```
import pandas as pd
df = pd.read_csv('train.csv')
df = df.dropna()
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
...	...	...	...	...	...	...	...	...	...	...	...	...
871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S
872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.0000	B51 B53 B55	S
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df = df.drop(columns=['Name', 'SibSp', 'Ticket', 'Fare', 'Cabin', 'Embarked'])

print(df)
```

```
     PassengerId  Survived  Pclass     Sex   Age  Parch
1              2         1       1  female  38.0      0
3              4         1       1  female  35.0      0
6              7         0       1    male  54.0      0
10            11         1       3  female   4.0      1
11            12         1       1  female  58.0      0
..           ...       ...     ...     ...   ...    ...
871          872         1       1  female  47.0      1
872          873         0       1    male  33.0      0
879          880         1       1  female  56.0      1
887          888         1       1  female  19.0      0
889          890         1       1    male  26.0      0

[183 rows x 6 columns]
```

```
passenger_ids = df['PassengerId']
df = df.drop(columns=['PassengerId'])

df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})


df['Age'] = (df['Age'] - df['Age'].min()) / (df['Age'].max() - df['Age'].min())
df
```

	Survived	Pclass	Sex	Age	Parch
1	1	1	1	0.468892	0
3	1	1	1	0.430956	0
6	0	1	0	0.671219	0
10	1	3	1	0.038948	1
11	1	1	1	0.721801	0
...	...	...	...	...	...
871	1	1	1	0.582701	1
872	0	1	0	0.405665	0
879	1	1	1	0.696510	1
887	1	1	1	0.228629	0
889	1	1	0	0.317147	0

183 rows × 5 columns

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
import numpy as np
```

```
df['sex_bias'] = np.where(df['Sex'] == 1, 0.3, 0)
```

```
print(df.head())
```

```
     Survived  Pclass  Sex       Age  Parch  sex_bias
1           1       1    1  0.468892      0       0.3
3           1       1    1  0.430956      0       0.3
6           0       1    0  0.671219      0       0.0
10          1       3    1  0.038948      1       0.3
11          1       1    1  0.721801      0       0.3
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

```
X = df.drop('Survived', axis=1)
y = df['Survived']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

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

```
y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy of the logistic regression model: {accuracy}")
```

```
Accuracy of the logistic regression model: 0.7837837837837838
```

```
new_df = pd.read_csv('test.csv')
new_df
```

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

Next steps: Generate code with new_df | View recommended plots | New interactive sheet

```
passenger_ids = new_df['PassengerId']
```

```
new_df = new_df.drop(['Ticket', 'Cabin', 'Embarked', 'Fare', 'SibSp', 'Name', 'PassengerId'], axis=1)
```

```
new_df['Sex'] = new_df['Sex'].replace({'male': 0, 'female': 1})
```

```
new_df['sex_bias'] = np.where(new_df['Sex'] == 1, 0.3, 0)

new_df['Age'] = new_df['Age'].fillna(new_df['Age'].mean())
new_df['Age'] = (new_df['Age'] - new_df['Age'].min()) / (new_df['Age'].max() - new_df['Age'].min())

new_df
```

```
<ipython-input-59-c9e538228526>:13: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future versi
  new_df['Sex'] = new_df['Sex'].replace({'male': 0, 'female': 1})
```

	Pclass	Sex	Age	Parch	sex_bias
0	3	0	0.452723	0	0.0
1	3	1	0.617566	0	0.3
2	2	0	0.815377	0	0.0
3	3	0	0.353818	0	0.0
4	3	1	0.287881	1	0.3
...	...	...	...	...	...
413	3	0	0.396975	0	0.0
414	1	1	0.512066	0	0.3
415	3	0	0.505473	0	0.0
416	3	0	0.396975	0	0.0
417	3	0	0.396975	1	0.0

418 rows × 5 columns

Next steps: Generate code with new_df | View recommended plots | New interactive sheet

```
predictions = model.predict(new_df)

new_df['predictions'] = predictions
```

```
submission_df = pd.DataFrame({'PassengerId': passenger_ids, 'Survived': predictions})

submission_df.to_csv('submission.csv', index=False)

print(submission_df.head())
```

```
   PassengerId  Survived
0          892         0
1          893         1
2          894         0
3          895         0
4          896         1
```