SAHYADRI

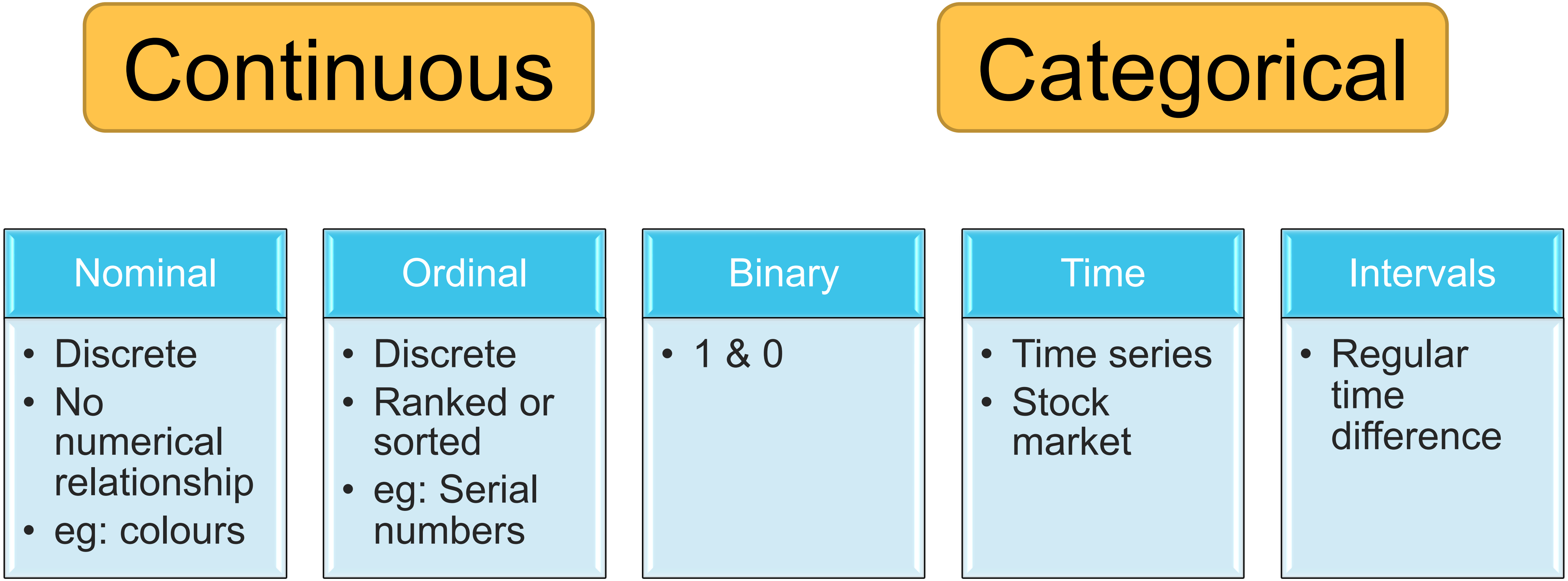
tce.

taplingua

TALENT COCO

AI/ML Projects/Internships

Types of Data



Continuous

Categorical

Nominal

- Discrete
- No numerical relationship
- eg: colours

Ordinal

- Discrete
- Ranked or sorted
- eg: Serial numbers

Binary

- 1 & 0

Time

- Time series
- Stock market

Intervals

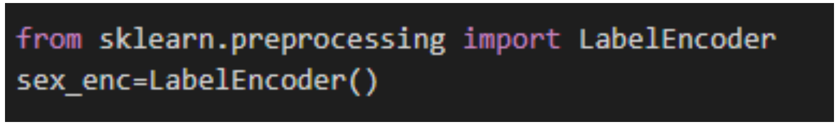
- Regular time difference

Encoding

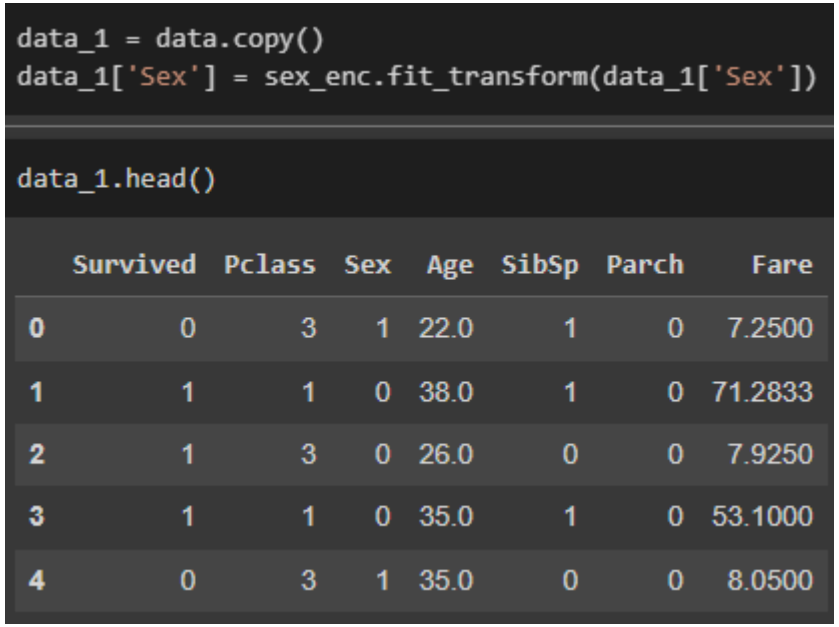
One-Hot Encoding

Label Encoding

Label Encoding



```
from sklearn.preprocessing import LabelEncoder
sex_enc=LabelEncoder()
```

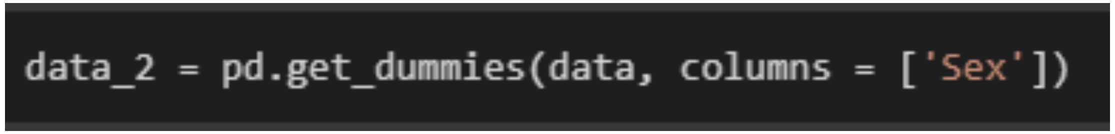


```
data_1 = data.copy()
data_1['Sex'] = sex_enc.fit_transform(data_1['Sex'])
```

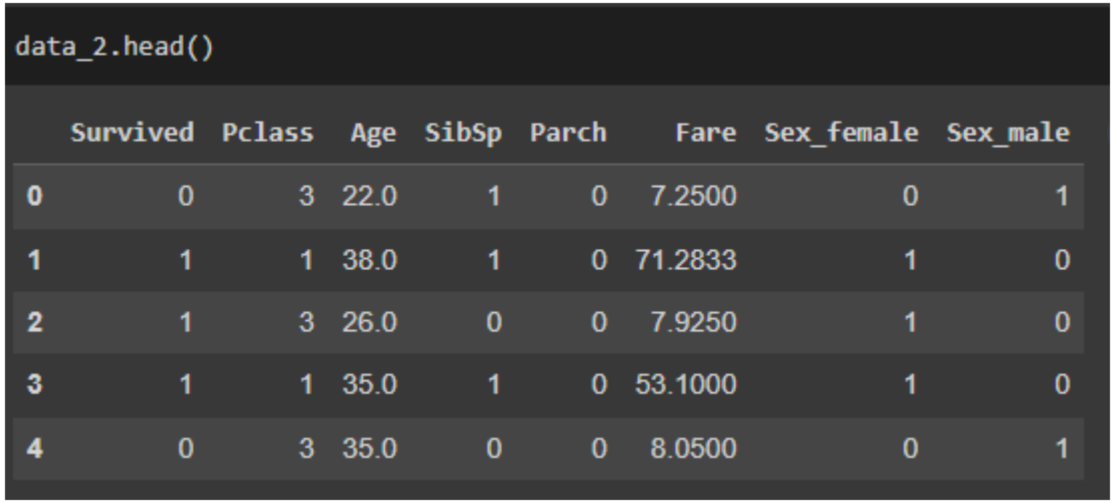
```
data_1.head()
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
0	0	3	1	22.0	1	0	7.2500
1	1	1	0	38.0	1	0	71.2833
2	1	3	0	26.0	0	0	7.9250
3	1	1	0	35.0	1	0	53.1000
4	0	3	1	35.0	0	0	8.0500

One-Hot Encoding



```
data_2 = pd.get_dummies(data, columns = ['Sex'])
```



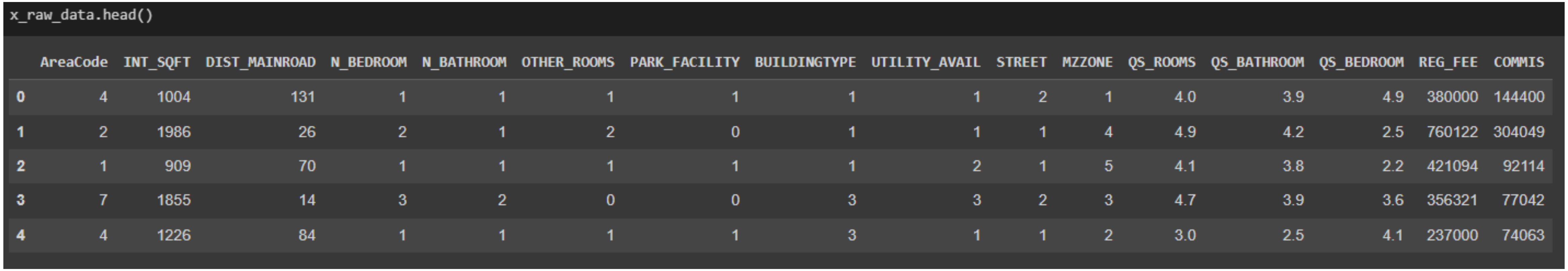
```
data_2.head()
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male
0	0	3	22.0	1	0	7.2500	0	1
1	1	1	38.0	1	0	71.2833	1	0
2	1	3	26.0	0	0	7.9250	1	0
3	1	1	35.0	1	0	53.1000	1	0
4	0	3	35.0	0	0	8.0500	0	1

Normalization

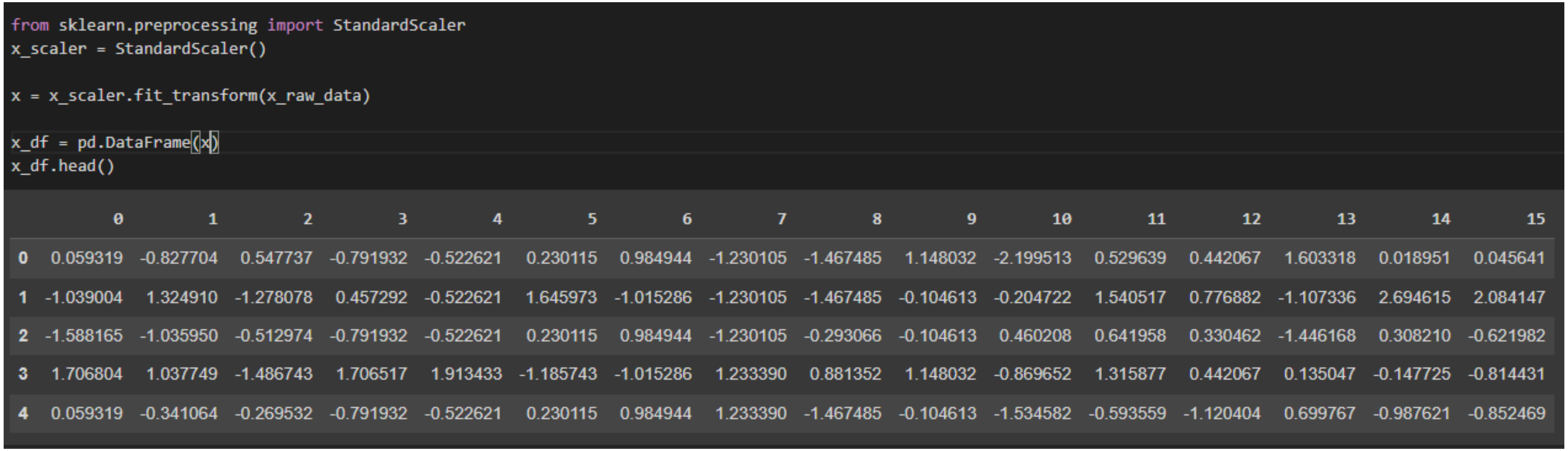
(-1 to 1)

