In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

Read Data

In [2]:
```
df_part1 = pd.read_csv('/content/arthritis_part1.csv')
df_part2 = pd.read_csv('/content/arthritis_part2.csv')
```

In [3]:
```
df = pd.concat([df_part1, df_part2], ignore_index=True)
df.head(10)
```

Out[3]:

	Unnamed: 0	id	y	sex	age	trt	baseline	time
0	1	1	4.0	2	54	2	2	1
1	2	1	5.0	2	54	2	2	3
2	3	1	5.0	2	54	2	2	5
3	4	2	4.0	1	41	1	3	1
4	5	2	4.0	1	41	1	3	3
5	6	2	4.0	1	41	1	3	5
6	7	3	3.0	2	48	2	3	1
7	8	3	4.0	2	48	2	3	3
8	9	3	4.0	2	48	2	3	5
9	10	4	4.0	2	40	1	3	1

Data Cleaning

In [4]:
```
df.drop(['Unnamed: 0'], axis=1, inplace=True)
df.head(10)
```

Out[4]:

	id	y	sex	age	trt	baseline	time
0	1	4.0	2	54	2	2	1
1	1	5.0	2	54	2	2	3
2	1	5.0	2	54	2	2	5
3	2	4.0	1	41	1	3	1
4	2	4.0	1	41	1	3	3
5	2	4.0	1	41	1	3	5
6	3	3.0	2	48	2	3	1
7	3	4.0	2	48	2	3	3
8	3	4.0	2	48	2	3	5
9	4	4.0	2	40	1	3	1

In [5]:
```
df.rename(columns={"id": "Patient_ID", "sex":"Gender","trt": "Treatment", "baseline": "Baseline", "y": "Swollen_Joints", "time
```

In [6]:
```
df.head(10)
```

Out[6]:

	Patient_ID	Swollen_Joints	Gender	age	Treatment	Baseline	Time
0	1	4.0	2	54	2	2	1
1	1	5.0	2	54	2	2	3
2	1	5.0	2	54	2	2	5
3	2	4.0	1	41	1	3	1
4	2	4.0	1	41	1	3	3
5	2	4.0	1	41	1	3	5
6	3	3.0	2	48	2	3	1
7	3	4.0	2	48	2	3	3
8	3	4.0	2	48	2	3	5
9	4	4.0	2	40	1	3	1

In [7]:
```
df['Treatment'] = df['Treatment'].map({1: "Prednisone", 2: "Placebo"})
```

In [8]:
```
df['Gender'] = df['Gender'].map({1: "Male", 2: "Female"})
```

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 906 entries, 0 to 905
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Patient_ID      906 non-null    int64
 1   Swollen_Joints  888 non-null    float64
 2   Gender          906 non-null    object
 3   age             906 non-null    int64
 4   Treatment       906 non-null    object
 5   Baseline        906 non-null    int64
 6   Time            906 non-null    int64
dtypes: float64(1), int64(4), object(2)
memory usage: 49.7+ KB
```

In [10]:
```
df.isna().sum()
```

Out[10]:

	0
Patient_ID	0
Swollen_Joints	18
Gender	0
age	0
Treatment	0
Baseline	0
Time	0

dtype: int64

In [11]:
```
df.describe()
```

Out[11]:

	Patient_ID	Swollen_Joints	age	Baseline	Time
count	906.000000	888.000000	906.000000	906.000000	906.000000
mean	151.500000	3.227477	50.377483	2.864238	3.000000
std	87.227565	0.963862	11.103377	0.927363	1.633895
min	1.000000	1.000000	21.000000	1.000000	1.000000
25%	76.000000	3.000000	42.000000	2.000000	1.000000
50%	151.500000	3.000000	54.000000	3.000000	3.000000
75%	227.000000	4.000000	60.000000	3.000000	5.000000
max	302.000000	5.000000	66.000000	5.000000	5.000000

