```
import pandas as pd
```

```
df = pd.read_excel('titanic.xlsx')
```

```
df.head(n = 10)
```

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	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	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
5	1	1	Anderson, Mr. Harry	male	48.0000	0	0	19952	26.5500	E12	S	3	NaN	New York, NY
6	1	1	Andrews, Miss. Kornelia Theodosia	female	63.0000	1	0	13502	77.9583	D7	S	10	NaN	Hudson, NY
7	1	0	Andrews, Mr. Thomas Jr	male	39.0000	0	0	112050	0.0000	A36	S	NaN	NaN	Belfast, NI
8	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0000	2	0	11769	51.4792	C101	S	D	NaN	Bayside, Queens, NY
9	1	0	Artagaveytia, Mr. Ramon	male	71.0000	0	0	PC 17609	49.5042	NaN	C	NaN	22.0	Montevideo, Uruguay

Next steps: Generate code with df View recommended plots

```
df.shape
```

```
(1309, 14)
```

```
df.columns
```

```
Index(['pclass', 'survived', 'name', 'sex', 'age', 'sibsp', 'parch', 'ticket',
       'fare', 'cabin', 'embarked', 'boat', 'body', 'home.dest'],
      dtype='object')
```

```
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
```

```
df.describe()
```

	pclass	survived	age	sibsp	parch	fare	body
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	2.294882	0.381971	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.837836	0.486055	14.413500	1.041658	0.865560	51.758668	97.696922
min	1.000000	0.000000	0.166700	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	21.000000	0.000000	0.000000	7.895800	72.000000
50%	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	155.000000
75%	3.000000	1.000000	39.000000	1.000000	0.000000	31.275000	256.000000
max	3.000000	1.000000	80.000000	8.000000	9.000000	512.329200	328.000000

Cleaning

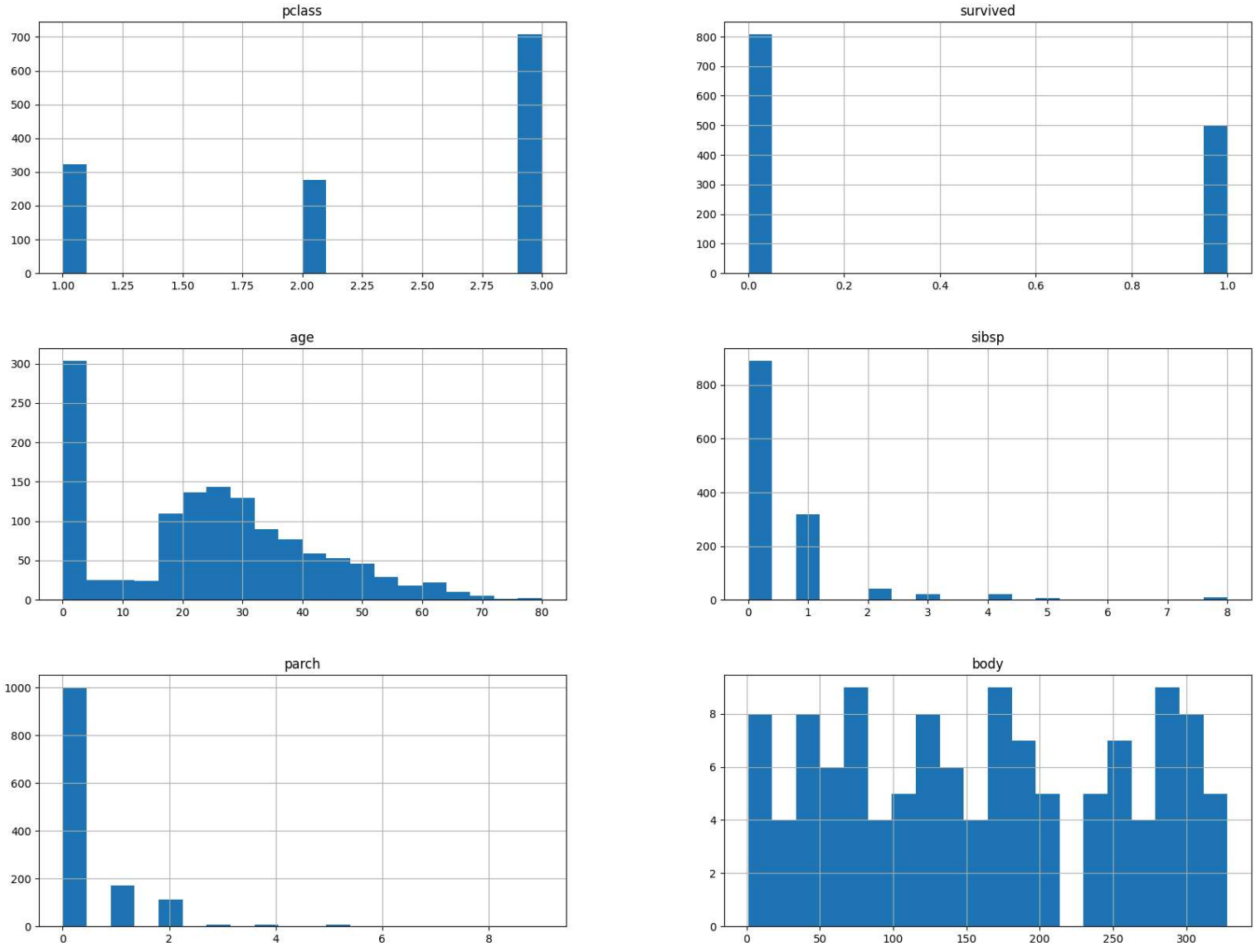
```
df.drop(['fare', 'home.dest', 'name'], axis = 1, inplace = True)
```

```
df['age'] = df['age'].fillna(0)
```

Histogram

```
import matplotlib.pyplot as plt
%matplotlib inline
```

```
df.hist(bins = 20, figsize = (20, 15))
plt.show()
```


pclass
survived
age
sibsp
parch
body

Outliers

```
df_copy = df.copy()
```

```
df_copy['age'].iloc[0:10] = 500
```

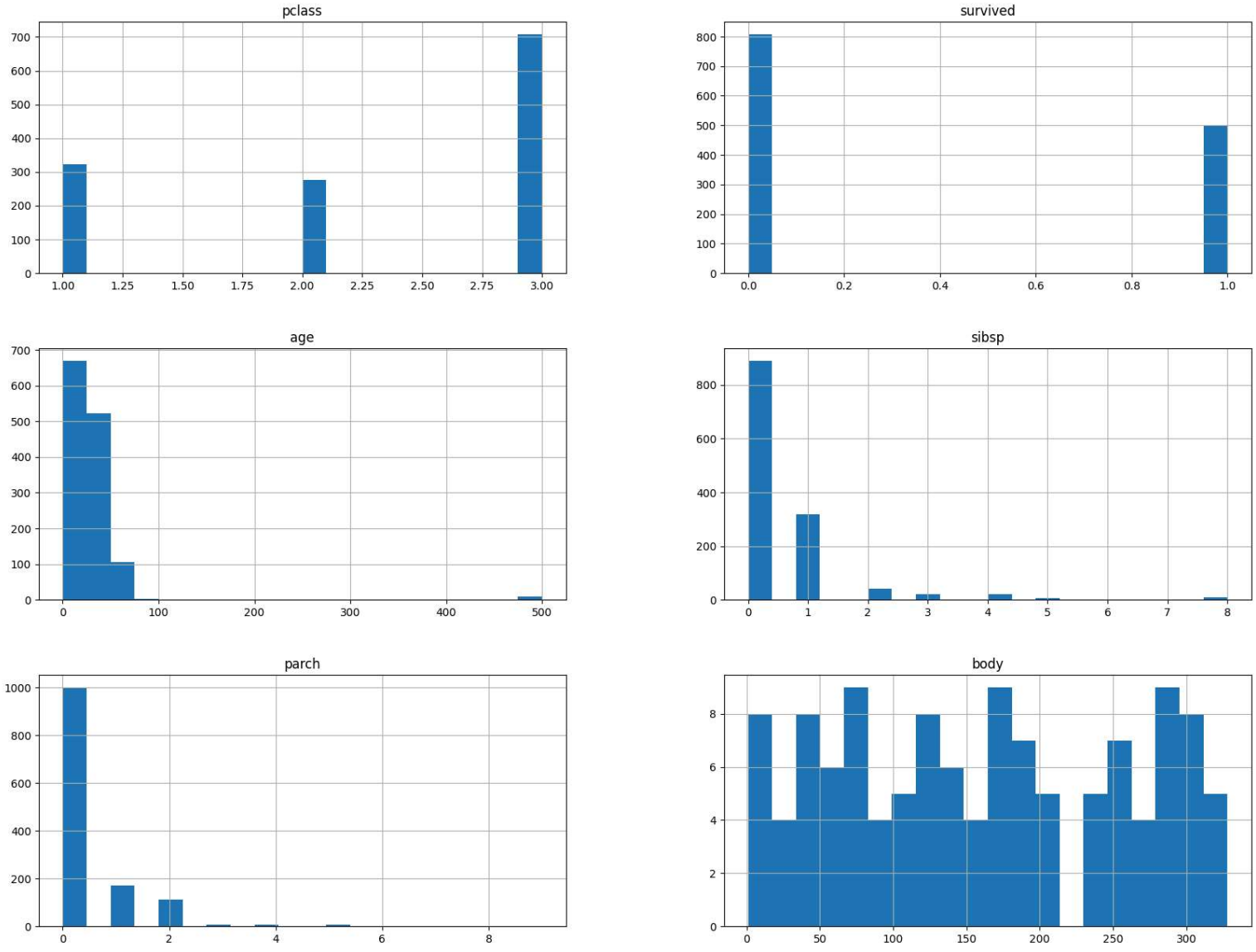
```
<ipython-input-26-16b8e2ddff47>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_copy['age'].iloc[0:10] = 500
```

```
df_copy.head(12)
```

	pclass	survived	sex	age	sibsp	parch	ticket	cabin	embarked	boat	body
0	1	1	female	500.0	0	0	24160	B5	S	2	NaN
1	1	1	male	500.0	1	2	113781	C22 C26	S	11	NaN
2	1	0	female	500.0	1	2	113781	C22 C26	S	NaN	NaN
3	1	0	male	500.0	1	2	113781	C22 C26	S	NaN	135.0
4	1	0	female	500.0	1	2	113781	C22 C26	S	NaN	NaN
5	1	1	male	500.0	0	0	19952	E12	S	3	NaN
6	1	1	female	500.0	1	0	13502	D7	S	10	NaN
7	1	0	male	500.0	0	0	112050	A36	S	NaN	NaN
8	1	1	female	500.0	2	0	11769	C101	S	D	NaN
9	1	0	male	500.0	0	0	PC 17609	NaN	C	NaN	22.0
10	1	0	male	47.0	1	0	PC 17757	C62 C64	C	NaN	124.0
11	1	1	female	18.0	1	0	PC 17757	C62 C64	C	4	NaN

Next steps: Generate code with df_copy View recommended plots

```
df_copy.hist(bins = 20, figsize = (20, 15))
plt.show()
```

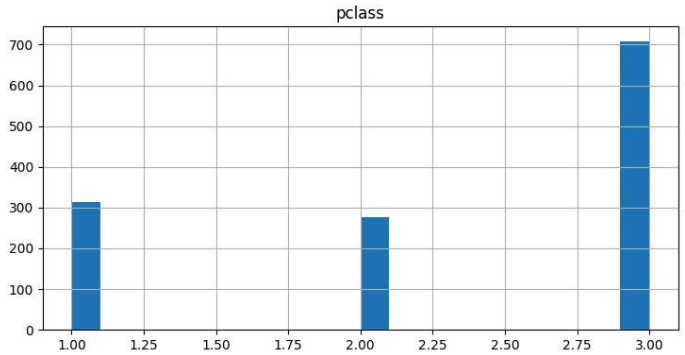

pclass
survived
age
sibsp
parch
body

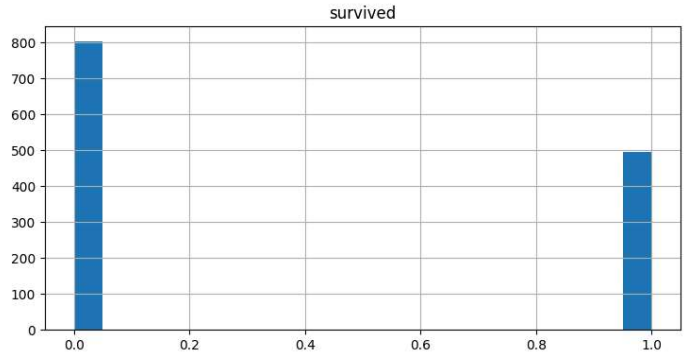
```
df_copy[df_copy['age'] > 100]['age'].index
```

```
Index([0, 1, 2, 3, 4, 5, 6, 7, 8, 9], dtype='int64')
```

```
df_copy.drop(df_copy[df_copy['age'] > 100]['age'].index, inplace = True)
```

```
df_copy.hist(bins = 20, figsize = (20, 15))
plt.show()
```






```
df_copy.shape
```

```
(1299, 11)
```

I'm Alive

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

```
sex
male      843
```

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.