In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [29]:

```
df= pd.read_csv('titanic3.csv')
df.head(
)
```

Out[29]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   pclass     1309 non-null   int64  
 1   survived   1309 non-null   int64  
 2   name       1309 non-null   object 
 3   sex        1309 non-null   object 
 4   age        1046 non-null   float64
 5   sibsp      1309 non-null   int64  
 6   parch      1309 non-null   int64  
 7   ticket     1309 non-null   object 
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object 
 10  embarked   1307 non-null   object 
 11  boat       486 non-null    object 
 12  body       121 non-null    float64
 13  home.dest  745 non-null    object 
dtypes: float64(3), int64(4), object(7)
memory usage: 143.3+ KB
```

In [5]:

```
df.select_dtypes(include = 'object').nunique()
```

Out[5]:

```
name         1307
sex             2
ticket        929
cabin         186
embarked        3
boat           27
home.dest     369
dtype: int64
```

In [6]:

```
df.drop(columns= ['cabin', 'boat','body','home.dest'], inplace = True)
```

In [7]:

```
df.head()
```

Out[7]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	embarked
0	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	S
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	S
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	S
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	S
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	S

In [8]:

```
df.isnull().sum()
```

Out[8]:

```
pclass        0
survived      0
name          0
sex           0
age         263
sibsp         0
parch         0
ticket        0
fare          1
embarked      2
dtype: int64
```

In [9]:

```
df[df['age'].isnull()]
```

Out[9]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	embarked
15	1	0	Baumann, Mr. John D	male	NaN	0	0	PC 17318	25.9250	S
37	1	1	Bradley, Mr. George ("George Arthur Brayton")	male	NaN	0	0	111427	26.5500	S
40	1	0	Brewe, Dr. Arthur Jackson	male	NaN	0	0	112379	39.6000	C
46	1	0	Cairns, Mr. Alexander	male	NaN	0	0	113798	31.0000	S
59	1	1	Cassebeer, Mrs. Henry Arthur Jr (Eleanor Genev...	female	NaN	0	0	17770	27.7208	C
...	...	...	...	...	...	...	...	...	...	...
1293	3	0	Williams, Mr. Howard Hugh "Harry"	male	NaN	0	0	A/5 2466	8.0500	S
1297	3	0	Wiseman, Mr. Phillippe	male	NaN	0	0	A/4. 34244	7.2500	S
1302	3	0	Yousif, Mr. Wazli	male	NaN	0	0	2647	7.2250	C
1303	3	0	Yousseff, Mr. Gerious	male	NaN	0	0	2627	14.4583	C
1305	3	0	Zabour, Miss. Thamine	female	NaN	1	0	2665	14.4542	C

263 rows × 10 columns

In [10]:

```
#fill Nan with median numbers
df.fillna(df.median(), inplace = True)
```

```
C:\Users\ASUS-PC\AppData\Local\Temp\ipykernel_6900\3097012583.py:1: FutureWarning: The default value of numeric_only in Dat
aFrame.median is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is
deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.fillna(df.median(), inplace = True)
```

In [11]:

```
df[df['embarked'].isnull()]
```

Out[11]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	embarked
168	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0	NaN
284	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.0	0	0	113572	80.0	NaN

In [12]:

```
df.iloc[284]
```

Out[12]:

```
pclass                                        1
survived                                      1
name        Stone, Mrs. George Nelson (Martha Evelyn)
sex                                        female
age                                          62.0
sibsp                                           0
parch                                           0
ticket                                     113572
fare                                         80.0
embarked                                      NaN
Name: 284, dtype: object
```

In [14]:

```
df.dropna(inplace = True)
```

In [15]:

```
df.iloc[284]
```

Out[15]:

```
pclass                                          1
survived                                        0
name        Straus, Mrs. Isidor (Rosalie Ida Blun)
sex                                        female
age                                          63.0
sibsp                                           1
parch                                           0
ticket                                   PC 17483
fare                                     221.7792
embarked                                        S
Name: 286, dtype: object
```

In [16]:

```
#our datas are clean and uniform
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1307 entries, 0 to 1308
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   pclass    1307 non-null   int64  
 1   survived  1307 non-null   int64  
 2   name      1307 non-null   object 
 3   sex       1307 non-null   object 
 4   age       1307 non-null   float64
 5   sibsp     1307 non-null   int64  
 6   parch     1307 non-null   int64  
 7   ticket    1307 non-null   object 
 8   fare      1307 non-null   float64
 9   embarked  1307 non-null   object 
dtypes: float64(2), int64(4), object(4)
memory usage: 112.3+ KB
```

In [17]:

