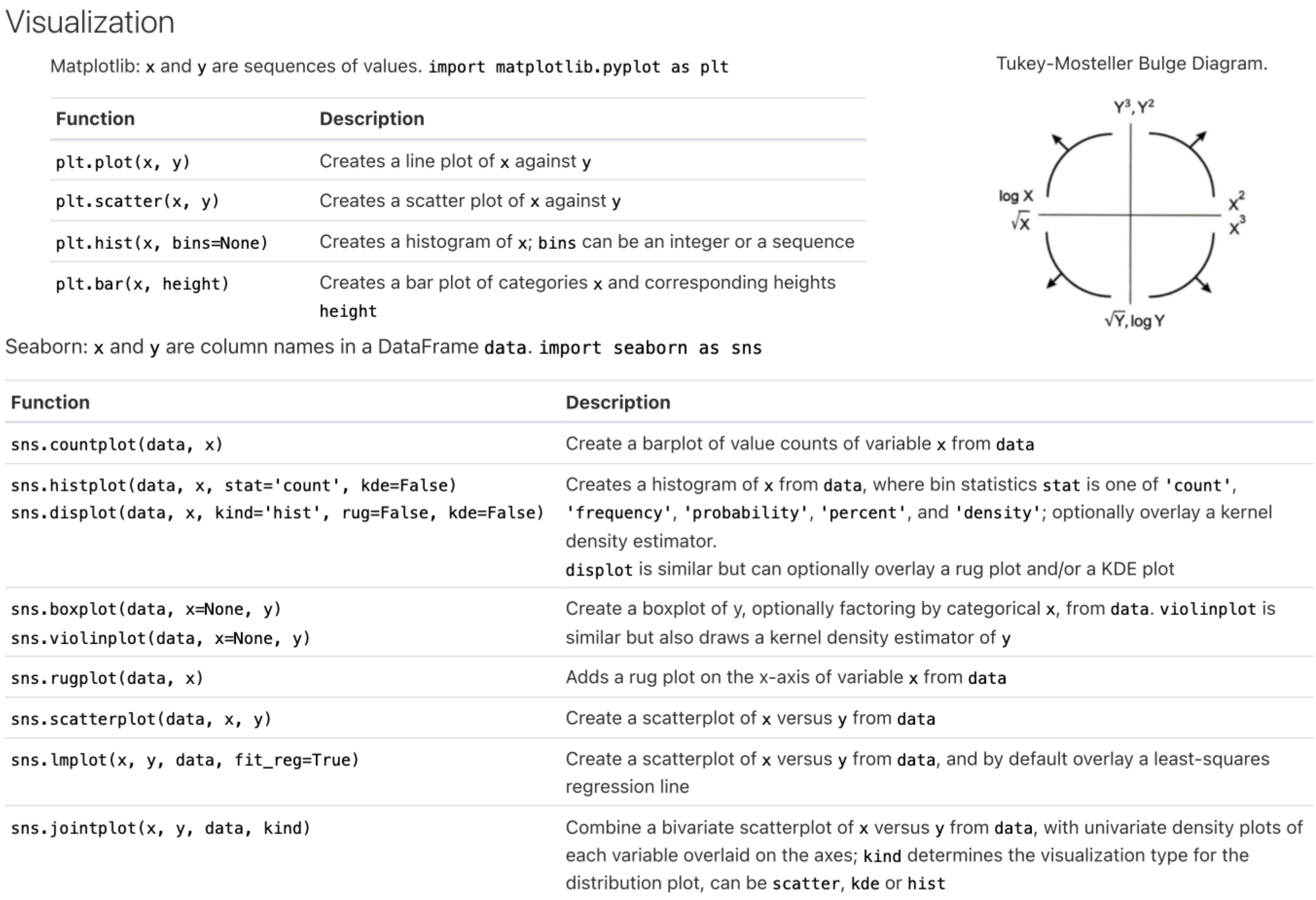
Data 100, Spring 2024

Discussion #4

Your TA will probably not cover all the problems. This is totally fine, the discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, labs, and homework.

Visualizations

Here's a snippet of the Visualization portion of the Fall 2023 Midterm Reference Sheet:



Visualization

Matplotlib: `x` and `y` are sequences of values. `import matplotlib.pyplot as plt`

Function	Description
`plt.plot(x, y)`	Creates a line plot of `x` against `y`
`plt.scatter(x, y)`	Creates a scatter plot of `x` against `y`
`plt.hist(x, bins=None)`	Creates a histogram of `x`; `bins` can be an integer or a sequence
`plt.bar(x, height)`	Creates a bar plot of categories `x` and corresponding heights `height`

Tukey-Mosteller Bulge Diagram.

Seaborn: `x` and `y` are column names in a DataFrame `data`. `import seaborn as sns`

Function	Description
`sns.countplot(data, x)`	Create a barplot of value counts of variable `x` from `data`

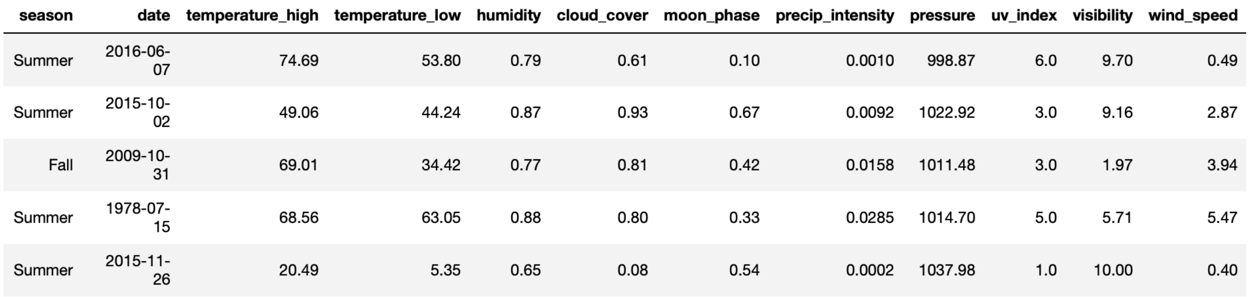
`sns.histplot(data, x, stat='count', kde=False)`
`sns.displot(data, x, kind='hist', rug=False, kde=False)`
	Creates a histogram of `x` from `data`, where bin statistics `stat` is one of `'count'`, `'frequency'`, `'probability'`, `'percent'`, and `'density'`; optionally overlay a kernel density estimator.
	`displot` is similar but can optionally overlay a rug plot and/or a KDE plot

`sns.boxplot(data, x=None, y)`
`sns.violinplot(data, x=None, y)`
	Create a boxplot of y, optionally factoring by categorical `x`, from `data`. `violinplot` is similar but also draws a kernel density estimator of `y`

`sns.rugplot(data, x)`	Adds a rug plot on the x-axis of variable `x` from `data`
`sns.scatterplot(data, x, y)`	Create a scatterplot of `x` versus `y` from `data`
`sns.lmplot(x, y, data, fit_reg=True)`	Create a scatterplot of `x` versus `y` from `data`, and by default overlay a least-squares regression line
`sns.jointplot(x, y, data, kind)`	Combine a bivariate scatterplot of `x` versus `y` from `data`, with univariate density plots of each variable overlaid on the axes; `kind` determines the visualization type for the distribution plot, can be `scatter`, `kde` or `hist`

Bigfoot is a mysterious ape-like creature that is said to live in North American forests. Most doubt its existence, but a passionate few swear that Bigfoot is real. In this discussion, you will be working with a dataset on Bigfoot sightings, visualizing variable distributions and combinations to understand better how/when/where Bigfoot is reportedly spotted and possibly either confirm or cast doubt on its existence. The Bigfoot data contains many variables about each reported Bigfoot spotting, including location information, weather, and moon phase.

This dataset is extremely messy, with observations missing many values across multiple columns. This is normally the case with data based on citizen reports (many do not fill out all required fields). For the purposes of this discussion, we will drop all observations with any missing values and some unneeded columns. However, note this is not a good practice, and you should almost never do this in real life!
Here are the first few entries of the `bigfoot` DataFrame:



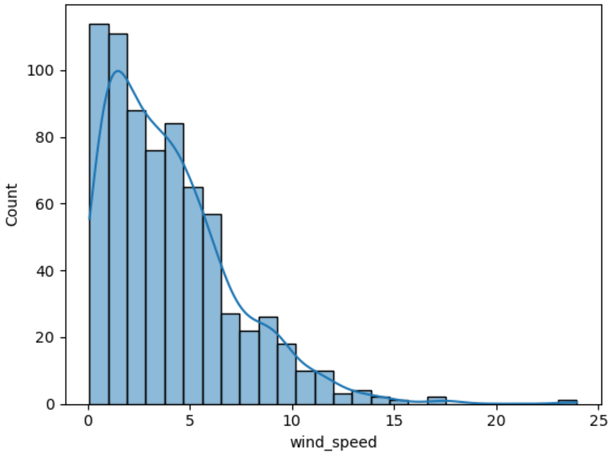
season	date	temperature_high	temperature_low	humidity	cloud_cover	moon_phase	precip_intensity	pressure	uv_index	visibility	wind_speed
Summer	2016-06-07	74.69	53.80	0.79	0.61	0.10	0.0010	998.87	6.0	9.70	0.49
Summer	2015-10-02	49.06	44.24	0.87	0.93	0.67	0.0092	1022.92	3.0	9.16	2.87
Fall	2009-10-31	69.01	34.42	0.77	0.81	0.42	0.0158	1011.48	3.0	1.97	3.94
Summer	1978-07-15	68.56	63.05	0.88	0.80	0.33	0.0285	1014.70	5.0	5.71	5.47
Summer	2015-11-26	20.49	5.35	0.65	0.08	0.54	0.0002	1037.98	1.0	10.00	0.40

1. Let's first look at distributions of individual quantitative variables. Let's say we're interested in `wind_speed`.

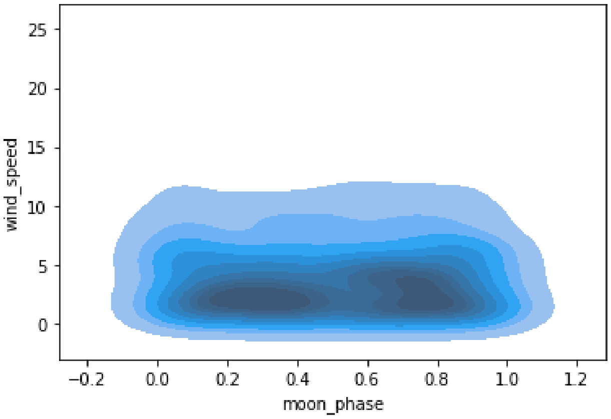
   (a) Which of the following are appropriate visualizations for plotting the distribution of a quantitative continuous variable?

   A. Pie chart
   B. Kernel Density Plot
   C. Scatter plot
   D. Box plot
   E. Histogram
   F. Hex plot

(b) Write a line of code that produces the visualization that depicts the variable's **distribution** (example shown below).

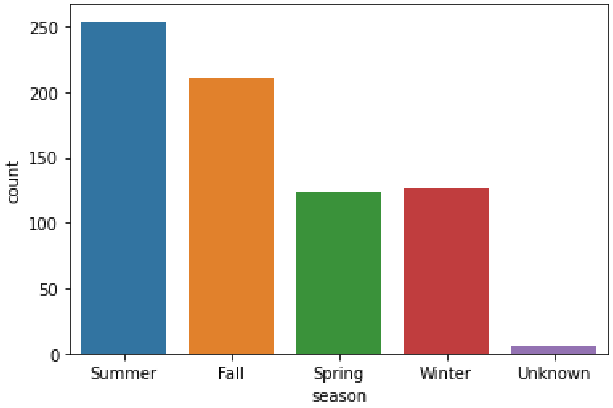


2. Now, let's see how two variables might relate to each other when Bigfoot is reportedly out. Fill in the function to produce a visualization that shows what combinations of values of `moon_phase` and `wind_speed` are most common when Bigfoot is spotted.

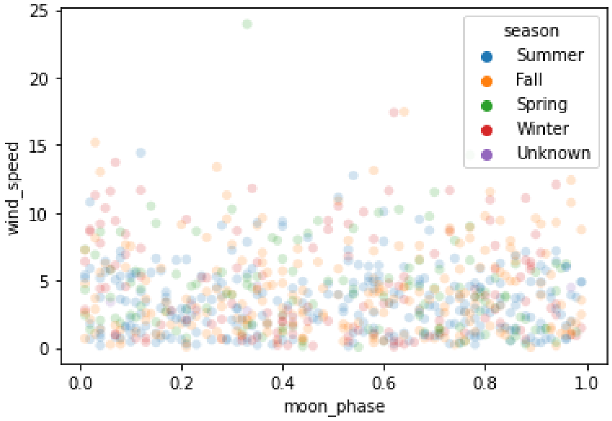


```
# type(variable1) == String
# type(variable2) == String
def plot(variable1,variable2):
        ____________________________
        ____________________________
        ____________________________
plot("moon_phase", "wind_speed")
```

3. Now, let's look at some qualitative variables. Write a line of code that produces a visualization that shows the distribution of Bigfoot sightings across the variable `season` (example shown below).



4. Produce a single visualization that showcases how the prevalence of bigfoot sightings at particular combinations of `moon_phase` and `wind_speed` vary across each season.
Hint: Think about color as the third information channel in the plot.



Kernel Density Estimation (KDE)

1. Kernel Density Estimation is used to estimate a probability density function (or density curve) from a set of data. A kernel with a bandwidth parameter α is placed on data observations x_i with $i \in \{1, ..., n\}$, and the density estimation is calculated by averaging all kernels. Below, Gaussian and Boxcar kernel equations are listed:

   - Gaussian Kernel: $K_\alpha(x, x_i) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-x_i)^2}{2\alpha^2}\right)$
   - Boxcar Kernel: $B_\alpha(x, x_i) = \begin{cases} \frac{1}{\alpha} & \text{if } -\frac{\alpha}{2} \leq x - x_i \leq \frac{\alpha}{2} \\ 0 & \text{else} \end{cases}$

   The KDE is calculated as follows: $f_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} K_\alpha(x, x_i)$.

   (a) Draw a KDE plot (by hand is fine) for data points `[1, 4, 8, 9]` using Gaussian Kernel and $\alpha = 1$. On the plot show x, x_i, α, and the KDE.

   (b) We wish to compare the results of KDE using a Gaussian kernel and a boxcar kernel. For $\alpha > 0$, which of the following statements is true? Choose all that apply.

      A. Decreasing α for a Gaussian kernel decreases the smoothness of the KDE.

      B. The Gaussian kernel is always better than the boxcar kernel for KDEs.

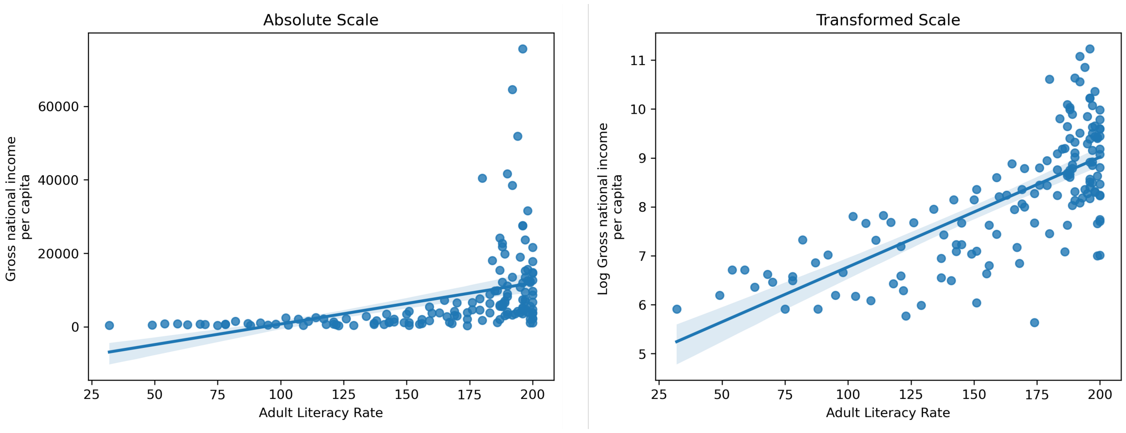
      C. Because the Gaussian kernel is smooth, we can safely use large α values for kernel density estimation without worrying about the actual distribution of data.

      D. The area under the boxcar kernel is 1, regardless of the value of α.

      E. None of the above.

Logarithmic Transformations

1. Ishani is a development economist interested in studying the relationship between literacy rates and gross national income in countries across the world. Originally, she plotted the data on a linear (absolute) scale, shown on the left. She noticed that the non-linear relationship between the variables with a lot of points clustered towards the larger values of literacy rate, so she consults the Tukey-Mosteller Bulge diagram and decides to do a $\log_{10}$ transformation of the y-axis, shown on the right. The solid blue line is a "line of best fit" (we'll formalize this later in the course).



(a) Instead of using the $\log_{10}$ transformation of the y-axis, what other transformations could Ishani have used to attempt to linearize the relatinsip between literacy rate (x) and gross national income per capita (y). Select all that apply.

A. $\log_e(y)$

B. 10^y

C. $\sqrt{x}$

D. x^2

E. y^2

(b) Let C and k be some constant values and x and y represent literacy rate and gross national income per capita, respectively. Based on the plots, which of the following best describes the pattern seen in the data?

○ A. $y = C + kx$ ○ B. $y = C \times 10^{kx}$ ○ C. $y = C + k\log_{10}(x)$ ○ D. $y = Cx^k$

(c) What parts of the plots could you use to make initial guesses on C and k?

(d) Ishani's friend, Yash, points to the solid line on the transformed plot and says "since this line is going up and to the right, we can say that, in general, the higher the literacy rate, the greater the gross national income per capita ". Is this a reasonable interpretation of the plot?

(e) Suppose that instead of plotting positive quantities, our data contained some zero and negative values. How can we reasonably apply a logarithmic transform to this data?