titanic-dataset-exploration

July 9, 2023

0.0.1 Assignment: Perform data cleaning and preprocessing on a given dataset.

- Step 1: Remove duplicates from the dataset.
- Step 2: Handle missing values by imputing or removing them.
- Step 3: Check and handle outliers in the data.
- Step 4: Normalize or standardize numerical features.
- Step 5: Encode categorical variables.

0.0.2 Titanic DataSet Details

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses	
parch	# of parents / children	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

0.0.3 Read CSV

[113]:
```
import pandas as pd
# df = pd.read_csv("train.csv",nrows=20)
trainData = pd.read_csv("train.csv")
trainData
```

[113]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
```

```
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
..                                                ...     …   …    …
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

0.0.4 Remove Duplicates

[114]:
```
trainData = trainData.drop_duplicates()
```

0.0.5 Fill Missing Data

[115]:
```
mean = trainData['Age'].mean()
mean
trainData['Age'] = trainData['Age'].fillna(mean)
trainData
```

[115]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex        Age  \
0                             Braund, Mr. Owen Harris    male  22.000000
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.000000
2                              Heikkinen, Miss. Laina  female  26.000000
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.000000
4                           Allen, Mr. William Henry    male  35.000000
..                                               …     …        …
886                            Montvila, Rev. Juozas    male  27.000000
887                     Graham, Miss. Margaret Edith  female  19.000000
888         Johnston, Miss. Catherine Helen "Carrie"  female  29.699118
889                            Behr, Mr. Karl Howell    male  26.000000
890                              Dooley, Mr. Patrick    male  32.000000

     SibSp  Parch            Ticket     Fare Cabin Embarked
0        1      0         A/5 21171   7.2500   NaN        S
1        1      0          PC 17599  71.2833   C85        C
2        0      0  STON/O2. 3101282   7.9250   NaN        S
3        1      0            113803  53.1000  C123        S
4        0      0            373450   8.0500   NaN        S
..     …    …                …      …   …      …
886      0      0            211536  13.0000   NaN        S
887      0      0            112053  30.0000   B42        S
888      1      2        W./C. 6607  23.4500   NaN        S
889      0      0            111369  30.0000  C148        C
890      0      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

[116]:
```
# removing cabins column as there are alot missing and are not going to be used
trainData.drop(['Cabin'], axis=1, inplace=True)  #inplace changes the original␣
 ↪dataset
trainData
```

[116]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex        Age  \
0                             Braund, Mr. Owen Harris    male  22.000000
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.000000
2                              Heikkinen, Miss. Laina  female  26.000000
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.000000
4                            Allen, Mr. William Henry    male  35.000000
..                                                 …     …        …
886                            Montvila, Rev. Juozas    male  27.000000
887                     Graham, Miss. Margaret Edith  female  19.000000
888          Johnston, Miss. Catherine Helen "Carrie"  female  29.699118
889                            Behr, Mr. Karl Howell    male  26.000000
890                              Dooley, Mr. Patrick    male  32.000000

     SibSp  Parch            Ticket     Fare Embarked
0        1      0         A/5 21171   7.2500        S
1        1      0          PC 17599  71.2833        C
2        0      0  STON/O2. 3101282   7.9250        S
3        1      0            113803  53.1000        S
4        0      0            373450   8.0500        S
..     …    …                 …      …      …
886      0      0            211536  13.0000        S
887      0      0            112053  30.0000        S
888      1      2        W./C. 6607  23.4500        S
889      0      0            111369  30.0000        C
890      0      0            370376   7.7500        Q