```
df.describe()
```

Out[17]:

	pclass	survived	age	sibsp	parch	fare
count	1307.000000	1307.000000	1307.000000	1307.000000	1307.000000	1307.000000
mean	2.296863	0.381025	29.471821	0.499617	0.385616	33.209595
std	0.836942	0.485825	12.881592	1.042273	0.866092	51.748768
min	1.000000	0.000000	0.170000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	22.000000	0.000000	0.000000	7.895800
50%	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200
75%	3.000000	1.000000	35.000000	1.000000	0.000000	31.275000
max	3.000000	1.000000	80.000000	8.000000	9.000000	512.329200

In [18]:

```
df.describe().T
```

Out[18]:

	count	mean	std	min	25%	50%	75%	max
pclass	1307.0	2.296863	0.836942	1.00	2.0000	3.0000	3.000	3.0000
survived	1307.0	0.381025	0.485825	0.00	0.0000	0.0000	1.000	1.0000
age	1307.0	29.471821	12.881592	0.17	22.0000	28.0000	35.000	80.0000
sibsp	1307.0	0.499617	1.042273	0.00	0.0000	0.0000	1.000	8.0000
parch	1307.0	0.385616	0.866092	0.00	0.0000	0.0000	0.000	9.0000
fare	1307.0	33.209595	51.748768	0.00	7.8958	14.4542	31.275	512.3292

In [19]:

```
df.corr()
```

```
C:\Users\ASUS-PC\AppData\Local\Temp\ipykernel_6900\1134722465.py:1: FutureWarning: The default value of numeric_only in Dat
aFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of
numeric_only to silence this warning.
  df.corr()
```

Out[19]:

	pclass	survived	age	sibsp	parch	fare
pclass	1.000000	-0.310412	-0.375811	0.059819	0.017304	-0.557915
survived	-0.310412	1.000000	-0.047090	-0.026931	0.083642	0.243109
age	-0.375811	-0.047090	1.000000	-0.189332	-0.125112	0.176554
sibsp	0.059819	-0.026931	-0.189332	1.000000	0.373383	0.161141
parch	0.017304	0.083642	-0.125112	0.373383	1.000000	0.222422
fare	-0.557915	0.243109	0.176554	0.161141	0.222422	1.000000

In [20]:

```
df.skew()
```

```
C:\Users\ASUS-PC\AppData\Local\Temp\ipykernel_6900\1665899112.py:1: FutureWarning: The default value of numeric_only in Dat
aFrame.skew is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is de
precated. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.skew()
```

Out[20]:

```
pclass     -0.602921
survived    0.490535
age         0.539590
sibsp       3.841281
parch       3.666038
fare        4.377390
dtype: float64
```

In [21]:

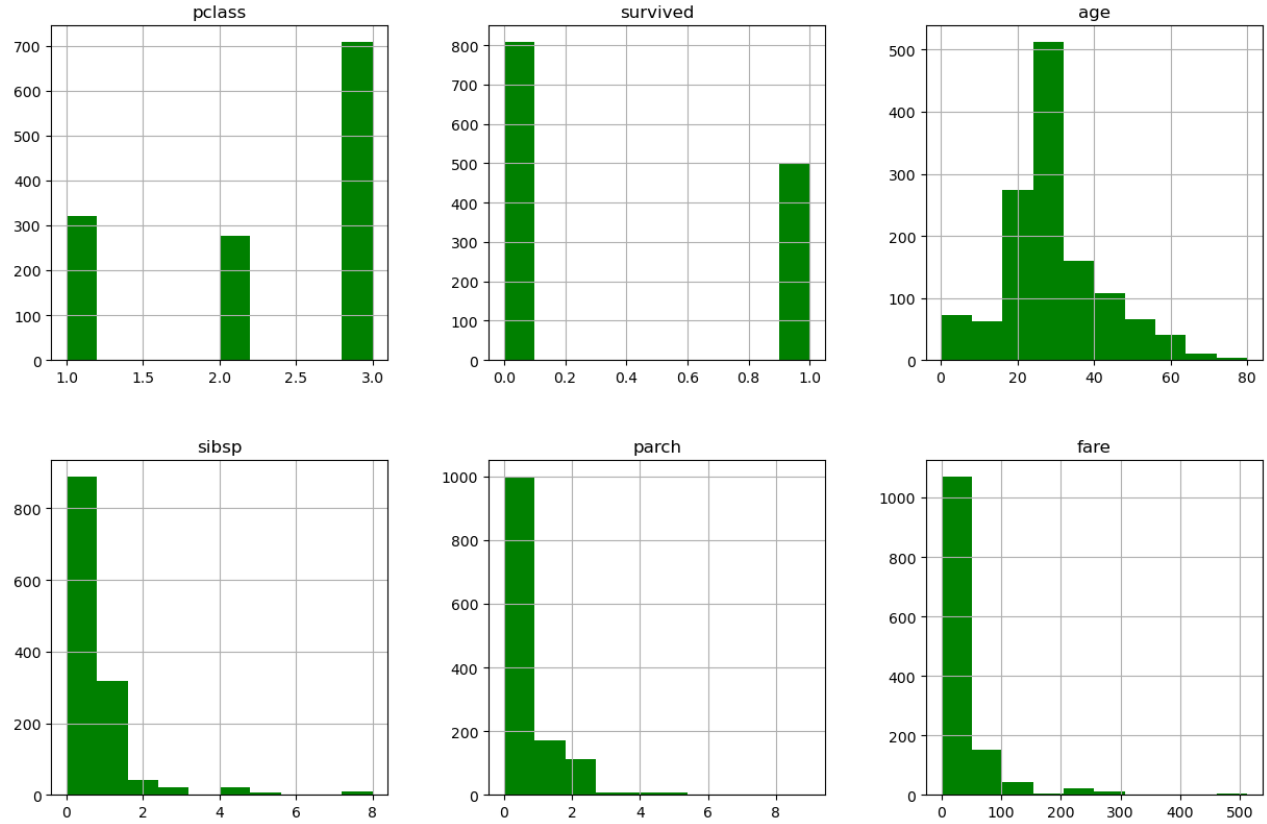
```
df['survived'].value_counts()
```

Out[21]:

```
0    809
1    498
Name: survived, dtype: int64
```

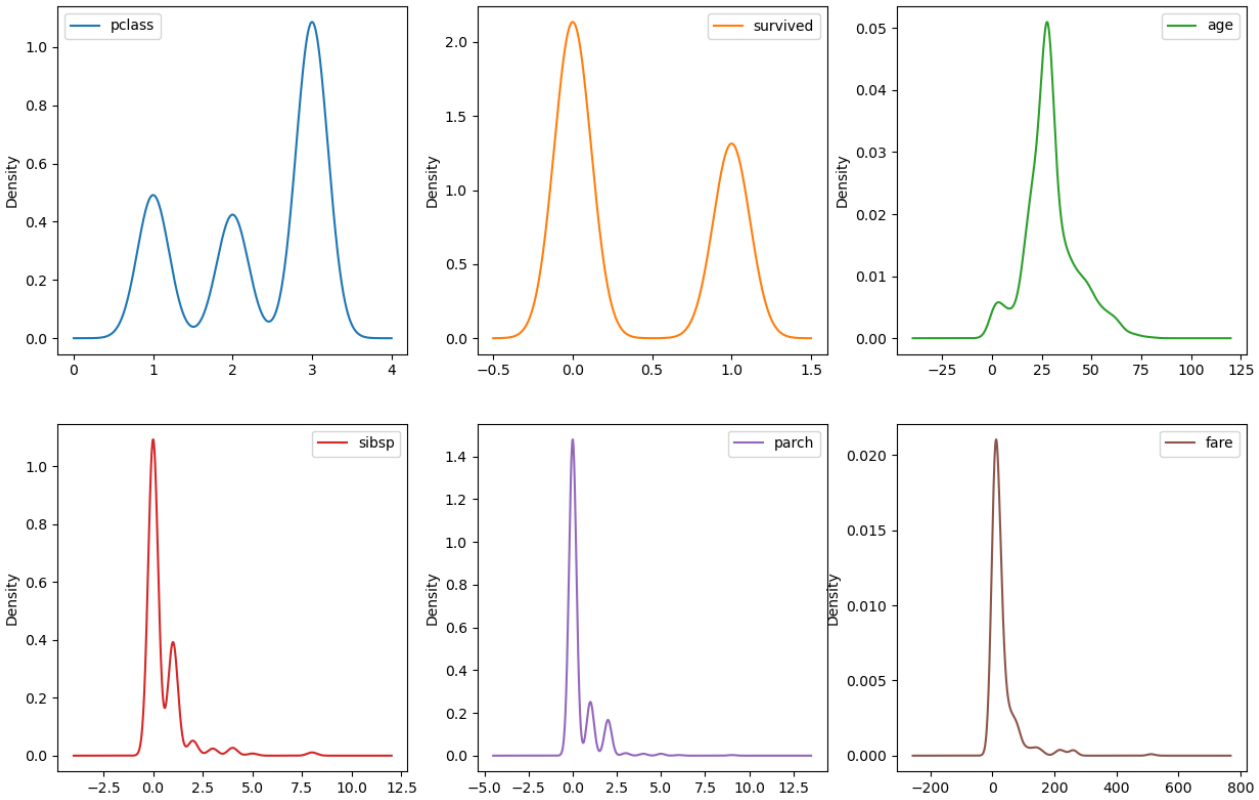
In [22]:

```
df.hist(figsize =(15,15), layout=(3,3), color= 'green')
plt.show()
```



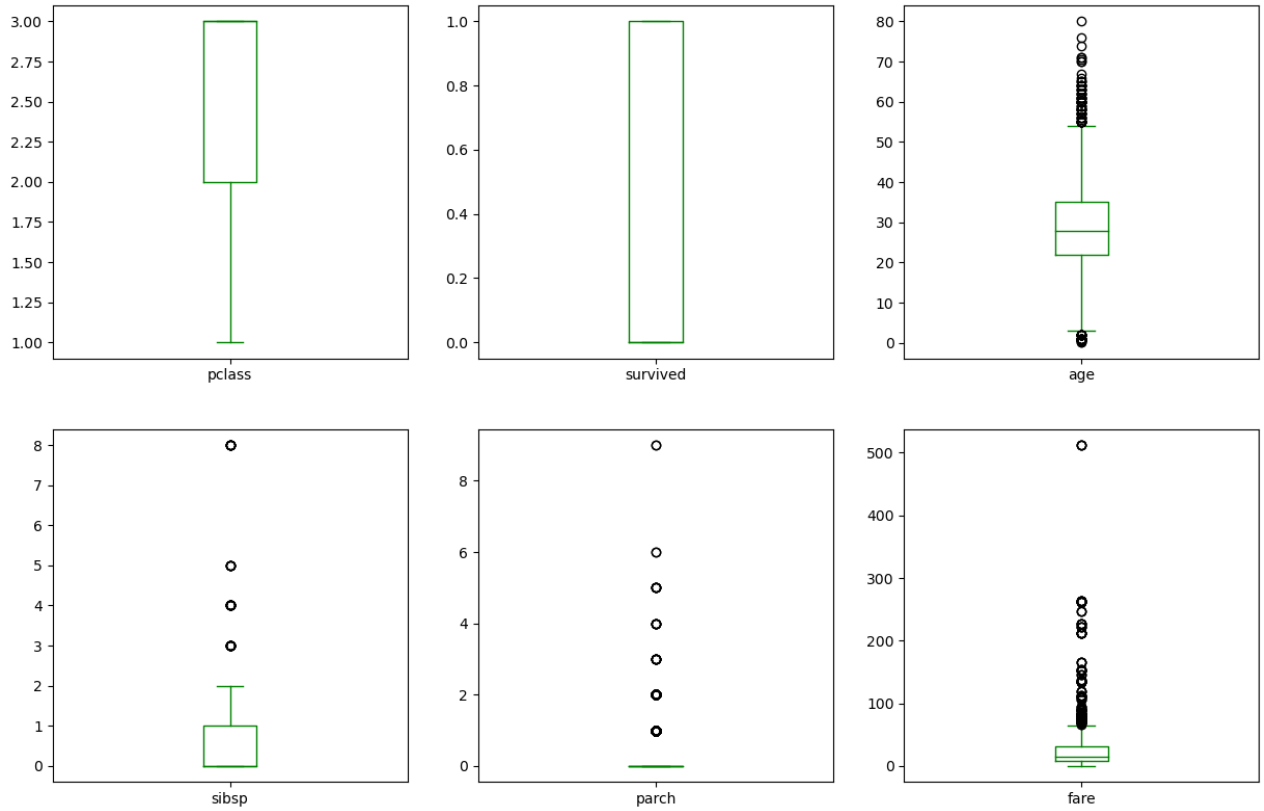
In [24]:

```
df.plot(kind='density', subplots=True, layout=(3,3), sharex=False, figsize=(15,15))
plt.show()
```



In [28]:

```
df.plot(kind='box', subplots=True, layout=(3,3), sharex=False, figsize=(15,15), color)
plt.show()
```



In [26]:

```
import seaborn as sns
plt.figure(figsize = (15,15))
sns.heatmap(df.corr(), annot = True)
```

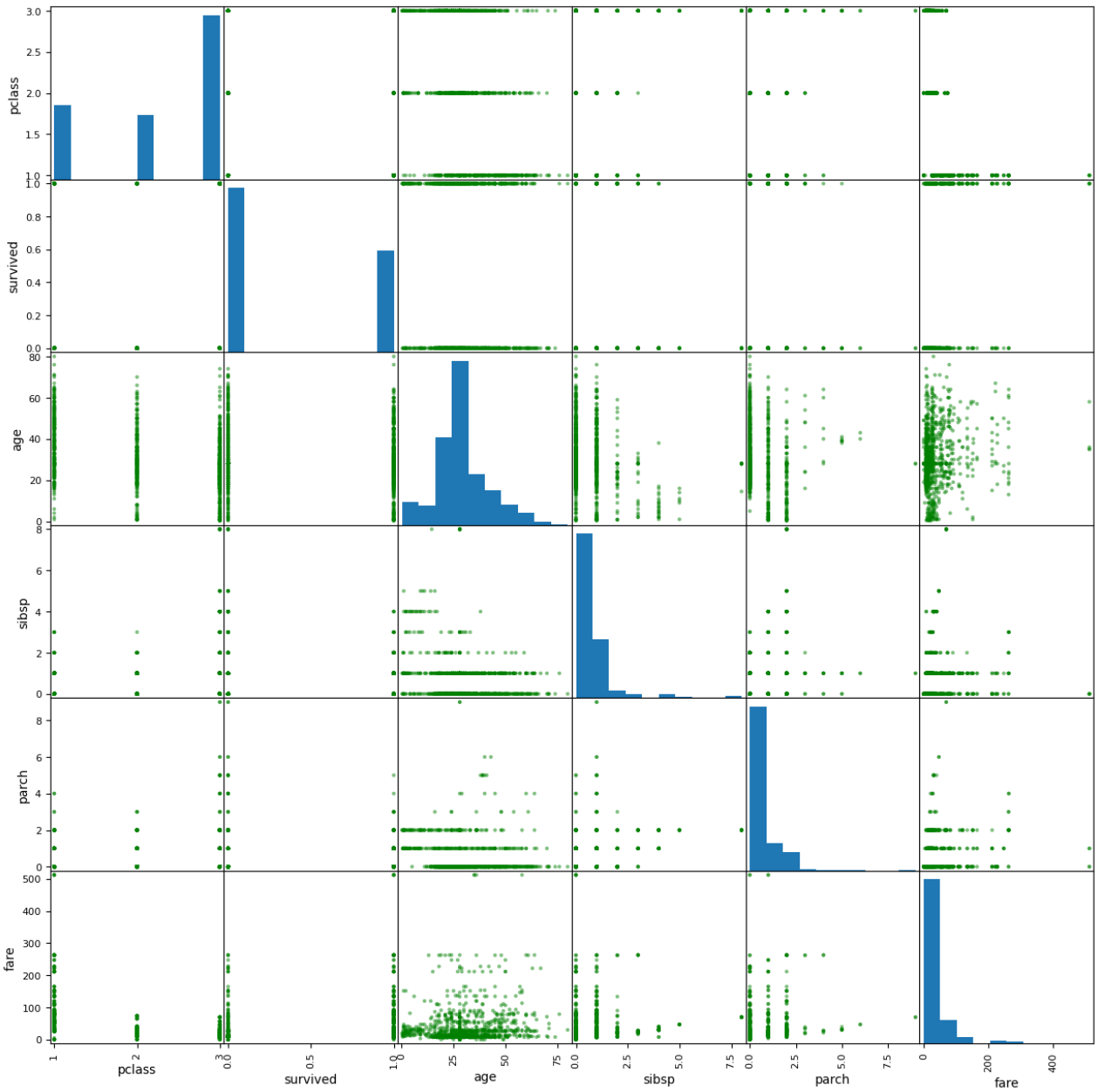
```
C:\Users\ASUS-PC\AppData\Local\Temp\ipykernel_6900\3227815026.py:3: FutureWarning: The default value of numeric_only in Dat
aFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of
numeric_only to silence this warning.
  sns.heatmap(df.corr(), annot = True)
```

Out[26]:

<Axes: >

In [27]:

```
pd.plotting.scatter_matrix(df, figsize=(15,15), color = 'green')
plt.show()
```



In [ ]: