Descriptive Statistics- Titanic Kaggle M.Alimi

September 23, 2020

1   # Week 1 Discussion

Download the train.csv dataset and run some basic descriptive statistics and graphs for two or
more variables of your choosing using Python. Provide your Python code here, perhaps as a
Jupyter notebook .html file on GitHub.

```
[1]: # Load packages
     import pandas as pd
     import matplotlib.pyplot as plt
```

```
[2]: # Read data
     titanicdata = pd.read_csv('train.csv',sep=',')
```

```
[3]: # Show a sample of the first 10 rows of data
     titanicdata.head(10)
```

```
[3]:    PassengerId  Survived  Pclass  \
     0            1         0       3
     1            2         1       1
     2            3         1       3
     3            4         1       1
     4            5         0       3
     5            6         0       3
     6            7         0       1
     7            8         0       3
     8            9         1       3
     9           10         1       2

                                                     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
     5                                     Moran, Mr. James    male   NaN      0
     6                              McCarthy, Mr. Timothy J    male  54.0      0
     7                       Palsson, Master. Gosta Leonard    male   2.0      3
     8    Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female  27.0      0
     9                   Nasser, Mrs. Nicholas (Adele Achem)  female  14.0      1
```

```
     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
5        0            330877   8.4583   NaN        Q
6        0             17463  51.8625   E46        S
7        1            349909  21.0750   NaN        S
8        2            347742  11.1333   NaN        S
9        0            237736  30.0708   NaN        C
```

[4]: ```
# Understand the data types of the data
titanicdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

[5]: ```
# Summarize the numerical fields in the data
titanicdata.describe()
```

[5]:
	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
```
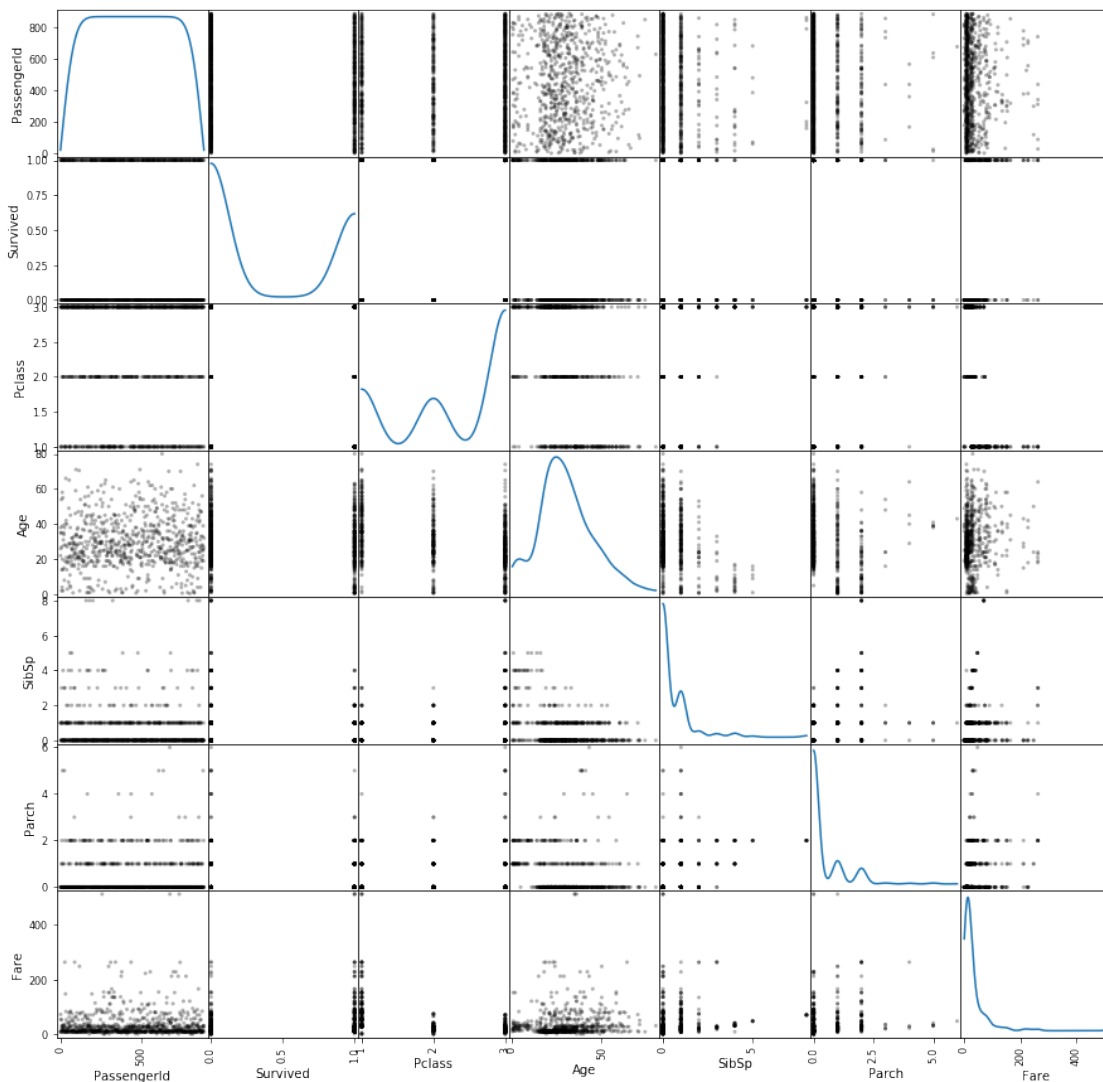
```
count   891.000000   891.000000
mean      0.381594    32.204208
std       0.806057    49.693429
min       0.000000     0.000000
25%       0.000000     7.910400
50%       0.000000    14.454200
75%       0.000000    31.000000
max       6.000000   512.329200
```

[6]:
```python
# Exploratory data analysis
pd.plotting.scatter_matrix(titanicdata, diagonal='kde', color='black', alpha=0.
 ↪3, figsize=(15,15))
```

[6]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1944240>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1985550>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC19AE908>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC19D6F98>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1A08668>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1A086A0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1A5F3C8>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1A87A58>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1AB7128>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1ADE7B8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1B06E48>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1B36518>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1B5FBA8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1B90278>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1BB7908>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1BE0F98>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1C0E668>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1C37CF8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1C683C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1C90A58>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1CC3128>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1CEA7B8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1D11E48>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1D42518>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1D6ABA8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1D9A278>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1DC3908>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1DEAF98>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1E1B668>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1E41CF8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1E733C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1E9CA58>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1ECE128>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1EF27B8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1F1DE48>],

```
[<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1F4D518>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1F74BA8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1FA8278>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1FCE908>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC1FF7F98>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC2024668>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC204ECF8>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC207E3C8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC20A6A58>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC20D9128>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC21017B8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC2127E48>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC2157518>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x0000020BC2182BA8>]],
dtype=object)
```

```
[8]: # Plot to show the Class and Frequency

     plt.figure(figsize=(10,10))
     plt.hist(titanicdata['Pclass'], 3, color='blue', alpha=0.3)
     plt.xticks(titanicdata['Pclass'])
     plt.title('Class Count')
     plt.xlabel('Class')
     plt.ylabel('Frequency')
```

[8]: Text(0,0.5,'Frequency')
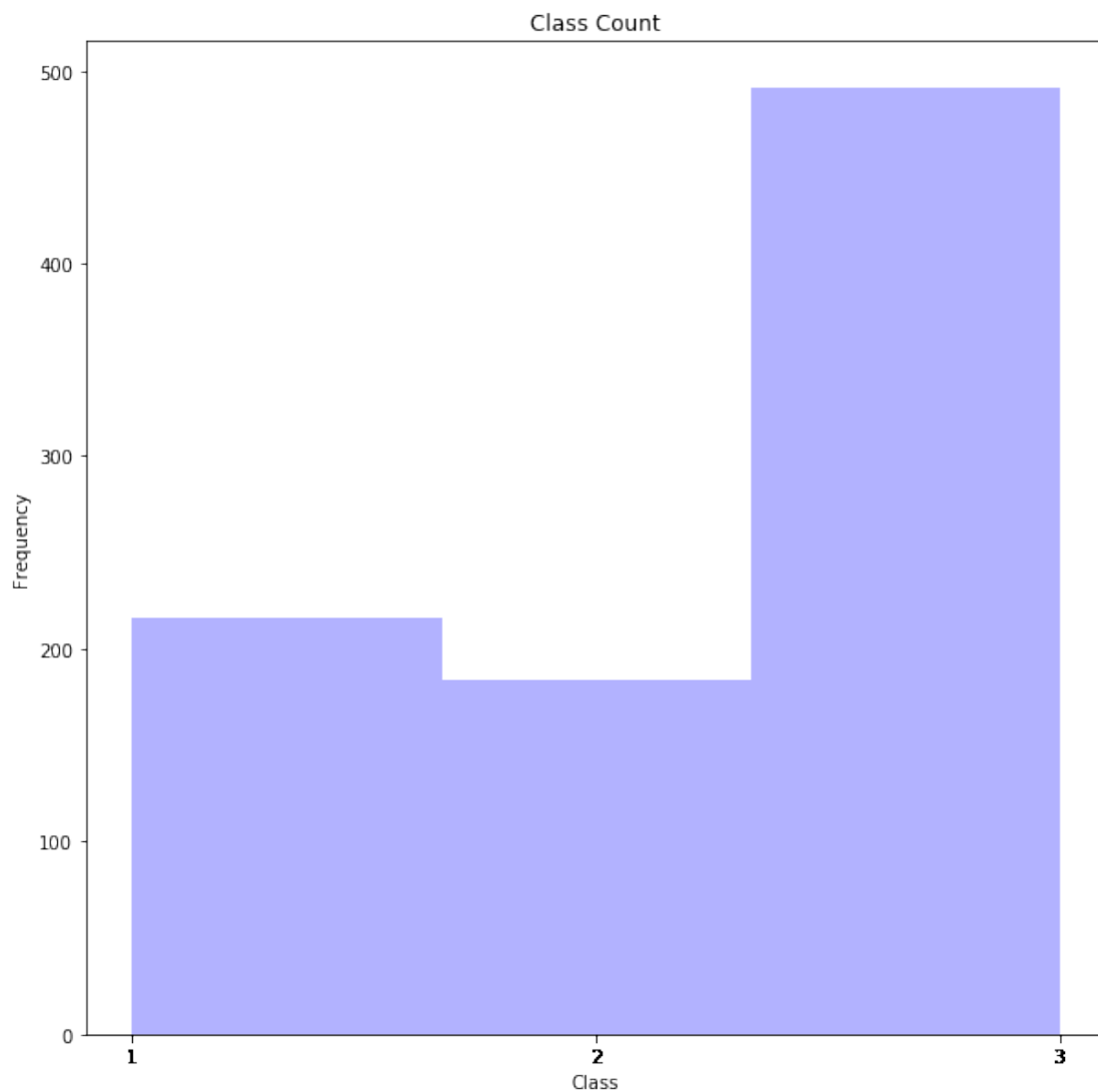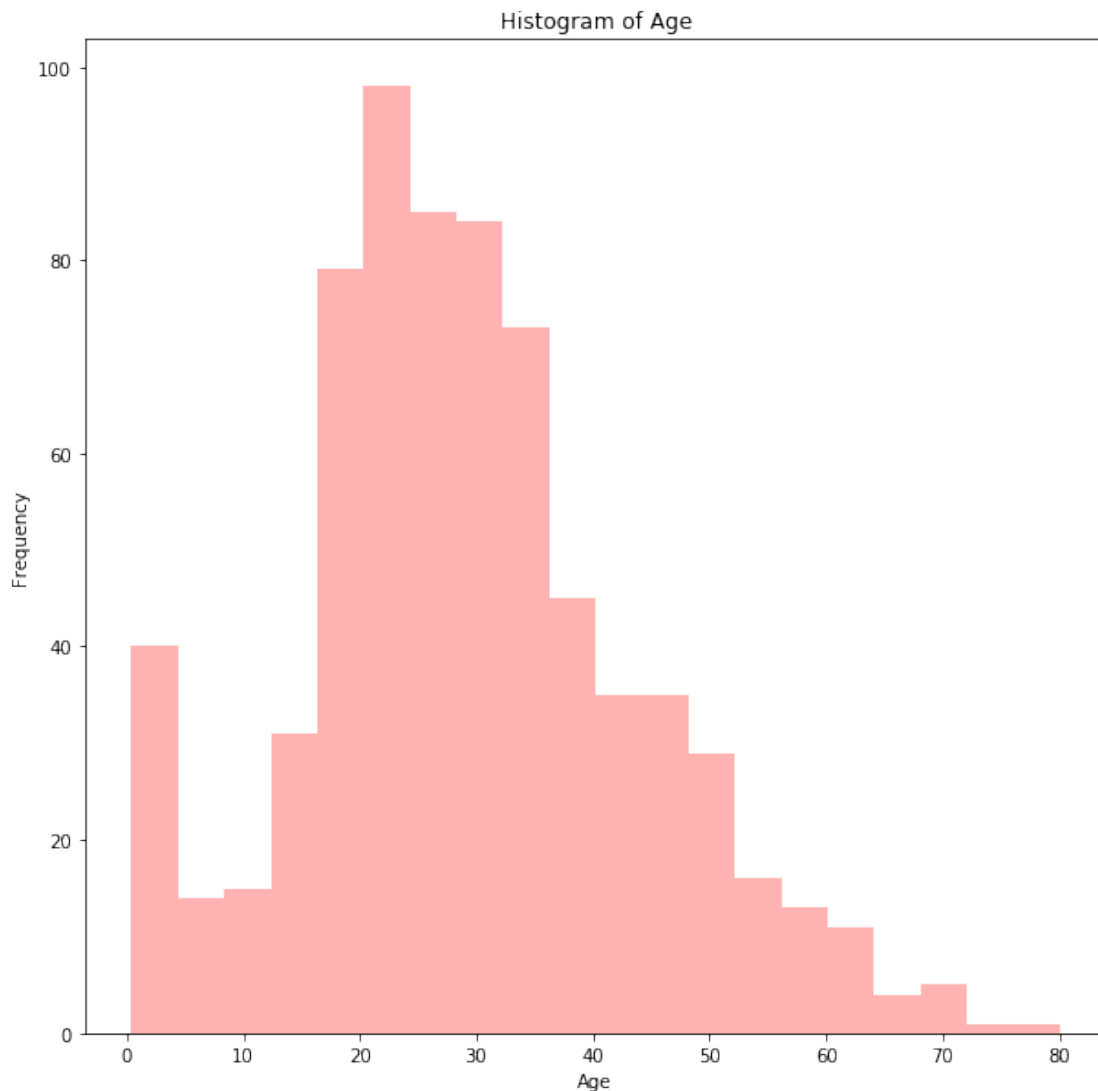


```
[9]: # Plot to show the Age and Frequency

     plt.figure(figsize=(10,10))
```

```
titanicage = titanicdata['Age']
plt.hist(titanicage.dropna(), 20, color='red', alpha=0.3)
plt.title('Histogram of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
```
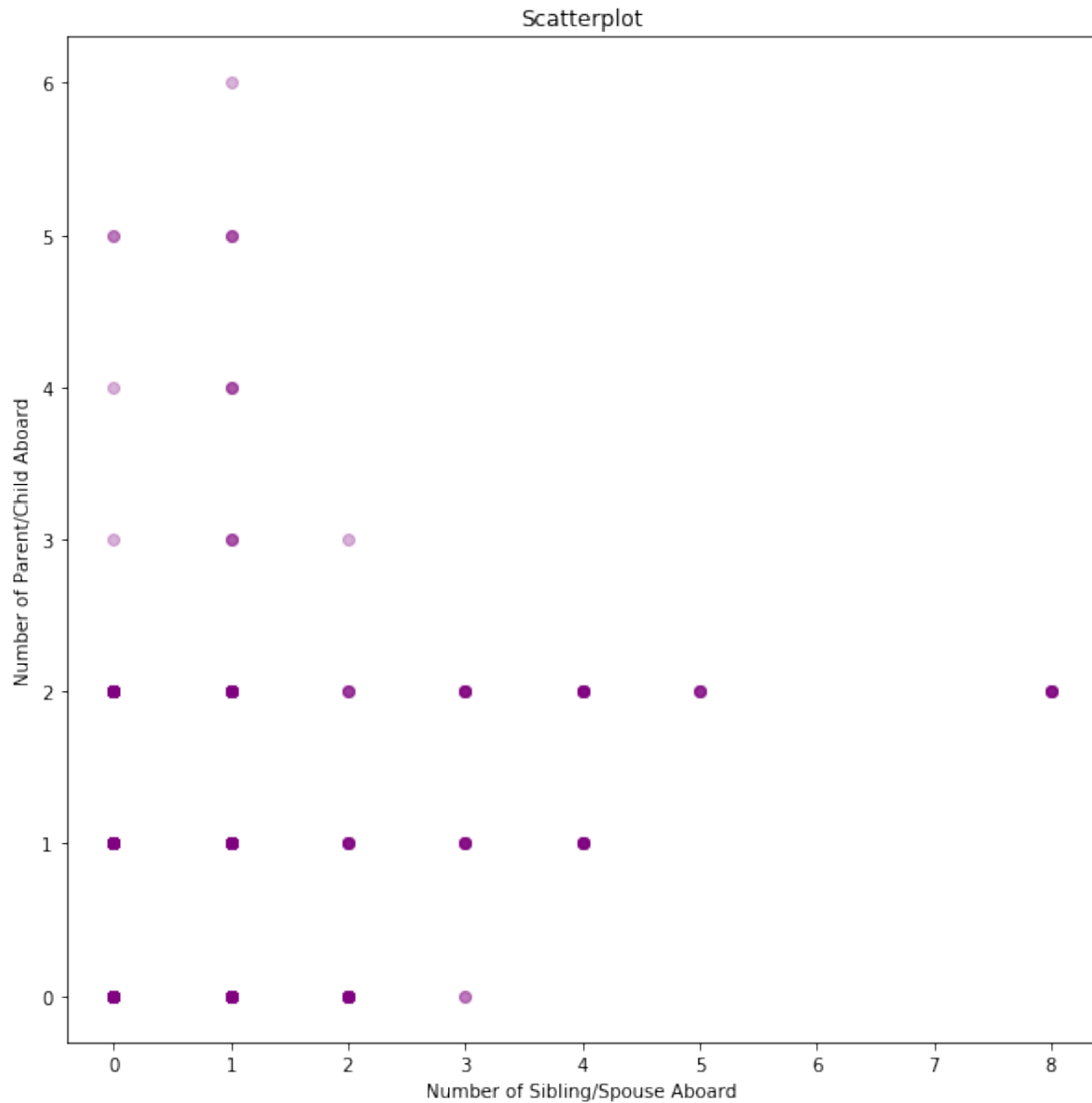
[9]: Text(0,0.5,'Frequency')

[10]:
```
# Plot to show the No of Sibling/Spouse Abroad and No of Parent/Child Abroad

plt.figure(figsize=(10,10))
plt.scatter(titanicdata['SibSp'], titanicdata['Parch'], color='purple', alpha=0.
 ↪3)
plt.title('Scatterplot')
```
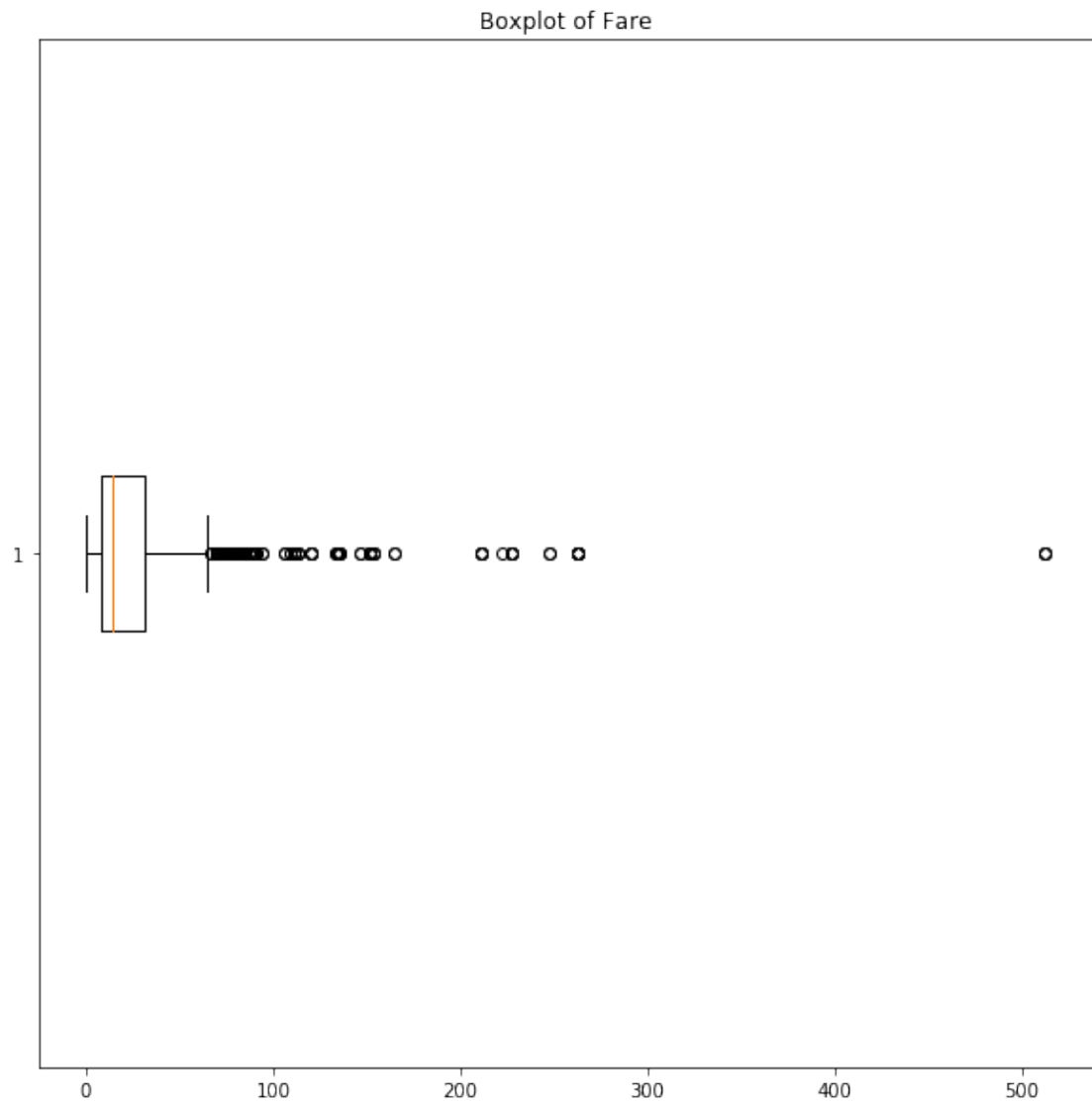
```
plt.xlabel('Number of Sibling/Spouse Aboard')
plt.ylabel('Number of Parent/Child Aboard')
```

[10]: Text(0,0.5,'Number of Parent/Child Aboard')

[11]:
```
# Plot to show the Fare
plt.figure(figsize=(10,10))
plt.boxplot(titanicdata['Fare'], vert=False)
plt.title('Boxplot of Fare')
```

[11]: Text(0.5,1,'Boxplot of Fare')

Boxplot of Fare

[ ]: