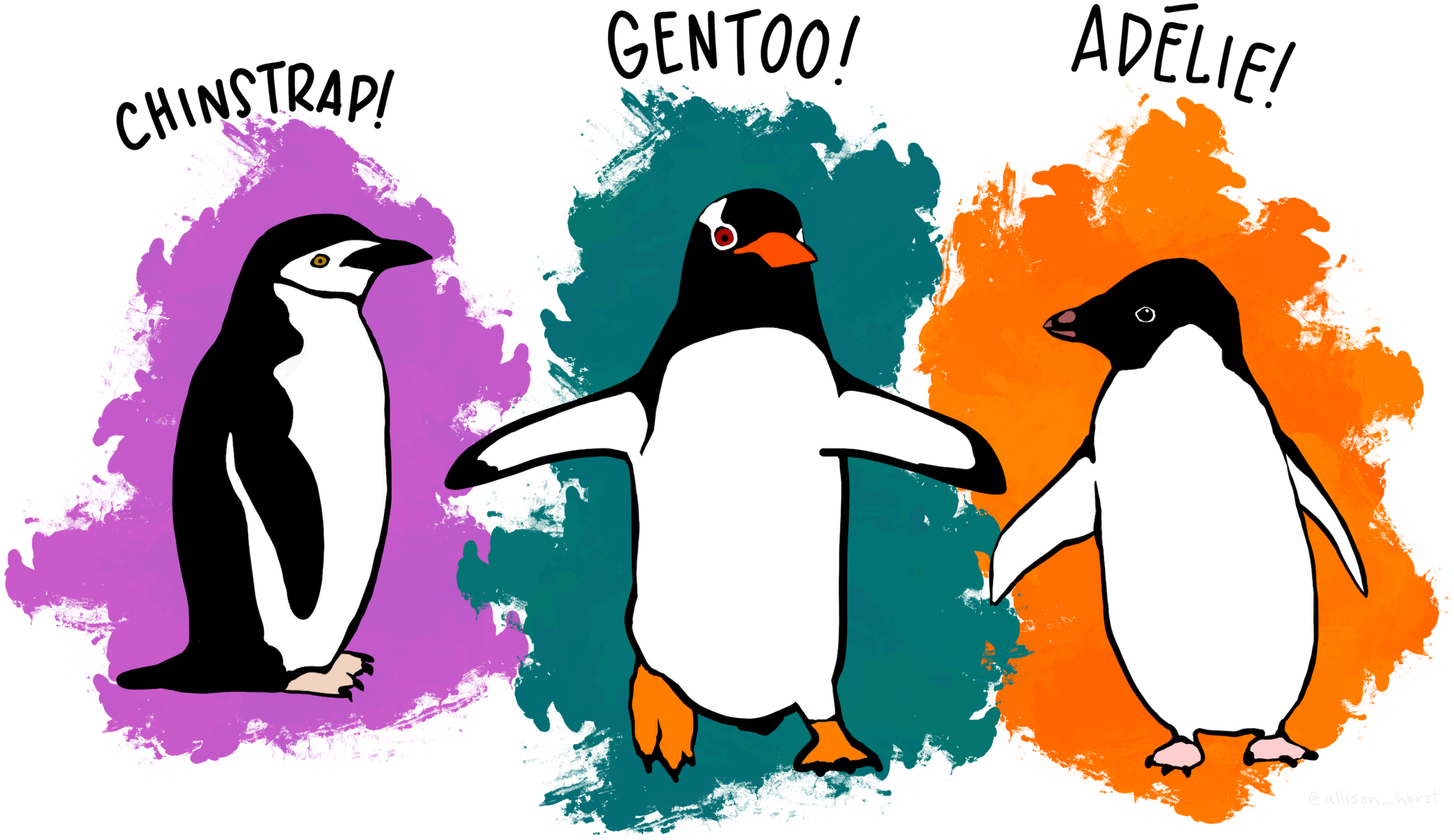
PROJECT: CLUSTERING ANTARCTIC PENGUIN SPECIES



source: @allison_horst https://github.com/allisonhorst/penguins

You have been asked to support a team of researchers who have been collecting data about penguins in Antartica! The data is available in csv-Format as `penguins.csv`

Origin of this data : Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

The dataset consists of 5 columns.

Column	Description
culmen_length_mm	culmen length (mm)
culmen_depth_mm	culmen depth (mm)
flipper_length_mm	flipper length (mm)
body_mass_g	body mass (g)
sex	penguin sex

Unfortunately, they have not been able to record the species of penguin, but they know that there are **at least three** species that are native to the region: **Adelie**, **Chinstrap**, and **Gentoo**. Your task is to apply your data science skills to help them identify groups in the dataset!

```python
# Import Required Packages
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Loading and examining the dataset
penguins_df = pd.read_csv("penguins.csv")
penguins_df.head()

# convert sex to 0 and 1
penguins_df['sex'] = pd.get_dummies(penguins_df['sex'], drop_first=True)
penguins_df.head()

# Transform the features
scaler = StandardScaler()
penguins_trans = scaler.fit_transform(penguins_df)
penguins_trans[:5]

# Cluster analysis
model = KMeans(n_clusters=3, random_state=42)
model.fit(penguins_trans)
labels = model.predict(penguins_trans)

# create the dataframe
label_df = pd.Series(labels, name='cluster')
new_penguin = pd.concat([penguins_df, label_df], axis=1)
stat_penguins = new_penguin.groupby('cluster').mean()
print(stat_penguins)
```

```
         culmen_length_mm  culmen_depth_mm  ...  body_mass_g       sex
cluster                                     ...
0               43.878302        19.111321  ...  4006.603774  1.000000
1               47.568067        14.996639  ...  5092.436975  0.512605
2               40.217757        17.611215  ...  3419.158879  0.000000
```

```
[3 rows x 5 columns]
```