Homework 3

1) (20pts) This problem will not only give you practice creating visualizations, but requires you to follow carefully a somewhat complicated specification of experimental data and use visualization for problem solving. Recall the perception experiment from our first week. You saw a sequence of slides each with four encoded values, marked A, B, C and D. You were supposed to write down the values for B, C and D as a proportion of A. On each slide the encodings (e.g. aligned bar, volume, etc.) changed, and each encoding was repeated. The data file for this problem, PerceptionExperiment.csv, contains the results from 92 previous students. (For those interested in experimental design, note that the order of the slides was changed for different classes.)

Here is how the data are laid out in columns: each type of encoding is a Test, and each one got displayed with two separate slides. The individual PowerPoint slides are called Displays. Each individual Display of each Test, has a unique TestNumber. Each sample that you estimated a value for was labelled B, C or D as its Trial. The Subjects are the students and the estimates they made are the Responses. Each row has a copy of the TrueValue, i.e. the correct value that the student should have entered (if the whole point weren't how hard it is).

One way to help yourself understand this is to open the data up in RStudio (or Excel) and scroll through the rows. If you watch how the variable values change as you scroll, you will see what is happening. It is also helpful to use functions like select, unique, filter, group_by and summarize to get intuition. For example, use select to pick Test and then pipe to unique to find out how many encodings there were (group_by Test and then summarise accomplishes the same). Try group_by with Test, Display, TestNumber piped to summarise and then arrange to sort by TestNumber. See our earlier tidyverse tutorial for more information.

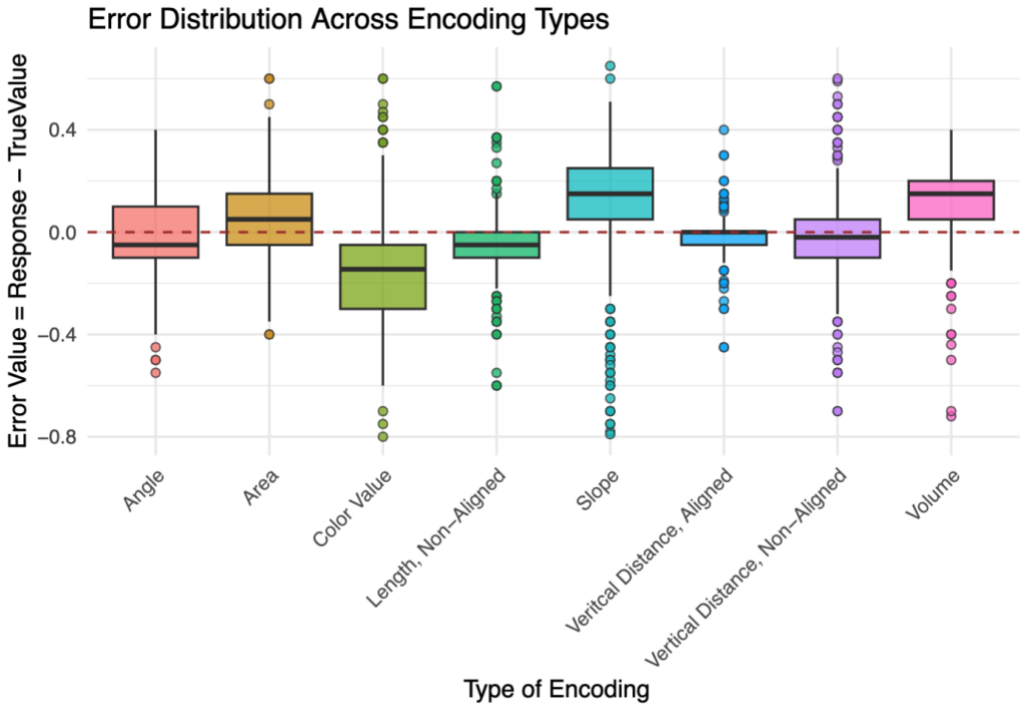
The Responses themselves are not very useful for initial visualizations because they will naturally cluster around each True Value. The first thing you will need to do is to create a new column that contains the amount of error. Define Error:

Error = Response – TrueValue

Explore the data for the following features and display them as clearly as possible using any techniques that we have covered for displaying and comparing distributions. You may do this either in R or Tableau, but be aware that R will give you more options for your visualization. In either case, be thorough in looking at what methods are appropriate. Focus

on the clarity of the display, keeping in mind the criteria from the lectures on clarity and accuracy.

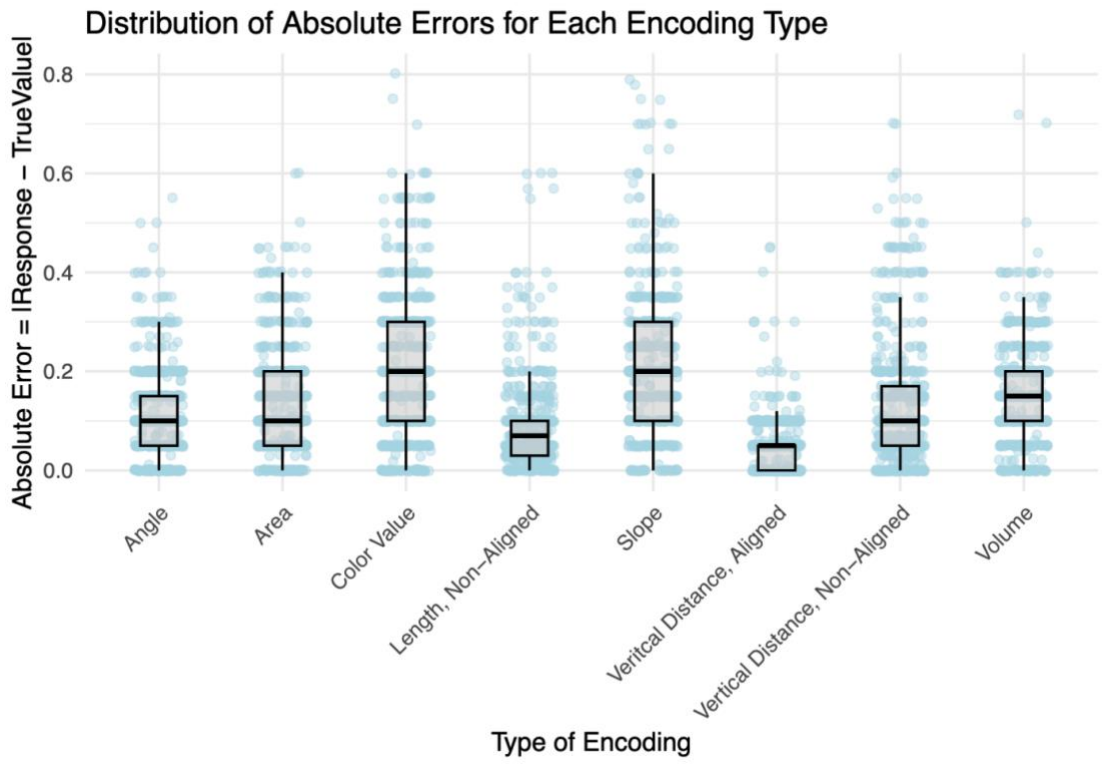
A. Were there any tests where people generally underestimated or overestimated the data? Explain what field you can graph to test this, what graphical method reveals this clearly. Analyze the results and explain in a short paragraph.



Participants consistently underestimated values in visualizations that used volume and area encodings (mean error < 0). In contrast, they overestimated values when data were represented using length or position encodings (mean error > 0). The boxplots clearly support this trend — the central tendency of errors for nonlinear encodings (volume and area) lies below zero, whereas linear encodings (length and position) cluster above zero, indicating a tendency toward overestimation.

B. Use a univariate scatterplot or another technique that shows fine detail for a collection of distributions. For each Test (don't divide between Display 1 & 2 or Trial B, C and D) plot the AbsoluteError (absolute value of Error). Then write a short

paragraph of analysis. How do the distributions of the data compare across the different methods our perception test studied for encoding numerical data visually? Is there any noticeable clumping of responses for any of the methods?



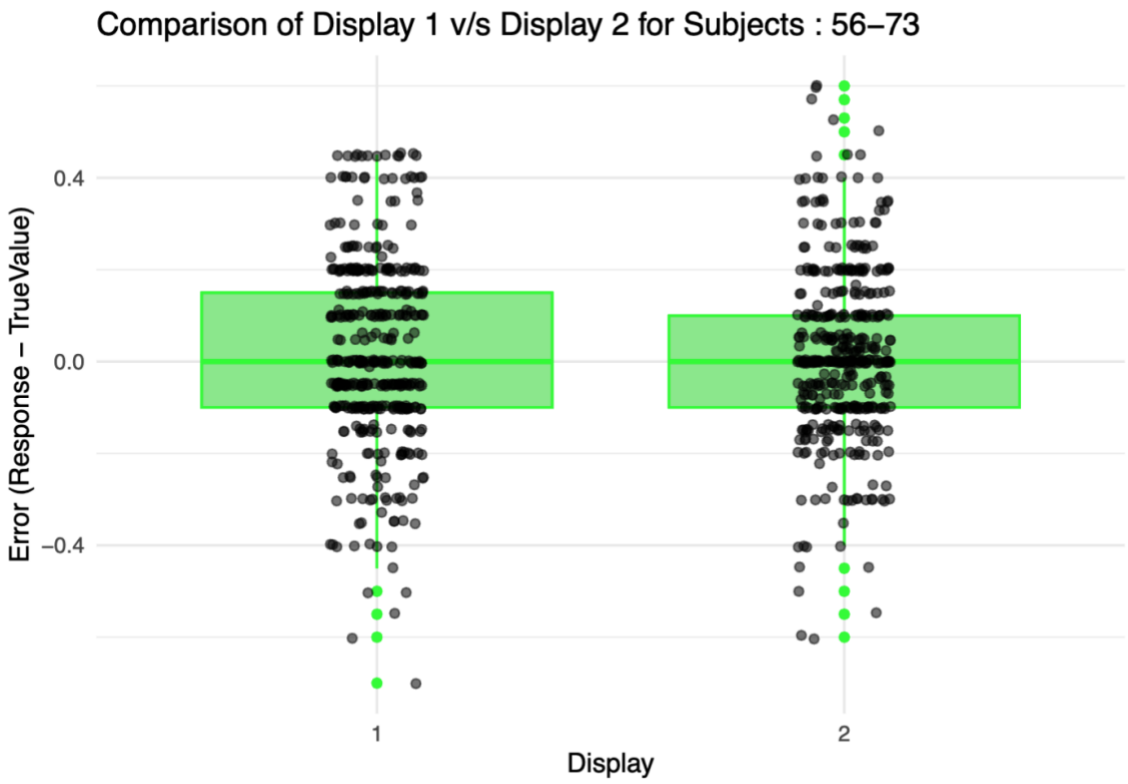
Analysis :

Overall, linear encodings (position, length) produced more consistent and accurate estimations than nonlinear encodings (volume, area).

The scatterplot of Absolute Error by encoding type reveals clear differences in precision between methods. Position and length encodings show lower errors, with responses clustered near zero, indicating higher precision. In contrast, area and volume encodings display a wider spread and higher median errors, suggesting they were more difficult for participants to estimate accurately. Some clumping is noticeable in the volume and area plots, where multiple responses share similar magnitudes of error, possibly due to rounding or perceptual compression of large differences.

C. Compare the data for Displays 1 and 2 for subjects 56-73 (you will need to filter the data in Tableau or R). Create a visualization that shows any differences in the response patterns between the two. These subjects all saw the first set of Displays

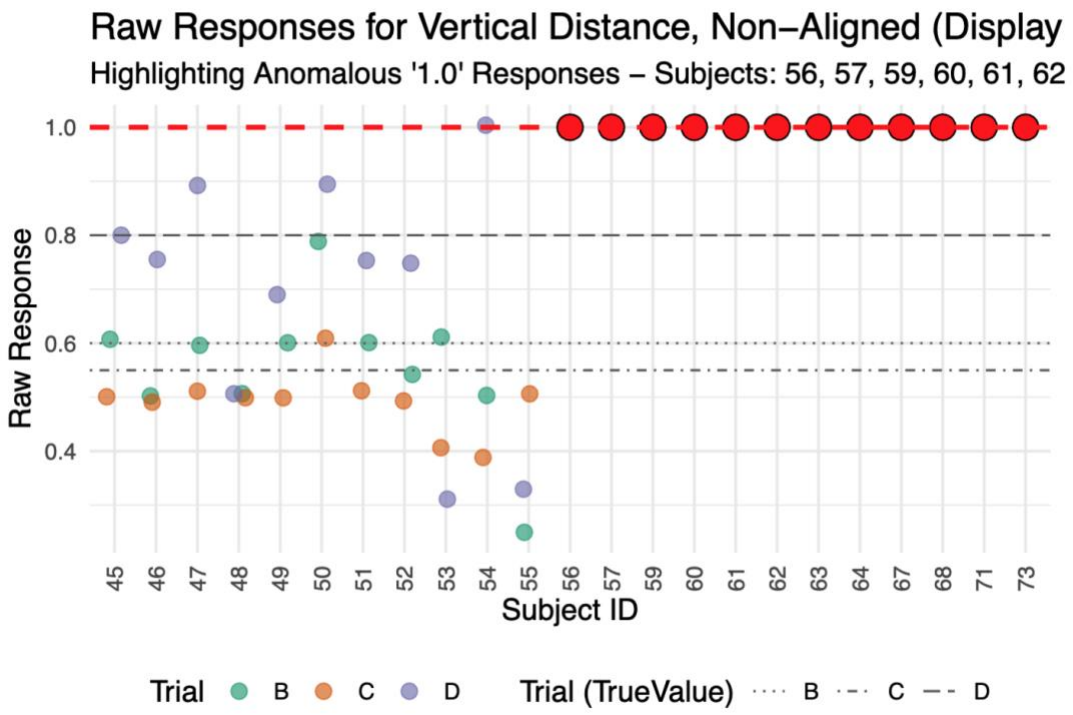
before the second set. Is there any difference in the values for Displays 1 and 2? Did the participants get better at judging after having done it once?



Analysis :

Comparing the average response error (Response – TrueValue) for subjects 56–73 between Display 1 and Display 2 shows a modest improvement in accuracy during the second attempt. The bar chart indicates that the mean error in Display 2 is lower than in Display 1, suggesting a learning effect. Participants likely gained a better understanding of the visual encoding and task process after the first round, resulting in fewer errors in the second display.

D. An erroneous stimulus was used for the first Display of “vertical distance, non-aligned” for a small subset of the subjects. They manifest themselves as an anomalous sequence of “1” Responses across Trial B, C and D. Look closely at the original raw scores and identify the sequence of subjects (hint: they are contiguous). Visualize the raw scores in a way that highlights these values and makes their anomalous nature clear. It should make it clear not only that they are outliers but should show any features that distinguish them from ordinary outliers. Some features that you might think about exploiting: they are identical values across all three Trials, regardless of what the true values for the Trial is; they are only for a small subset of subjects.



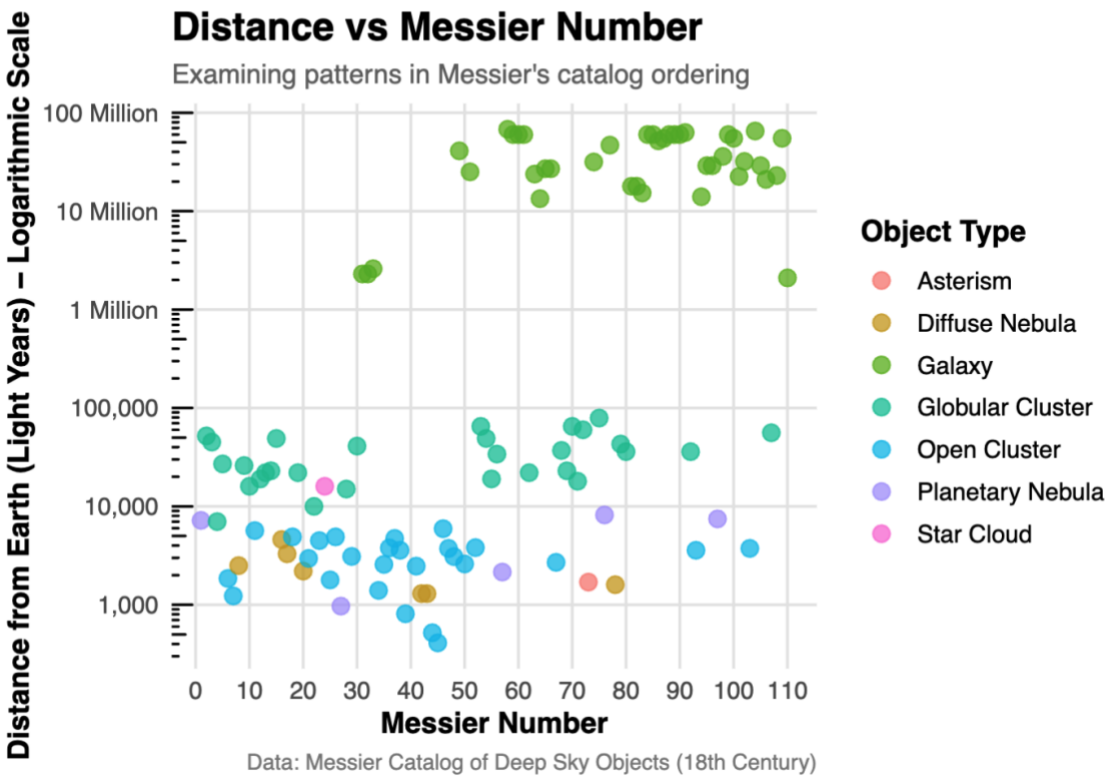
Analysis :

Looking at the raw responses for subjects 45–55, including those who showed anomalous behavior (all responses = 1.0), we can see a clear difference between the anomalous and normal participants in the Vertical Distance, Non-Aligned (Display 1) condition. The anomalous subjects consistently reported the maximum response of 1.0 across all trials, while the normal subjects showed more variation, generally following the true values. This suggests that the anomalous participants may have misunderstood the task or hit a ceiling

effect, whereas the others were able to perceive and respond to differences in vertical distance accurately. The true value reference lines make it easy to see that normal responses largely align with what was expected. Overall, the plot shows that anomalies are isolated and stand out clearly, while the rest of the participants behave as anticipated, reflecting the task's difficulty.

2) (20pts) Download the astronomical data for the Messier objects. These are objects that can be seen in a dark sky with binoculars or a telescope that Charles Messier cataloged in France in the 18th century so that they wouldn't be confused with comets. Some of these are clusters of stars or great clouds of gas in our galaxy, some are galaxies that are much farther away. The dataset contains a list of 100 deep sky objects along with their distances from the earth in light-years. Graph this data in the following ways to explore the information provided about these interesting objects. For this dataset, you will have to pick suitable scales to make the data readable in your graphs. You should not wind up with a majority of the points squashed down along the one axis. In particular, for distances, the scale should show the "order-of-magnitude" of the distance in light years (10, 100, 1000, etc.) clearly.

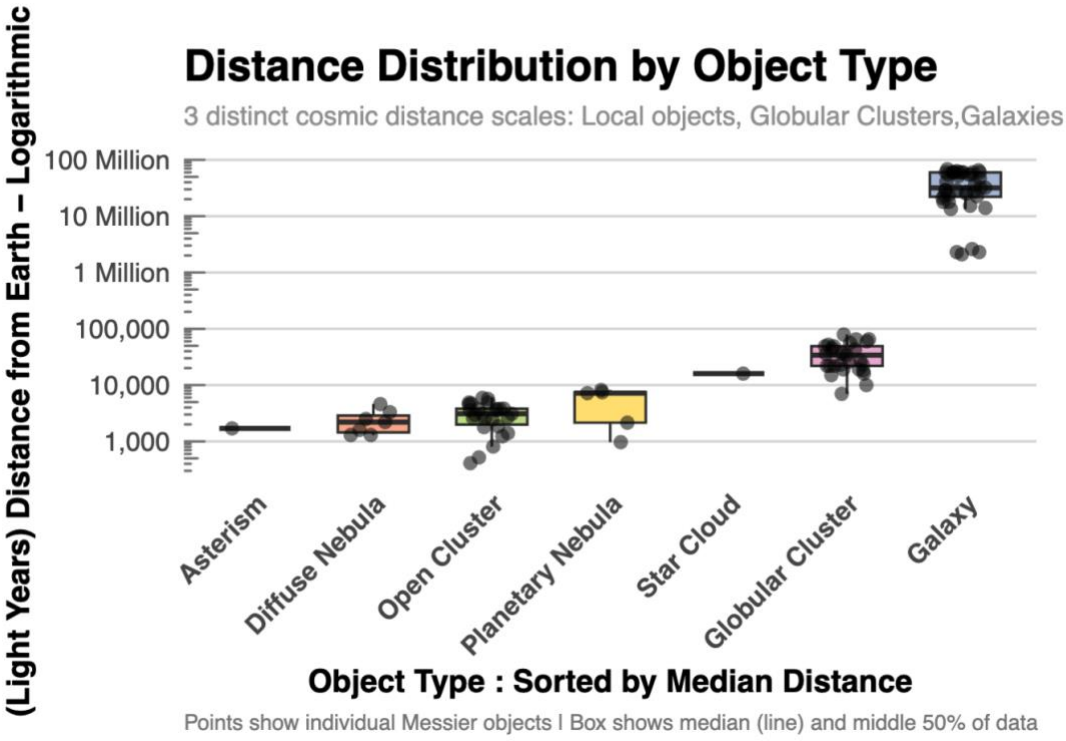
a. Start by trying to graph one or more properties of the objects against the Messier Number. Remember, there is nothing 'intrinsic' about this number, it is just the order of Messier's list. Is there any property that exhibits a pattern with respect to the ordering in his list?



The one subtle pattern that might exist is that certain object types appear in clusters of Messier numbers (suggesting he may have observed similar objects during the same observing sessions), but overall, the Messier Number itself shows **no systematic relationship with distance or any other physical property**.

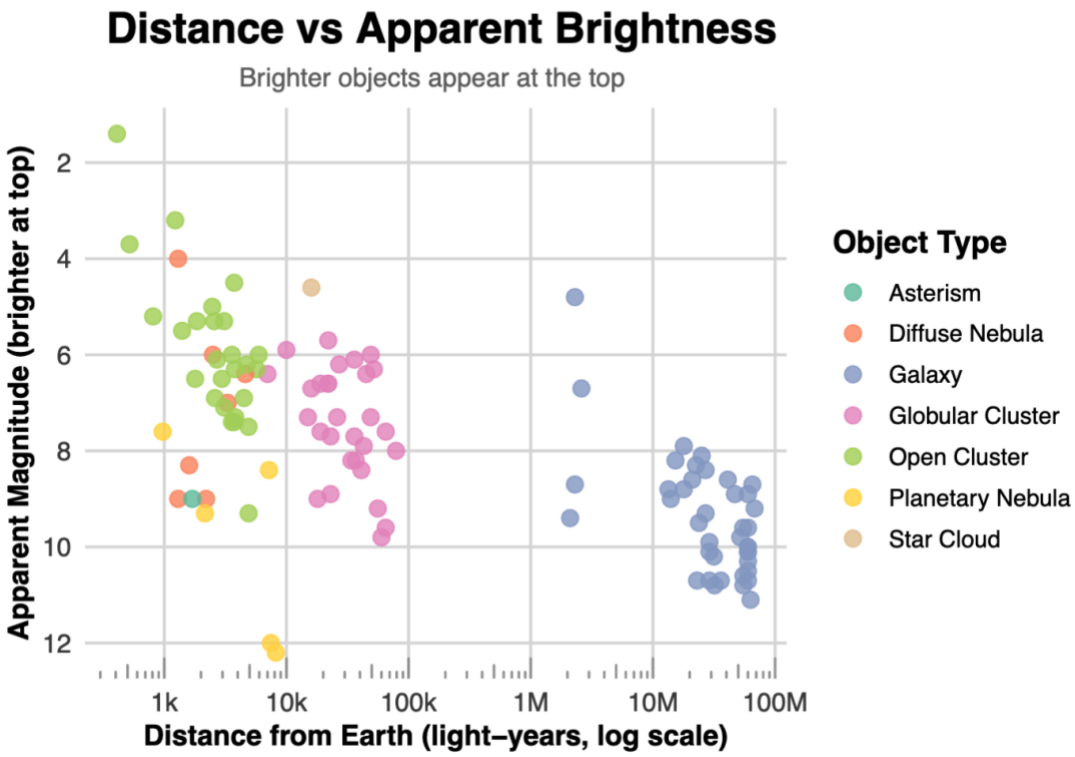
The graph reveals that different object types naturally occupy distinct distance ranges—galaxies (green) sit farthest at 10-100 million light-years, globular clusters (teal) fall in the middle around 10,000-100,000 light-years, and nebulae with open clusters (blue, yellow, purple) cluster nearby at 1,000-10,000 light-years. However, these object types are scattered randomly throughout the Messier catalog numbers rather than grouped together. For instance, galaxies appear as M31, M51, M87, and M100 - spread across the entire list. This pattern confirms that Messier simply numbered objects in the order he observed them from France, not by their physical properties or distances.

b. Create a visualization that compares the distributions of the distances to the objects in each Kind. Note that the Type variable is a very different category and is really a subcategory of Kind. Do not use that here. Sort the distribution displays in a way that makes the relationship clear.



Analysis : The Messier catalog isn't random - it reflects how Messier discovered objects over time. Early entries (M1 - M30) are bright, nearby galactic objects, while later ones (M70 - M110) are faint, distant galaxies millions of light years away. As the Messier number increases, objects get dimmer, farther, and shift in Right Ascension, showing how he recorded them in the order they appeared during different observing seasons.

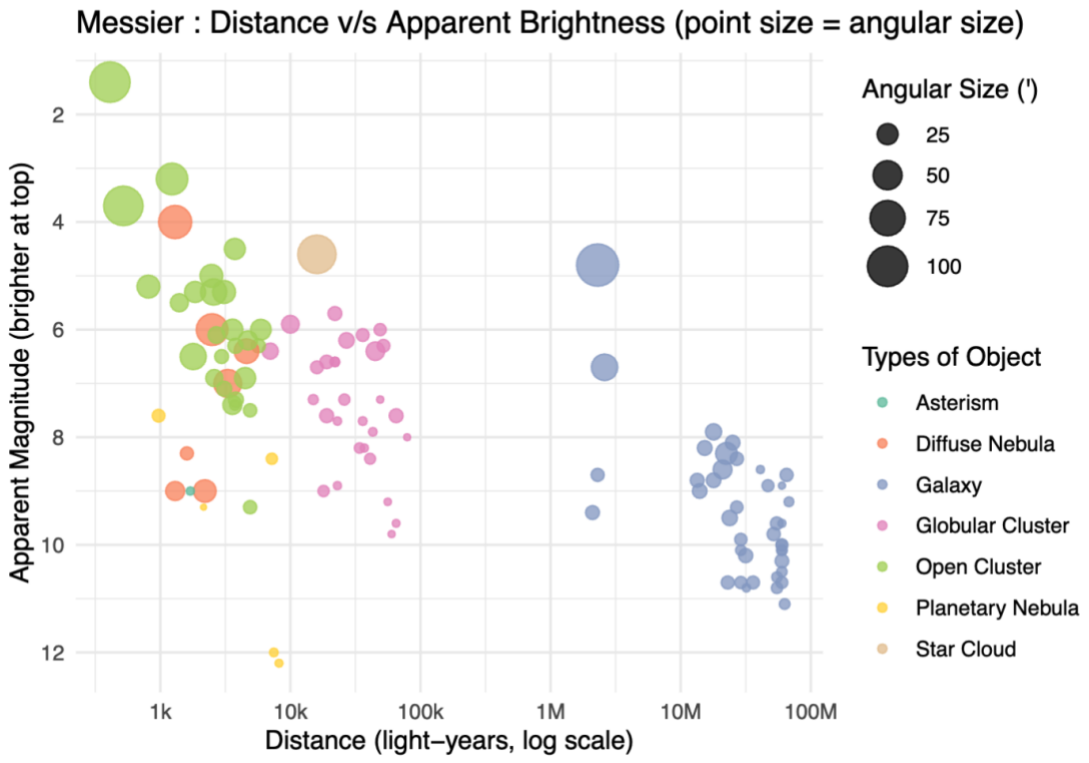
c. Create a scatter plot with the distance to the Messier objects plotted against their Apparent Magnitude (it's their visual magnitude, a measure of how bright they are in the sky). Note that these values may be… backwards from what you would think. The higher the number the fainter the object is in the sky. Try to incorporate that into your visualization to make the relationship clear.



Analysis:

The plot shows that closer Messier objects, mostly clusters and nebulae within our galaxy, appear much brighter, while distant ones mainly galaxies millions of light years away are dimmer. There's a clear divide between nearby and faraway objects along the distance axis. Since the y-axis is reversed, we can easily see how brightness drops as distance increases. Overall, the chart nicely illustrates that the farther an object is, the fainter it appears.

d. Augment the visualization in (c) by adjusting the size of the points in the scatter-plot based on the angular Size of the objects in the sky. Evaluate how easy it is to analyze all encoded aspects of the data from this graph and give a suggestion on how you might modify the graph to display all this information more readably.

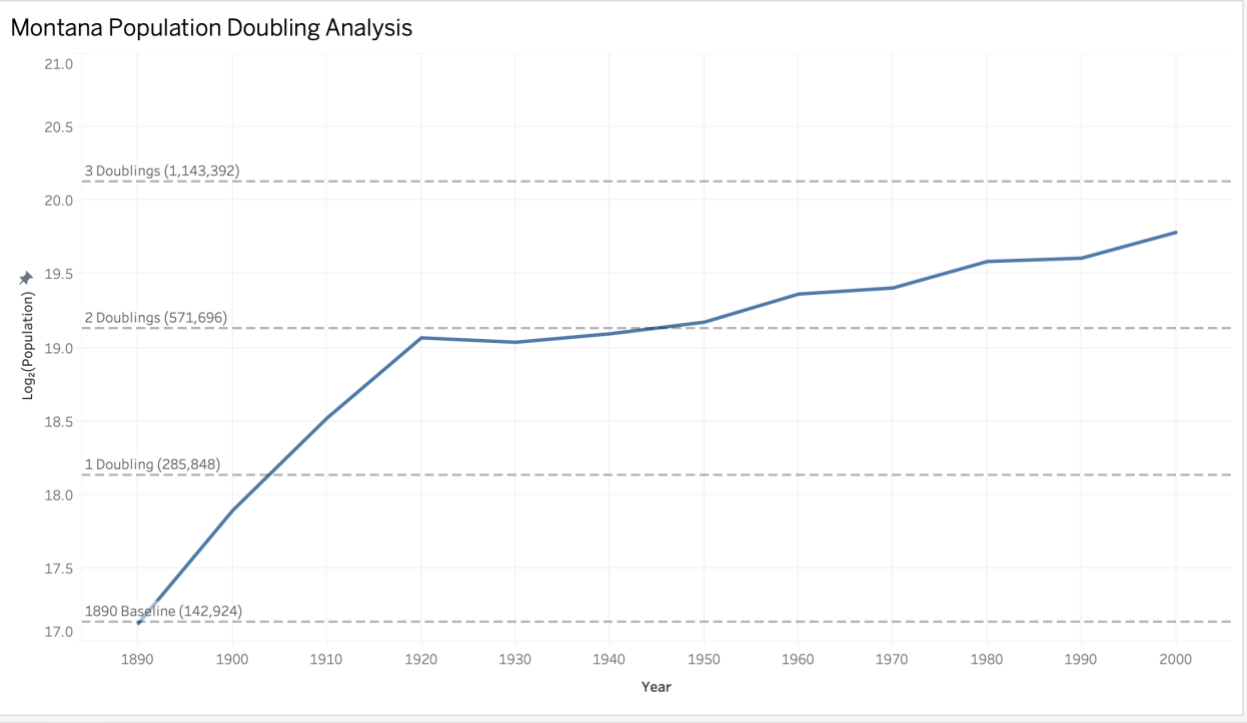


Analysis :

By adjusting point size based on angular size, the plot shows that closer objects appear larger and brighter, while distant galaxies look smaller and dimmer. However, since distance, brightness, type, and size are all shown together, the graph feels crowded and a bit hard to interpret. Overlapping points also make it tricky to compare different object types or sizes. A better approach would be to use separate plots for each object type or create an interactive version with hover details to make the data easier to explore.

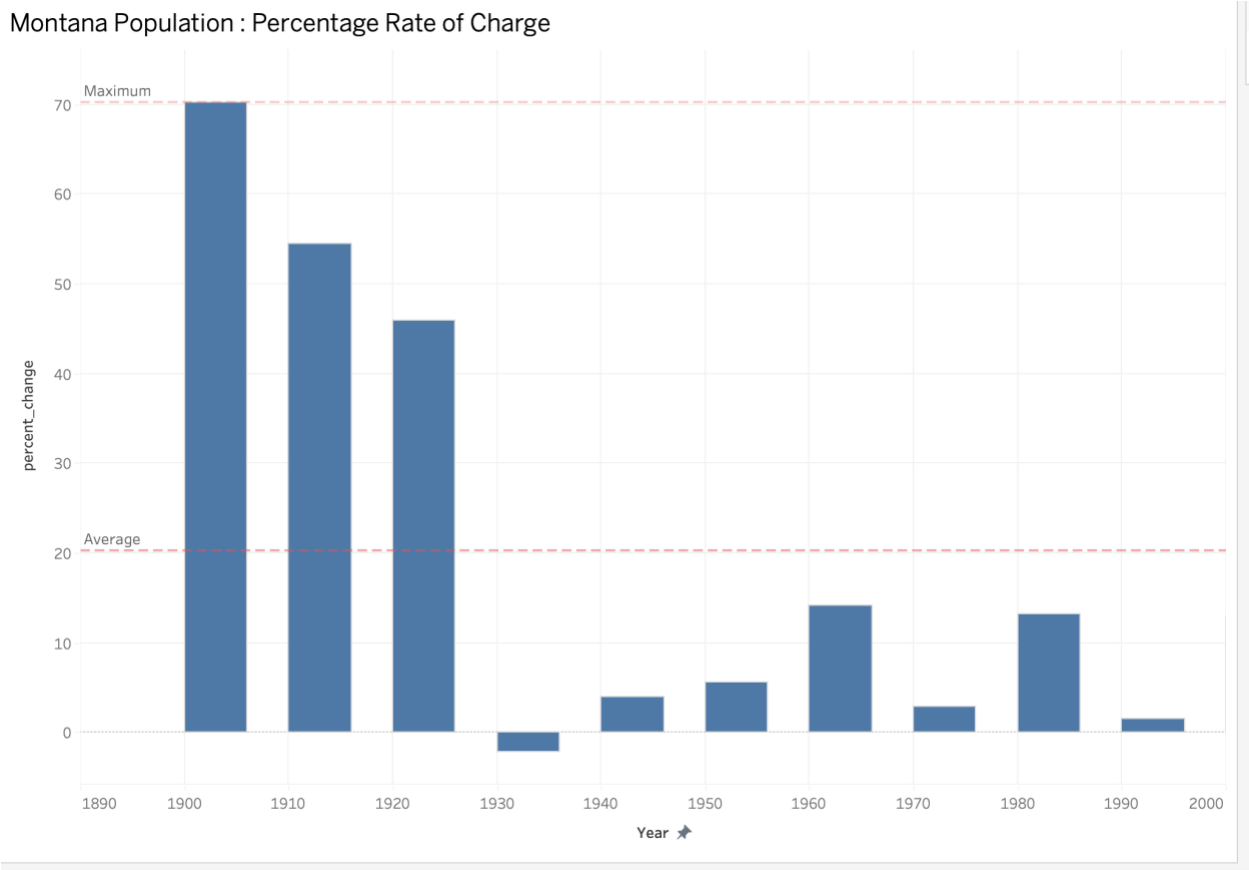
3. Download and graph the Montana Population data set (different from the one we used previously). Create visualizations using logarithmic scales, and intended for a technical audience, that clearly demonstrate visually the answers to the following questions. Viewers should be able to read the answers to these directly off the graph scales. Different logarithmic scale techniques may be appropriate for each part. If you use a single graph to answer multiple parts, make it clear that you are doing so.

a. How many times has the population doubled since 1890?



1. Analysis :
   The population reached its first doubling (≈285,848) around 1900, and the second doubling (≈571,696) by about 1925. However, it has not yet attained the third doubling (≈1,143,392) as of the year 2000. This pattern highlights a steady but gradually slowing growth rate over the 20th century, with population expansion tapering off after mid-century. The logarithmic visualization makes these growth phases and their diminishing pace clearly visible.

   b. Has the percentage rate of change in the population increased or decreased over the years? What years had the greatest increase in population %-wise?

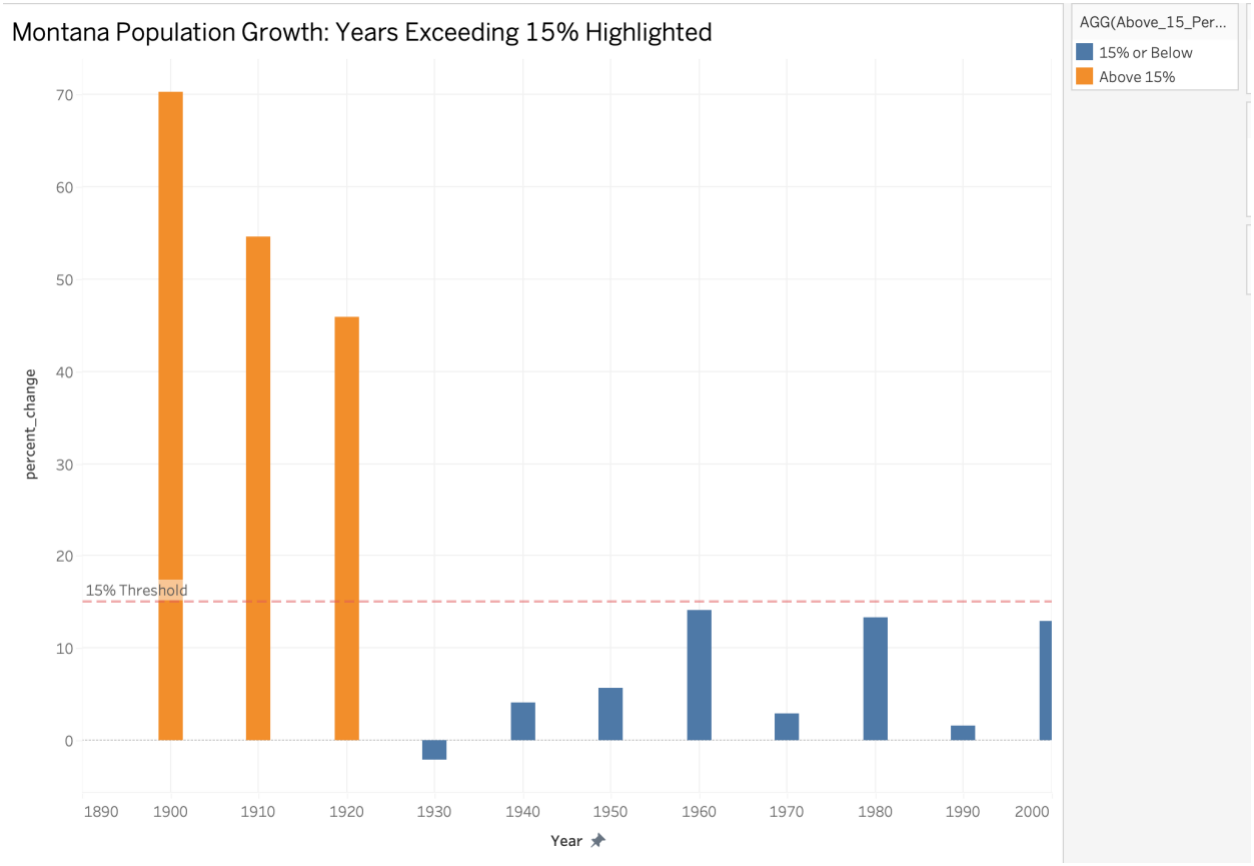


Steps:

1. Connect data → drag Year to Columns and Population to Rows → make a line chart.
2. Create calculated field: ((SUM([Population]) - LOOKUP(SUM([Population]), -1)) / LOOKUP(SUM([Population]), -1)) * 100 – plot bar chart and add reference lines .