In [12]:
```
df['Gender'].value_counts()
```

Out[12]:

Gender	count
Female	657
Male	249

dtype: int64

In [13]:
```
df['Swollen_Joints'].value_counts()
```

Out[13]:

Swollen_Joints	count
3.0	345
4.0	275
2.0	159
5.0	76
1.0	33

dtype: int64

In [14]:
```
df['Treatment'].value_counts()
```

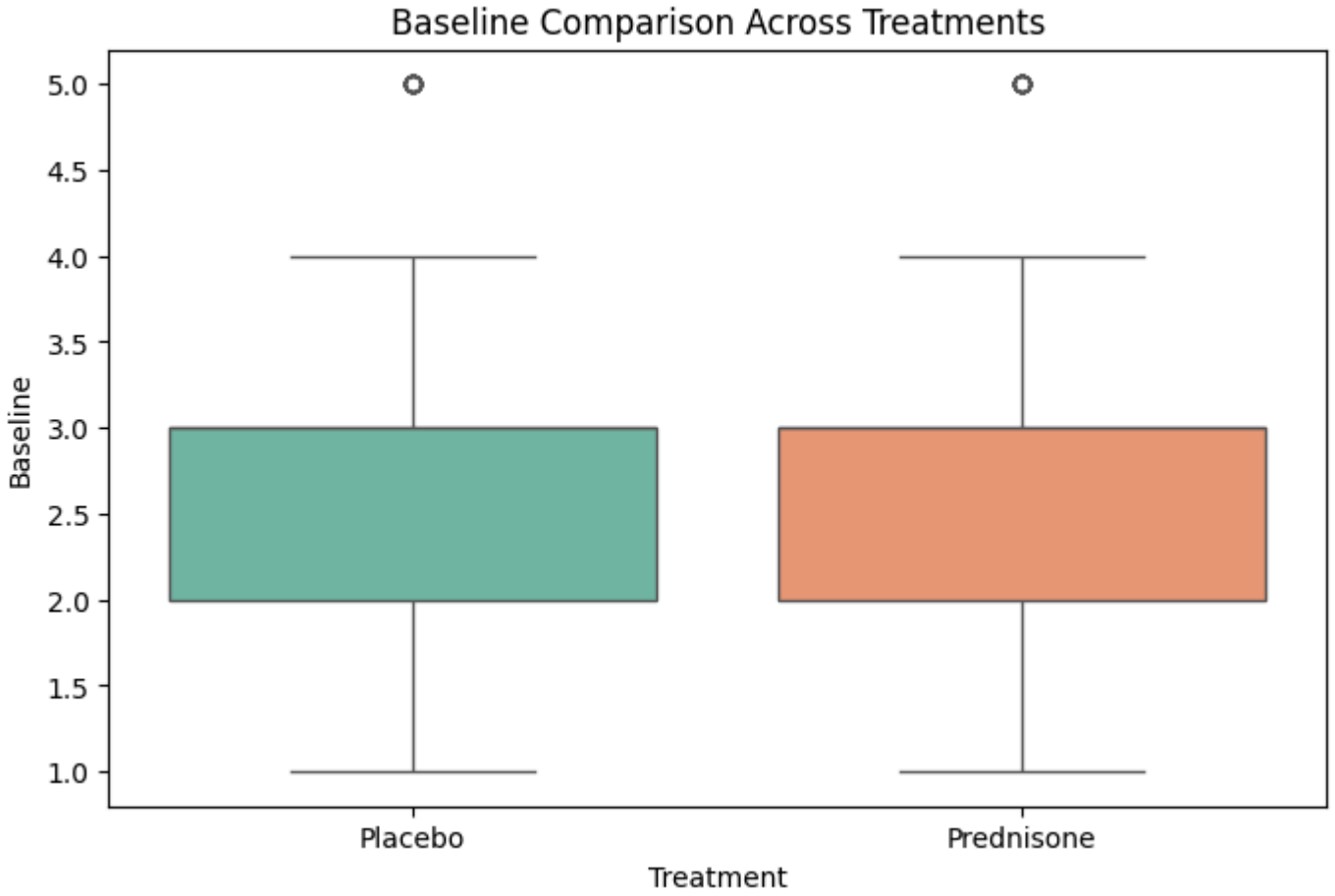
Out[14]:

Treatment	count
Placebo	459
Prednisone	447

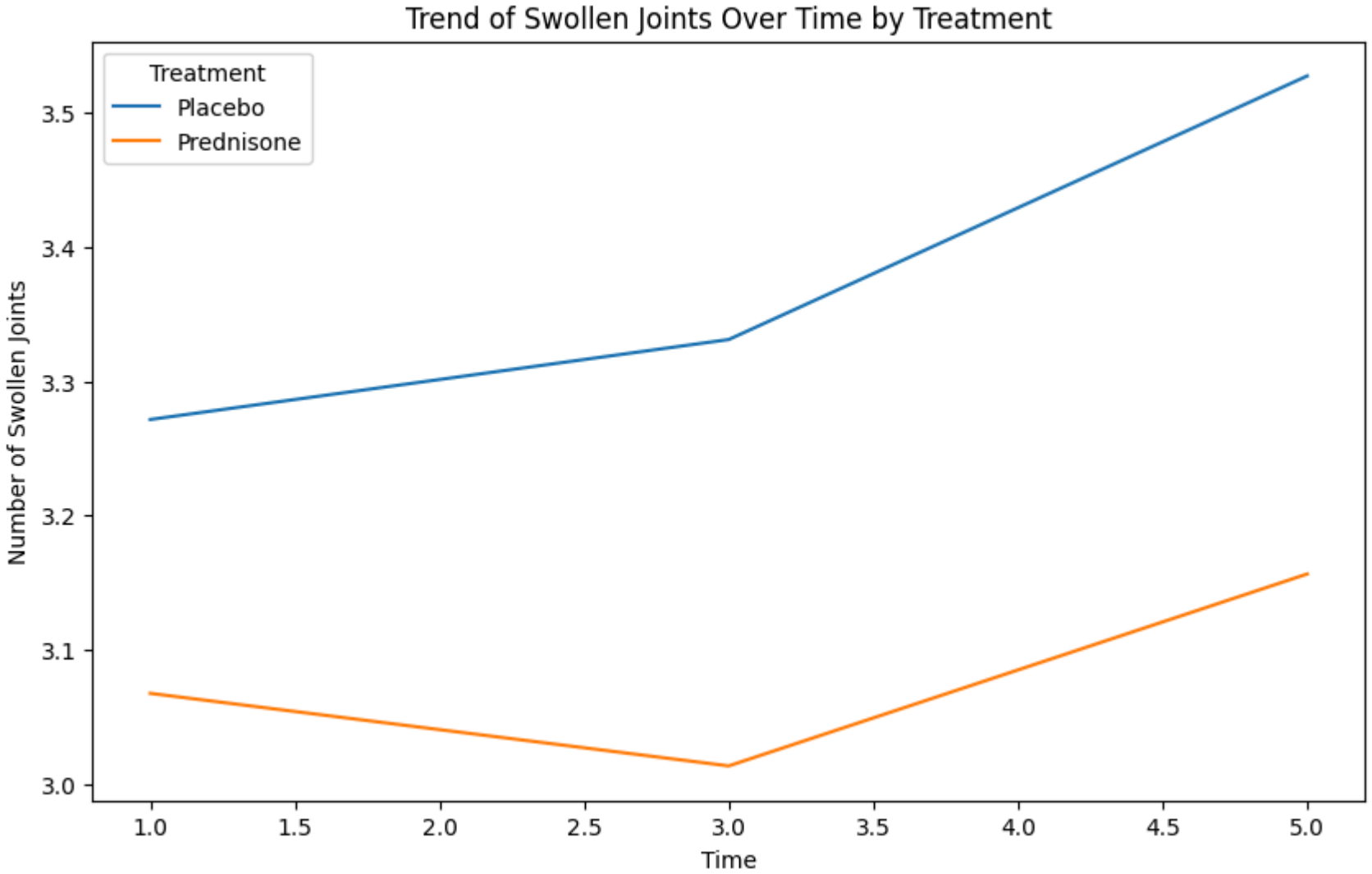
dtype: int64

Exploratory Data Analysis (EDA)

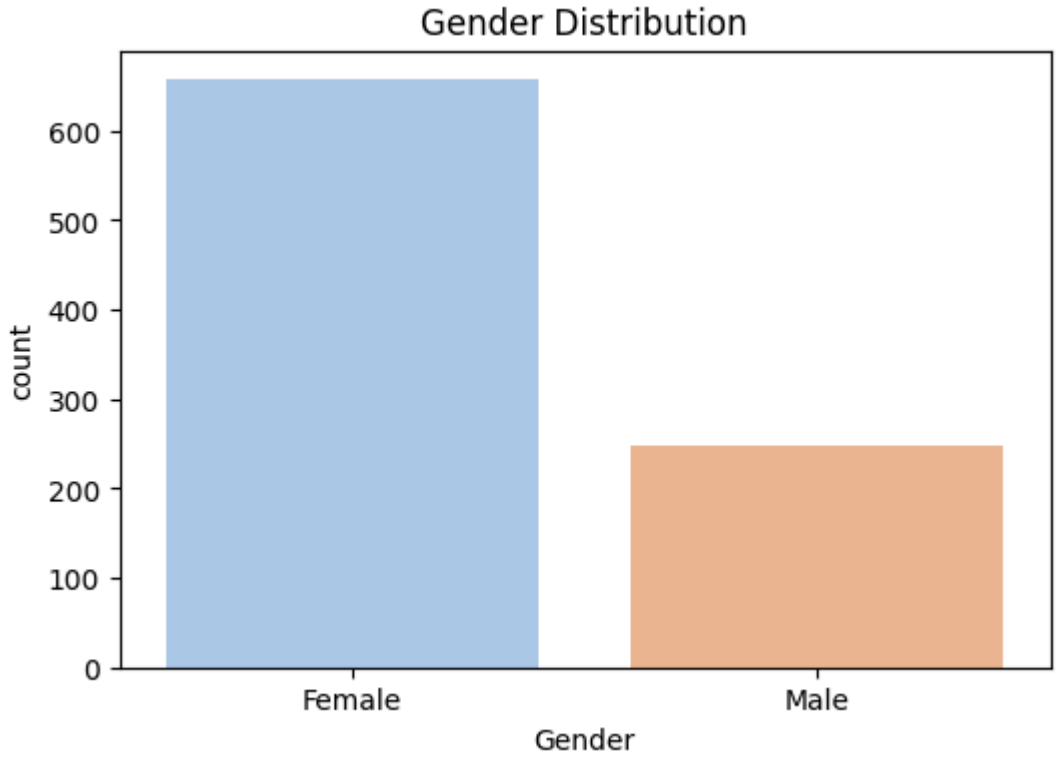
In [21]:
```
# Baseline comparison between treatments
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x="Treatment", y="Baseline", palette="Set2")
plt.title("Baseline Comparison Across Treatments")
plt.show()
```



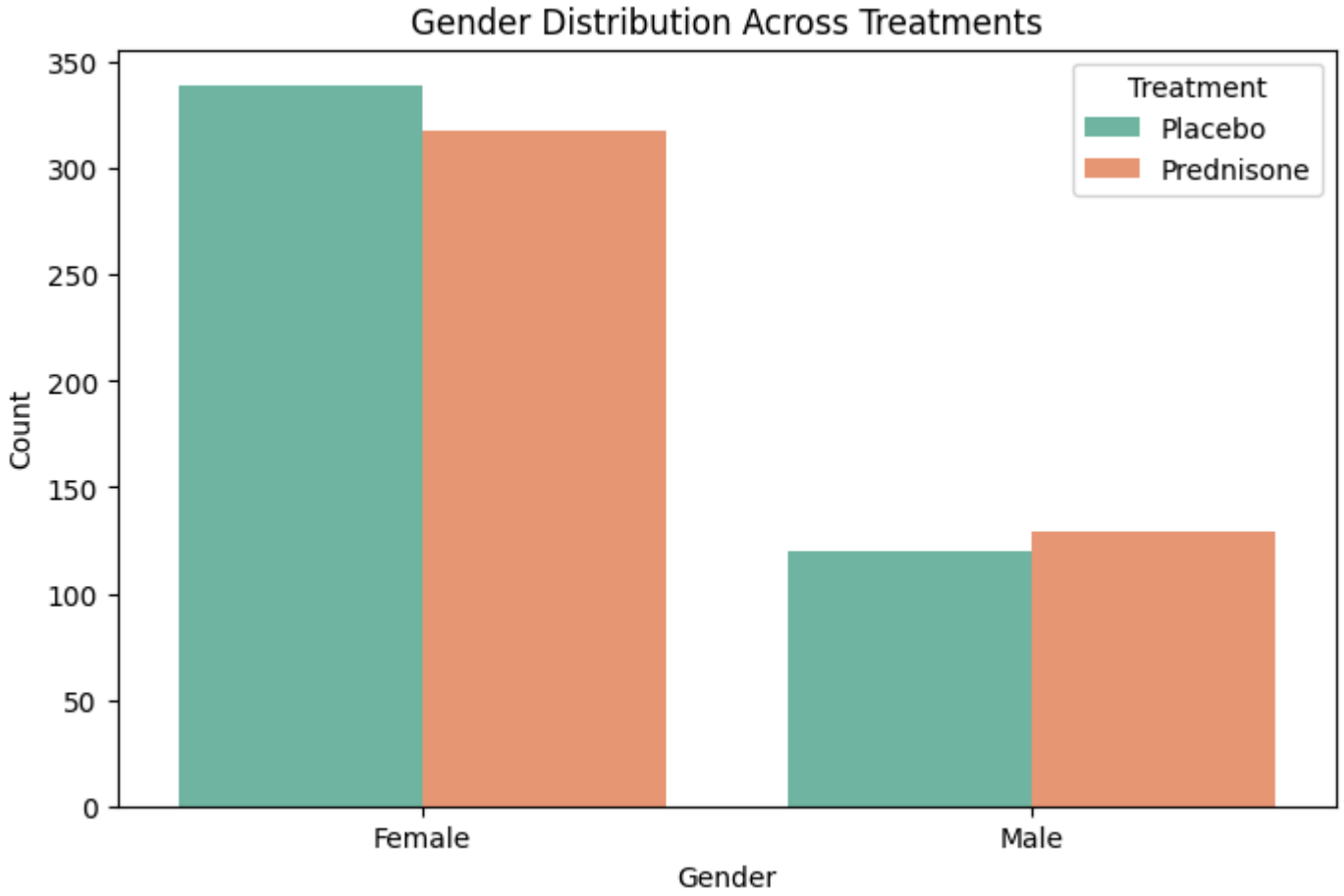
In [16]:
```
# Swollen joints (y) trend over time for each treatment
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x="Time", y="Swollen_Joints", hue="Treatment", ci=None)
plt.title("Trend of Swollen Joints Over Time by Treatment")
plt.xlabel("Time")
plt.ylabel("Number of Swollen Joints")
plt.legend(title="Treatment")
plt.show()
```



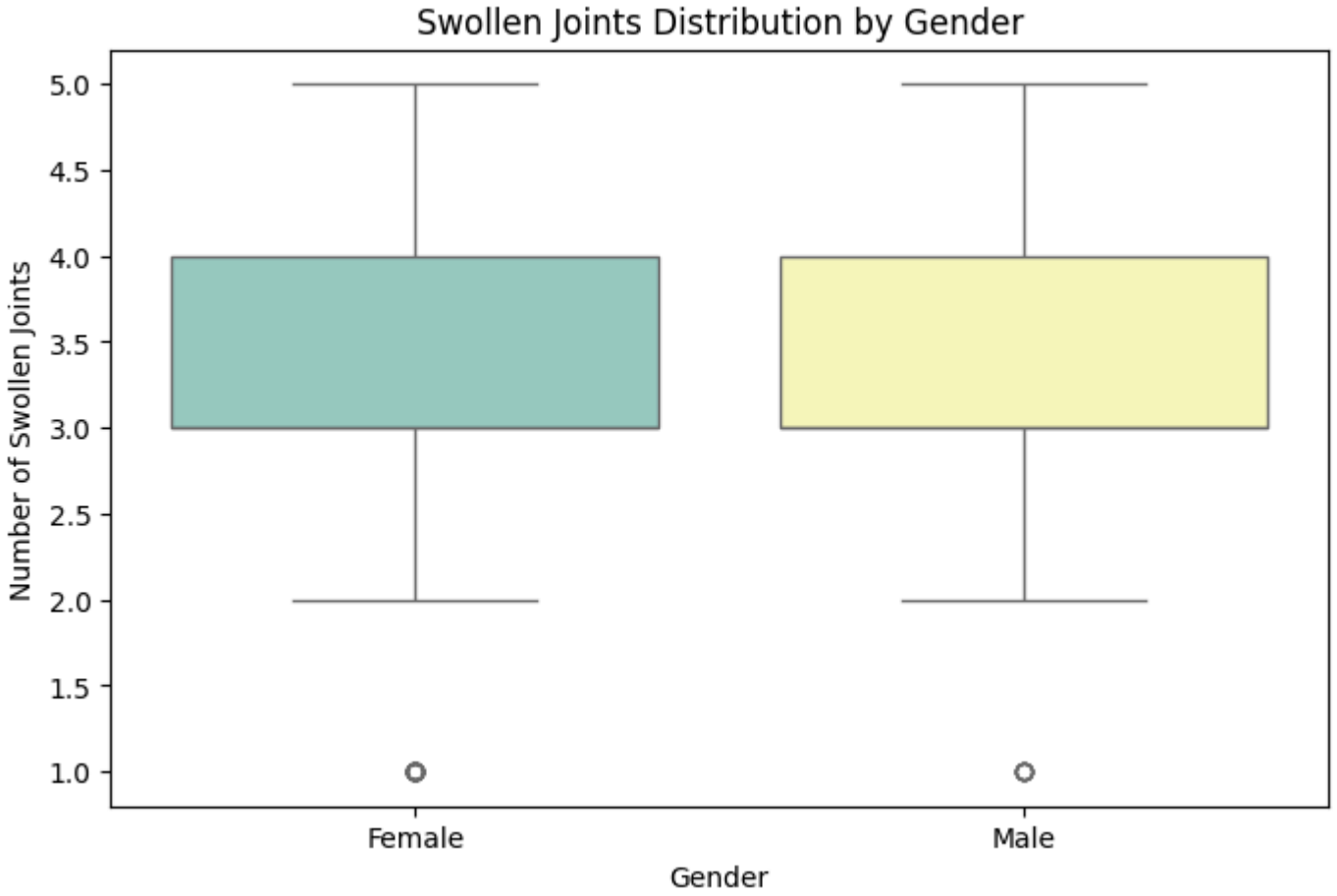
In [17]:
```
# Gender distribution in the dataset
plt.figure(figsize=(6, 4))
sns.countplot(data=df, x="Gender", palette="pastel")
plt.title("Gender Distribution")
plt.show()
```



In [35]:
```
# Count plot for Gender Distribution by Treatment
plt.figure(figsize=(8, 5))
sns.countplot(data=df, x="Gender", hue="Treatment", palette="Set2")
plt.title("Gender Distribution Across Treatments")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.legend(title="Treatment")
plt.show()
```

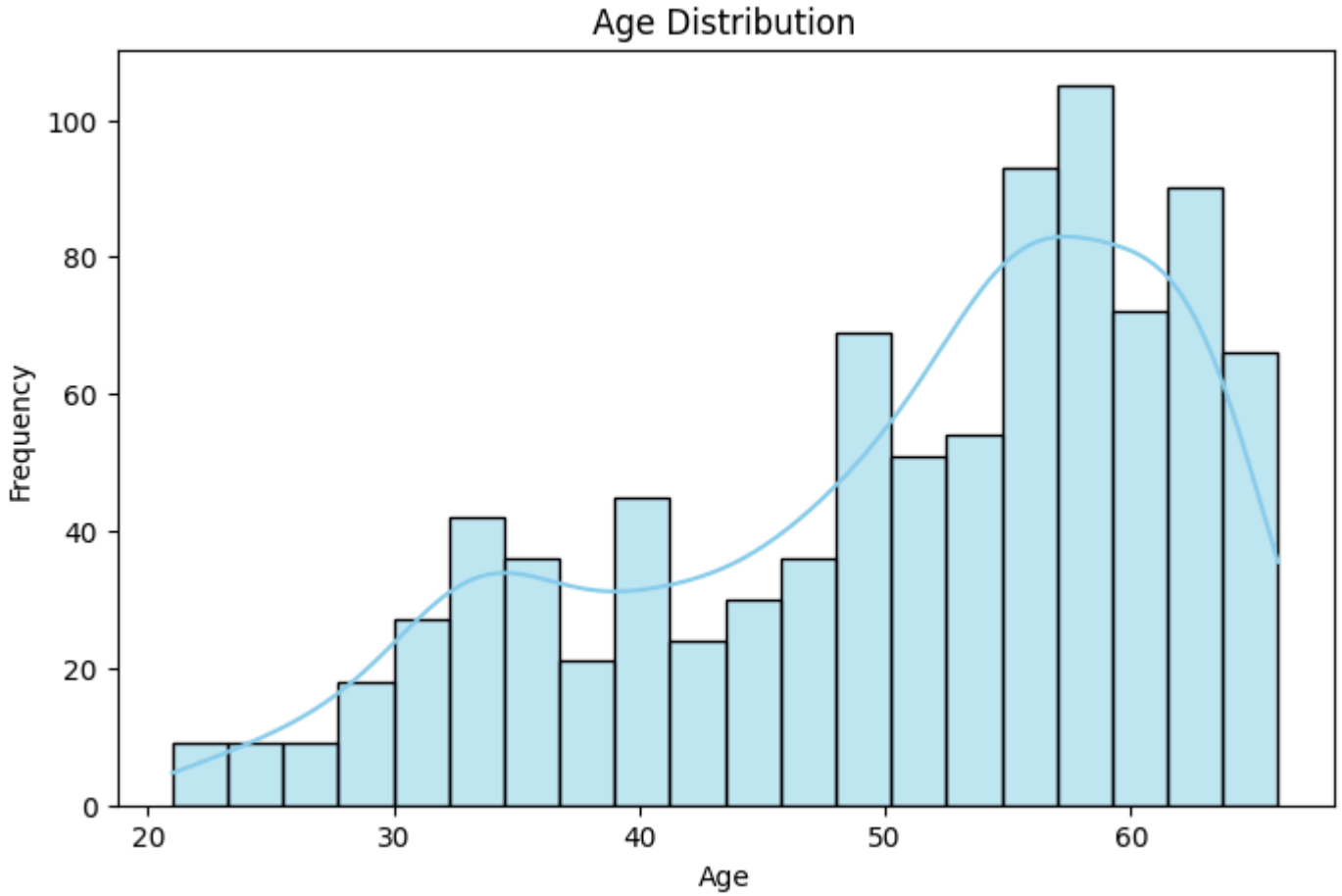


In [36]:
```
# Box plot of Swollen Joints by Gender
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x="Gender", y="Swollen_Joints", palette="Set3")
plt.title("Swollen Joints Distribution by Gender")
plt.xlabel("Gender")
plt.ylabel("Number of Swollen Joints")
plt.show()
```

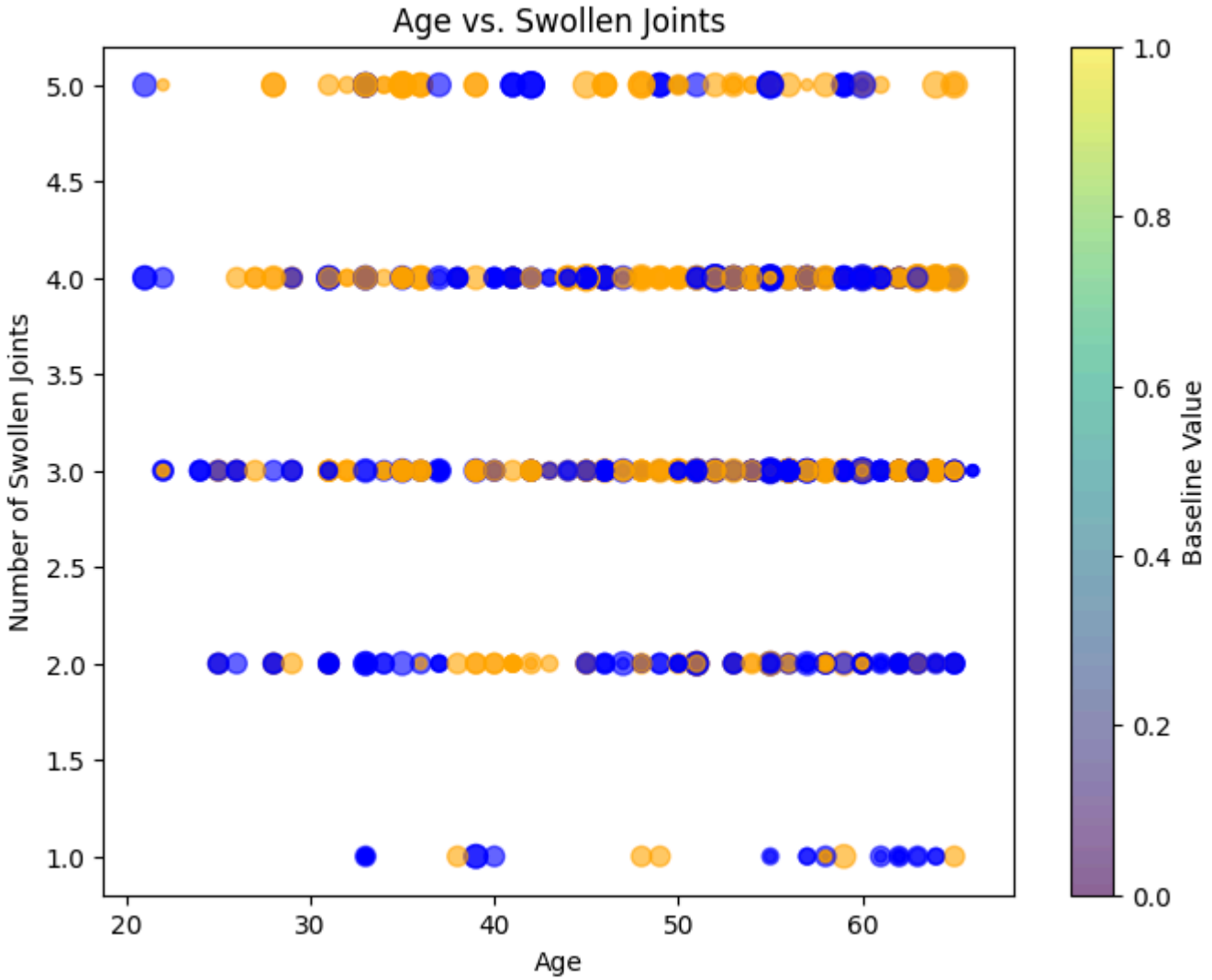


In [19]:
```
# Compare swollen joints by treatment group
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x="Treatment", y="Swollen_Joints", palette="pastel")
plt.title("Swollen Joints Across Treatments")
plt.show()
```

In [25]:
```
# Age distribution
plt.figure(figsize=(8, 5))
sns.histplot(df['age'], bins=20, kde=True, color='skyblue')
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```



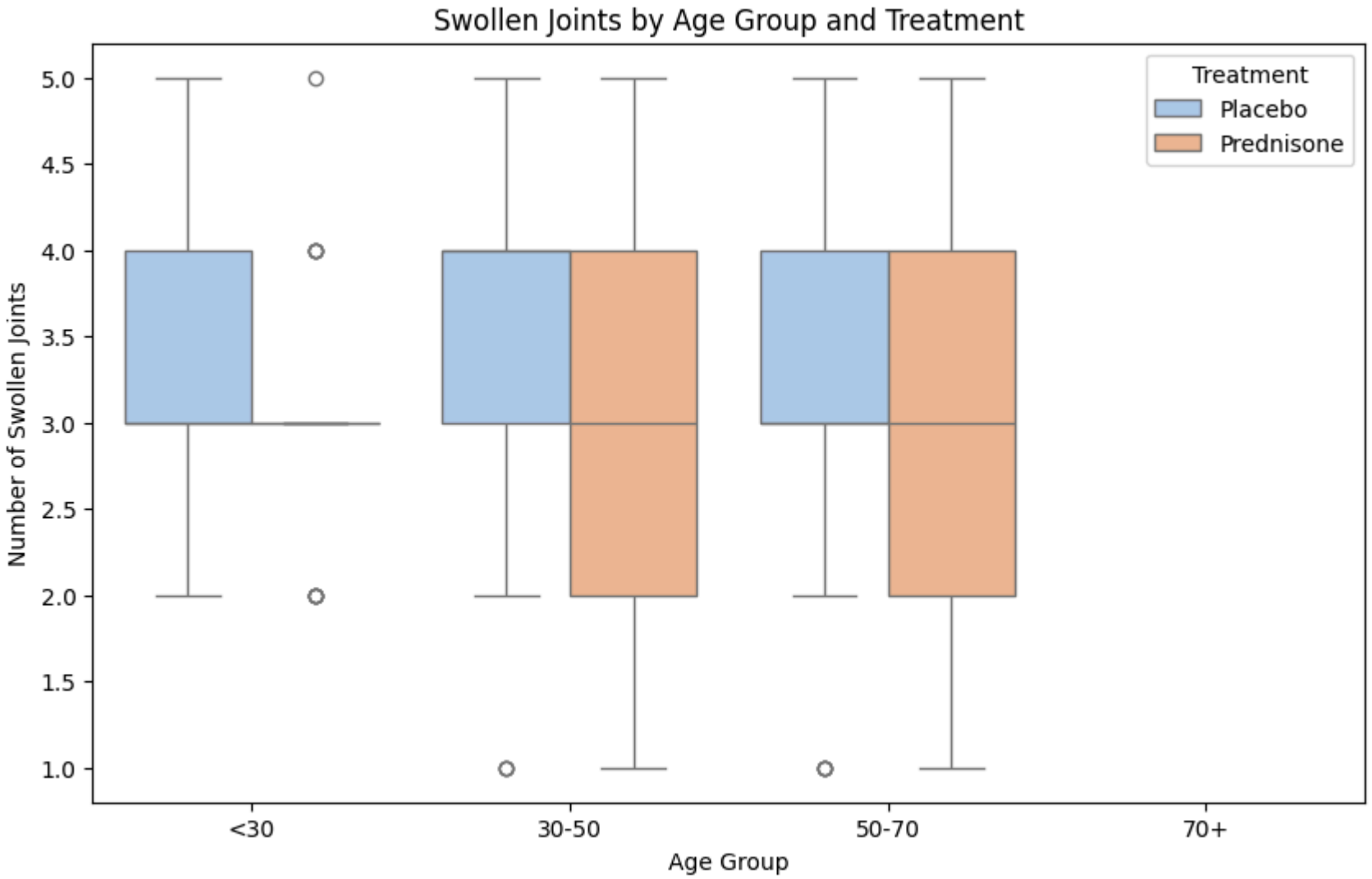
In [33]:
```
# Age vs. Swollen Joints with Baseline as Bubble Size
plt.figure(figsize=(8, 6))
bubble_size = df['Baseline'] / df['Baseline'].max() * 100  # Scale the bubble sizes
plt.scatter(df['age'], df['Swollen_Joints'], s=bubble_size, c=df['Treatment'].map({"Prednisone": "blue", "Placebo": "orange"})
plt.title("Age vs. Swollen Joints")
plt.xlabel("Age")
plt.ylabel("Number of Swollen Joints")
plt.colorbar(label="Baseline Value")
plt.show()
```



In [34]:
```
# Create age groups
df['Age_Group'] = pd.cut(df['age'], bins=[0, 30, 50, 70, 90], labels=["<30", "30-50", "50-70", "70+"])

# Swollen Joints by Age Group and Treatment
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x="Age_Group", y="Swollen_Joints", hue="Treatment", palette="pastel")
plt.title("Swollen Joints by Age Group and Treatment")
```

```
plt.xlabel("Age Group")
plt.ylabel("Number of Swollen Joints")
plt.legend(title="Treatment")
plt.show()
```



In [ ]: