Palmer Archipelago (Antarctica) Penguins Dataset

Source

I found this dataset ok Kaggle here: https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data It is originally collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Key attributes/ Dimensions

I will be using the penguin size dataset. There are several attributes for this dataset. Some key attributes are:

- species: penguin species (Chinstrap, Adélie, or Gentoo)
- flipper_length_mm: flipper length (mm)
- body_mass_g: body mass (g)

Goals and Tasks

1. Why is a task pursued? (goal)
   - I am pursuing the penguin dataset as a fun project to understand altair visualization and in the process learn about penguins. I would like to understand the relations between different penguin species, and identify patterns in their physical attributes. I can also explore the distribution of penguins across different locations.
2. How is a task conducted? (means)
   - I will conduct the task with different visualizations such as charts, and graphs.
3. What does a task seek to learn about the data? (characteristics)
   - A task may seek to learn about various characteristics of the data such as the distribution of penguins, the relationship between their physical attributes (body mass, flipper dimensions, beak dimensions).
4. Where does the task operate? (target data)
   - The target data of this task using the penguin dataset can be found here: https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data. This includes all the information I need about Penguin species and their physical attributes.
5. When is the task performed? (workflow)
   - The workflow of this task may involve data cleaning and preparation, perhaps some exploratory data analysis to identify patterns and relationships and finally visualizations to communicate insights.
6. Who is executing the task? (roles)
   - The roles involved in task execution could be data analysts, data scientists, researchers studying climate change or environmental impacts on animals in Antarctica. It could also be biologists or ecologists.

Import Pandas and Dataset

In [1]:

```
import pandas as pd
import altair as alt

penguins_df = pd.read_csv("data/penguins_size.csv")
penguins_df.head()
```

Out[1]:

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

In [2]:

```
# check if there are non-null values in columns
penguins_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   culmen_length_mm   342 non-null    float64
 3   culmen_depth_mm    342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                334 non-null    object 
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

In [3]:

```
# for this project, I will drop all na's
penguins_df = penguins_df.dropna()
```

In [4]:

```
# check for unique values in the sex column:
penguins_df['sex'].unique()
```

Out[4]:

```
array(['MALE', 'FEMALE', '.'], dtype=object)
```

In [5]:

```
# drop rows containing "." value for sex
penguins_df.drop(penguins_df.loc[penguins_df['sex'] == '.'].index, inplace=True)
```

In [6]:

```
penguins_df['sex'].unique()
```

Out[6]:

```
array(['MALE', 'FEMALE'], dtype=object)
```

In [7]:

```
penguins_df.head()
```

Out[7]:

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Adelie	Torgersen	39.3	20.6	190.0	3650.0	MALE

In [8]:

```
# now the dataset should not have any non-null values
penguins_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 333 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            333 non-null    object 
 1   island             333 non-null    object 
 2   culmen_length_mm   333 non-null    float64
 3   culmen_depth_mm    333 non-null    float64
 4   flipper_length_mm  333 non-null    float64
 5   body_mass_g        333 non-null    float64
 6   sex                333 non-null    object 
dtypes: float64(4), object(3)
memory usage: 20.8+ KB
```

Visualizations

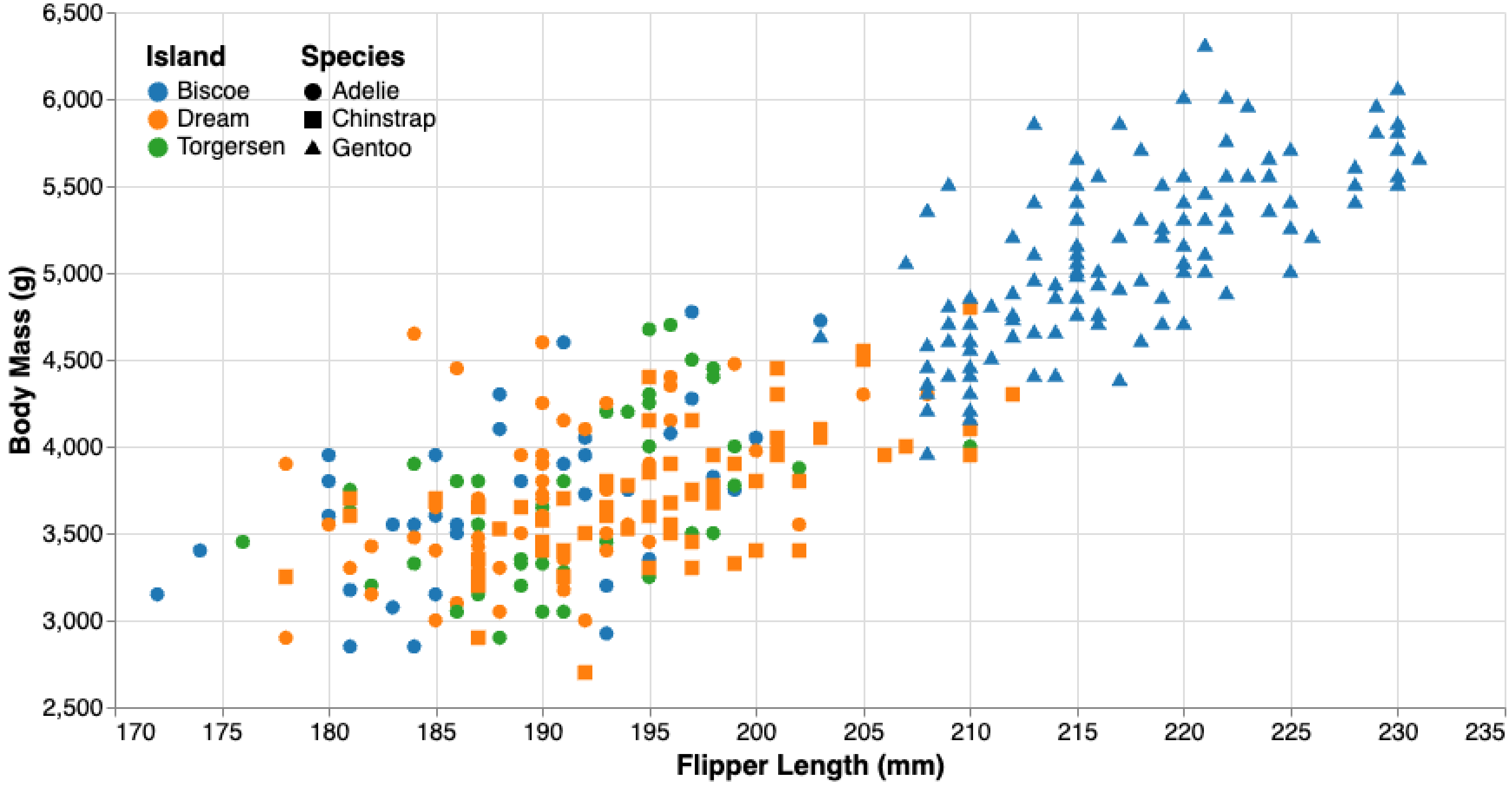
Going back to my goal, I will try to understand the relationship between a penguin's physical characteristics and how they differ for differet species.

In [9]:

```
# Define the color palette
colors = alt.Scale(domain=['Biscoe', 'Dream', 'Torgersen'],
                   range=['#1f77b4', '#ff7f0e', '#2ca02c'])

# Define the species dropdown selector
dropdown = alt.binding_select(options=penguins_df['species'].unique(), name='Select a species:')
selection = alt.selection(type='single', fields=['species'], bind=dropdown)

# Create the scatter plot
scatter = alt.Chart(penguins_df, title='Relationship between Flipper Length and Body Mass of Penguins across Different Species and Islands').mark_point(size=40).encode(
    x=alt.X('flipper_length_mm', scale=alt.Scale(domain=[170, 235]), title='Flipper Length (mm)'),
    y=alt.Y('body_mass_g', scale=alt.Scale(domain=[2500, 6500]), title='Body Mass (g)'),
    color=alt.Color('island:N', scale=colors, legend=alt.Legend(title='Island')),
    shape=alt.Shape('species:N', scale=alt.Scale(range=['circle', 'square', 'triangle']), legend=alt.Legend(title='Species')),
    fill=alt.Fill('island:N', scale=colors),
    tooltip=['island', 'body_mass_g'],
    opacity=alt.condition(selection, alt.value(1), alt.value(.2))
).add_selection(selection)

# Format the chart
chart = scatter.properties(
    width=800,
    height=400
).configure_axis(
    labelFontSize=14,
    titleFontSize=16
).configure_legend(
    titleFontSize=16,
    labelFontSize=14,
    orient='top-left'
).configure_title(
    fontSize=18,
    fontWeight='bold')


chart.configure_view(stroke=None)
```

Out[9]:

Relationship between Flipper Length and Body Mass of Penguins across Different Species and Islands



Select a species: Adelie

From the plot above, it can be noted that as the Flipper length increaes, the body mass also increases. This makes sense because flippers are essentially the wings of the penguin and the larger the wings, the greater the body mass. Furthermore, it can be seen that the Species Gentoo has the largest Flipper Length and Body Mass while the species Adelie has the smallest Flipper Length. Further differences can be seen across different islands with Biscoe mostly having the species Gentoo and Dream mostly having Chinstrap. The Adelie species are spread across the islands Biscoe and Dream.

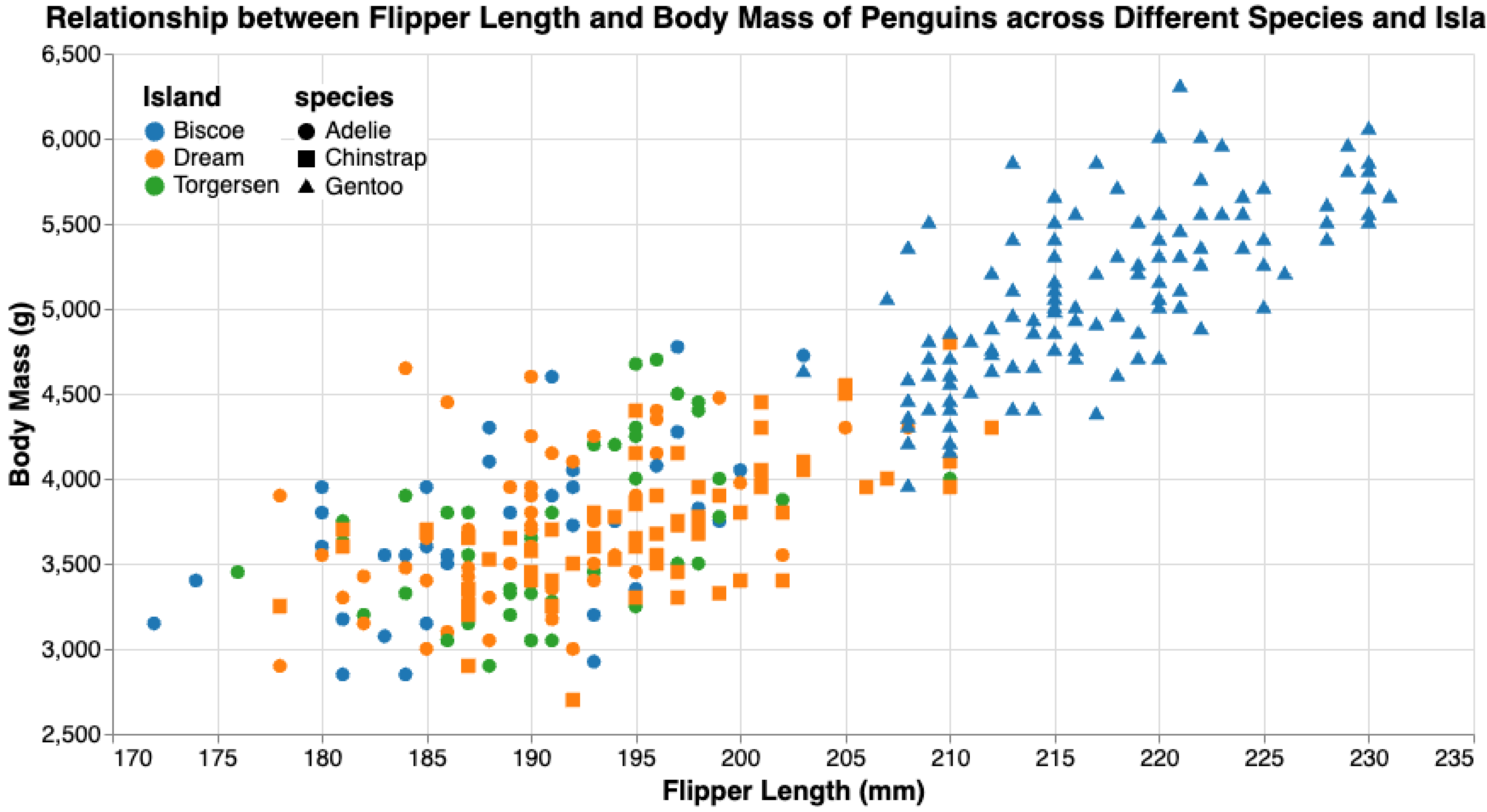
Visualizing Penguin Culmen Dimensions

I will use similar methodology as the previous plot to understand the relationships between Culmen dimensions and differnt penguins.

In [10]:

```
# Create the scatter plot
scatter = alt.Chart(penguins_df, title='Relationship between Flipper Length and Body Mass of Penguins across Different Species and Islands').mark_point(size=40).encode(
    x=alt.X('flipper_length_mm', scale=alt.Scale(domain=[170, 235]), title='Flipper Length (mm)'),
    y=alt.Y('body_mass_g', scale=alt.Scale(domain=[2500, 6500]), title='Body Mass (g)'),
    color=alt.Color('island:N', scale=colors, legend=alt.Legend(title='Island')),
    shape=alt.Shape('species:N', scale=alt.Scale(range=['circle', 'square', 'triangle'])),
    fill=alt.Fill('island:N', scale=colors),
    tooltip=['island', 'body_mass_g'],
    opacity=alt.condition(selection, alt.value(1), alt.value(.2))
).add_selection(selection)

# Format the chart
chart = scatter.properties(
    width=800,
    height=400
).configure_axis(
    labelFontSize=14,
    titleFontSize=16
).configure_legend(
    titleFontSize=16,
    labelFontSize=14,
    orient='top-left',
    fillColor='#FFFFFF',
).configure_title(
    fontSize=18,
    fontWeight='bold')


chart.configure_view(stroke=None)
```

Out[10]:



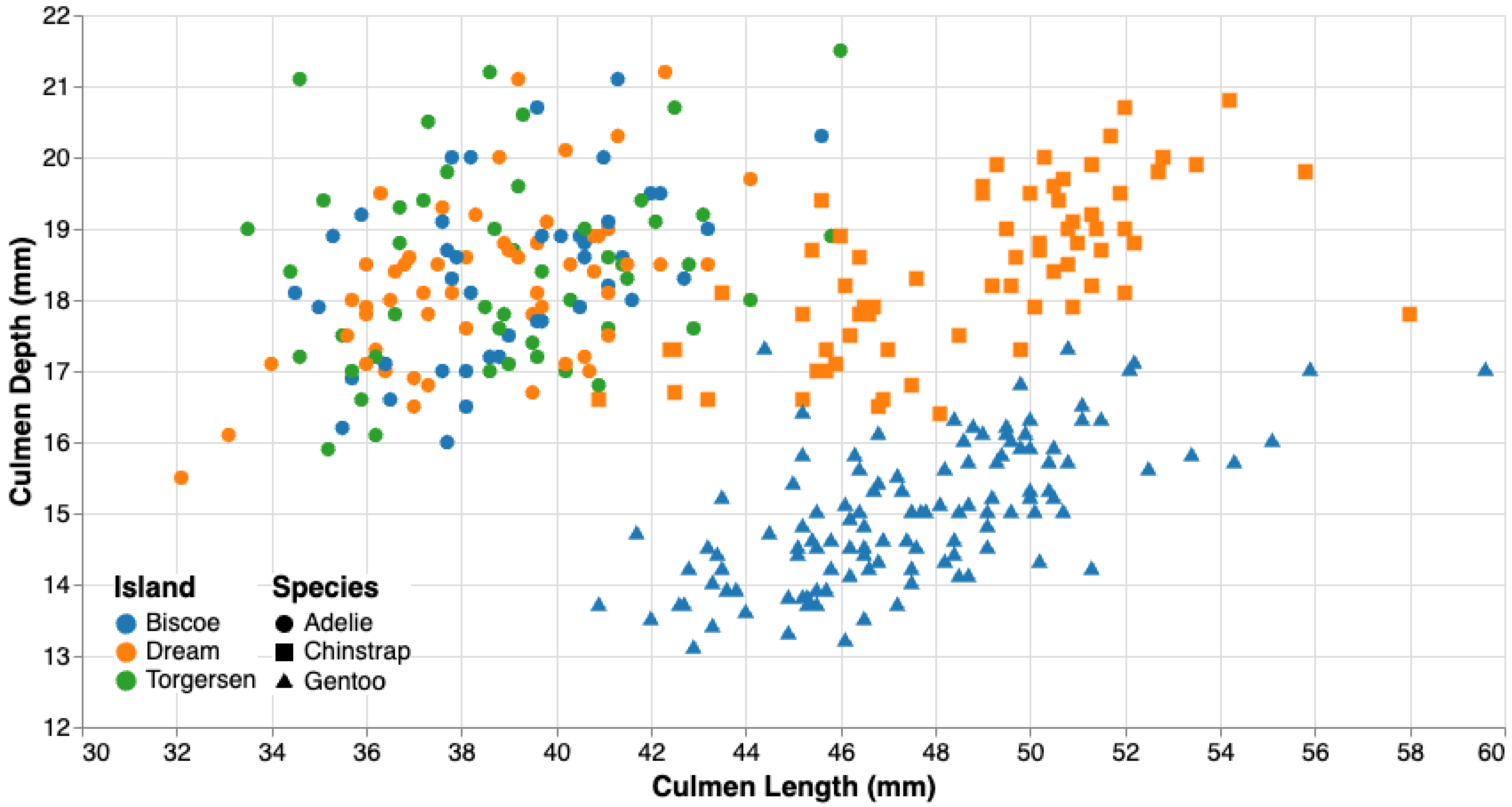
Select a species: Adelie

In [11]:

```
dropdown = alt.binding_select (options=penguins_df["species"].unique(), name="Select a species:")
selection = alt.selection(type='single', fields=['species'], bind=dropdown)

scatter = alt.Chart(penguins_df, title='Relationship between Culmen length and depth of Penguins across Different Species and Islands').mark_point(size=40).encode(
    x=alt.X('culmen_length_mm', scale=alt.Scale(domain=[30, 60]), title='Culmen Length (mm)'),
    y=alt.Y('culmen_depth_mm', scale=alt.Scale(domain=[12, 22]), title='Culmen Depth (mm)'),
    color=alt.Color('island', scale=colors, legend=alt.Legend(title='Island')),
    shape=alt.Shape('species', scale=alt.Scale(range=['circle', 'square', 'triangle']), legend=alt.Legend(title='Species')),
    fill=alt.Fill('island:N', scale=colors),
    tooltip=['island', 'body_mass_g'],
    opacity=alt.condition(selection, alt.value(1), alt.value(.2))
).add_selection(selection)

chart = scatter.properties(
    width=800,
    height=400
).configure_axis(
    labelFontSize=14,
    titleFontSize=16
).configure_legend(
    titleFontSize=16,
    labelFontSize=14,
    orient='bottom-left',
    fillColor='#FFFFFF',
).configure_title(
    fontSize=18,
    fontWeight='bold')


chart.configure_view(stroke=None)
```

Out[11]:

Relationship between Culmen length and depth of Penguins across Different Species and Islands



Select a species: Adelie

From the plot above, it can be noted that there is a mild linear relationship between Culmen length and Culmen depth. If you select different species, you will notice that:

- For the Gentoo species, as the Culmen length increases, so does the Culmen depth. The Gentoo species is mostly found on the Biscoe island.
- For the Chinstrap species, as the Culmen length increases, do does the delth. The Gentoo species is mostly found on the Dream island.
- For the Adelie species, as the culmen length increases, the depth does not necessarily increase but there is a somewhat upwards trend. The Adelie species is found across all different islands but they are mostly on the Torgersen island.

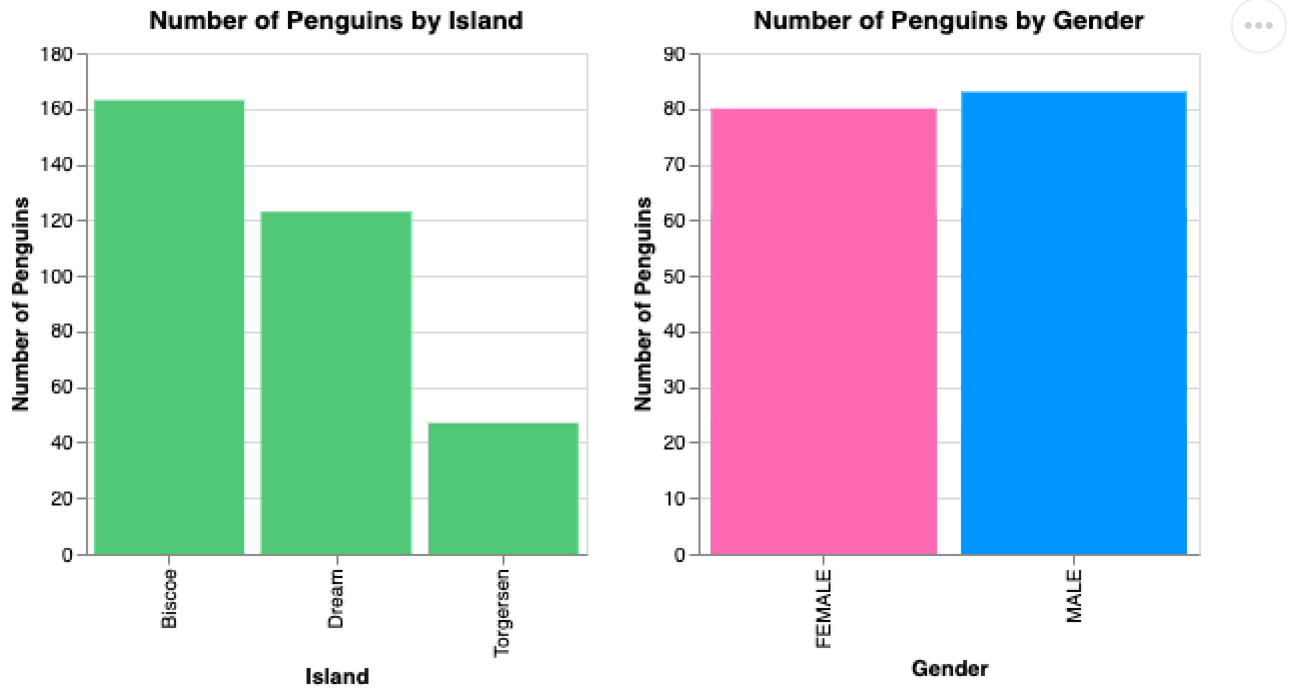
Penguin Populations

In [12]:

```
# Implement filtering using dynamic queries.
selection = alt.selection(type="multi", fields=["island"])
```

```
# Create a container for two different views
base = alt.Chart(penguins_df).properties(width=500, height=250)

# Specify our overview chart
overview = alt.Chart(penguins_df, title="Number of Penguins by Island").mark_bar().encode(
    x=alt.X('island:N', title='Island'),
    y=alt.Y('count()', title='Number of Penguins'),
    color=alt.condition(selection, alt.value("#50C878"), alt.value("lightgrey")),
    tooltip=["island:N", "count()"],
).add_selection(selection).properties(height=250, width=250)

# Create a detail chart
detail = base.mark_bar().encode(
    x=alt.Y('sex:N', title="Gender", axis=alt.Axis(labels=True, ticks=True)),
    y=alt.Y('count()', title='Number of Penguins'),
    tooltip=["island:N", "count()"],
    color=alt.Color('sex:N', scale=alt.Scale(range=["#FF69B4", "#0096FF"]), legend=None)
).properties(title="Number of Penguins by Gender", height=250, width=250).transform_filter(selection)

overview | detail
```

Out[12]:



The bar char above illustrates the population of Penguins across different islands. By clicking each island, you can see the female and male population on each island on the right side plot. Island Biscoe has the greatest population, Dream has the second most population, and Torgersen has the least.

All islands have a fairly erual number of males and females, with very small differences.

Violin Plot

In [13]:

```
# Find Unique penguins
penguins_df['species'].unique()
```

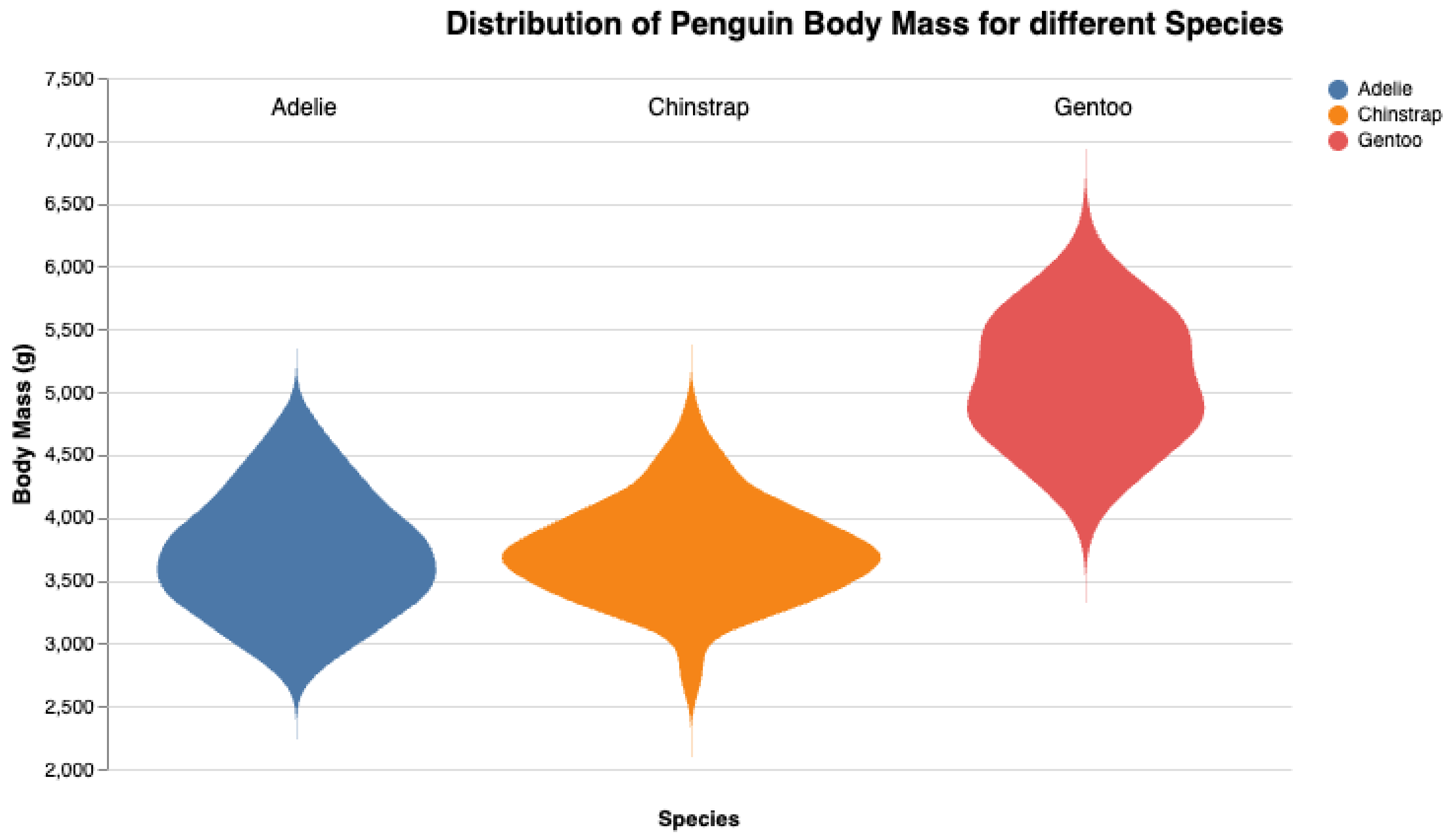
Out[13]:

```
array(['Adelie', 'Chinstrap', 'Gentoo'], dtype=object)
```

In [14]:

```
alt.Chart(penguins_df).transform_density(
    'body_mass_g',
    as_=['body_mass_g', 'density'],
    extent=[2000, 7500],
    groupby=['species']
).mark_area(orient='horizontal').encode(
    y=alt.Y('body_mass_g:Q', title="Body Mass (g)", axis=alt.Axis(titleFontSize=12, labelFontSize=10)),
    color=alt.Color('species:N', legend=alt.Legend(title=None, titleFontSize=12, labelFontSize=10)),
    x=alt.X(
        'density:Q',
        stack='center',
        impute=None,
        title=None,
        axis=alt.Axis(titleFontSize=12, labelFontSize=12, labels=False, values=[0],grid=False, ticks=False),
    ),
    column=alt.Column(
        'species:N',
         sort=['Adelie', 'Chinstrap', 'Gentoo'],
        header=alt.Header(
            titleOrient='bottom',
            labelOrient='bottom',
            title="Species",
            labelFontSize=12,
            labelPadding=10
        ),
    )
).properties(
    width=200, height=350
).configure_facet(
    spacing=0
).configure_view(
    stroke=None
).properties(
    title={
        "text": "Distribution of Penguin Body Mass for different Species",
        "dx": 70,
        "fontSize": 16,
        "fontWeight": "bold",
        "anchor": "middle",
        "color": "black",
        "subtitleFontSize": 12,
        "subtitleColor": "gray",
        "subtitlePadding": 10,
        "dy": -10
    }
)
```

Out[14]:



The plot above illustrates the body mass of different penguin species using a violin plot. A violin plot is useful to understand the density of Body Mass and the shape indicates how skewed the data is. We can see how different species have different distribution of body mass.

Adelie and Chinstrap species of penguins have similar distributions with most penguins having a mass between 3000 and 4500 grams. Gentoo species is different such that the distribution is between 4500 and 6000. Furthermore, density of body mass is greatest in the middle for all species.

Systhesis of Findings

I can improve my visualizations in the following ways:

In [ ]: