Group Contributions Statement.

Our group members are Cody Lejang, Megan Tieu, and Arely Perez. All three of us wrote the data acquisition and preparation for our data. Arely created the first table with Culmen Length and Depth. Cody led the first figure which is a scatter plot, Arely led the second figure which is a histogram and Megan led the third figure which is a box plot. We each did our own explanations for our respective figures. Megan led the ideas for our figure selection process. We all worked on the models together and the inital code to set these up, Arely worked on our Logistic Regression model, Cody on our Random Forest model, and Megan on our Support Vector model. Cody led the ideas for our discussion portion and we all made sure to check each other’s work and made revisions to each other's code and explanations

```
import pandas as pd
import numpy as np
import urllib
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
#importing and accessing all of our tools

url =
"https://philchodrow.github.io/PIC16A/datasets/palmer_penguins.csv"
##we imported the data from the link

penguins = pd.read_csv(url)
#reading the data in our url link

train, test = train_test_split(penguins, test_size = 0.3, random_state
= 0)
from sklearn import preprocessing
def prep_penguins_data(data):

    """
    function called prep data
    making sure to take in the data from the url
    creating a dataframe
    """

    df = data.copy()
    le = preprocessing.LabelEncoder()
    df = df[df.Sex != '.']
    df = df.sort_values(by=['Sex'])
    X=penguins[["Culmen Length (mm)","Culmen Depth (mm)"]]
    y=penguins['Species']
    return(X, y)
X_train, y_train = prep_penguins_data(train)
X_test,  y_test  = prep_penguins_data(test)
## we're splitting the data into train and test groups where the test
group is 30% of the data and the train group is 70% of the data.
```

```
penguins
#printing our table
```

```
    studyName  Sample Number                              Species
Region  \
0     PAL0708              1  Adelie Penguin (Pygoscelis adeliae)
Anvers
1     PAL0708              2  Adelie Penguin (Pygoscelis adeliae)
Anvers
2     PAL0708              3  Adelie Penguin (Pygoscelis adeliae)
Anvers
3     PAL0708              4  Adelie Penguin (Pygoscelis adeliae)
Anvers
4     PAL0708              5  Adelie Penguin (Pygoscelis adeliae)
Anvers
..        ...            ...                                  ...
...
339   PAL0910            120    Gentoo penguin (Pygoscelis papua)
Anvers
340   PAL0910            121    Gentoo penguin (Pygoscelis papua)
Anvers
341   PAL0910            122    Gentoo penguin (Pygoscelis papua)
Anvers
342   PAL0910            123    Gentoo penguin (Pygoscelis papua)
Anvers
343   PAL0910            124    Gentoo penguin (Pygoscelis papua)
Anvers

        Island               Stage Individual ID Clutch Completion
Date Egg  \
0    Torgersen  Adult, 1 Egg Stage          N1A1               Yes
11/11/07
1    Torgersen  Adult, 1 Egg Stage          N1A2               Yes
11/11/07
2    Torgersen  Adult, 1 Egg Stage          N2A1               Yes
11/16/07
3    Torgersen  Adult, 1 Egg Stage          N2A2               Yes
11/16/07
4    Torgersen  Adult, 1 Egg Stage          N3A1               Yes
11/16/07
..         ...                 ...           ...               ...
...
339     Biscoe  Adult, 1 Egg Stage         N38A2                No
12/1/09
340     Biscoe  Adult, 1 Egg Stage         N39A1               Yes
11/22/09
341     Biscoe  Adult, 1 Egg Stage         N39A2               Yes
11/22/09
342     Biscoe  Adult, 1 Egg Stage         N43A1               Yes
```

```
11/22/09
343     Biscoe  Adult, 1 Egg Stage         N43A2               Yes
11/22/09

     Culmen Length (mm)  Culmen Depth (mm)  Flipper Length (mm)  \
0                  39.1               18.7                 181.0
1                  39.5               17.4                 186.0
2                  40.3               18.0                 195.0
3                   NaN                NaN                   NaN
4                  36.7               19.3                 193.0
..                  ...                ...                   ...
339                 NaN                NaN                   NaN
340                46.8               14.3                 215.0
341                50.4               15.7                 222.0
342                45.2               14.8                 212.0
343                49.9               16.1                 213.0

     Body Mass (g)     Sex  Delta 15 N (o/oo)  Delta 13 C (o/oo)  \
0           3750.0    MALE                NaN                NaN
1           3800.0  FEMALE            8.94956          -24.69454
2           3250.0  FEMALE            8.36821          -25.33302
3              NaN     NaN                NaN                NaN
4           3450.0  FEMALE            8.76651          -25.32426
..             ...     ...                ...                ...
339            NaN     NaN                NaN                NaN
340         4850.0  FEMALE            8.41151          -26.13832
341         5750.0    MALE            8.30166          -26.04117
342         5200.0  FEMALE            8.24246          -26.11969
343         5400.0    MALE            8.36390          -26.15531

                           Comments
0    Not enough blood for isotopes.
1                               NaN
2                               NaN
3                Adult not sampled.
4                               NaN
..                              ...
339                             NaN
340                             NaN
341                             NaN
342                             NaN
343                             NaN

[344 rows x 17 columns]
```

Here we have our penguins data set. Now that we can view it we are able to use specific columns and rows in order to predict the species of a penguin. We also are able to split the data into train and test groups where the test group is 30% of the data and the train group is 70% of the data. We also cleaned the data by making a duplicate data set and dropping the values under sex that were a "." value, because we are only looking for values that are male and female. We then translated that into number values using the preprocessing label encoder. After we sorted the values and then specified that we are only looking for specific column values for y and x.

```
penguins.groupby(["Species","Island"])[["Culmen Length (mm)", "Body
Mass (g)" ,"Culmen Depth (mm)"]].mean()
#based on our data table grouping by species island culmen length and
body mass
#only looking at the mean of these columns
```

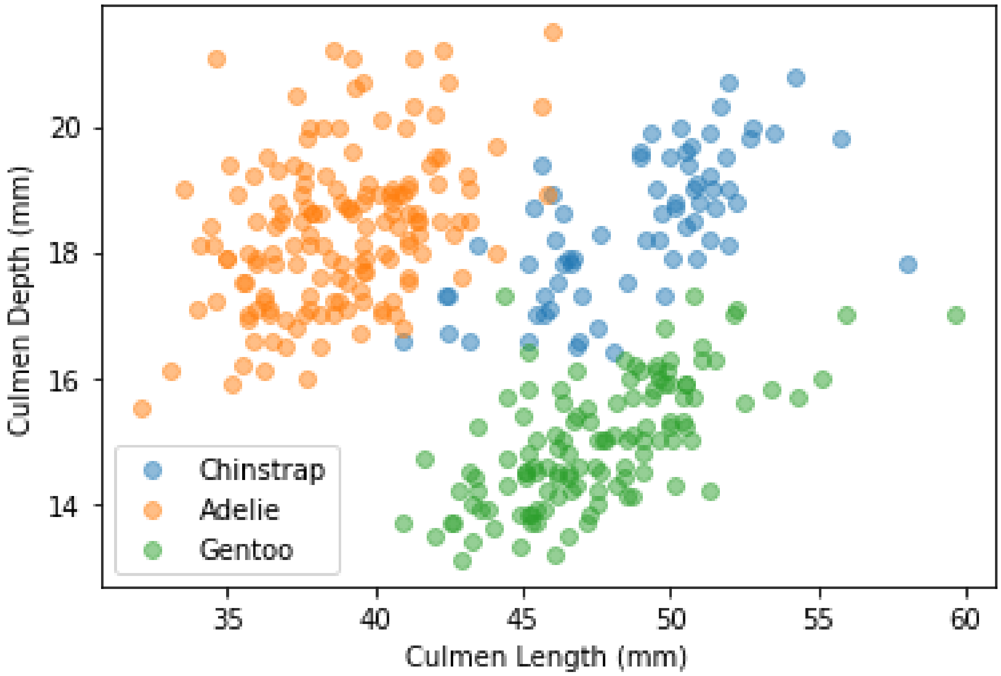
```
                                                     Culmen Length
(mm)  \
Species                                   Island

Adelie Penguin (Pygoscelis adeliae)       Biscoe
38.975000
                                          Dream
38.501786
                                          Torgersen
38.950980
Chinstrap penguin (Pygoscelis antarctica) Dream
48.833824
Gentoo penguin (Pygoscelis papua)         Biscoe
47.504878

                                                     Body Mass (g)  \
Species                                   Island
Adelie Penguin (Pygoscelis adeliae)       Biscoe       3709.659091
                                          Dream        3688.392857
                                          Torgersen    3706.372549
Chinstrap penguin (Pygoscelis antarctica) Dream        3733.088235
Gentoo penguin (Pygoscelis papua)         Biscoe       5076.016260

                                                     Culmen Depth (mm)

Species                                   Island

Adelie Penguin (Pygoscelis adeliae)       Biscoe             18.370455

                                          Dream              18.251786

                                          Torgersen          18.429412

Chinstrap penguin (Pygoscelis antarctica) Dream              18.420588

Gentoo penguin (Pygoscelis papua)         Biscoe             14.982114
```

This table shows the penguins grouped by species, and then broken down by island. The columns display the penguins culmen length, body mass, and culmen depth. It's apparent that the adelie penguins have a pretty consistent culmen length mean of around 38 mm across all islands, whereas the chinstrap and biscoe penguins' culmen length is a lot higher at around 47-

48mm. This means that culmen length could be a good identifier to distinguish whether the penguins belong to either adelie or chinstrap/gentoo. In terms of body mass, the mean for adelie and chinstrap penguins are pretty similar at around 3700 g, whereas the average for gentoo is a lot higher at 5000 g. This means that culmen length could be a good identifier to distinguish whether the penguins belong to either adelie/chinstrap or gentoo. If we were to observe culmen depth, we would see that the mean for adelie and chinstrap penguins are pretty similar at around 18 mm, whereas the average for gentoo is a lot lower at 15 . If we were to only choose two features out of the three, this table demonstrates to us that we should either choose a combination of culmen length and body mass or a combination of culmen length and culmen depth, because this would allow us to differentiate across all three penguin species. We shouldn’t choose a combination of body mass and culmen depth because that would only be useful in differentiating adelie/chinstrap vs gentoo penguins; there is too much overlap which prevents us from separating adlie and chinstrap.

```
fig, ax = plt.subplots(1)
#creating an empty figure
ax.set (xlabel= "Culmen Length (mm)", #length vs. depth graph
        ylabel= "Culmen Depth (mm)")
#creating a length vs. depth graph
#setting the x and y axis

species = set(penguins['Species'])
#only looking at our data for each species

for s in species:
  sub = penguins[penguins["Species"]==s]
  ax.scatter(sub['Culmen Length (mm)'], sub ['Culmen Depth (mm)'],
label=s.split(' ')[0], alpha=0.5)
#plotting the points for each species
#generate scatterplot color coded by species
#making sure points are a bit lighter

ax.legend()
#creating our legend and labels
```

```
<matplotlib.legend.Legend at 0x1d3c772f0a0>
```

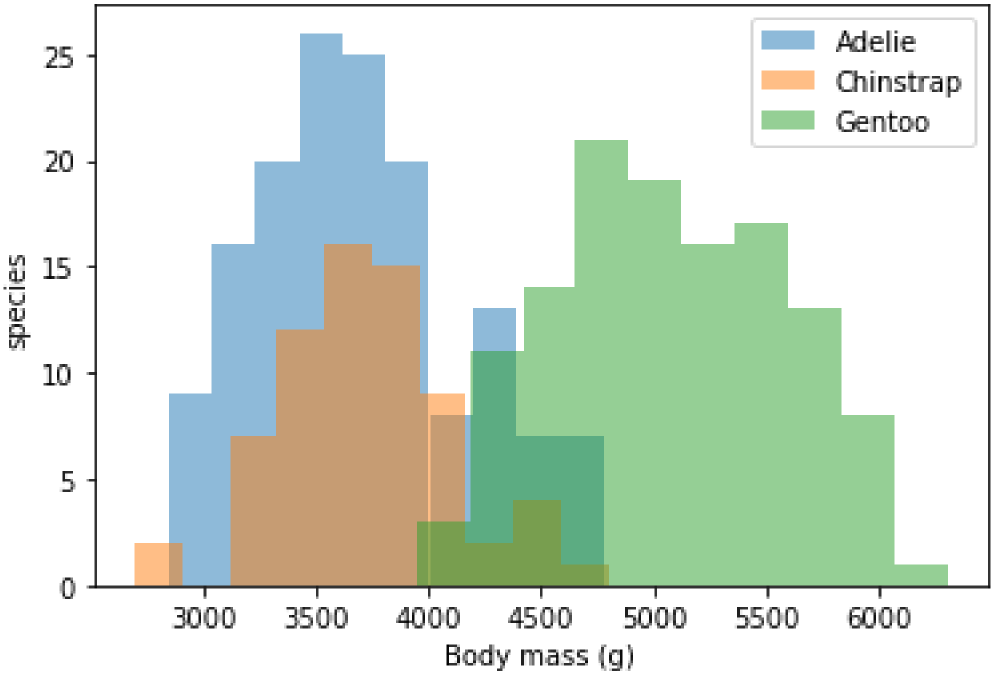


Explanation: The data shows a slight positive correlation between Culmen Length vs. Culmen Depth in a given species. The data shows that the Gentoo has the lowest culmen depth and the Adelie and Chinstrap pebguins have equally high culmen depths. The Adelie penguins have a shorter culmen length in comparison to Gentoo and Chinstrap penguins.

Next steps: Now that we have our first figure we can see the positive correlations between culmen length and depth for each species and can start to notice patterns between each species like Gentoo having the lowest culmen depth, adelie and chinstrap having equally high culmen depth and adelie penguins having a shorter culmen length in comparison to gentoo. Now we notice these patterns and can continue comparing other variables so that we can ultimatelty predict the species after noticing more patterns

```
fig,ax=plt.subplots(1)
#creating a figure


def plot_hist(df,colname,alpha):

    """
    function called plot_hist is created
    creating a histogram on the body mass
    for each species

    """


    ax.hist(df[colname],alpha=alpha)
    #creating a histogram
    ax.set(xlabel="Body mass (g)")
    #creating an x axis label

    ax.set(ylabel="species")
    #creating a y axis label

    ax.legend(('Adelie', 'Chinstrap','Gentoo'), loc='upper right');
    #creating a legend on the upper right of our figure


penguins.groupby("Species").apply(plot_hist,'Body Mass (g)',0.5)
#grouping by species and looking at the body mass for each species
#making sure bars are 0.5 lighter
```

```
Empty DataFrame
Columns: []
Index: []
```

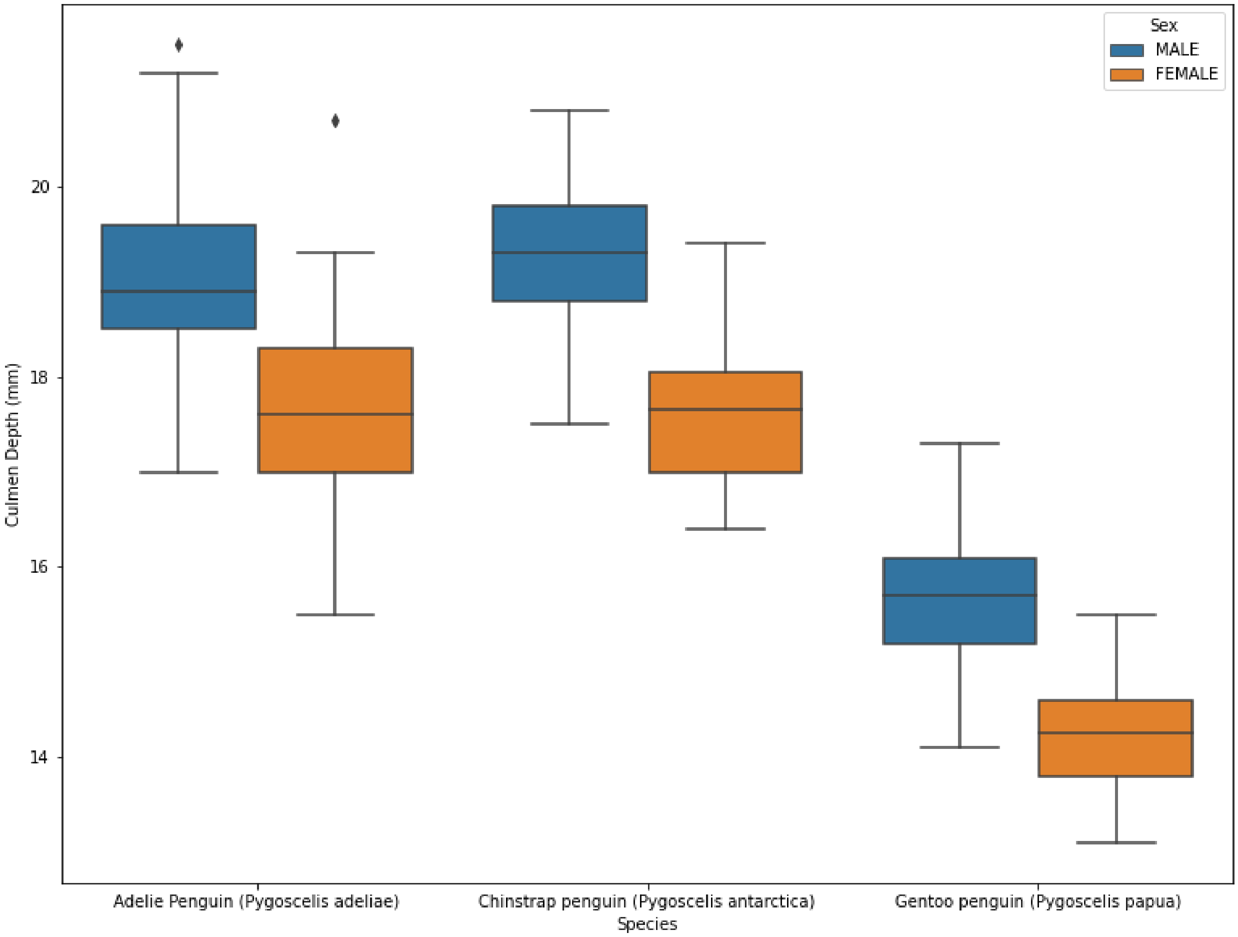


Explanation: This figure shows the correlation between all three species and their body mass in grams. Our three species are Gentoo, Adelie, and Chinstrap. Our bar graph has Body mass on our x axis and number of species pertaining to this body mass on the y axis. We can see that our blue bars have to do with the Adelie species, while the orange bars have to do with the Chinstrap species, and our green bars have to do with the Gentoo species. The higher you go on the y axis the more species there are with the specified body mass. In the histogram we can see that there is a very high amount of Adelie species that have a body mass of about 3500 grams as the graph shows approximately 25 of the penguins have said body mass. We can see that the Chinstrap species has about 17 penguins that have about 3750 grams of body mass, while the Gentoo penguins have about 20 of them which have approximately 4500 grams of body mass. Overall in this histogram we can see all three species in comparison to each other and see the most popular body masses for each species. The last thing I noticed was that Chinstrap has the penguins with the largest body mass while Adelie has the penguins with the smallest body mass.

Next steps: Now that we can look at two more variables like body mass and species we can continue to notice patterns between our species and choose two more variables to graph so we can ultimately predict a species.

```
import seaborn as sns
# box plot of penguins culmen depth grouped by species and seperated
by sex
from matplotlib import pyplot as plt
fig, ax = plt.subplots(figsize = (13,10))
df = penguins.copy()
 #copy penguins dataframe

df = df[df.Sex != '.']
#drop rows where sex is listed as '.'
sns.boxplot(x = "Species", y = "Culmen Depth (mm)",data = df, hue =
"Sex")
#setting the x and y axis labels and color coding based on sex
```

```
<AxesSubplot:xlabel='Species', ylabel='Culmen Depth (mm)'>
```



Explanation: This figure shows that across all three species, males have a longer Culmen Depth than females. Gentoo penguins have shorter Culmen Depths than both Adelie and Chinstrap penguins, regardless of if the penguins are male or female. From these observations, we can conclude that observing Culmen Depth can help determine which species the penguins belong to because typically male penguins with a culmen depth of greater than 18 mm belong to the adelie or chinstrap, whereas male penguins with a culmen depth less than 18 mm are more likely to be gentoo. For females, penguins with a culmen depth greater than 16 mm are likely to be adelie or chinstrap, whereas females who are under 16 mm are likely to be gentoo.

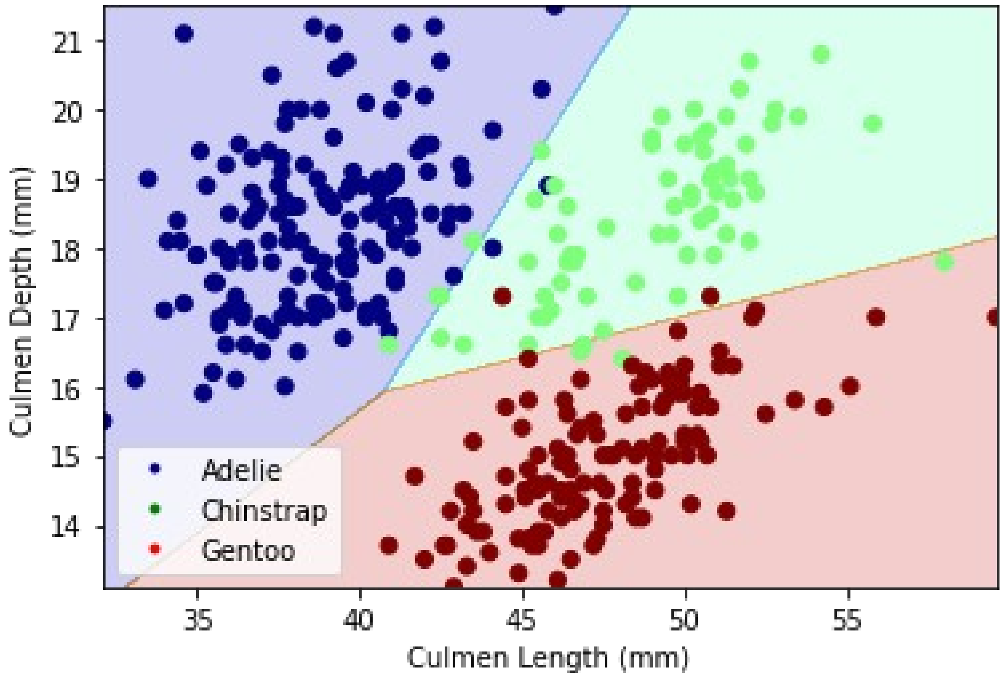
Next Steps: Now we can begin to notice more patterns specifically in regards to the difference between males and females. Given that we have looked at three figures with different pairs of variables we have enough information to go on to feature selection where we can decide which feature to choose based on the patterns we have seen above with our figures.

Feature Selection

Based on the 3 figures above, we decided to choose Culmen Length (mm) and Culmen Depth (mm). We decided these were the best features used to predict the species, because we rated it the best out of the 3 figures above. The first figure compared culmen length and depth, while highlighting the data in different colors depending on the species. It is clear that there are different sections for each species with enough space in between each species to make distinctions. Using culmen length and culmen depth would be the best predictors because there is a big enough contrast amongst the species to outline them into seperate boxes. While this is not entirely accurate because there are some data points that overlap into a species different than its true form, it is the best combination of features to predict the species. The second figure compared body mass (g) of the different species and while it is evident that there is some distinction between the three of them since they have different centers (mean and median), the difference in body mass across the three species is not large enough to easily make clear boundaries against each other. As for figure three, which compares culmen depth while seperating it into sex, it is evident that the center of the Adelie and the Chinstrap penguins are too similar, and therefore, are not reliable predictors for species. The figure is only useful in predicting if a penguin belongs to either Gentoo or Chinstrap/Adelie. It can't differentiate between 3 different species accurately, which is why figure three and it's features were not chosen for best feature selection. In conclusion, the features of figure two, which were culmen length and culmen depth, were the most accurate in predicting species because of how clearly grouped all the species are. It is easy to draw invisible boundaries on the figure that help us make accurate predictions of what species a penguin is.

```
from sklearn import preprocessing

#recode the labels
le=preprocessing.LabelEncoder()
penguins["Species"]=le.fit_transform(penguins["Species"])

penguins=penguins.dropna(subset=["Culmen Length (mm)","Culmen Depth
(mm)"])

X=penguins[["Culmen Length (mm)","Culmen Depth (mm)"]]
y=penguins['Species']

def plot_regions(c,X,y):
    """
    function called plot_regions is created
    so that we can

    """
    c.fit(X,y)

    x0=X["Culmen Length (mm)"]
    x1=X["Culmen Depth (mm)"]
    #setting X0 and X1 as new objects
    #each equal to the data corresponing in the culmen length and
depth column

    grid_x=np.linspace(x0.min(),x0.max(),501)
    grid_y=np.linspace(x1.min(),x1.max(),501)
    #fixing up the row and column line space

    xx,yy=np.meshgrid(grid_x,grid_y)
    np.shape(xx),np.shape(yy)
    #getting the shape of our x and y axis

    XX=xx.ravel()
    YY=yy.ravel()
    #looking at an array


    p=c.predict(np.c_[XX,YY])
    #is is the new object equal to our predictions

    p=p.reshape(xx.shape)
    #reshaping our new object p

    fig,ax=plt.subplots(1)
    #creating an empty figure with no plots yet


    ax.contourf(xx,yy,p,cmap="jet",alpha=.2)
    #plot the decision regions

    ax.scatter(x0,x1,c=y,cmap="jet")
    #setting a scatter of our points

    ax.set(xlabel="Culmen Length (mm)",ylabel="Culmen Depth (mm)")
    #labeling our x and y axis

    custom = [Line2D([], [], marker='.', color='darkblue',
linestyle='None'),
    Line2D([], [], marker='.', color='green', linestyle='None'),
    Line2D([], [], marker='.', color='red', linestyle='None')]
    ax.legend(custom, ['Adelie',
    'Chinstrap',
    'Gentoo'], loc='lower left')
    #legend created with color coded species

from sklearn.linear_model import LogisticRegression
from matplotlib.lines import Line2D
LR=LogisticRegression()
#creating our logistic regression
model=plot_regions(LR,X,y)
#plotting our regions

from sklearn.model_selection import cross_val_score


cv_scores=cross_val_score(LR,X,y,cv=5)
#cross validating our scores
cv_scores.mean()
#getting the mean of every score
```

```
C:\Users\arace\anaconda3\lib\site-packages\sklearn\base.py:450:
UserWarning: X does not have valid feature names, but
LogisticRegression was fitted with feature names
  warnings.warn(

0.9618925831202046
```

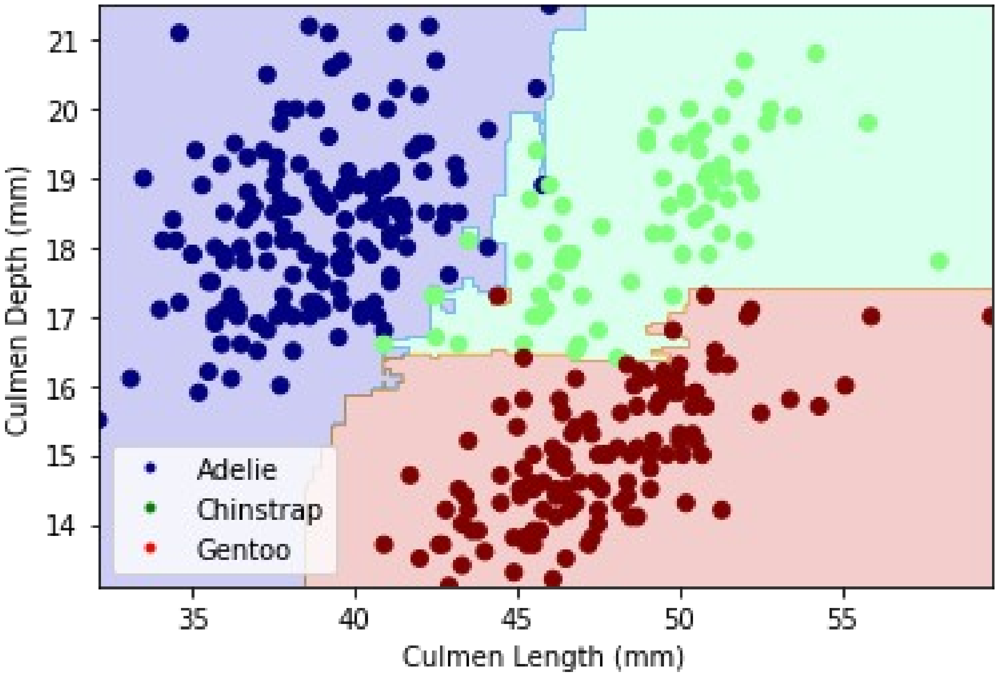


Our first model is a logistic regression model. It has a cross value score of 0.961 and it compares Culmen Length and Depth for each species.

```
from sklearn.ensemble import RandomForestClassifier
RF=RandomForestClassifier()
#creating random forest

plot_regions(RF,X,y)
#plotting our regions

from sklearn.model_selection import cross_val_score
from sklearn import tree

T=tree.DecisionTreeClassifier(max_depth=3)

cv_scores=cross_val_score(RF,X,y,cv=5)
#cross validation scores
cv_scores.mean()
#mean of all of the cross validation scores
```

```
C:\Users\arace\anaconda3\lib\site-packages\sklearn\base.py:450:
UserWarning: X does not have valid feature names, but
```

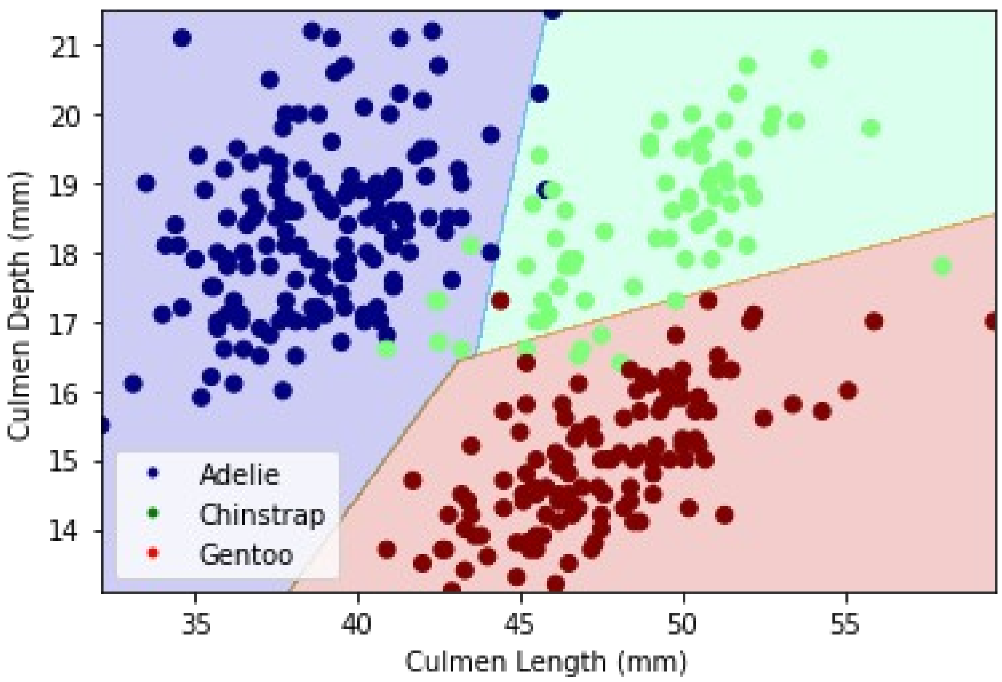
```
RandomForestClassifier was fitted with feature names
  warnings.warn(

0.9647911338448424
```



Our second model is a random forest model. It has a cross value score of 0.964 and it compares Culmen Length and Depth for each species.

```
from sklearn import svm

SVM=svm.SVC()
plot_regions(SVM,X,y)
from sklearn.model_selection import cross_val_score
cv_scores=cross_val_score(SVM,X,y,cv=5)
cv_scores.mean()
```

```
C:\Users\arace\anaconda3\lib\site-packages\sklearn\base.py:450:
UserWarning: X does not have valid feature names, but SVC was fitted
with feature names
  warnings.warn(

0.9531116794543906
```



Our final model is a support vector machine. It has a cross value score of 0.953 and it compares Culmen Length and Depth for each species.

Discussion. Describe the overall performance of your models, state which combination of model and features (measurements) you recommend. Discuss how the model could be improved if more or different data were available. **

Model Performance

The cross value scores of the Logistic Regression model is roughly .961. The Random Forest classifier model's cross value score is approximately .964. The SVM model's score is lowest at .953. Based on cross value scores, the Logistic Regression and Random Forest classifier models are more accurate than the SVM model.

Model/Feature Recommendations

We decided to use the Culmen Length vs. Culmen Depth relationship to generate a species predicting model. The features were selected through analysis of measures of center and overall trends in the dataset. This relationship had a clear distinction, in terms of spread, between each of the species. The body mass vs. species relationship was not selected because there was too little of a difference in body mass across each of the species. In the final figure involving species vs culmen depth with sex as a qualitative variable, it was determined that the center of the Adelie and the Chinstrap penguins are too similar. Consequently, this relationship would be a poor choice for predicting species.

All machine learning models have a high cross value score and therefore would be accurate predictors for species based on the two quantitative variables/features we selected.

Introduction of more data

The introduction of more data would bolster the accuracy of the model. With additional data, there is mroe training data, leading to a greater amount of accuracy. If different data was given, there is a possibility that there are two different variables that are an even better predictor of species.