Standard Scalar



```
x_raw_data.head()
```

	AreaCode	INT_SQFT	DIST_MAINROAD	N_BEDROOM	N_BATHROOM	OTHER_ROOMS	PARK_FACILITY	BUILDINGTYPE	UTILITY_AVAIL	STREET	MZZONE	QS_ROOMS	QS_BATHROOM	QS_BEDROOM	REG_FEE	COMMIS
0	4	1004	131	1	1	1	1	1	1	2	1	4.0	3.9	4.9	380000	144400
1	2	1986	26	2	1	2	0	1	1	1	4	4.9	4.2	2.5	760122	304049
2	1	909	70	1	1	1	1	1	2	1	5	4.1	3.8	2.2	421094	92114
3	7	1855	14	3	2	0	0	3	3	2	3	4.7	3.9	3.6	356321	77042
4	4	1226	84	1	1	1	1	3	1	1	2	3.0	2.5	4.1	237000	74063

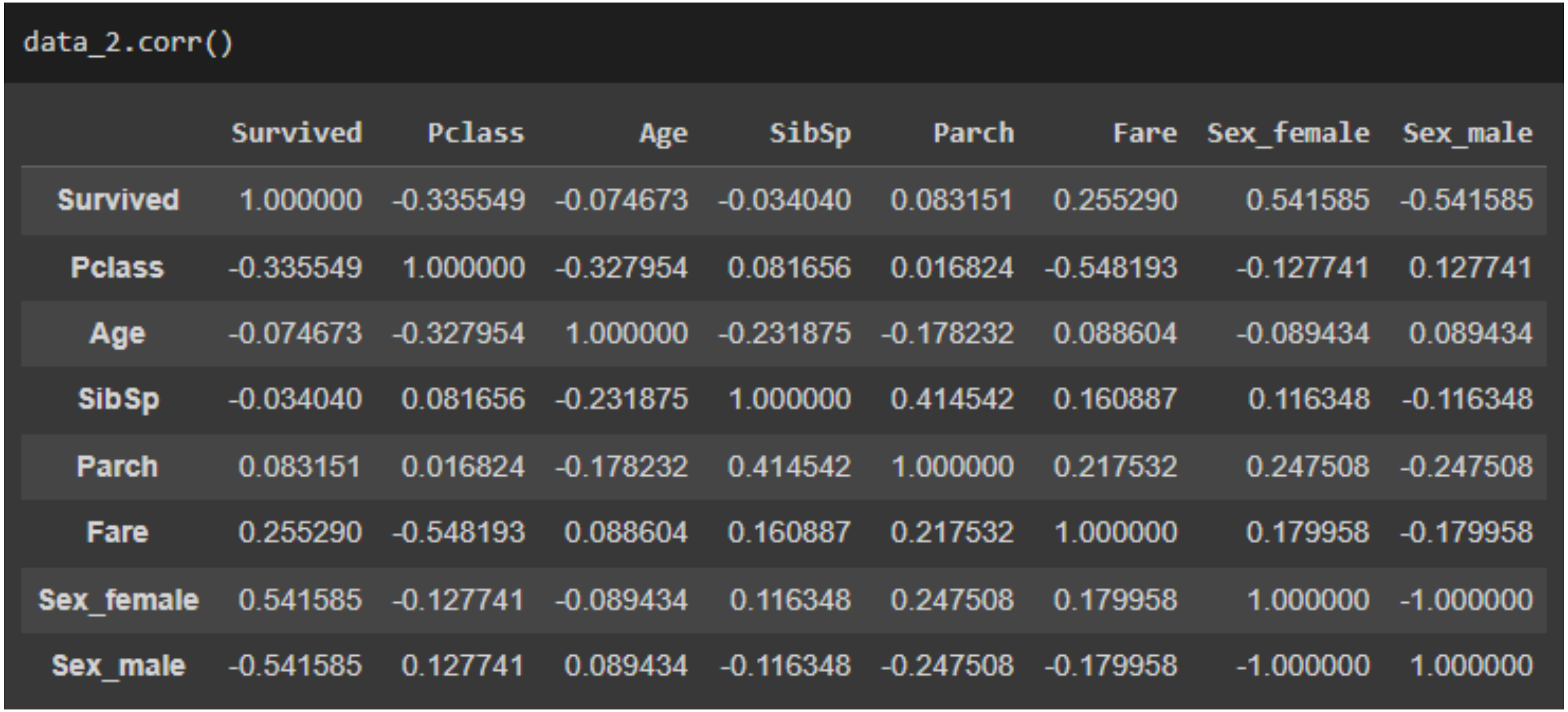


```
from sklearn.preprocessing import StandardScaler
x_scaler = StandardScaler()

x = x_scaler.fit_transform(x_raw_data)

x_df = pd.DataFrame(x)
x_df.head()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0.059319	-0.827704	0.547737	-0.791932	-0.522621	0.230115	0.984944	-1.230105	-1.467485	1.148032	-2.199513	0.529639	0.442067	1.603318	0.018951	0.045641
1	-1.039004	1.324910	-1.278078	0.457292	-0.522621	1.645973	-1.015286	-1.230105	-1.467485	-0.104613	-0.204722	1.540517	0.776882	-1.107336	2.694615	2.084147
2	-1.588165	-1.035950	-0.512974	-0.791932	-0.522621	0.230115	0.984944	-1.230105	-0.293066	-0.104613	0.460208	0.641958	0.330462	-1.446168	0.308210	-0.621982
3	1.706804	1.037749	-1.486743	1.706517	1.913433	-1.185743	-1.015286	1.233390	0.881352	1.148032	-0.869652	1.315877	0.442067	0.135047	-0.147725	-0.814431
4	0.059319	-0.341064	-0.269532	-0.791932	-0.522621	0.230115	0.984944	1.233390	-1.467485	-0.104613	-1.534582	-0.593559	-1.120404	0.699767	-0.987621	-0.852469

Co-relation

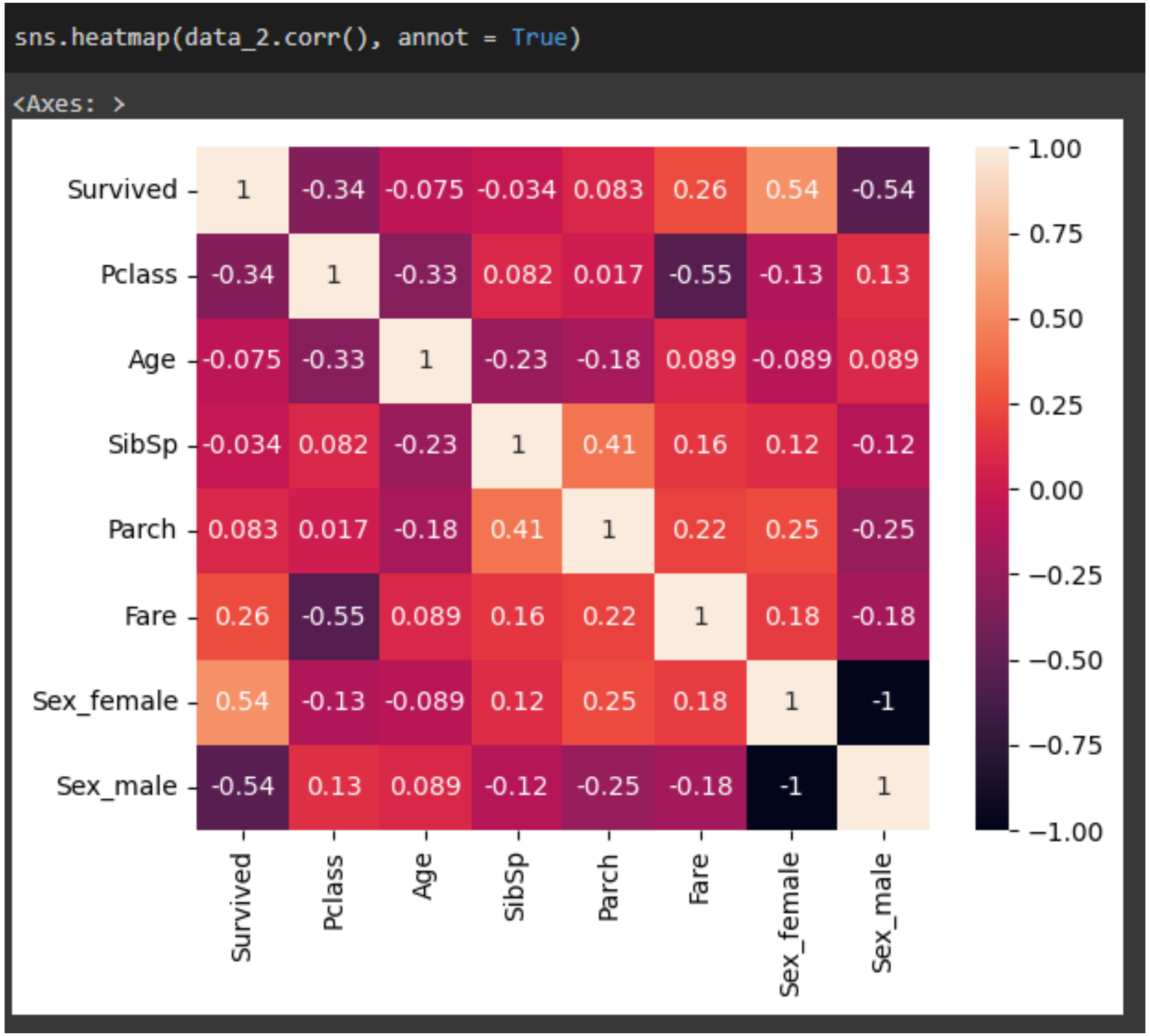
data.corr()



```
data_2.corr()
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male
Survived	1.000000	-0.335549	-0.074673	-0.034040	0.083151	0.255290	0.541585	-0.541585
Pclass	-0.335549	1.000000	-0.327954	0.081656	0.016824	-0.548193	-0.127741	0.127741
Age	-0.074673	-0.327954	1.000000	-0.231875	-0.178232	0.088604	-0.089434	0.089434
SibSp	-0.034040	0.081656	-0.231875	1.000000	0.414542	0.160887	0.116348	-0.116348
Parch	0.083151	0.016824	-0.178232	0.414542	1.000000	0.217532	0.247508	-0.247508
Fare	0.255290	-0.548193	0.088604	0.160887	0.217532	1.000000	0.179958	-0.179958
Sex_female	0.541585	-0.127741	-0.089434	0.116348	0.247508	0.179958	1.000000	-1.000000
Sex_male	-0.541585	0.127741	0.089434	-0.116348	-0.247508	-0.179958	-1.000000	1.000000

Heat map



```
sns.heatmap(data_2.corr(), annot = True)
```

```
<Axes: >
```

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male
Survived	1	-0.34	-0.075	-0.034	0.083	0.26	0.54	-0.54
Pclass	-0.34	1	-0.33	0.082	0.017	-0.55	-0.13	0.13
Age	-0.075	-0.33	1	-0.23	-0.18	0.089	-0.089	0.089
SibSp	-0.034	0.082	-0.23	1	0.41	0.16	0.12	-0.12
Parch	0.083	0.017	-0.18	0.41	1	0.22	0.25	-0.25
Fare	0.26	-0.55	0.089	0.16	0.22	1	0.18	-0.18
Sex_female	0.54	-0.13	-0.089	0.12	0.25	0.18	1	-1
Sex_male	-0.54	0.13	0.089	-0.12	-0.25	-0.18	-1	1

Cluster map

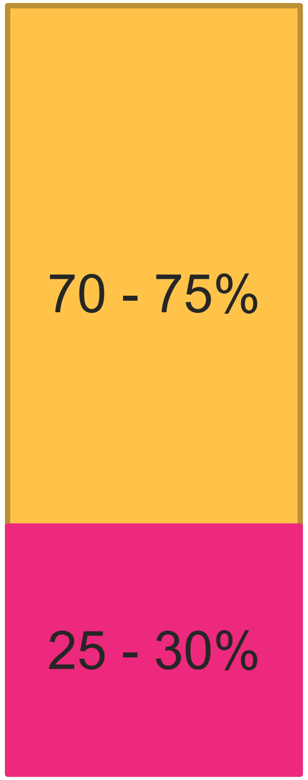
Pairplot

Feature selection

Statistical Inferences

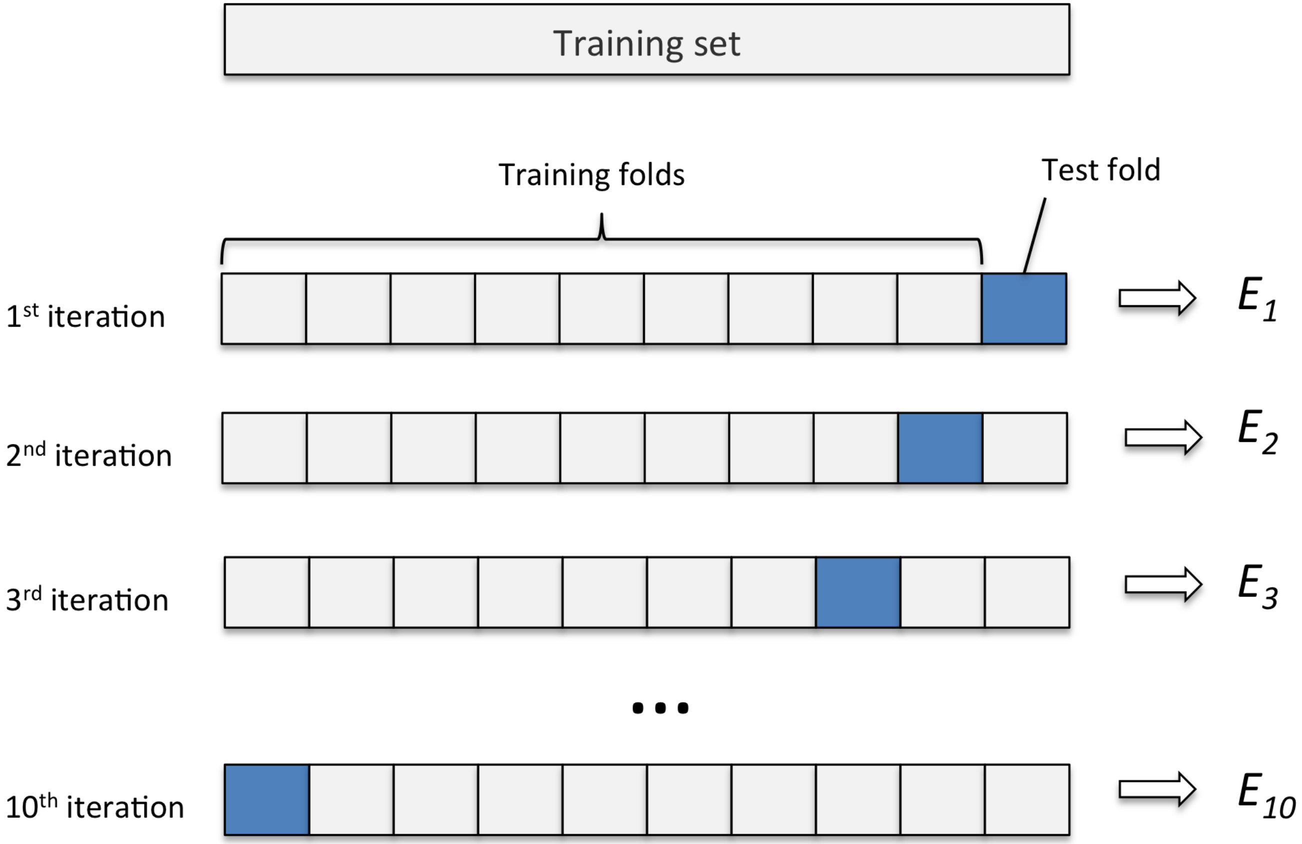
Train – Test Split

Split Validation vs Cross Validation



70 - 75%

25 - 30%



Training set

Training folds

Test fold

1st iteration $\Rightarrow E_1$

2nd iteration $\Rightarrow E_2$

3rd iteration $\Rightarrow E_3$

...

10th iteration $\Rightarrow E_{10}$

Class Balance/Imbalance

Assignment

- Check the need for Encoding, Normalization
- Implement and Justify the above the steps
- Perform EDA using Scatterplot, Heatmaps and Pair plots
- Note down your observation
- Drop or Keep features based your Statistical Inference
- Check the class balance

Outliers