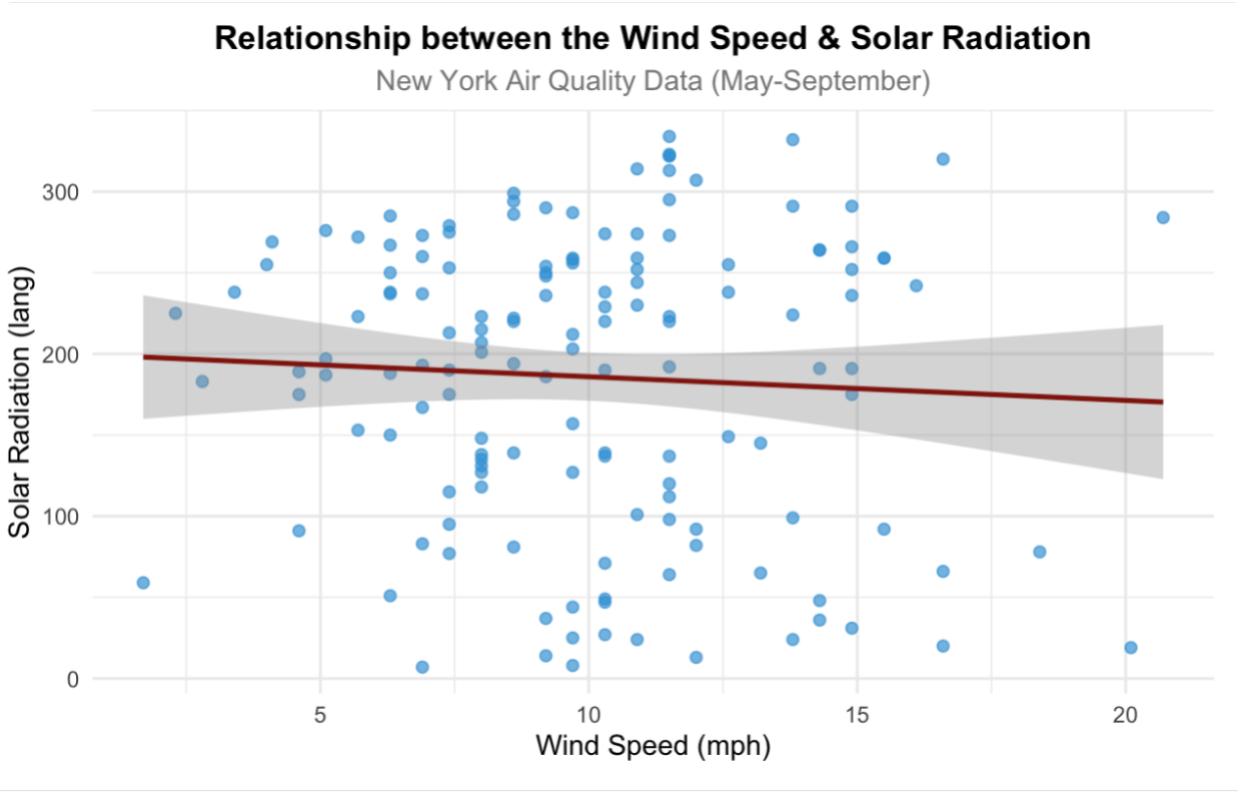
Analysis :

The greatest percentage increase occurred between 1890-1900 with approximately 70% growth (142,924 to 243,329), followed by 1900-1910 with 54.6% growth (243,329 to 376,053), representing the most dramatic expansion period.



Steps:
Added color filter : Above 15% orange and 15% or below other color and added reference line on Y axis for 15% threshold .

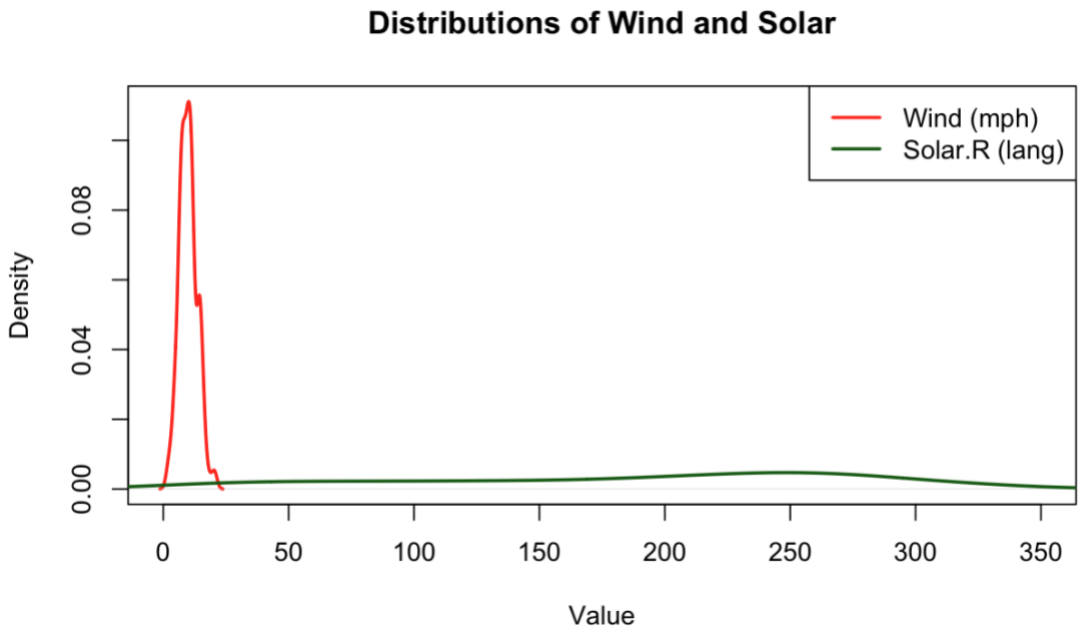
4) (20 pts) We will look at data on air quality, captured from May to September in New York. This is actually built into R, but not as a data frame. There is a copy on the D2L site.
a. Use a scatterplot to look at the relationship between Wind and Solar.R (solar radiation). Show a fit line. Make sure to produce a clean visualization with emphasis on the trend. This provides one view of the relationship. For help doing this in R, see Tutorial 5. In Tableau, this is available from the Analysis tab. It is one of the tabs along with Data for the panel on the far left (i.e. look at the top of the panel from which you drag variables).



Analysis :
The scatterplot shows a clear negative relationship between wind speed and solar radiation : as wind speed increases, solar radiation tends to decrease. This suggests that in New York, during May to September, windier days are often accompanied by cloudier conditions that reduce the amount of sunlight reaching the ground.

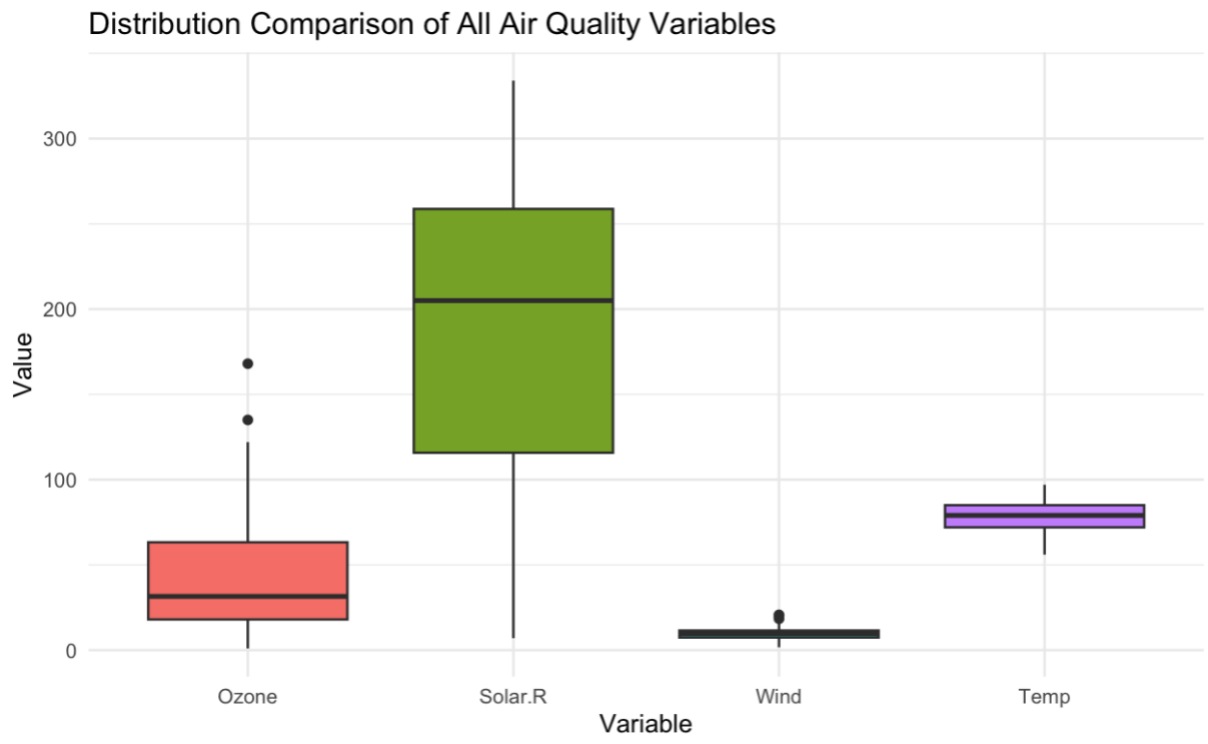
b. Use a plot that will show the distributions of Wind and Solar.R and allow you to compare with fine detail.

Analysis :
Overall, wind speed is more consistent and predictable, while solar radiation varies much more from day to day over the observation period. The density plots show that wind speed follows a fairly symmetric, bell-shaped pattern centered around 10 mph, with most values between 7 and 15 mph. In contrast, solar radiation has a wider and more uneven distribution ranging from 0 to 350 lang, with several peaks that reflect changing sunlight conditions.



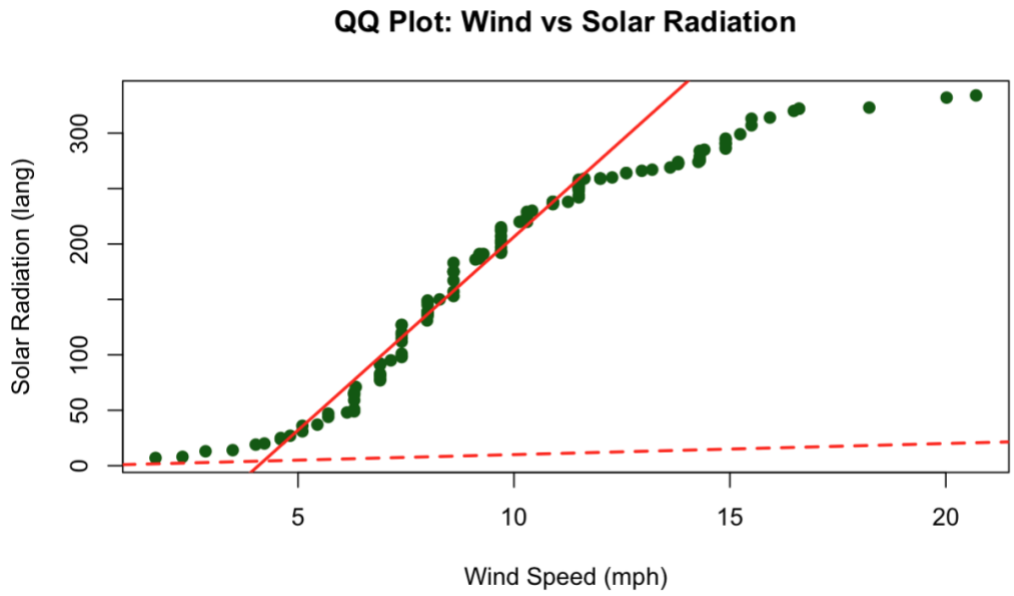
c. Finally, show these distributions in context of the rest of the variables by using a technique for comparing multiple distributions. Note: you will need to transform the data in a particular way that we have studied. I it showed in the Tableau tutorial and in an R tutorial.

Hint – you need to collapse the current variables into two: (1) stores the original variable name, and (2) stores the corresponding original value.



Analysis :
The boxplots show that the four air quality variables span very different ranges. Temperature (56 - 97°F) and Ozone (1 - 168 ppb) have moderate spreads, Wind Speed (2–21 mph) is tightly clustered with the smallest range, and Solar Radiation (7–334 lang) varies the most. Looking at the distributions, Wind Speed and Temperature are fairly symmetric with few outliers, while Ozone and Solar Radiation are right skewed with occasional extreme values, highlighting that these two measures are more variable and sometimes reach very high levels during the observation period.

d. For extra credit, compare Wind and Solar.R again with a QQ plot. What does this tell you?



Analysis :

The QQ plot highlights noticeable deviations from the diagonal line, showing that Wind Speed and Solar Radiation have quite different distributions and scales. Solar Radiation has a wider range and heavier tails, while Wind Speed is more normally distributed. The curve in the plot confirms that these variables behave differently: Wind Speed is fairly consistent, whereas Solar Radiation is right skewed with occasional extreme values, reflecting the distinct nature of the atmospheric phenomena they represent.