Your final project post should include:

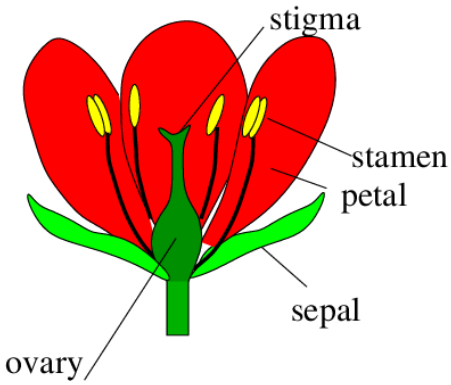
- A brief recap of your data, goals, and tasks, focusing on those that most directly influence your design
- Screenshots of and/or a link to your visualization implementation (see below for additional guidance)
- A summary of the key elements of your design and accompanying justification
- A discussion of your final evaluation approach, including the procedure, people recruited, and results. Note that, due to the difficulty of recruiting experts, you can use colleagues, friends, classmates, or family to evaluate your designs if experts or others from your target population are unavailable.
- A synthesis of your findings, including what elements of your approach worked well and what elements you would refine in future iterations.

The link for the code generate these plots：
https://github.com/carsonxie/vital_skills/tree/main/DS%20as%20a%20field%20Final%20Projects/Data%20Visualization

Final project report for Data visualization

1. Description of data, goals and task:

The data set used in this project is called the iris dataset(https://en.wikipedia.org/wiki/Iris_flower_data_set). The iris dataset is a classic data used in machine learning for multi-class classification problems. According to wikipedia it was first introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. This dataset contains 150 rows and 4 columns, the 4 columns represent iris flowers' petal width and length, and sepal width and length. (See below picture, sepal and petal are the green and red parts of a flower.) And there are 150 samples containing 3 types of iris in this dataset: etosa, versicolor and virginica.



And the goals of this project can we show visualization plots to our 3 participants who are not familiar to this dataset and let them answer following questions:

- Can we distinguish these 3 types of iris flower by sepal length/width and petal length/width?
- Among these 4 features, which one makes iris most distinguishable from others? Or say if you want to tell the difference between these 3 iris in a garden which of petal/sepal length/width you would use to make the decision?
- After seeing 2 types of plots in the same group, which one is easier to read?

2. **Link to visualization implementation: And some screen shots:**
   https://github.com/carsonxie/vital_skills/tree/main/DS%20as%20a%20field%20Final%20Projects/Data%20Visualization）

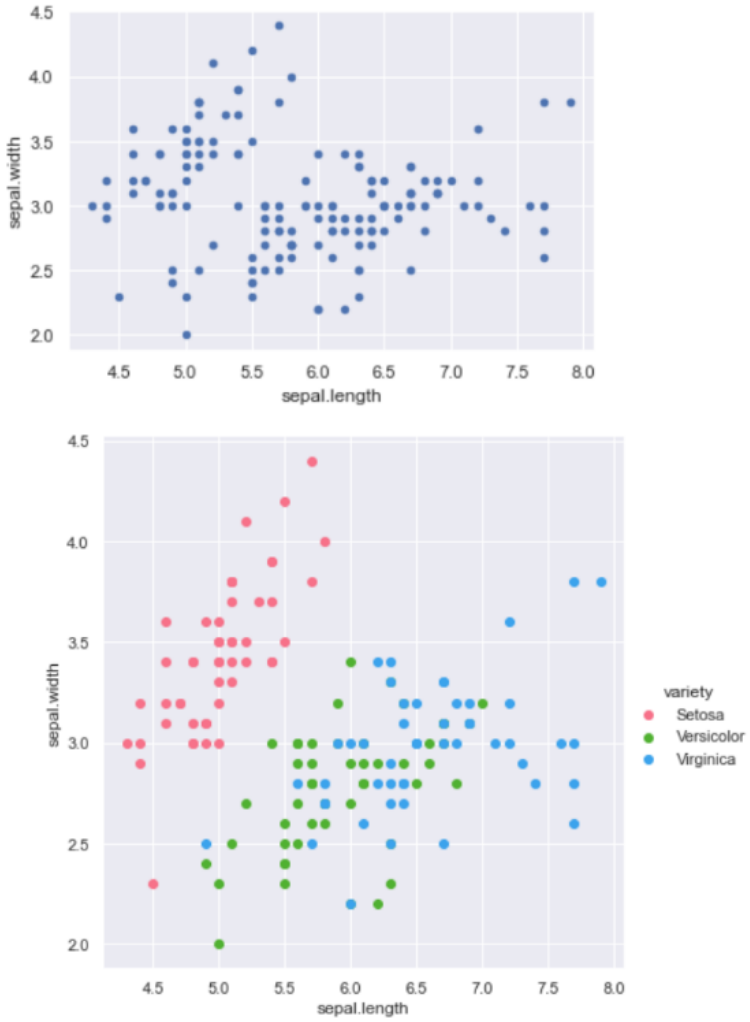


Fig 1. Species difference in sepal length vs width.

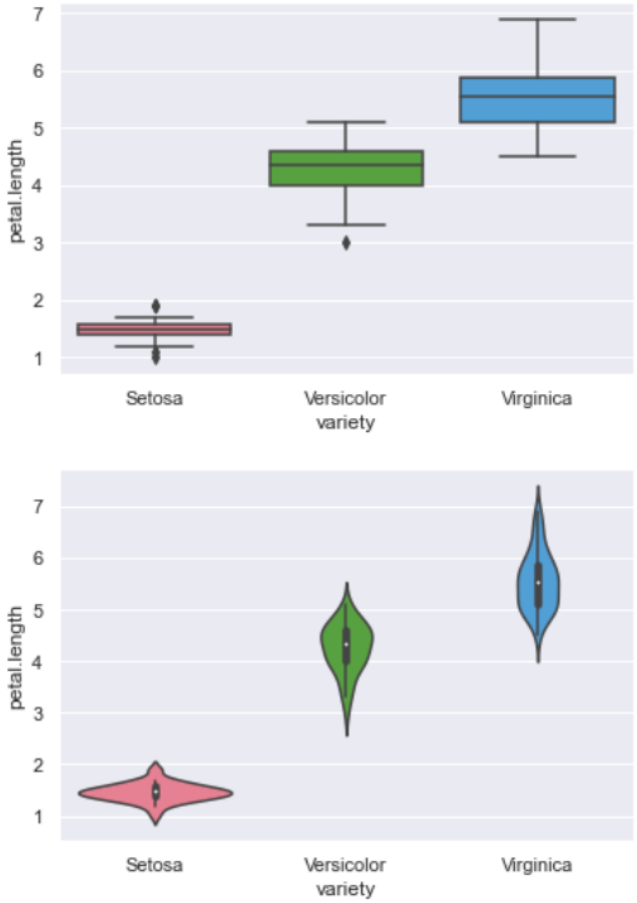


Fig 2. BoxPlot vs Violin Plot


sepal.length
sepal.width
petal.length
petal.width
variety
Setosa
Versicolor
Virginica


sepal.length
sepal.width
petal.length
petal.width
variety
Setosa
Versicolor
Virginica

Fig 3. Pair plot, shows Iris-setosa species is separated from the other two across all feature combinations

3. Summary of the key elements of your design and accompanying justification:

Color: follow the principles of good color choices discussed in class, in the first and third plot, using distinct colors, ensure semantically meaningful and roughly equally salient these 3 principles. Choosing categorical encodings because iris varieties are discrete variables.

Overview vs detail. Fig 1 is an overview of length/width of 3 varieties of iris, and fig 2 shows more detailed information about petal length of 3 kinds of iris including median, variance, range of the data by boxplot.

4. A discussion of your final evaluation approach, including the procedure, people recruited, and results:

For part 4, I separated this question into 4 parts and explained how I did this evaluation approach.

Set the target question: Before handing plots to the audience, I design 3 questions that they will need to answer after they see the plot to check if any valuable information were gained through these plots.

1. Can we distinguish these 3 types of iris flower by sepal length/width and petal length/width? (If respond Yes, need to provide reasons, like fig1, red dots looks more away from others, fig2 setosa iris have very different mean value of petal length from virginica iris)
2. Among these 4 features, which one makes iris most distinguishable from others? Or say if you want to tell the difference between these 3 iris in a garden which of petal/sepal length/width you would use to make the decision? (plot 3, look at the distribution plot and find which feature makes 3 color-distribution plot most separated from each other, in this case sepal width/length)
3. After seeing 2 types of plots in the same group, which one is easier to read?

People recruit to answer questions: I pick 3 people, one is my colleague who works as an accountant, second one is my mother with very little knowledge in quantitative analysis, and finally my 7 years old niece.

Measure and Procedure: Similar to an experiment but more casual providing them 3 plots and making sure they have no idea about this dataset. After showing them these plots, I will ask them the 3 questions above and what amount of info they can get from 3 plots. Also ask them how they feel about these plots after I explain the ideas behind each one, like confused/messy/easy to understand/so on.

Result:

- Color: distinct and meaningful choice of color are welcome for all 3 participants. Fig1 3 distinct colors are better than just 1 color for all 3 types of iris.
- Shape and type of plots better to be common. I personally think violin plots are great, not only providing what box plots can provide, but also showing how data is distributed in the horizontal direction. But my participants think the box plot is better, less confusing.
- Design different types of visualization based on your audience type. Are they your stakeholders? Clients know nothing about statistics, or a personal video/blog face to wider public. For example you are doing an academic report about data science, maybe the audience are all studying/working in this field, then you can make more complex plots and pack more information in one single visualization plot.

5. **A synthesis of your findings, including what elements of your approach worked well and what elements you would like to refine in the future.**

- Elements that worked well: as mentioned above, the choice of color, shape of some plots works well. Also the way we get feedback, like an insights based evaluation is effective, I get what I need to know about the pros and cons of plots provided to them. Based on the time they need on the plots, the number of insights, importance of insights also how these discovered insights correspond to the designer's purpose.
- Things to refine: I can think of one case related to work. If our team wants to make a data report to our clients, we can first make a prototype for team members with different backgrounds then collect feedback and brainstorm any refinement to make. One thing to refine is, this process is still slow in an actual work environment, so maybe we can build a “visualization prototype warehouse” for different types of datasets, audience. Including the R or Python function that generates the plots. When there is a new requirement, we can quickly use one of the prototypes and make some modifications to generate new reports.

And the jupyter notebook code I used:

Your final project post should include:

A brief recap of your data, goals, and tasks, focusing on those that most directly influence your design

Screenshots of and/or a link to your visualization implementation (see below for additional guidance)

A summary of the key elements of your design and accompanying justification

A discussion of your final evaluation approach, including the procedure, people recruited, and results. Note that, due to the difficulty of recruiting experts, you can use colleagues, friends, classmates, or family to evaluate your designs if experts or others from your target population are unavailable.

A synthesis of your findings, including what elements of your approach worked well and what elements you would refine in future iterations.

In [7]:
```
import pandas as pd

# data= pd.read_csv("WHR_2016.csv")
# data.head()
import warnings
warnings.filterwarnings("ignore")
```

In [5]:
```
from sklearn.datasets import load_iris
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
sns.set(color_codes=True)
# iris = load_iris(as_frame=True)

# df = pd.DataFrame(data=iris.data, columns=iris.feature_names)

# df['species'] = iris.target
iris= pd.read_csv("iris.csv")

iris.head()
```

Out[5]:

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [10]:
```
iris.target_names
```

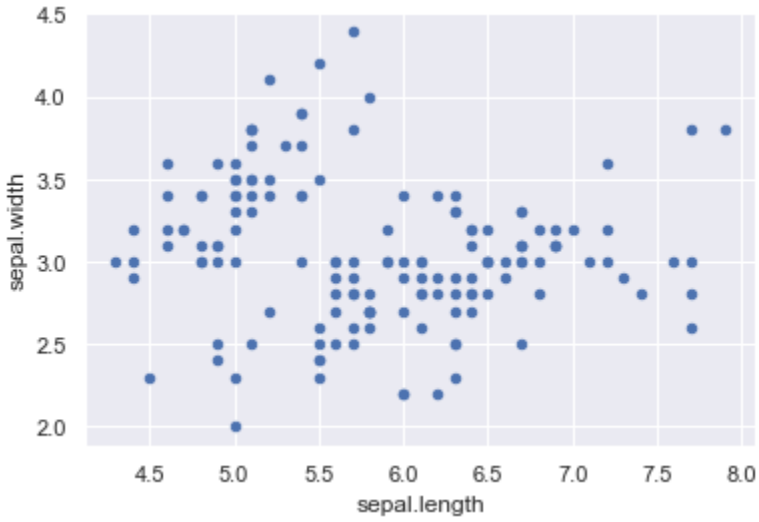
Out[10]:
```
array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

In [9]:
```
iris.plot(kind="scatter", x="sepal.length", y="sepal.width")
plt.show()
```



```
ooks like a single numeric RGB or RGBA sequence, which should be avoided as
```

```
value-mapping will have precedence in case its length matches with *x* & *y*.  Please use
the *color* keyword-argument or provide a 2D array with a single row if you intend to spec
ify the same RGB or RGBA value for all points.
```

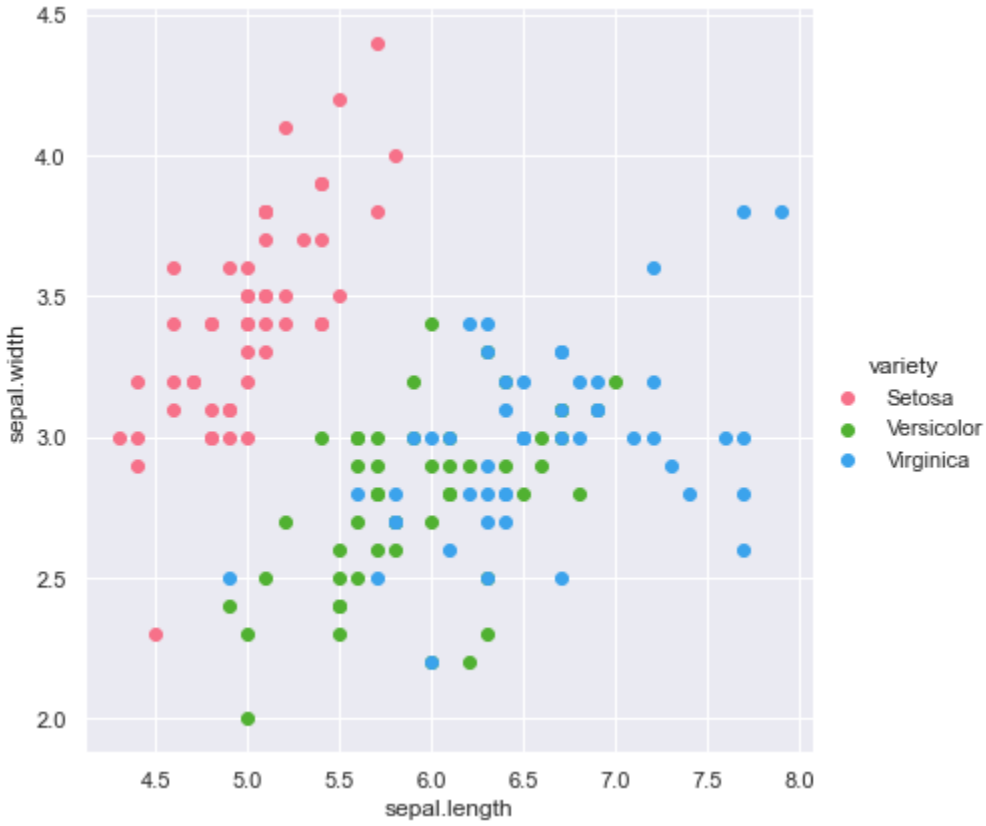


In [19]:

```
sns.FacetGrid(iris, hue="variety", palette="husl", size=6) \
   .map(plt.scatter, "sepal.length", "sepal.width") \
   .add_legend()
```

Out[19]:

```
<seaborn.axisgrid.FacetGrid at 0x25b1880bf10>
```

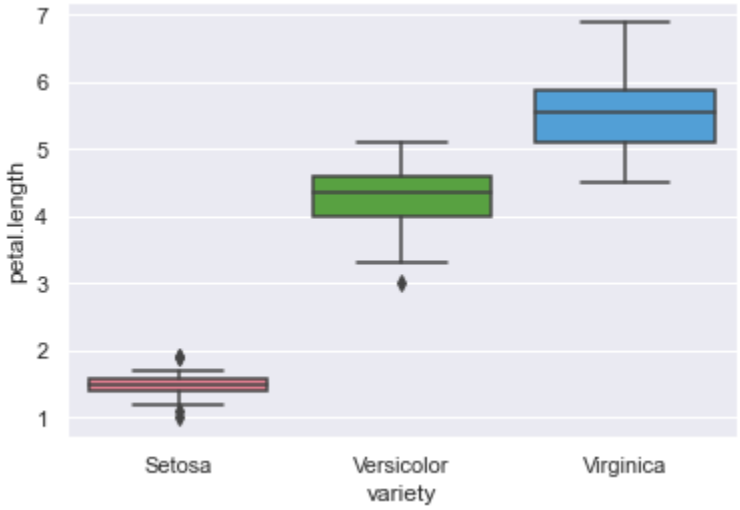


In [20]:

```
sns.boxplot(x="variety", y="petal.length", palette="husl", data=iris)
```

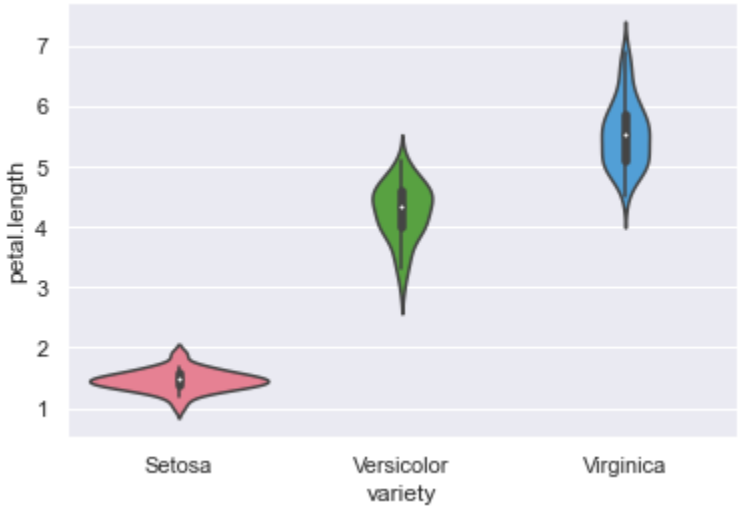
Out[20]:

```
<AxesSubplot:xlabel='variety', ylabel='petal.length'>
```



In [21]:
```
sns.violinplot(x="variety", y="petal.length", palette="husl", data=iris)
```

Out[21]:
```
<AxesSubplot:xlabel='variety', ylabel='petal.length'>
```



In [31]:
```
# Another useful seaborn plot is the pairplot, which shows the bivariate relation
# between each pair of features
#
# From the pairplot, we'll see that the Iris-setosa species is separataed from the other
# two across all feature combinations
#
# Note that we want to drop the ID variable because it has no correlation with any other
# variable. The ID number assigned to a particular observation has no bearing on the analy
# and would only mess up the plots

sns.pairplot(iris, hue="variety", palette="husl", size=3, diag_kind="hist")
```

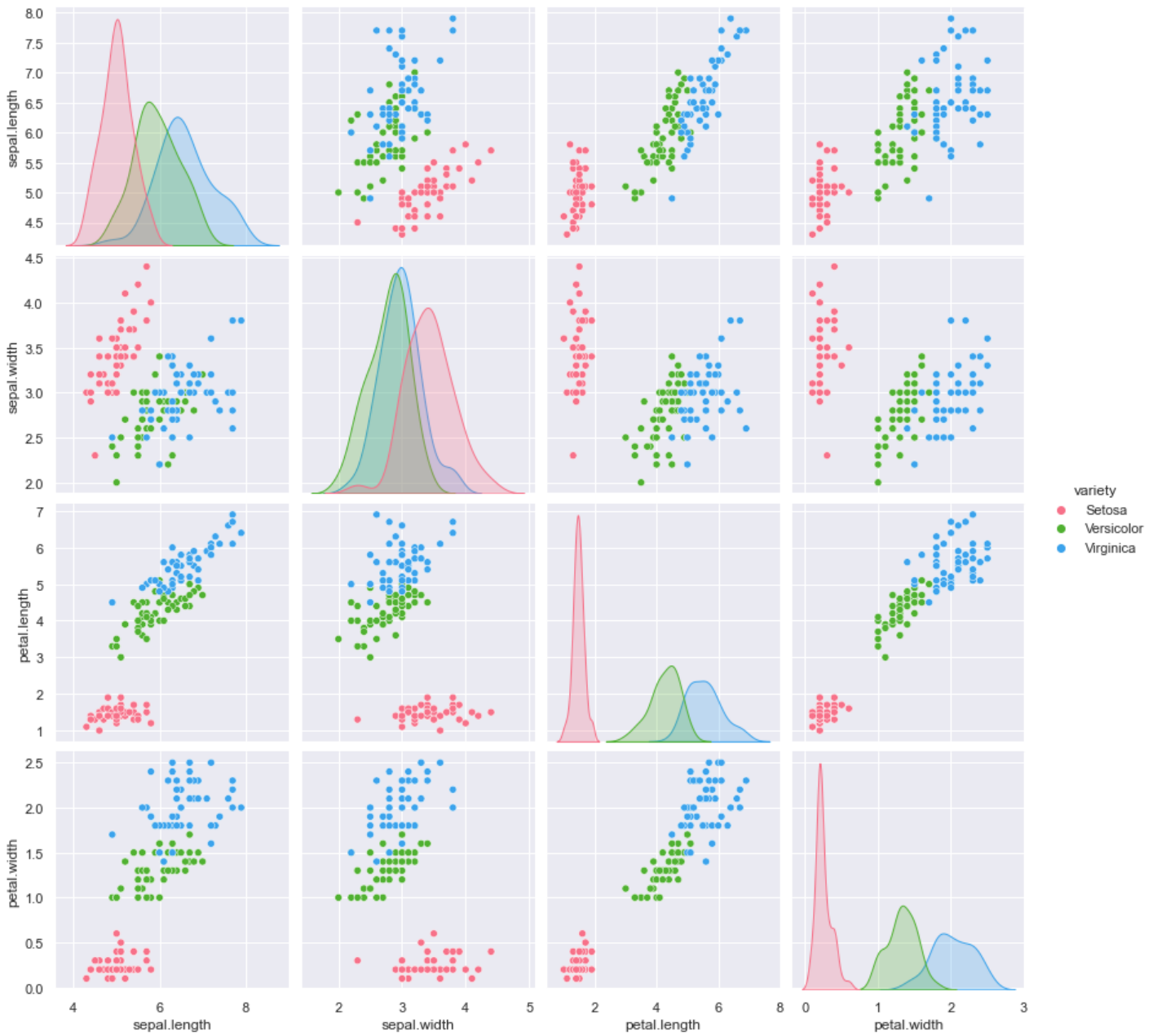
Out[31]:
```
<seaborn.axisgrid.PairGrid at 0x25b1c04eaf0>
```

In [29]:

```
# The diagonal elements in a pairplot show the histogram by default, as you can see above.
# We can update these elements to show other things, such as a kde
sns.pairplot(iris, hue="variety", palette="husl", size=3, diag_kind="kde")
```

Out[29]:

```
<seaborn.axisgrid.PairGrid at 0x25b199c04c0>
```



In [ ]: