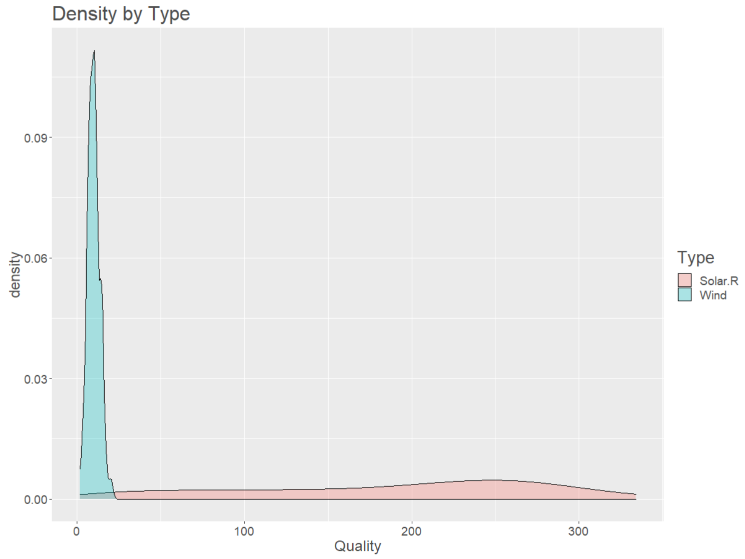


Looking at solar and wind separately shows a large magnitude difference. Additionally, the LOESS smoothing shows a pattern for each throughout the month: Solar waxes at the beginning and wanes near the end while wind is relatively stable with a slight increase mid-month.

B. Wind & Solar Density Plots

Corresponding Lines in Code File: 752 to 812

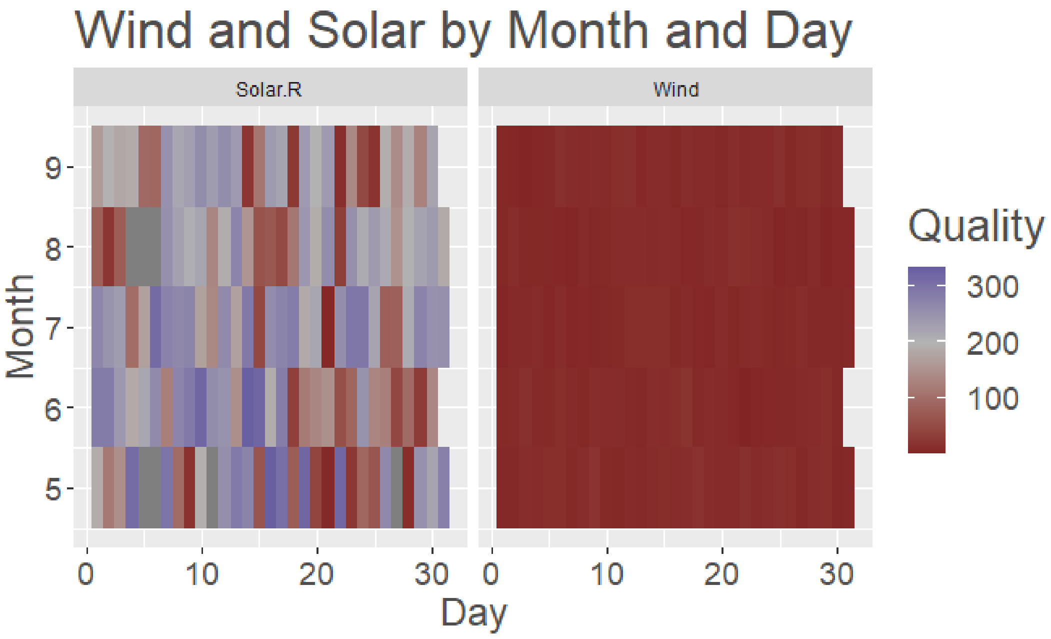


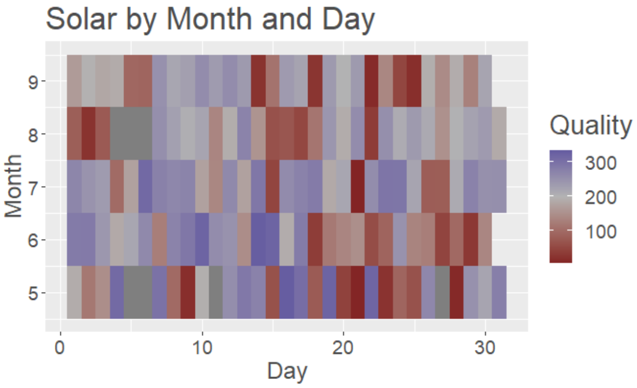


The above density plots show the same distribution as from problem 4A – Solar has a wide range while wind has a very narrow range.

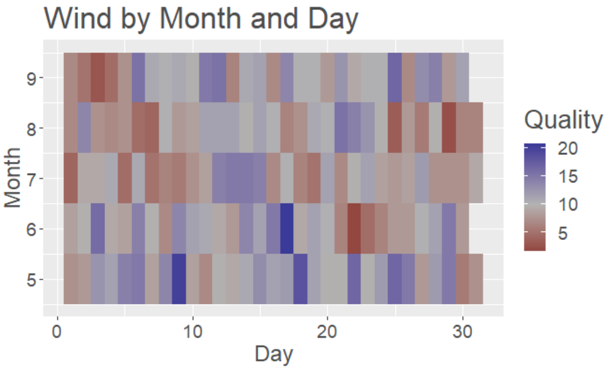
C. Distributions in context of all variables

Corresponding Lines in Code File: 814 to 905





Solar by Month and Day



Wind by Month and Day