[891 rows x 11 columns]
```

0.0.6 Handling Outliers

[60]:
```
pip install seaborn
```

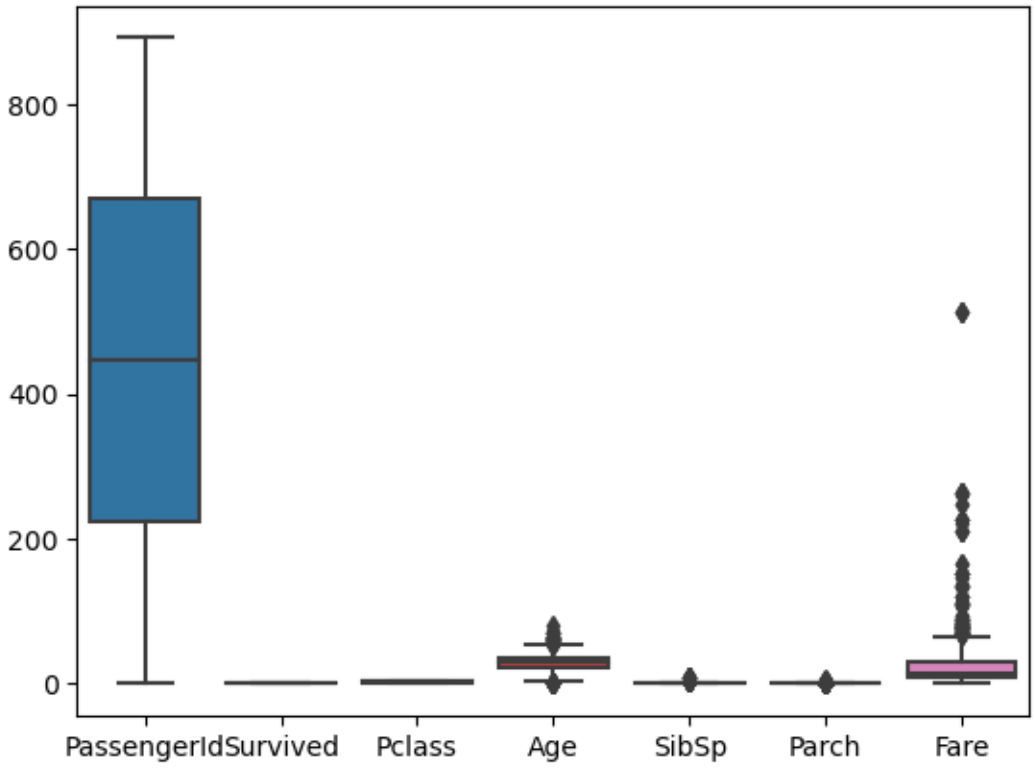
```
Requirement already satisfied: seaborn in d:\apps\anaconda\files\lib\site-
packages (0.12.2)Note: you may need to restart the kernel to use updated
```

```
packages.

Requirement already satisfied: numpy!=1.24.0,>=1.17 in
d:\apps\anaconda\files\lib\site-packages (from seaborn) (1.23.5)
Requirement already satisfied: pandas>=0.25 in d:\apps\anaconda\files\lib\site-
packages (from seaborn) (1.4.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in
d:\apps\anaconda\files\lib\site-packages (from seaborn) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(1.0.5)
Requirement already satisfied: cycler>=0.10 in d:\apps\anaconda\files\lib\site-
packages (from matplotlib!=3.6.1,>=3.1->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(1.4.4)
Requirement already satisfied: packaging>=20.0 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(23.0)
Requirement already satisfied: pillow>=6.2.0 in d:\apps\anaconda\files\lib\site-
packages (from matplotlib!=3.6.1,>=3.1->seaborn) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(2.8.2)
Requirement already satisfied: importlib-resources>=3.2.0 in
d:\apps\anaconda\files\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn)
(5.2.0)
Requirement already satisfied: pytz>=2020.1 in d:\apps\anaconda\files\lib\site-
packages (from pandas>=0.25->seaborn) (2022.7)
Requirement already satisfied: zipp>=3.1.0 in d:\apps\anaconda\files\lib\site-
packages (from importlib-resources>=3.2.0->matplotlib!=3.6.1,>=3.1->seaborn)
(3.11.0)
Requirement already satisfied: six>=1.5 in d:\apps\anaconda\files\lib\site-
packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.1->seaborn) (1.12.0)
```

[62]:
```
import seaborn as sb
```

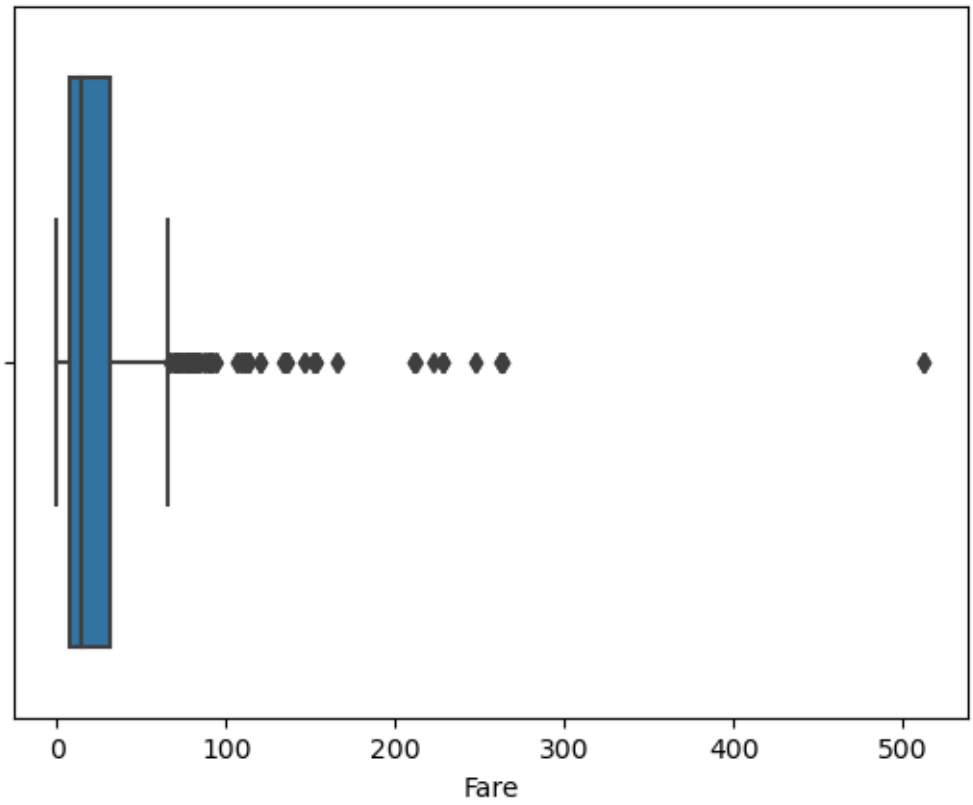
[117]:
```
sb.boxplot(trainData)
```

[117]:
```
<Axes: >
```



[118]:
```
sb.boxplot(x = trainData['Fare'])
```

[118]:
```
<Axes: xlabel='Fare'>
```



[119]:
```
# We will use the IQR or turkey's rule for handling outliers
trainData[['Fare']].describe()
```

[119]:

	Fare
count	891.000000
mean	32.204208
std	49.693429
min	0.000000
25%	7.910400
50%	14.454200
75%	31.000000
max	512.329200

0.0.7 Calculate IQR (Inter Quartile Range)

[120]:
```
# 25%
Q1 = trainData['Fare'].quantile(0.25)
Q3 = trainData['Fare'].quantile(0.75)
IQR = Q3-Q1
IQR
```

[120]:
```
23.0896
```

[121]:
```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
```

[122]:
```
lower_limit
```

[122]:
```
-26.724
```

[123]:
```
upper_limit
```

[123]:
```
65.6344
```

[124]:
```
# DROPPING OUTLIERS
outliers_low = (trainData['Fare']<lower_limit)
outliers_up = (trainData['Fare']>upper_limit)
```

[125]:
```
trainData[(outliers_low | outliers_up)]
```

[125]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.000000	1	0	PC 17599	71.2833	C
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.000000					
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	29.699118					
34	35	0	1	Meyer, Mr. Edgar Joseph	male	28.000000					
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.000000					
..	…	…	…	…	…	…					
846	847	0	3	Sage, Mr. Douglas Bullen	male	29.699118					
849	850	1	1	Goldenberg, Mrs. Samuel L (Edwiga Grabowska)	female	29.699118					
856	857	1	1	Wick, Mrs. George Dennick (Mary Hitchcock)	female	45.000000					
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	29.699118					
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.000000					

```
27       3      2     19950  263.0000        S
31       1      0  PC 17569  146.5208        C
34       1      0  PC 17604   82.1708        C
52       1      0  PC 17572   76.7292        C
..     …    …       …       …       …
846      8      2  CA. 2343   69.5500        S
849      1      0     17453   89.1042        C
856      1      1     36928  164.8667        S
863      8      2  CA. 2343   69.5500        S
879      0      1     11767   83.1583        C

[116 rows x 11 columns]
```

0.0.8 Drop Outliers

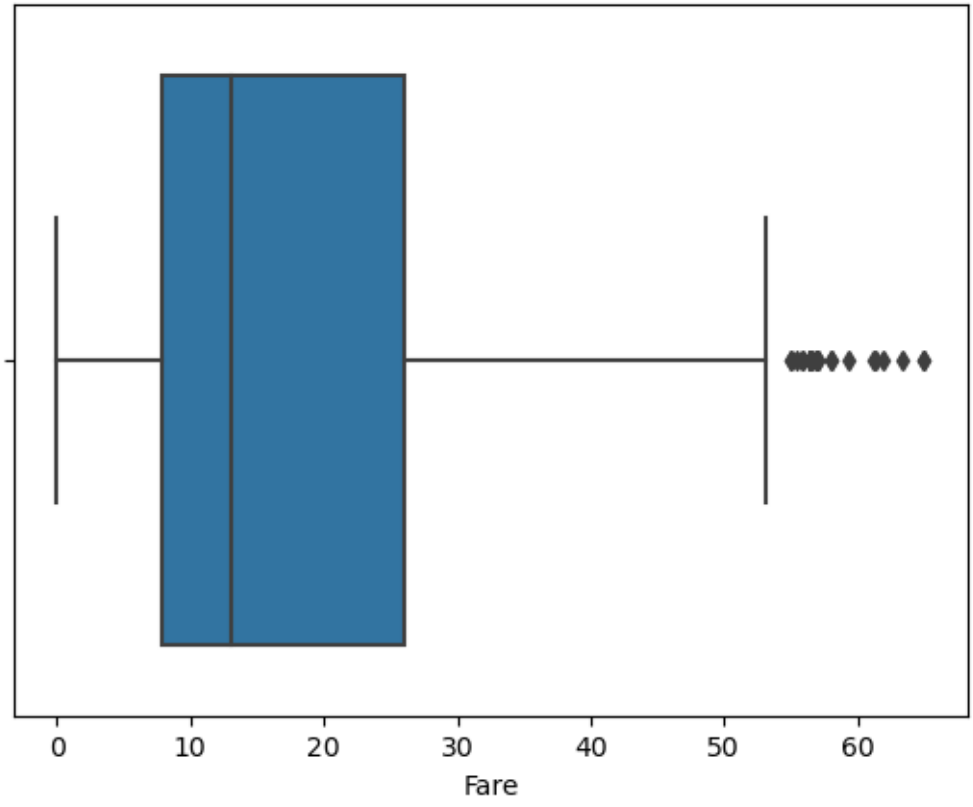
[127]:
```
trainData = trainData[~(outliers_low | outliers_up)]
trainData

# as you can see we have reduced 891 to 775 by removing outliers in Fare column
```

[127]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
2              3         1       3
3              4         1       1
4              5         0       3
5              6         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex         Age  SibSp  \
0                             Braund, Mr. Owen Harris    male   22.000000      1
2                              Heikkinen, Miss. Laina  female   26.000000      0
3    Futrelle, Mrs. Jacques Heath (Lily May Peel)  female   35.000000      1
4                           Allen, Mr. William Henry    male   35.000000      0
5                                   Moran, Mr. James    male   29.699118      0
..                                                ...     …         …    …
886                            Montvila, Rev. Juozas    male   27.000000      0
887                     Graham, Miss. Margaret Edith  female   19.000000      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   29.699118      1
889                            Behr, Mr. Karl Howell    male   26.000000      0
890                              Dooley, Mr. Patrick    male   32.000000      0

     Parch            Ticket     Fare Embarked
```

```
0        0         A/5 21171   7.2500        S
2        0  STON/O2. 3101282   7.9250        S
3        0            113803  53.1000        S
4        0            373450   8.0500        S
5        0            330877   8.4583        Q
..     ...               ...      ...      ...
886      0            211536  13.0000        S
887      0            112053  30.0000        S
888      2        W./C. 6607  23.4500        S
889      0            111369  30.0000        C
890      0            370376   7.7500        Q

[775 rows x 11 columns]
```

[130]:
```
sb.boxplot(x=trainData['Fare'])
```

[130]:
```
<Axes: xlabel='Fare'>
```



[131]:
```
trainData
```

[131]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	S
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	S
5	6	0	3	Moran, Mr. James	male	29.699118	0	0	330877	8.4583	Q
..	…	…	…	…	…	…	…	…	…	…	…
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	Q

[775 rows x 11 columns]

[ ]:

0.0.9 Data Details

[18]:
```
# Exploratory data analysis
trainData.head()
```

[18]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

[19]:
```
# Data info
trainData.info
```

[19]:
```
<bound method DataFrame.info of      PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
```

```
..                                                 …     …   …    …
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …       …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]>
```

[20]:
```
trainData.describe()
```

[20]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[21]:
```
testData = pd.read_csv("test.csv")
testData
```

[21]:
```
     PassengerId  Pclass                                          Name  \
0            892       3                              Kelly, Mr. James
1            893       3              Wilkes, Mrs. James (Ellen Needs)
2            894       2                     Myles, Mr. Thomas Francis
3            895       3                              Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..           …     …                                             …
413         1305       3                           Spector, Mr. Woolf
414         1306       1                 Oliva y Ocana, Dona. Fermina
415         1307       3                 Saether, Mr. Simon Sivertsen
416         1308       3                        Ware, Mr. Frederick
417         1309       3                   Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0      male  34.5      0      0              330911    7.8292   NaN        Q
1    female  47.0      1      0              363272    7.0000   NaN        S
2      male  62.0      0      0              240276    9.6875   NaN        Q
3      male  27.0      0      0              315154    8.6625   NaN        S
4    female  22.0      1      1             3101298   12.2875   NaN        S
..      …   …    …    …                 …        …   …       …
413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0            PC 17758  108.9000  C105        C
415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0              359309    8.0500   NaN        S
417    male   NaN      1      1                2668   22.3583   NaN        C

[418 rows x 11 columns]
```

[22]:
```
testData.head()
```

[22]:
```
   PassengerId  Pclass                                          Name     Sex  \
0          892       3                              Kelly, Mr. James    male
1          893       3              Wilkes, Mrs. James (Ellen Needs)  female
2          894       2                     Myles, Mr. Thomas Francis    male
3          895       3                              Wirz, Mr. Albert    male
4          896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female

    Age  SibSp  Parch   Ticket     Fare Cabin Embarked
0  34.5      0      0   330911   7.8292   NaN        Q
1  47.0      1      0   363272   7.0000   NaN        S
2  62.0      0      0   240276   9.6875   NaN        Q
3  27.0      0      0   315154   8.6625   NaN        S
4  22.0      1      1  3101298  12.2875   NaN        S
```

[23]:
```
testData.info
```

[23]:
```
<bound method DataFrame.info of      PassengerId  Pclass
Name  \
0            892       3                                Kelly, Mr. James
1            893       3                Wilkes, Mrs. James (Ellen Needs)
2            894       2                       Myles, Mr. Thomas Francis
3            895       3                               Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..           …     …                                               …
413         1305       3                            Spector, Mr. Woolf
414         1306       1                  Oliva y Ocana, Dona. Fermina
415         1307       3                  Saether, Mr. Simon Sivertsen
416         1308       3                           Ware, Mr. Frederick
417         1309       3                      Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0      male  34.5      0      0              330911    7.8292   NaN        Q
1    female  47.0      1      0              363272    7.0000   NaN        S
2      male  62.0      0      0              240276    9.6875   NaN        Q
3      male  27.0      0      0              315154    8.6625   NaN        S
4    female  22.0      1      1             3101298   12.2875   NaN        S
..      …   …    …    …                 …        …   …      …
413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0            PC 17758  108.9000  C105        C
415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0              359309    8.0500   NaN        S
417    male   NaN      1      1                2668   22.3583   NaN        C

[418 rows x 11 columns]>
```

[24]:
```
testData.describe()
```

[24]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

[26]:
```
trainData
```

[26]:

	PassengerId	Survived	Pclass
0	1	0	3
1	2	1	1
2	3	1	3

```
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
..                                                ...     …   …    …
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

0.0.10 Calculating Numerical Data

[36]:
```
# trainData.loc[trainData['Sex']=='male'][trainData['Survived']==1]
survived = trainData.loc[trainData['Sex']=='male']['Survived']
percent = sum(survived)/len(survived)
percent
```

[36]:
```
0.18890814558058924
```

[37]:
```
# trainData.loc[trainData['Sex']=='female'][trainData['Survived']==1]
survived = trainData.loc[trainData['Sex']=='female']['Survived']
percent = sum(survived)/len(survived)
percent
```

[37]:
```
0.7420382165605095
```

[40]:
```
survived = trainData.loc[trainData['Age']>20]['Survived']
percent = sum(survived)/len(survived)
percent
```

[40]:
```
0.38878504672897196
```

[41]:
```
survived = trainData.loc[trainData['Age']>75]['Survived']
percent = sum(survived)/len(survived)
percent
```

[41]:
```
1.0
```

0.0.11 Handling Categorical Data

[135]:
```
# Categorical Data
trainData['Sex'].value_counts()
```

[135]:
```
male      531
female    244
Name: Sex, dtype: int64
```

[141]:
```
trainData['Sex'] = trainData['Sex'].astype('category')
trainData['Sex_categorical'] = trainData['Sex'].astype('category').cat.codes

trainData
```

```
C:\Users\Hp\AppData\Local\Temp\ipykernel_9000\3129464243.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  trainData['Sex'] = trainData['Sex'].astype('category')
C:\Users\Hp\AppData\Local\Temp\ipykernel_9000\3129464243.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  trainData['Sex_categorical'] = trainData['Sex'].astype('category').cat.codes
```

[141]:

	PassengerId	Survived	Pclass	\
0	1	0	3	
2	3	1	3	
3	4	1	1	
4	5	0	3	
5	6	0	3	
..	…	…	…	
886	887	0	2	
887	888	1	1	
888	889	0	3	
889	890	1	1	
890	891	0	3	

	Name	Sex	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.000000	1	
2	Heikkinen, Miss. Laina	female	26.000000	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	
4	Allen, Mr. William Henry	male	35.000000	0	
5	Moran, Mr. James	male	29.699118	0	
..	…	…	…	…	
886	Montvila, Rev. Juozas	male	27.000000	0	
887	Graham, Miss. Margaret Edith	female	19.000000	0	
888	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	
889	Behr, Mr. Karl Howell	male	26.000000	0	
890	Dooley, Mr. Patrick	male	32.000000	0	

	Parch	Ticket	Fare	Embarked	Sex_categorical
0	0	A/5 21171	7.2500	S	1
2	0	STON/O2. 3101282	7.9250	S	0
3	0	113803	53.1000	S	0
4	0	373450	8.0500	S	1
5	0	330877	8.4583	Q	1
..	…	…	…	…	…
886	0	211536	13.0000	S	1
887	0	112053	30.0000	S	0
888	2	W./C. 6607	23.4500	S	0
889	0	111369	30.0000	C	1
890	0	370376	7.7500	Q	1

[775 rows x 12 columns]