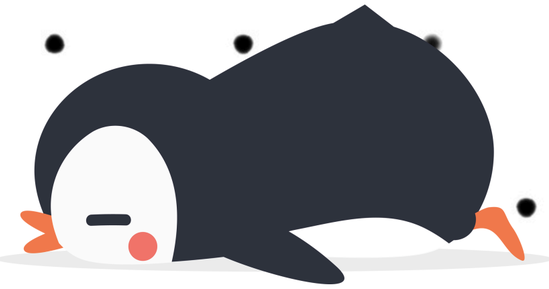
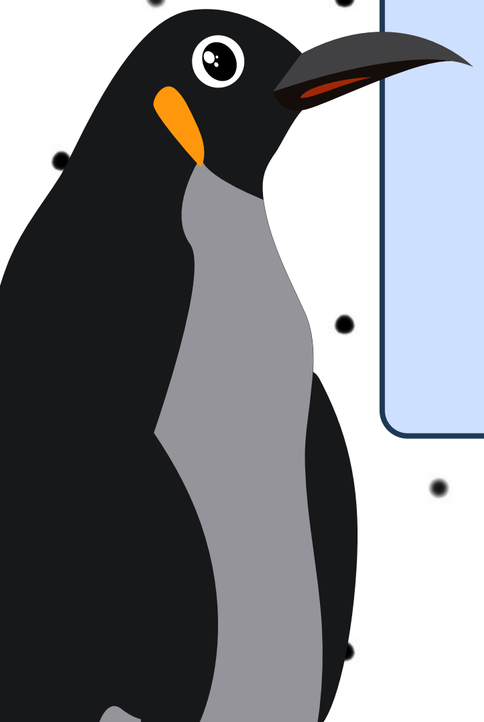


PREDICTING PENGUIN BODY MASS



GROUP 5

SAMIR BARAKAT - PEDRO ALEJANDRO MEDELLIN

THOMAS RENWICK - NOUR SEWILAM - JOY ZHONG

AGENDA

THE DATASET

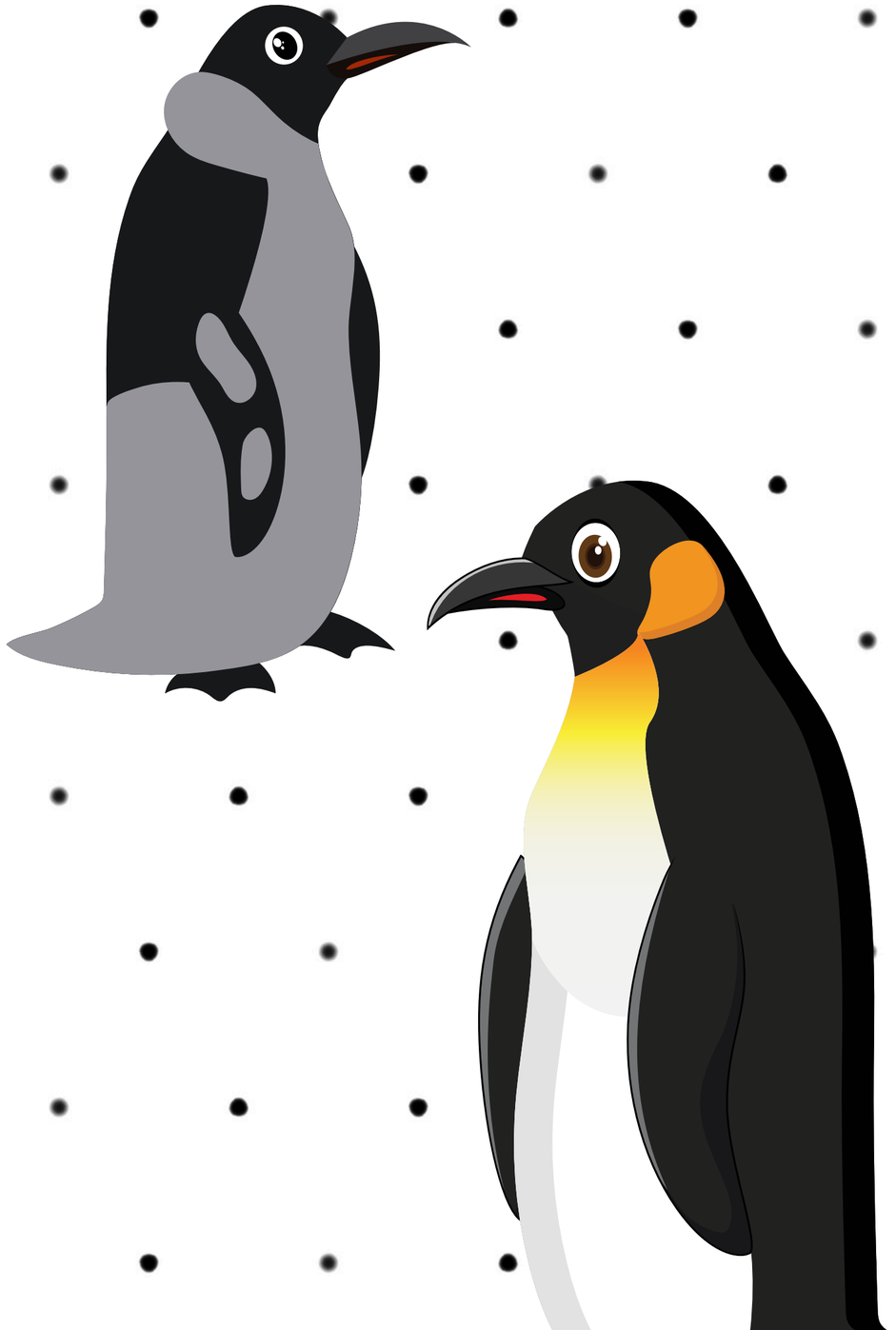
EDA & MODEL SELECTION

MODEL DEPLOYMENT PROCESS

STREAMLIT APP

AREAS OF IMPROVEMENT

VALUE ADDED



THE DATASET

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
Adelie	Torgersen	39.1	18.7	181	3750	male
Adelie	Torgersen	39.5	17.4	186	3800	female
Adelie	Torgersen	40.3	18.0	195	3250	female
Adelie	Torgersen	36.7	19.3	193	3450	female
Adelie	Torgersen	39.3	20.6	190	3650	male

