

Discussion #4 Solutions

Visualizations

Many of you have probably heard of Bigfoot before. It's 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 thereof to better understand how/when/where bigfoot is reportedly spotted, and possibly either confirm or cast doubt on its existence. The bigfoot data contains a ton of 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!

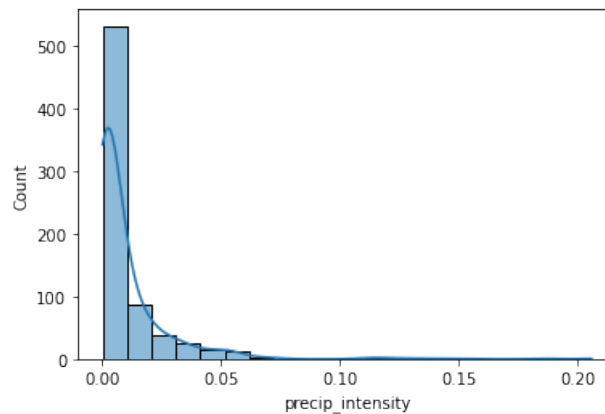
Here are the first few entries of the **bigfoot** table:

| 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 |

- Let's first look at distributions of individual quantitative variables. Let's say we're interested in `wind_speed`.
 - Which of the following are appropriate visualizations for plotting the distribution of a quantitative continuous variable?
 - Pie charts
 - Kernel Density Plot**
 - Scatter plot
 - Boxplot**
 - Histogram**
 - Hexplots

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

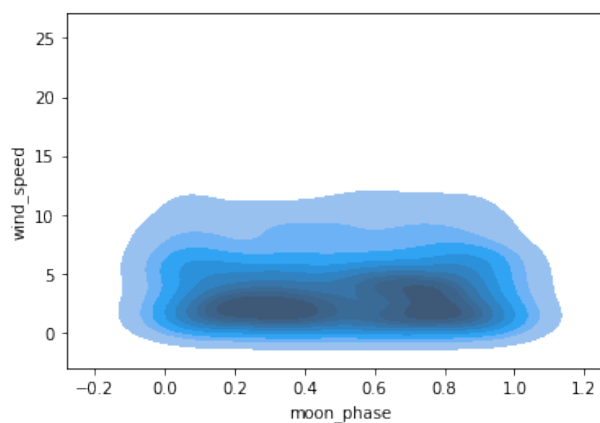
hint: Use `seaborn(sns.histplot)/matplotlib(plt.hist)`



Solution:

```
sns.histplot(data = bigfoot, x = "wind_speed")  
OR  
plt.hist(x=bigfoot["wind_speed"])
```

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. **hint:** Use `seaborn(sns.kdeplot)`

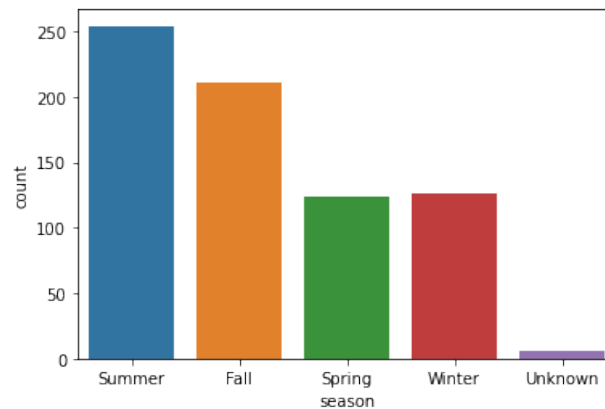


```
#type(variable1) == String
#type(variable2) == String
def plot(variable1,variable2):
    _____
    _____
    _____
plot("moon_phase", "wind_speed")
```

Solution:

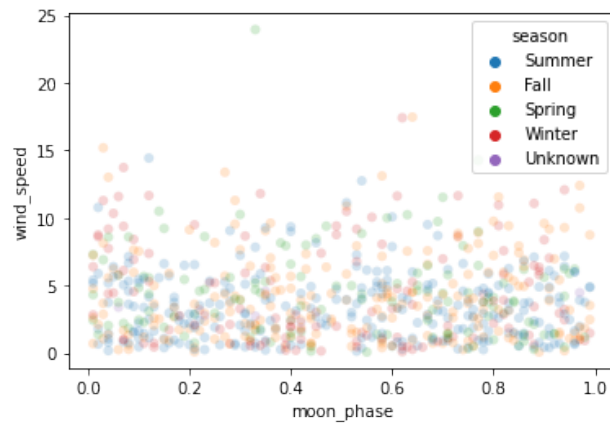
```
def plot(variable1,variable2):
    sns.kdeplot(data = bigfoot, x = variable1,
                y = variable2,
                shade = True)
```

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). **hint:** Use `seaborn(sns.countplot)/matplotlib(plt.bar)`.

**Solution:**

```
sns.countplot(data = bigfoot, x = 'season');
```

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 on the plot.

**Solution:**

```
sns.scatterplot(data = bigfoot,
                 x = 'moon_phase',
                 y = 'wind_speed',
                 hue = 'season',
                 alpha = 0.2);
```

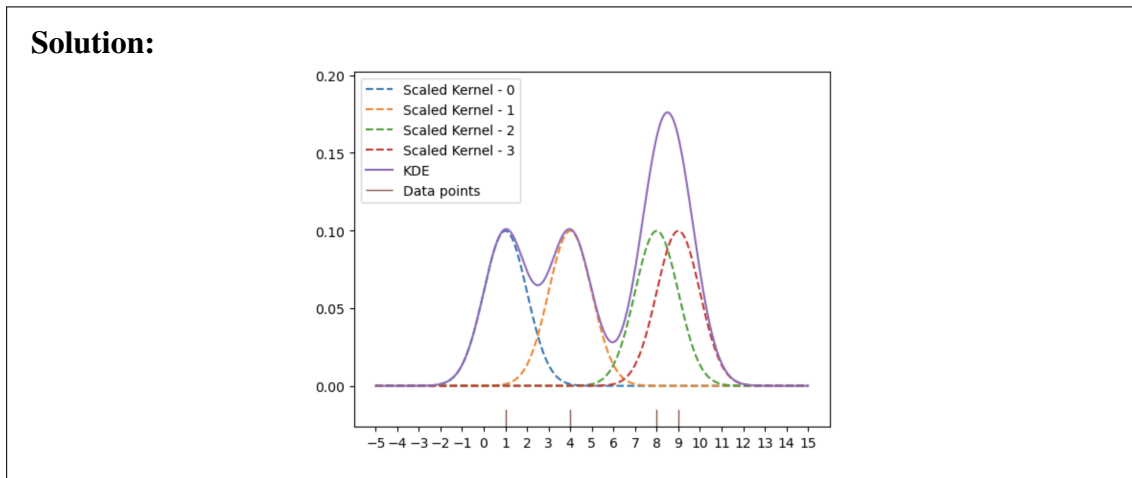
Kernel Density Estimation (KDE)

- 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 $i \in \{1, \dots, n\}$, 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: $f_\alpha(x) = \frac{1}{n} \sum_{i=1}^n K_\alpha(x, x_i)$.

- 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.



- 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.

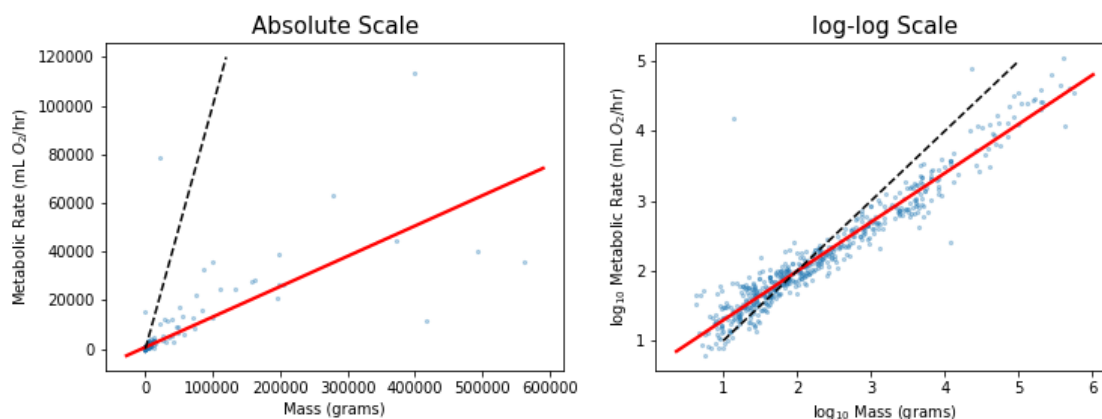
- A. Decreasing α for a Gaussian kernel decreases the smoothness of the KDE.
- B. The Gaussian kernel is always better than the boxcar kernel for KDEs.
- C. Because the Gaussian kernel is smooth, we can safely use large α values for kernel density estimation without worrying about the actual distribution of data.
- D. The area under the boxcar kernel is 1, regardless of the value of α .
- E. None of the above.

Solution:

- A. True.
- B. False; if the α values are not carefully selected for the Gaussian kernel, the boxcar kernel can provide a better kernel density estimate.
- C. False; if we set α too high, we potentially risk including too many points in our estimate, resulting in a flatter curve.
- D. True.

Logarithmic Transformations

1. One of your friends at a biology lab asks you to help them analyze panTHERIA, a database of mammals. They are interested in the relationship between mass, measured in grams, and metabolic rate ("energy expenditure"), measured by oxygen use per hour. Originally, they showed you the data on a linear (absolute) scale, shown on the left. You notice that the values on both axes vary over a large range with many data points clustered around the smaller values, so you suggest that they instead plot the data on a log-log scale, shown on the right. The solid red line is a "line of best fit" (we'll formalize this later in the course) while the black dashed line represents the identity line $y = x$.



- (a) Let C and k be some constant values and x and y represent mass and metabolic rate, respectively. Based on the plots, which of the following best describes the pattern seen in the data? **Reminder:** $\log(a \times b) = \log(a) + \log(b)$.

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

Solution: Starting with $y = Cx^k$, we can take the \log_{10} of both sides to find the relationship between $\log_{10}(y)$ and $\log_{10}(x)$.

$$\begin{aligned}\log_{10}(y) &= \log_{10}(Cx^k) \\ &= \log_{10}(C) + \log_{10}(x^k) \\ &= \log_{10}(C) + k \log_{10}(x)\end{aligned}$$

Thus, $\log_{10}(y)$ and $\log_{10}(x)$ are linearly related, which matches what the log-log plot shows above.

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

Solution:

- $C: 10^b$, where b is the y-intercept of the solid red line in the log-log plot.
- k : the slope of the solid red line log-log plot.

- (c) Your friend points to the solid line on the log-log plot and says "since this line is going up and to the right, we can say that, in general, the bigger a mammal is, the greater its metabolic rate". Is this a reasonable interpretation of the plot?

Solution: Yes, the observation is equivalent to saying that the slope is positive, which means increases in x correspond to increases in y .

- (d) 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?

Solution: We can add the magnitude (or absolute value) of the most negative number in our data, then add a small positive number (e.g., 1). As a concrete example, suppose we had the dataset: $-3, -2, 4$. The most negative number is -3 , so we add $|-3| = 3$ to all numbers to obtain: $0, 1, 7$. By adding a small positive number such as 1, we are able to take the \log of the numbers: $1, 2, 8$. Why is this last step necessary?