CSC 465

Homework 3

QUESTION 1

(20 pts) 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.

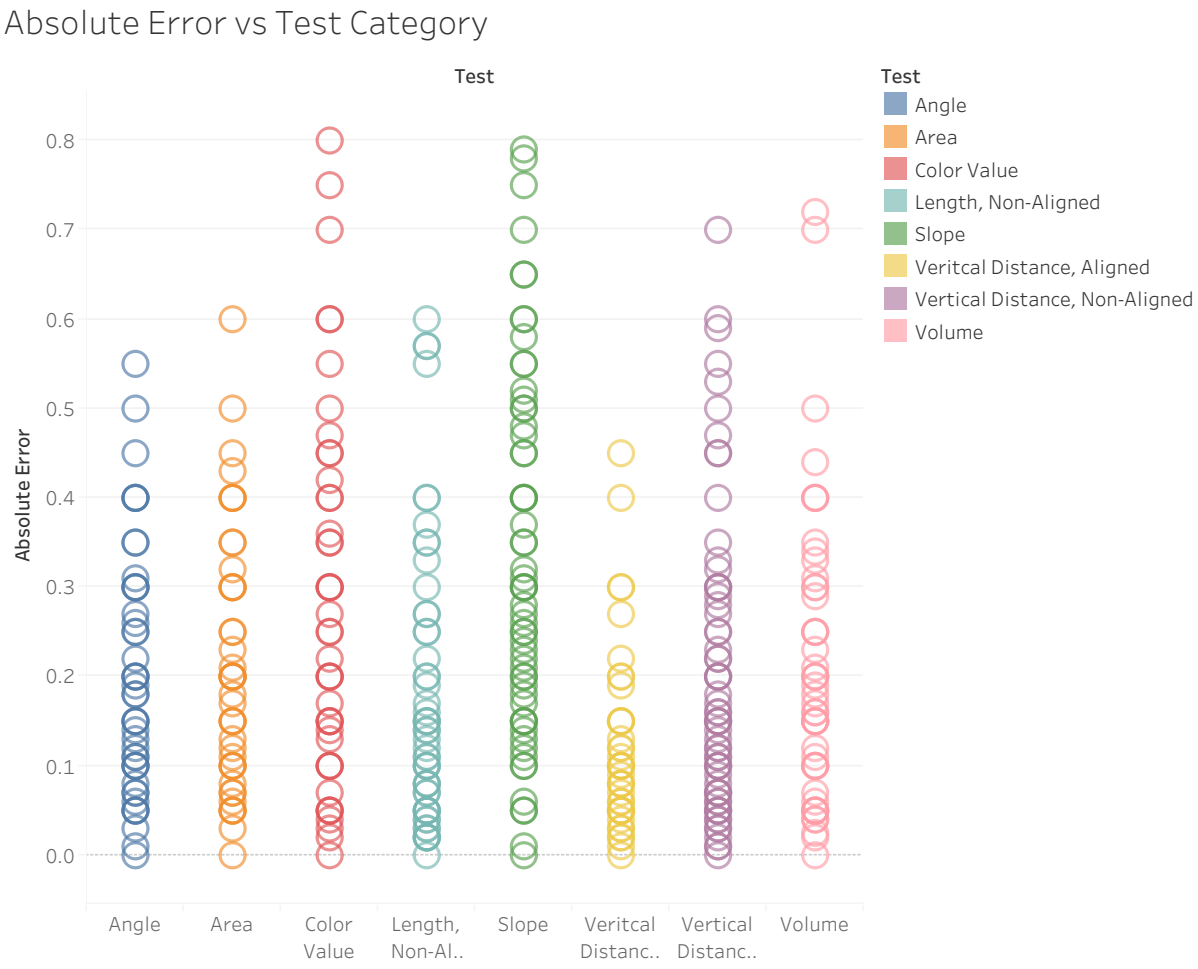
a. **Were there any tests where people generally underestimated or overestimated the data? Explain what field you can graph to test this, what graphical method reveals this clearly. Analyse the results and explain in a short paragraph.**

Sum of Error for each Test.

To analyse the accuracy of student estimates, I used a bar graph to visualize the difference between their responses and the true values for each test category. This is the error. The categories included angle, area, color value, length (non-aligned), slope, vertical distance (aligned and non-aligned), and volume. For example, in the "Volume" test, the students tended to overestimate the value, with an average error of approximately over 60. Conversely, in the "color value" test, they consistently underestimated the true value, with an average error of around over -80. Overall, this graph provides insights into the patterns of overestimation and underestimation across various test categories.

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?



Absolute Error for each Test. Color shows details about Test.

The provided scatter plot illustrates the absolute error distribution for different methods of visually encoding numerical data. The x-axis represents the test categories (angle, area, color value, length non-aligned, slope, vertical distance aligned and non-aligned, and volume), while the y-axis indicates the absolute error.

The data points reveal distinct patterns in the accuracy of estimations. For angle and area representations, the absolute error is generally low, and the points are clustered near the true values. This suggests that participants were able to accurately estimate these quantities. In contrast, color value and length non-aligned methods exhibited higher variability in absolute error, with no clear clustering of responses. This indicates that these methods were less effective in conveying the numerical information.

Interestingly, vertical distance and volume representations showed specific clusters of responses at particular error levels. This suggests that participants consistently made similar errors when using these methods. Overall, the scatter plot provides valuable insights into the

effectiveness of different visual encoding methods in facilitating accurate numerical estimations.

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?

I've filtered the data to focus on subjects 56-73, as specified in the prompt. The resulting visualization, a scatter plot, illustrates the responses for these subjects across Display 1 and Display 2.

Key Observations

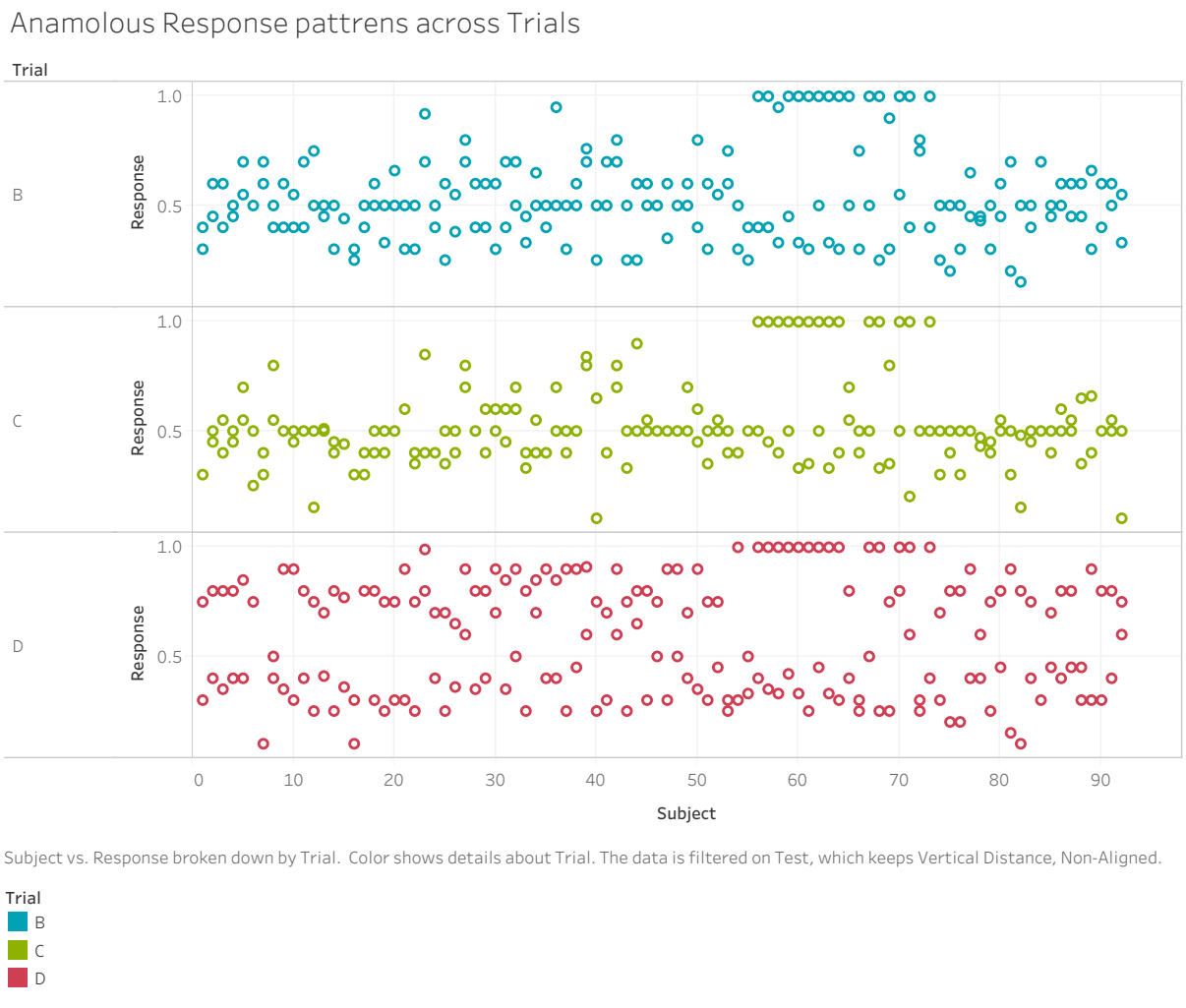
1. **Shift in Response Distribution:**
    - **Display 1:** Responses are more spread out, indicating a wider range of values.
    - **Display 2:** Responses are clustered towards higher values, suggesting a potential shift in participants' perception or judgment.
2. **Improved Accuracy or Consistency:**
    - The narrower distribution of responses for Display 2 suggests that participants may have become more accurate or consistent in their judgments after completing the task once.

Conclusion

Based on the analysis of responses for subjects 56-73, there is evidence to suggest that participants experienced a shift in their response patterns between Display 1 and Display 2. The responses for Display 2 were more concentrated towards higher values, indicating a potential improvement in accuracy or consistency. This finding supports the hypothesis that participants may benefit from prior experience in making judgments on similar tasks.

Note: To provide a more definitive conclusion, further statistical analysis, such as a paired t-test, could be conducted to assess the significance of the observed differences between Display 1 and Display 2.

d. An erroneous stimulus was used for the first Display of "vertical distance, non-aligned" for a small subset of the xsubjects. 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.



Observation of Consistent Outliers

The scatter plot clearly highlights an anomalous sequence of "1" responses across Trials B, C, and D for a specific subset of subjects. These outliers are distinct from ordinary outliers due to their consistent values across all three trials, regardless of the true values.

Visualizing the Anomalous Sequence

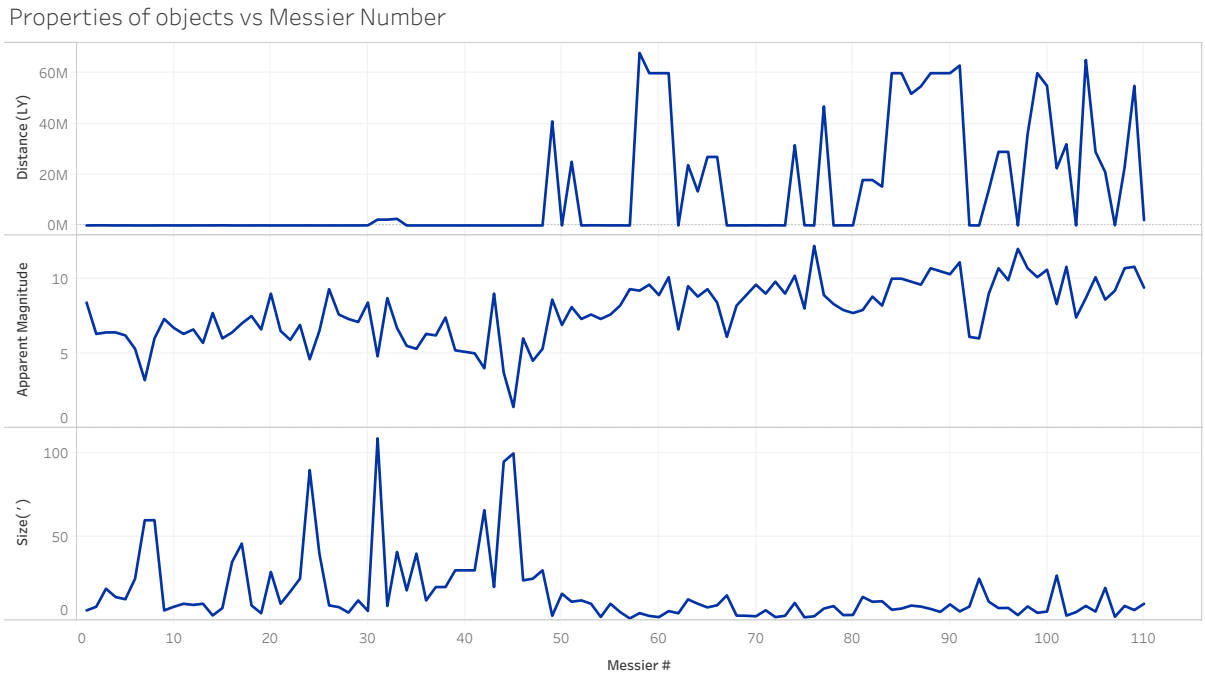
To further emphasize the anomalous nature of these responses, we can create a visualization that focuses solely on the subjects exhibiting this pattern. A line chart or bar chart could be used to illustrate the consistent "1" values across the trials for these subjects. Additionally, the visualization could include a reference line indicating the expected range of responses, making the outliers even more apparent.

QUESTION 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 18^{th} 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 trends of sum of Distance (LY), sum of Apparent Magnitude and sum of Size(') for Messier #. The view is filtered on sum of Apparent Magnitude, which keeps non-Null values only.

Distance (LY)

There is a general tendency for distance to increase with higher Messier numbers, though the trend isn't perfectly linear. Significant fluctuations suggest that factors beyond the Messier number affect object location, with no clear pattern in distances between Messier numbers 0 and 50.

Apparent Magnitude

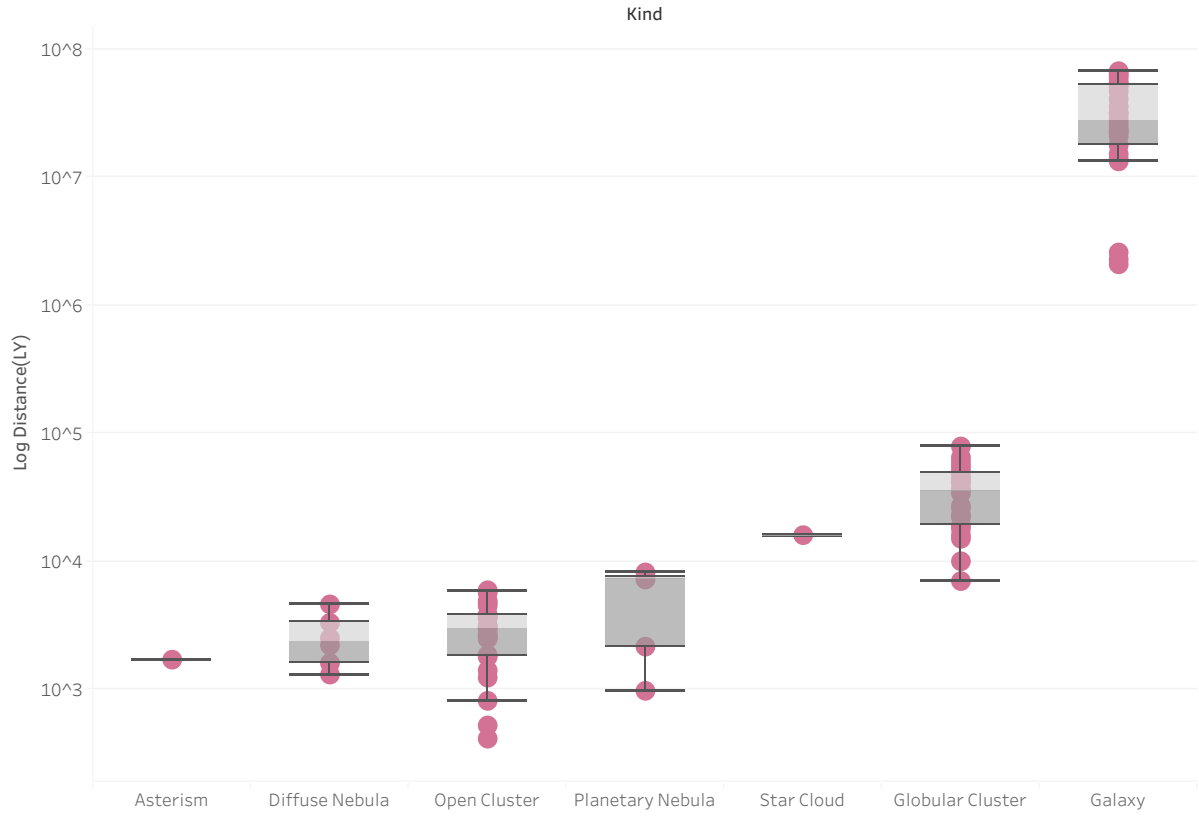
Apparent magnitude generally increases with higher Messier numbers, but noticeable fluctuations indicate other influences on brightness.

Size

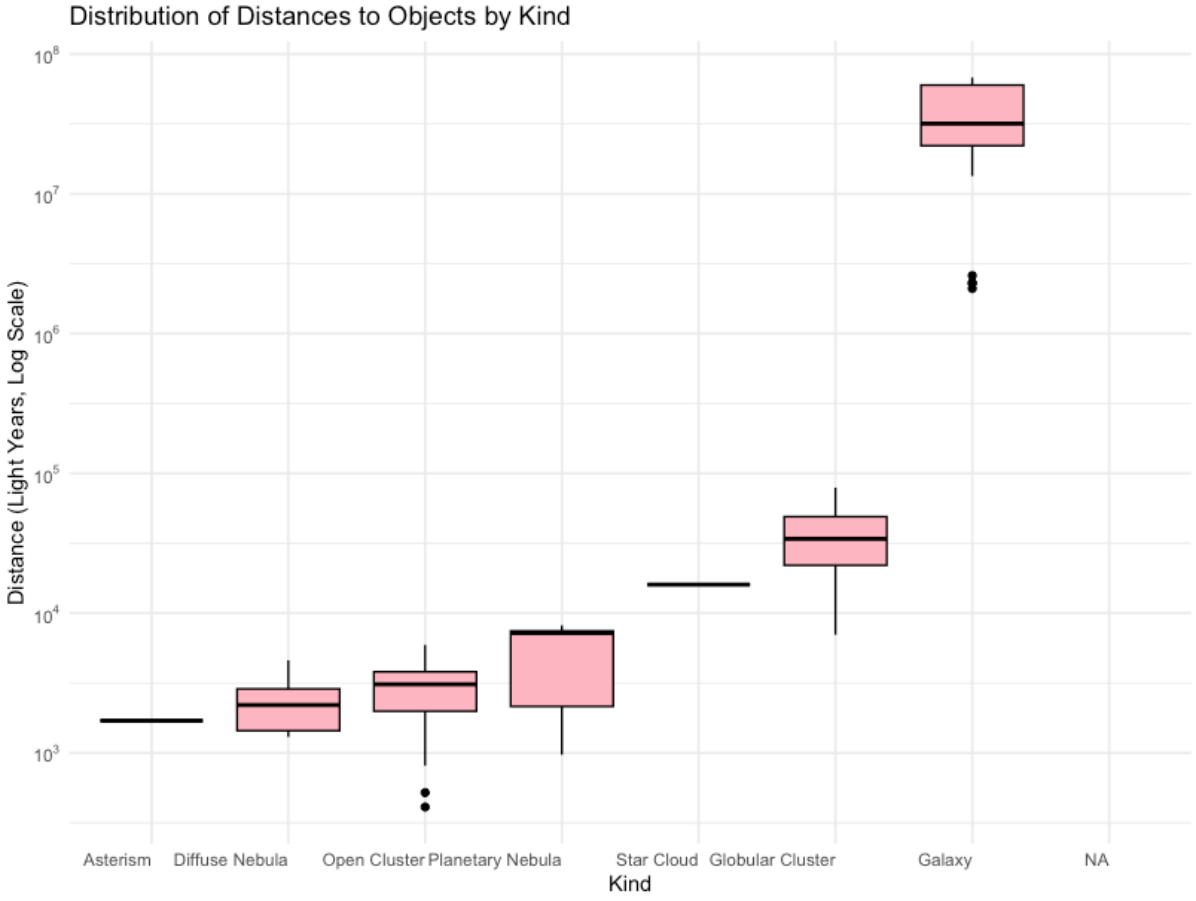
The sizes of Messier objects vary widely, with no clear correlation to their Messier numbers. Some objects stand out as outliers, reflecting the diverse characteristics within the catalog.

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 sub- category of Kind. Do not use that here. Sort the distribution displays in a way that makes the relationship clear.

Distribution of disances by kind



Distance(LY) for each Kind. Details are shown for Kind. The data is filtered on Distance (LY), which ranges from 410 to 68000000.

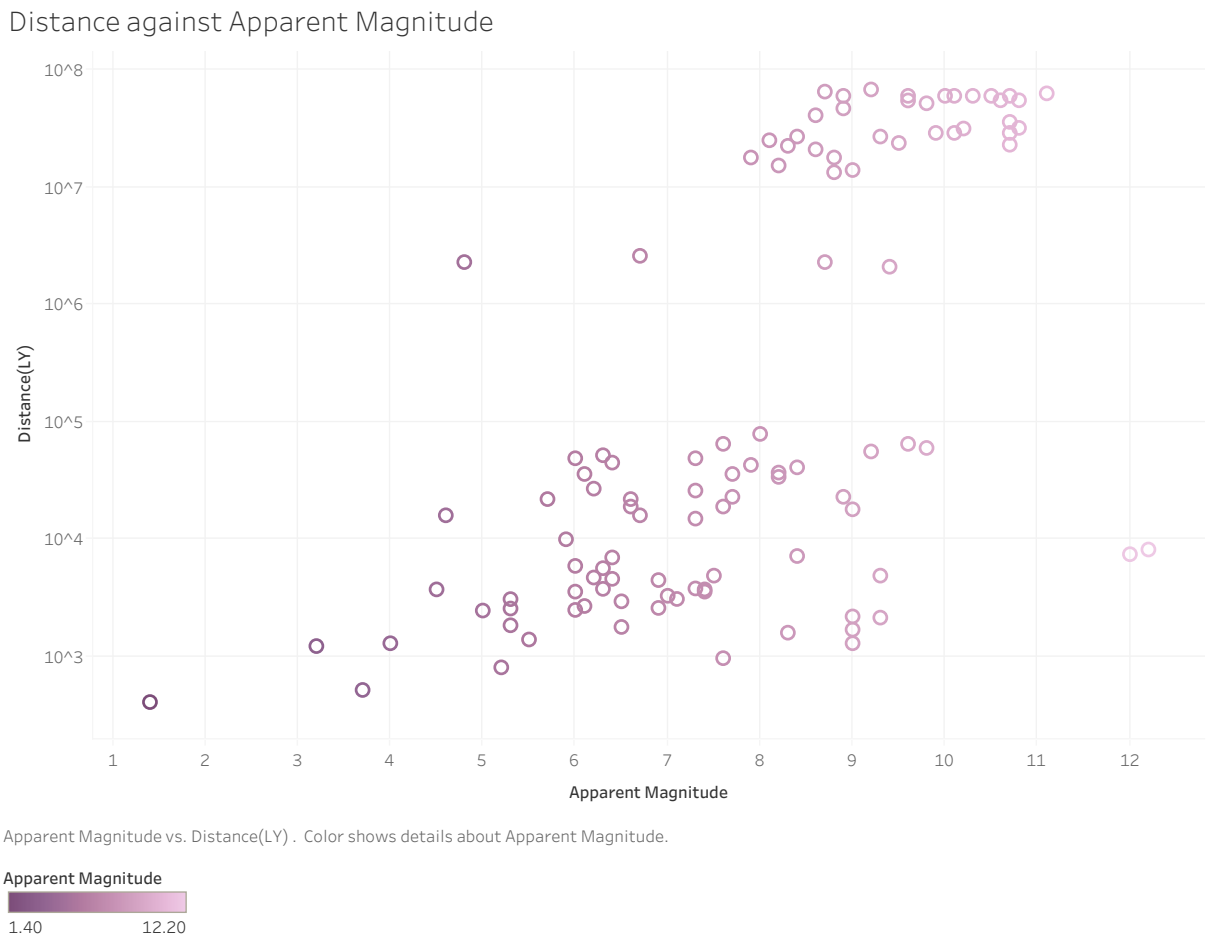


The box plot effectively illustrates the distribution of distances across various types of celestial objects. By organizing the categories in ascending order based on their median distance, the plot emphasizes differences in both scale and variability among them.

- **Distance Scale:** The logarithmic y-axis highlights the vast range of distances, making it easier to compare objects located at significantly different distances.
- **Distribution Patterns:** The box plots display the spread of distances within each type. Some categories, like Star Clouds, show a tighter distribution, while others, such as Galaxies, have a broader range.
- **Outliers:** Data points lying beyond the whiskers identify outliers, offering insight into unusually distant or unique objects within each group.
- **Median Distance:** The median, marked by the line inside each box, provides a quick reference for comparing the typical distance across categories.

This visualization effectively captures the unique distribution characteristics of different celestial objects, offering meaningful insights into their spatial arrangement.

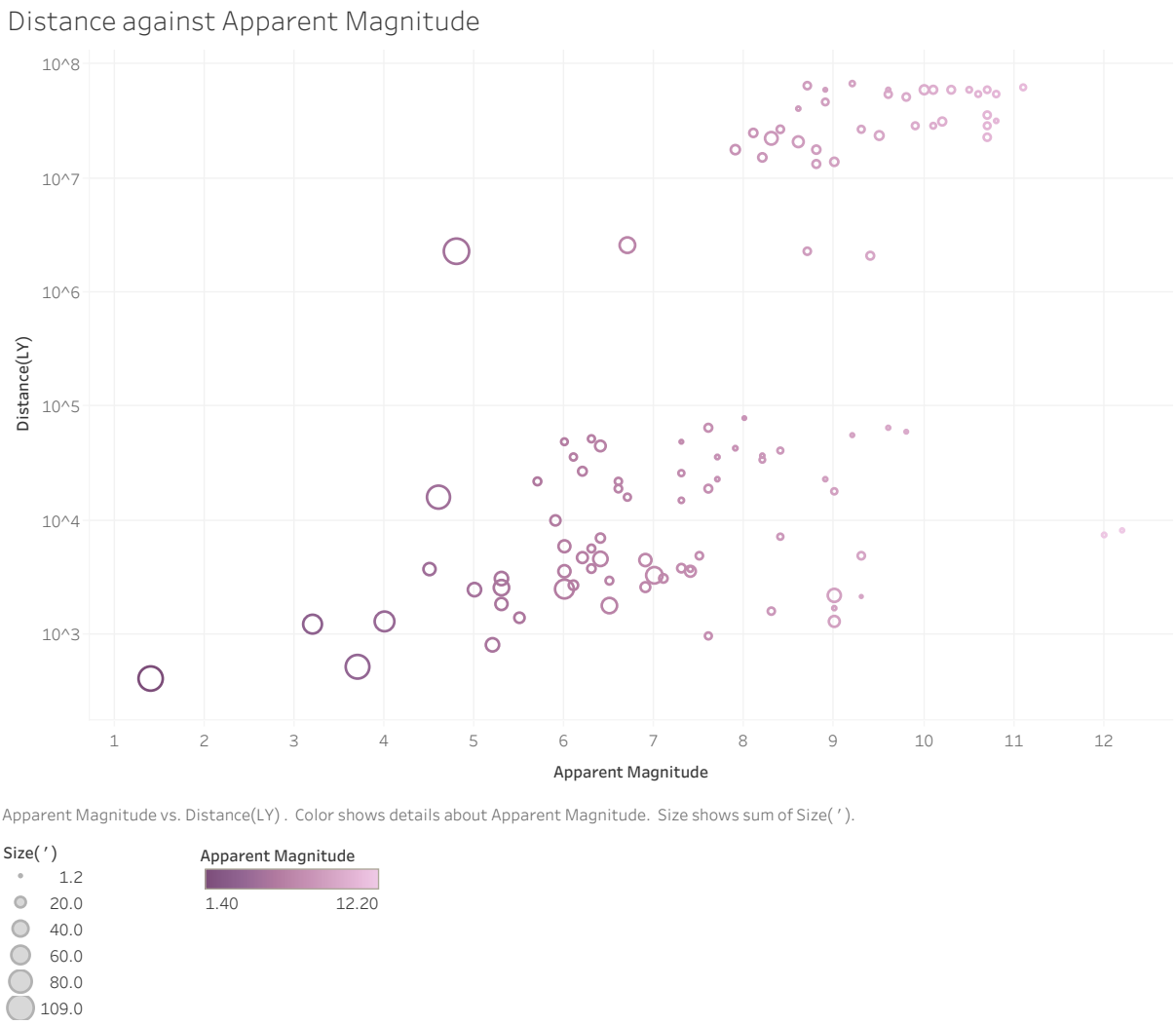
c. Create a scatter plot with the distance to the Messier objects plotted against their Apparent Magnitude (it's their visual magnitude, a measure of how bright they are in the sky). Note that these values may be… backwards from what you would think. The higher the number the fainter the object is in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



The scatter plot highlights the relationship between distance and apparent magnitude for Messier objects. Using a logarithmic scale for distance and inverting the y-axis for apparent magnitude (where higher values indicate fainter objects) captures the expected trend accurately. As distance increases, the general pattern shows that objects tend to appear dimmer.

While the overall trend is clear, some scatter around the line suggests that factors beyond distance influence an object's brightness. A few outliers also stand out, possibly due to unusual properties or measurement inaccuracies. This visualization offers meaningful insights into the spatial distribution and brightness of Messier objects.

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.



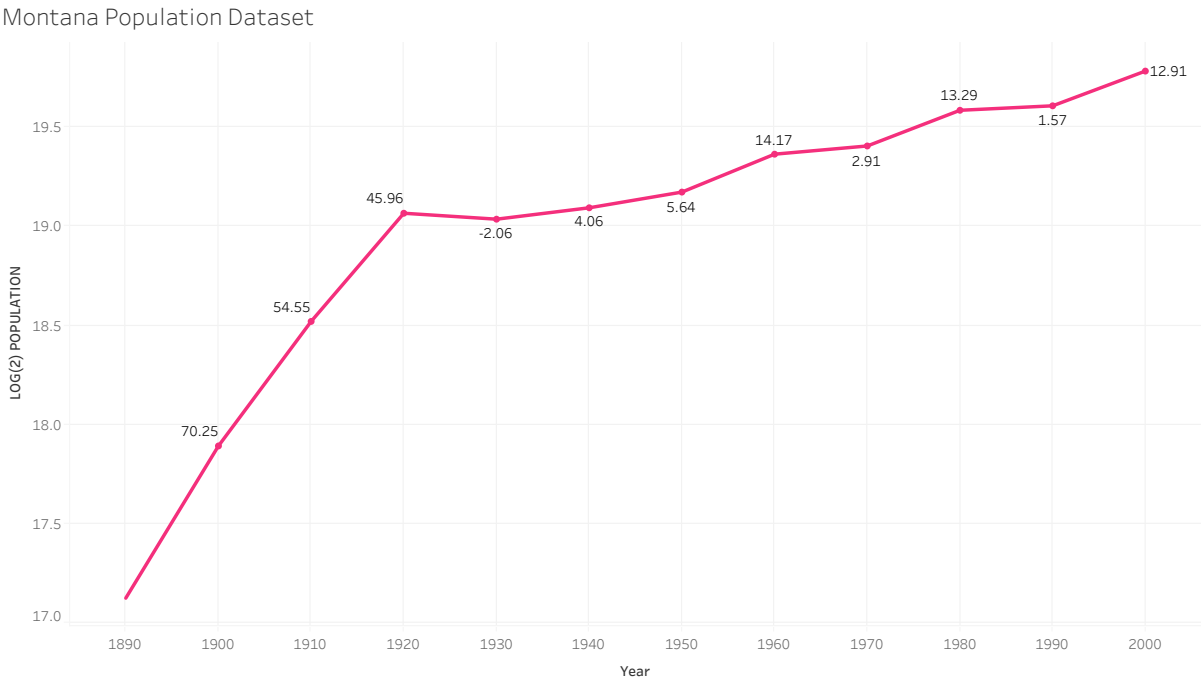
The augmented scatter plot combines distance, apparent magnitude, and angular size by scaling point sizes based on angular size, offering a comprehensive view of the data. The relationship between distance and apparent magnitude remains visible, with farther objects generally appearing dimmer, and larger points indicating objects with greater angular size. However, overlapping points with similar values can make it difficult to differentiate objects by size alone. Improvements like color coding for angular size, applying transparency to reduce overlap, or adding interactive elements (e.g., tooltips) could enhance readability and provide deeper insights into the data.

QUESTION 3

(15pts) 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.

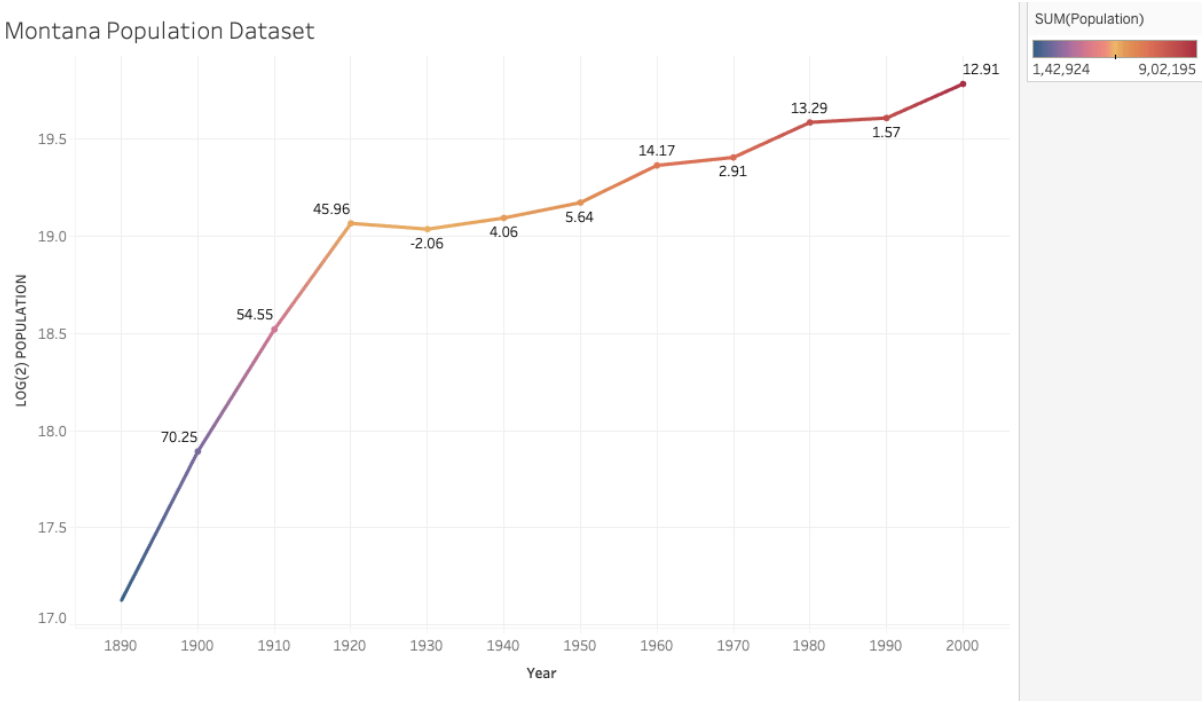
a. How many times has the population doubled since 1890?



In the given graph, the use of a logarithmic scale allows us to easily identify the points where the population has doubled. By looking for points that are 1 unit apart on the y-axis, we can determine the approximate years of population doubling. It appears the population doubled on three occasions:

- Around 1910, the population was roughly twice the size it was in 1890.
- Between 1940 and 1950, another doubling took place.
- A third doubling occurred between the late 1970s and early 1980s.

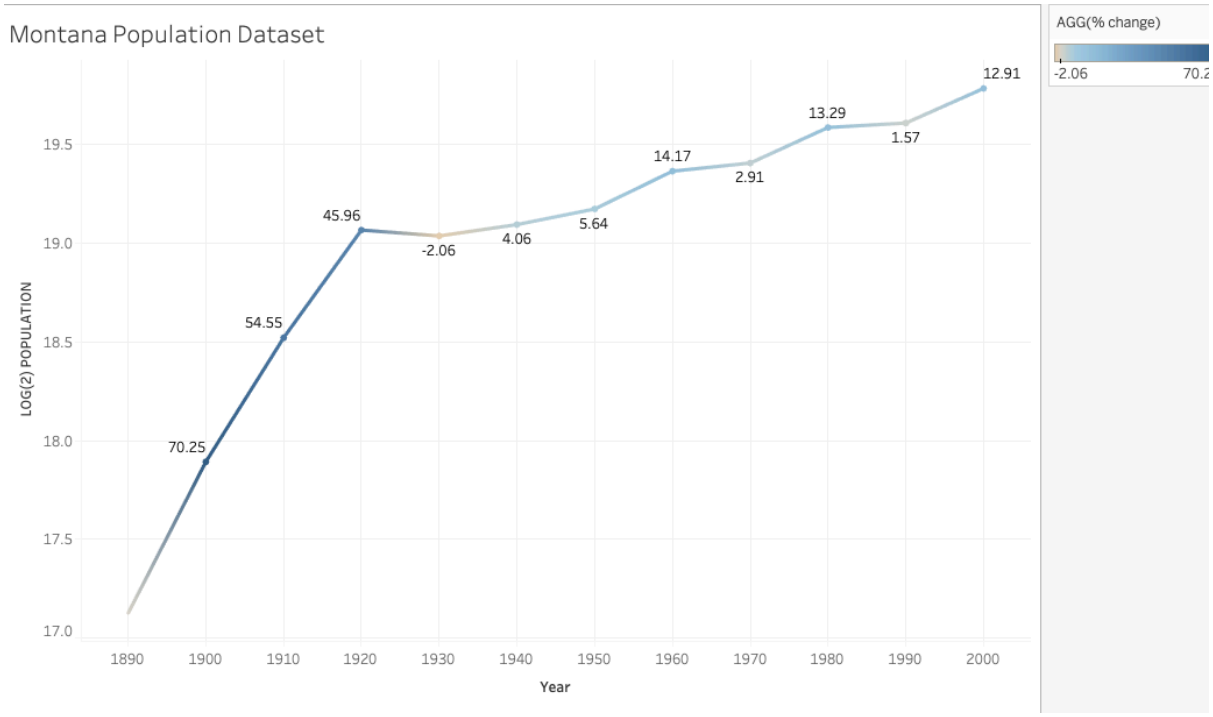
b. Has the percentage rate of change in the population increased or decreased over the years? What years had the greatest increase in population %-wise?



The annual growth rate has fluctuated considerably over the years. In 1920, there was a decline of -2.06%. Significant increases were observed in the early 1900s, as well as the late 1970s and early 1980s. The highest growth rates occurred in 1900 (70.25%) and 1980 (13.29%).

c. What years was the population percentage increase greater than 15%?

Year	% change
1890	17.124888671
1900	70.250622709
1910	54.545080940
1920	45.960542796
1930	-2.055606871
1940	4.064314758
1950	5.642624264
1960	14.169136956
1970	2.910930736
1980	13.289142278
1990	1.573046562
2000	12.906334278



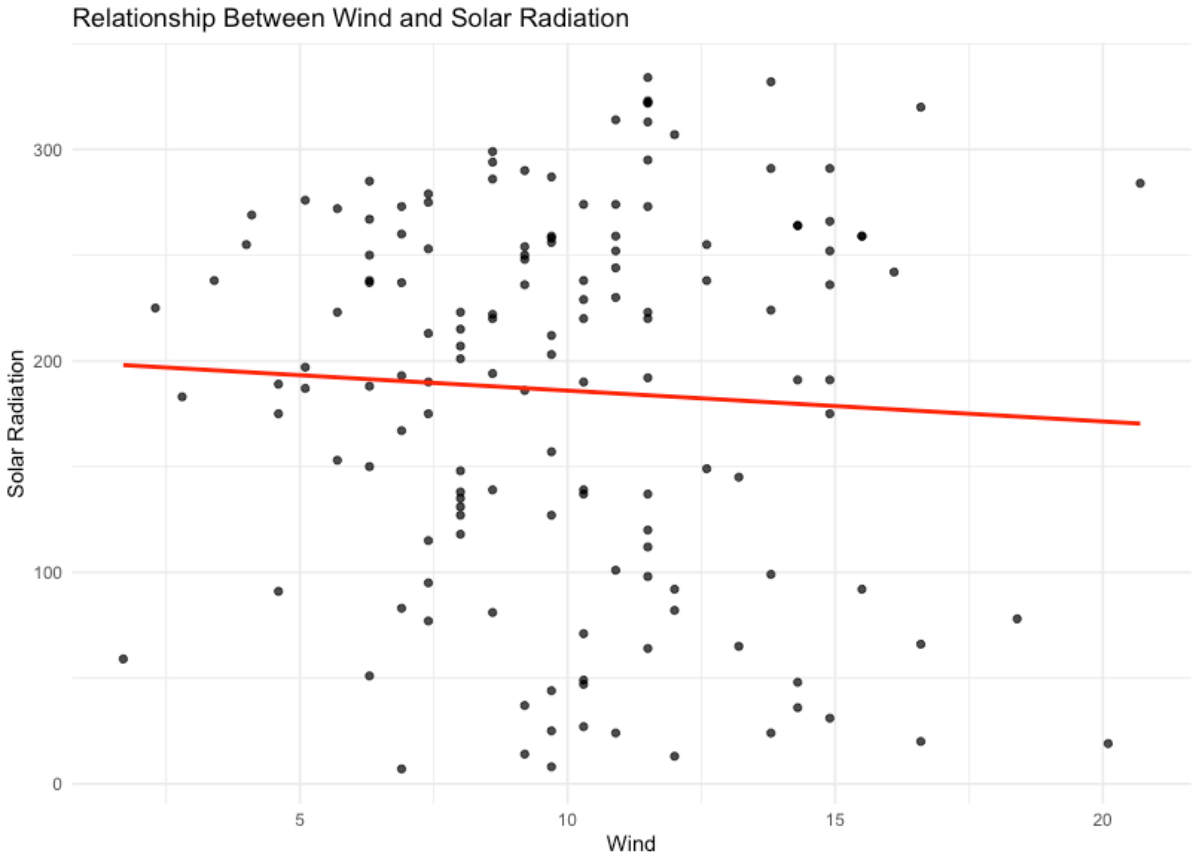
From the image above, the dark blue gradient represents the years 1900,1910 and 1920 that have the percentage increase greater than 15%.

QUESTION 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.

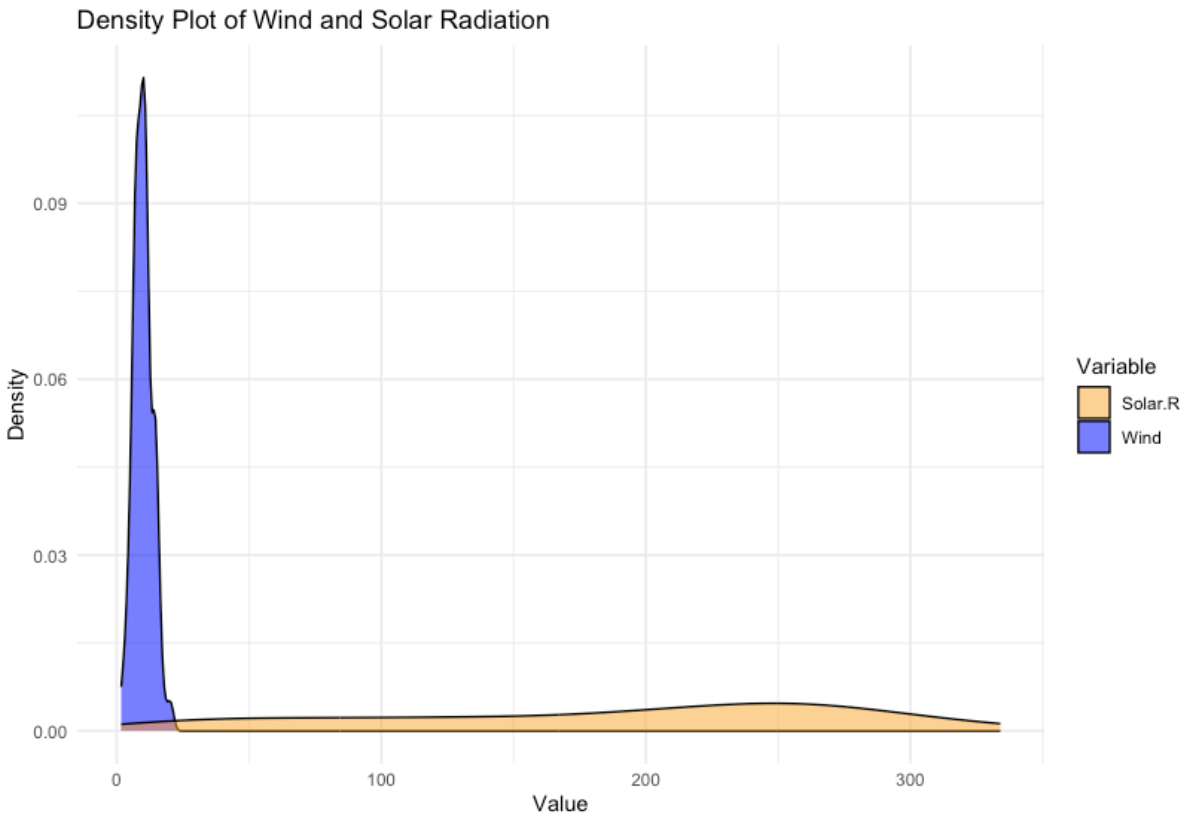
a. Use a scatterplot to look at the relationship between Wind and Solar.R (solar radiation). Show a fit line. Make sure to produce a clean visualization with emphasis on the trend. This provides one view of the relationship.

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).



The above graph is a scatterplot, which illustrates the relationship between wind speed and solar radiation. Here each point is representing a measurement at a specific moment. The trend line red colour highlights any potential patterns, such as an increase in solar radiation with higher wind speeds. This visualization helps us explore how the two factors may interact. A clean and straightforward design ensures the focus remains on the relationship without unnecessary distractions.

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

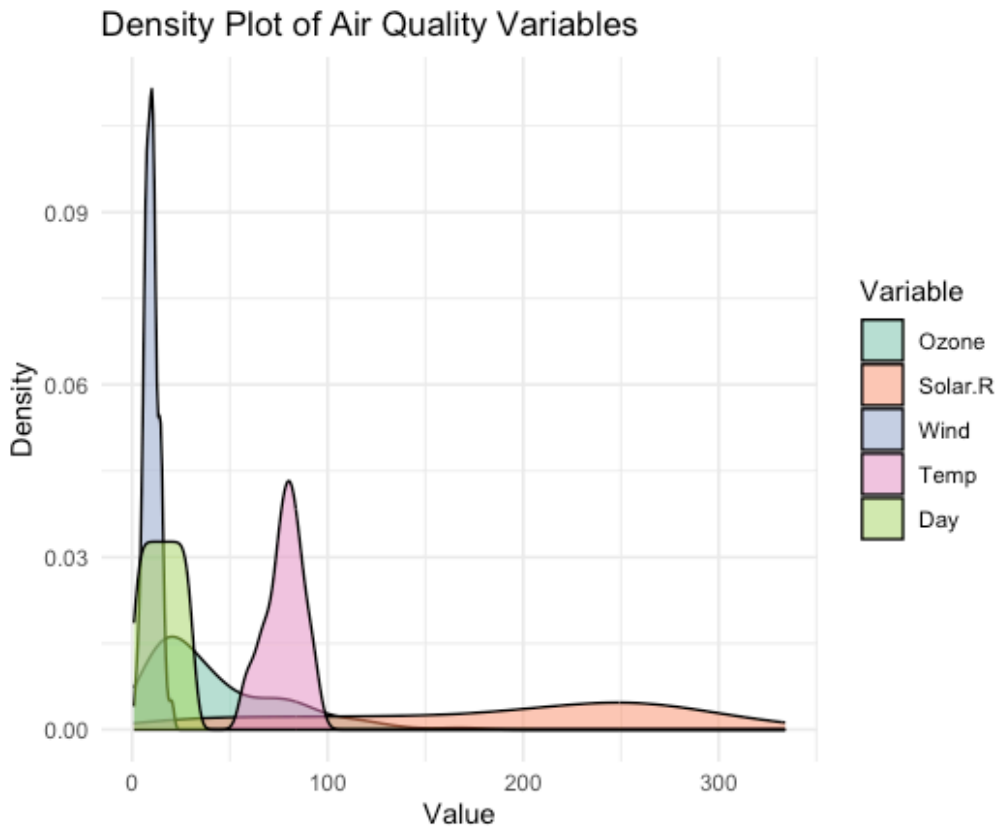


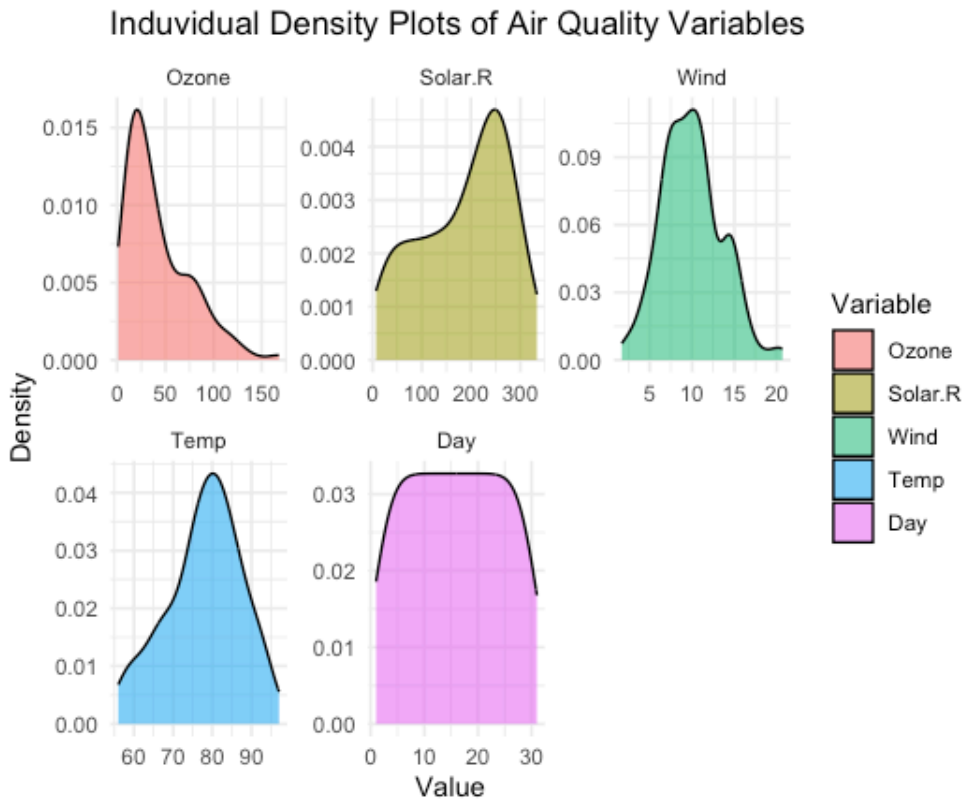
This plot illustrates the density of wind speed and solar radiation data, with each curve representing the frequency of various values. Different colors are used to distinguish between the two variables, making comparisons straightforward. The transparency highlights areas where the curves overlap or diverge.

The horizontal axis displays the range of wind speed and solar radiation values, while the vertical axis indicates how frequently each value occurs. This visualization provides a clear understanding of how both factors vary within the dataset.

c. Finally, show these distributions in context of the rest of the variables by using a technique for comparing multiple distributions.

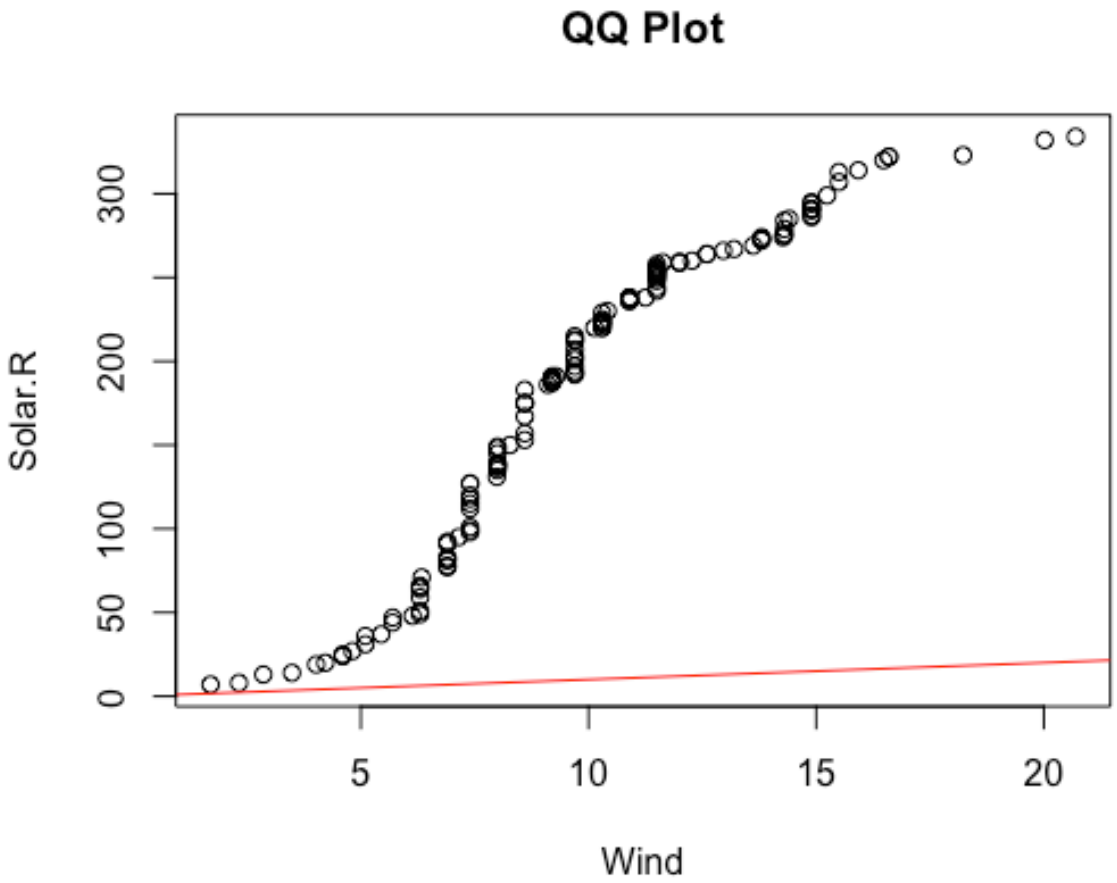
Note: you will need to transform the data in a particular way that we have studied. I it showed in the Tableau tutorial and in an R tutorial. Hint – you need to collapse the current variables into two: (1) stores the original variable name, and (2) stores the corresponding original value.


Density Plot of Air Quality Variables
Density
0.09
0.06
0.03
0.00
0
100
200
300
Value
Variable
Ozone
Solar.R
Wind
Temp
Day


Induvidual Density Plots of Air Quality Variables
Ozone
Solar.R
Wind
Temp
Day
Density
0.015
0.010
0.005
0.000
0.004
0.003
0.002
0.001
0.09
0.06
0.03
0
50
100
150
200
300
5
10
15
20
60
70
80
90
0.04
0.02
0.01
30
Value
Variable
Ozone
Solar.R
Wind
Temp
Day

The code restructures the data by pairing each variable with its corresponding value. It then generates a plot to visualize the distribution of these values. Different colors are used to represent each variable, while the height of the shaded areas indicates the frequency of the values. This approach makes it easy to compare the spread of values across variables, providing insight into their patterns.

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



The scatter plot offers a visual comparison of the distribution patterns of wind speed and solar radiation. If all points align perfectly along the diagonal, it indicates that both variables share identical distribution patterns. However, any deviations from this line reveal differences in their distributions.

For instance, points falling below the diagonal suggest that solar radiation is generally lower than wind speed, while points above the line indicate the opposite. This visualization helps assess whether the two variables exhibit similar or differing patterns.

In this case, the plot reveals a positive correlation, indicating that higher wind speeds are typically associated with increased solar radiation levels.