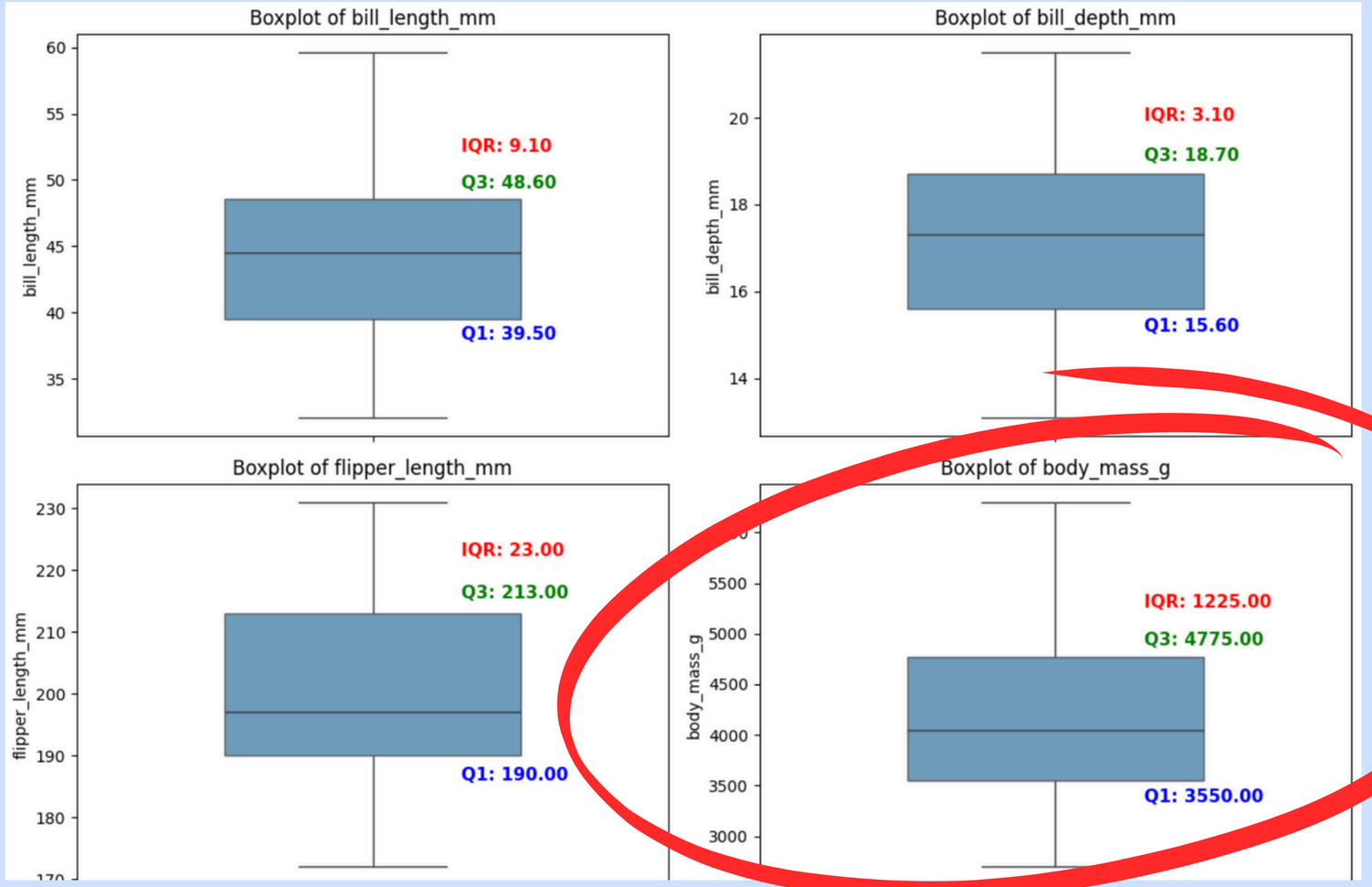
NUMERICAL VARIABLES

BILL LENGTH
BILL DEPTH
FLIPPER LENGTH
BODY MASS

CATEGORICAL VARIABLES

SEX
ISLAND
SPECIES

EDA & MODEL SELECTION



MODEL CHOSEN

LINEAR REGRESSION
LABEL = BODY MASS

WHY?

LARGEST RANGE
&
POTENTIAL
CORRELATION TO OTHER
FEATURES

STEP 1: DATA PREPARATION

1: CHECK FOR DUPLICATES

```
df.duplicated().any()
```

```
False
```

2: CHECK FOR NULLS & CORRECT DATA TYPES

```
Column             Non-Null Count  Dtype
------             --------------  -----
species            333 non-null    object
island             333 non-null    object
bill_length_mm     333 non-null    float64
bill_depth_mm      333 non-null    float64
flipper_length_mm  333 non-null    int64
body_mass_g        333 non-null    int64
sex                333 non-null    object
```

3: FIND UNIQUE VALUES FOR CATEGORICAL VARIABLES

```
island
Biscoe        163
Dream         123
Torgersen      47
```

```
sex
male          168
female        165
```

```
species
Adelie        146
Gentoo        119
Chinstrap      68
```

STEP 2: FEATURE ENGINEERING

1: ONE HOT ENCODING

	species_Chinstrap	species_Gentoo	island_Dream	island_Torgersen	sex_male
0	False	False	False	True	True
1	False	False	False	True	False
2	False	False	False	True	False
3	False	False	False	True	False
4	False	False	False	True	True

2: TRANSFORMED DATAFRAME

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	species_Chinstrap	species_Gentoo	island_Dream	island_Torgersen	sex_male
0	39.1	18.7	181	3750	False	False	False	True	True
1	39.5	17.4	186	3800	False	False	False	True	False
2	40.3	18.0	195	3250	False	False	False	True	False
3	36.7	19.3	193	3450	False	False	False	True	False
4	39.3	20.6	190	3650	False	False	False	True	True

STEP 3: SPLITTING LABELS & FEATURES

X : FEATURES

```
X.head()
```

	bill_length_mm	bill_depth_mm	flipper_length_mm	species_Chinstrap	species_Gentoo	island_Dream	island_Torgersen	sex_male
0	39.1	18.7	181	False	False	False	True	True
1	39.5	17.4	186	False	False	False	True	False
2	40.3	18.0	195	False	False	False	True	False
3	36.7	19.3	193	False	False	False	True	False
4	39.3	20.6	190	False	False	False	True	True

Y : LABEL [BODY MASS]

```
y.head()
```

```
0    3750
1    3800
2    3250
3    3450
4    3650
```

STEP 4: FEATURE SCALING

1: TRAIN TEST SPLIT

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

TRAIN = 80% & TEST = 20%

2: SCALING THE FEATURES

```
scaler = MinMaxScaler()
```

```
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

STEP 5: TRAINING THE MODEL

1: INSTANTIATING THE MODEL

```
model = LinearRegression()
```

2: TRAINING THE MODEL

```
model.fit(X_train,y_train)
```

LinearRegression ℹ ?

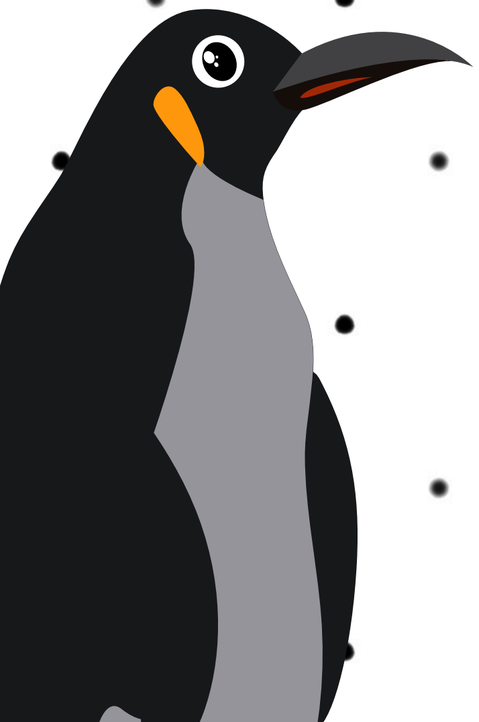
LinearRegression()

MODEL EVALUATION

MAE = 225.409204

MSE = 78,777.650800

RMSE = 280.673566



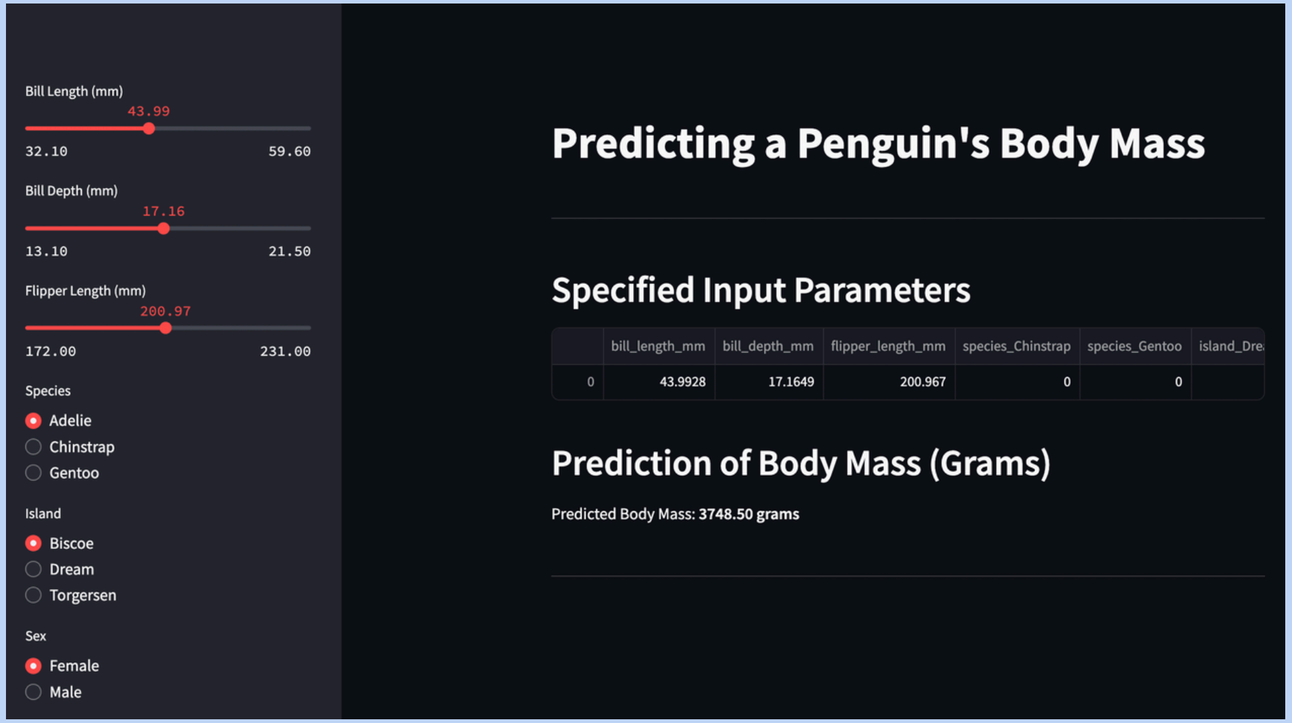
MODEL DEPLOYMENT : STREAMLIT APP

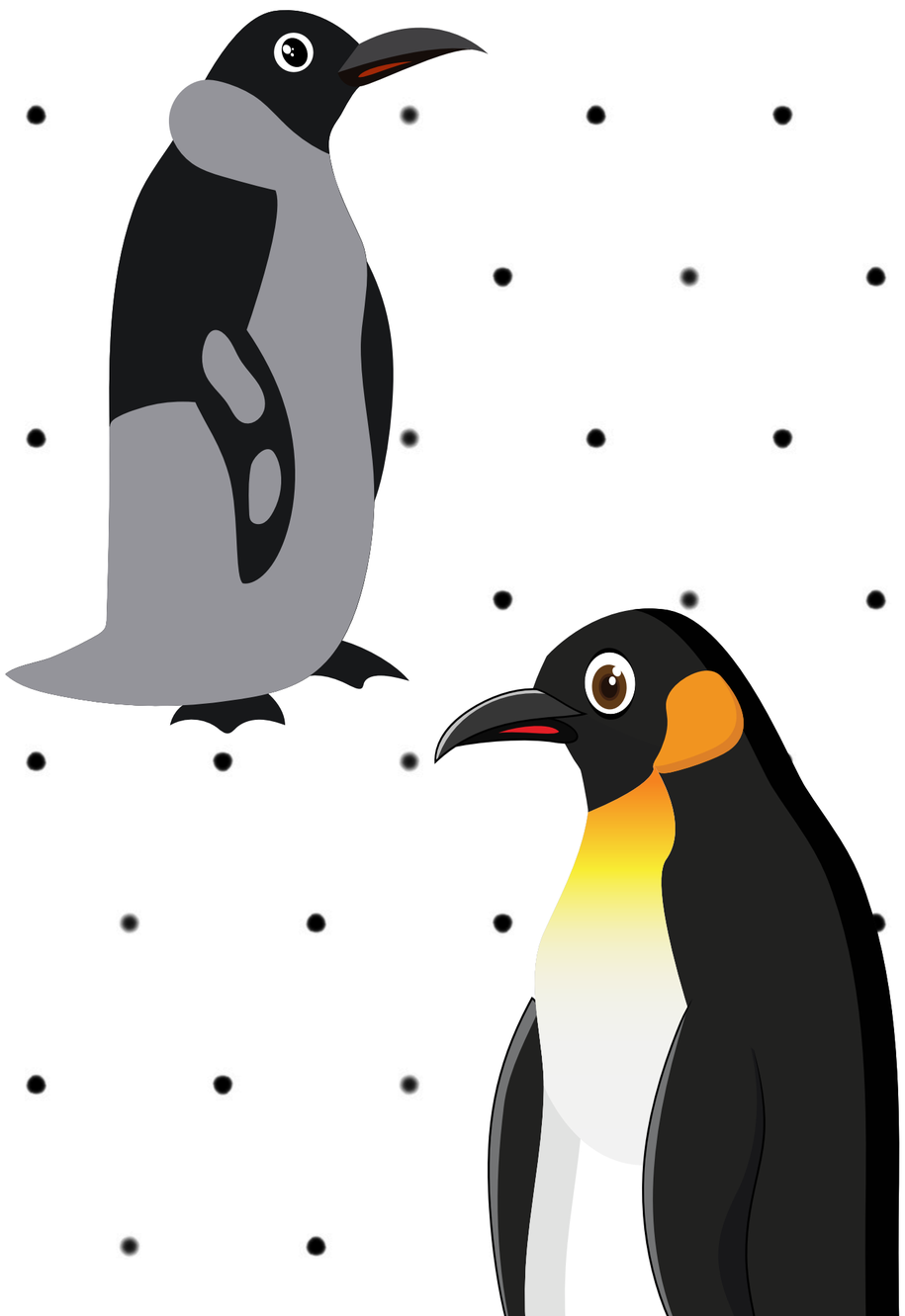
```python
# Streamlit function to collect user input
def user_input_features():
    bill_length_mm = st.sidebar.slider('Bill Length (mm)', float(X['bill_length_mm'].min()), float(X['bill_length_mm'].max()), float(X['bill_length_mm'].mean()))
    bill_depth_mm = st.sidebar.slider('Bill Depth (mm)', float(X['bill_depth_mm'].min()), float(X['bill_depth_mm'].max()), float(X['bill_depth_mm'].mean()))
    flipper_length_mm = st.sidebar.slider('Flipper Length (mm)', float(X['flipper_length_mm'].min()), float(X['flipper_length_mm'].max()), float(X['flipper_length_mm'].mean()))

    # Categorical feature selection using radio buttons
    species = st.sidebar.radio('Species', ['Adelie', 'Chinstrap', 'Gentoo'])
    island = st.sidebar.radio('Island', ['Biscoe', 'Dream', 'Torgersen'])
    sex = st.sidebar.radio('Sex', ['Female', 'Male'])

    # Encoding categorical variables based on dummy encoding
    species_Chinstrap = 1 if species == 'Chinstrap' else 0
    species_Gentoo = 1 if species == 'Gentoo' else 0
    island_Dream = 1 if island == 'Dream' else 0
    island_Torgersen = 1 if island == 'Torgersen' else 0
    sex_male = 1 if sex == 'Male' else 0  # Female is the reference category (0)

    # Creating the user input dataframe
    data = {
        'bill_length_mm': bill_length_mm,
        'bill_depth_mm': bill_depth_mm,
        'flipper_length_mm': flipper_length_mm,
        'species_Chinstrap': species_Chinstrap,
        'species_Gentoo': species_Gentoo,
        'island_Dream': island_Dream,
        'island_Torgersen': island_Torgersen,
        'sex_male': sex_male
    }
    features = pd.DataFrame(data, index=[0])
    return features
```

STREAMLIT APP

SCAN TO CHECK OUT OUR STREAMLIT APP !

PREVIEW





AREAS OF IMPROVEMENT

POLYNOMIAL REGRESSION FOR A BETTER MODEL FIT

SERIALIZING THE MODEL FOR SCALABILITY

FEATURE SCALING IMPROVEMENT

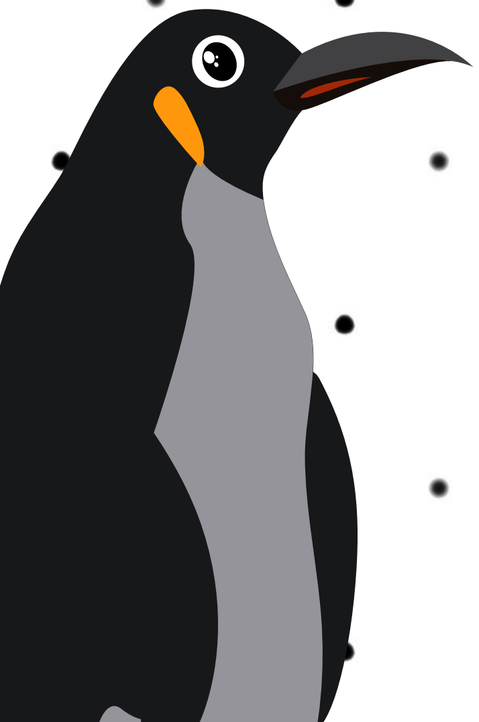
VALUES & BENEFITS OF THIS MODEL

MINIMIZES ANIMAL STRESS & IMPROVES ETHICAL RESEARCH

ENHANCES CONSERVATION EFFORTS

APPLIES TO OTHER WILDLIFE & RESEARCH AREAS

INCREASES ACCESSIBILITY & USABILITY



THANK YOU