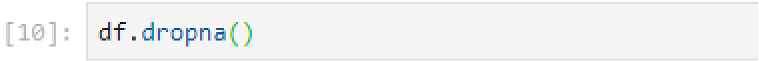
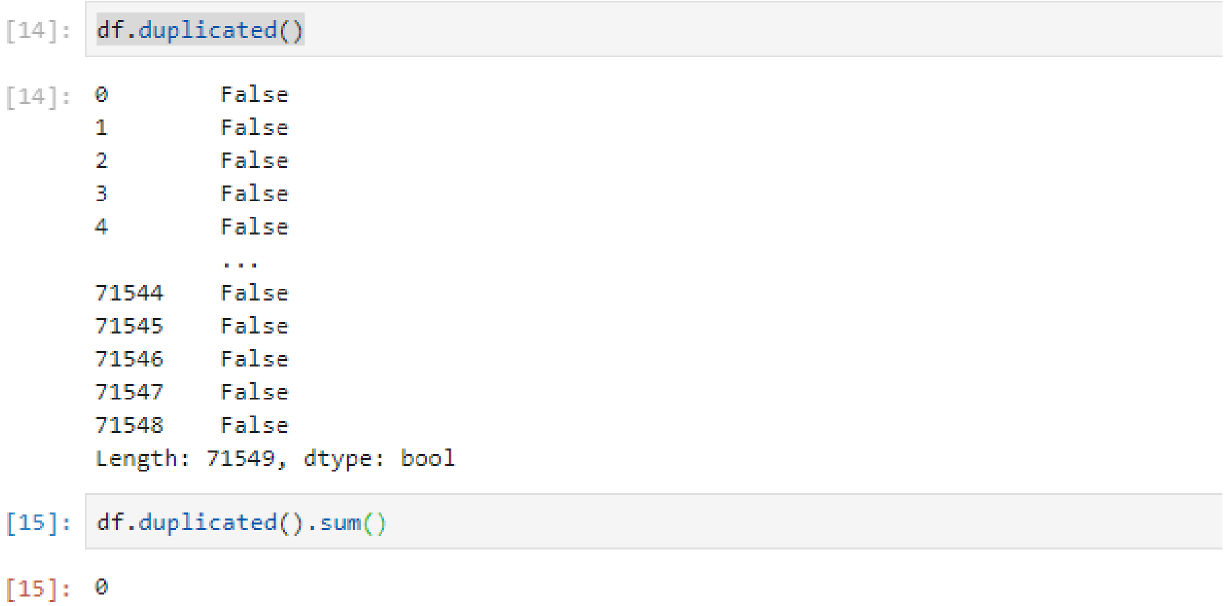
- Drop missing data



```
[10]: df.dropna()
```

- Removing Duplicated rows



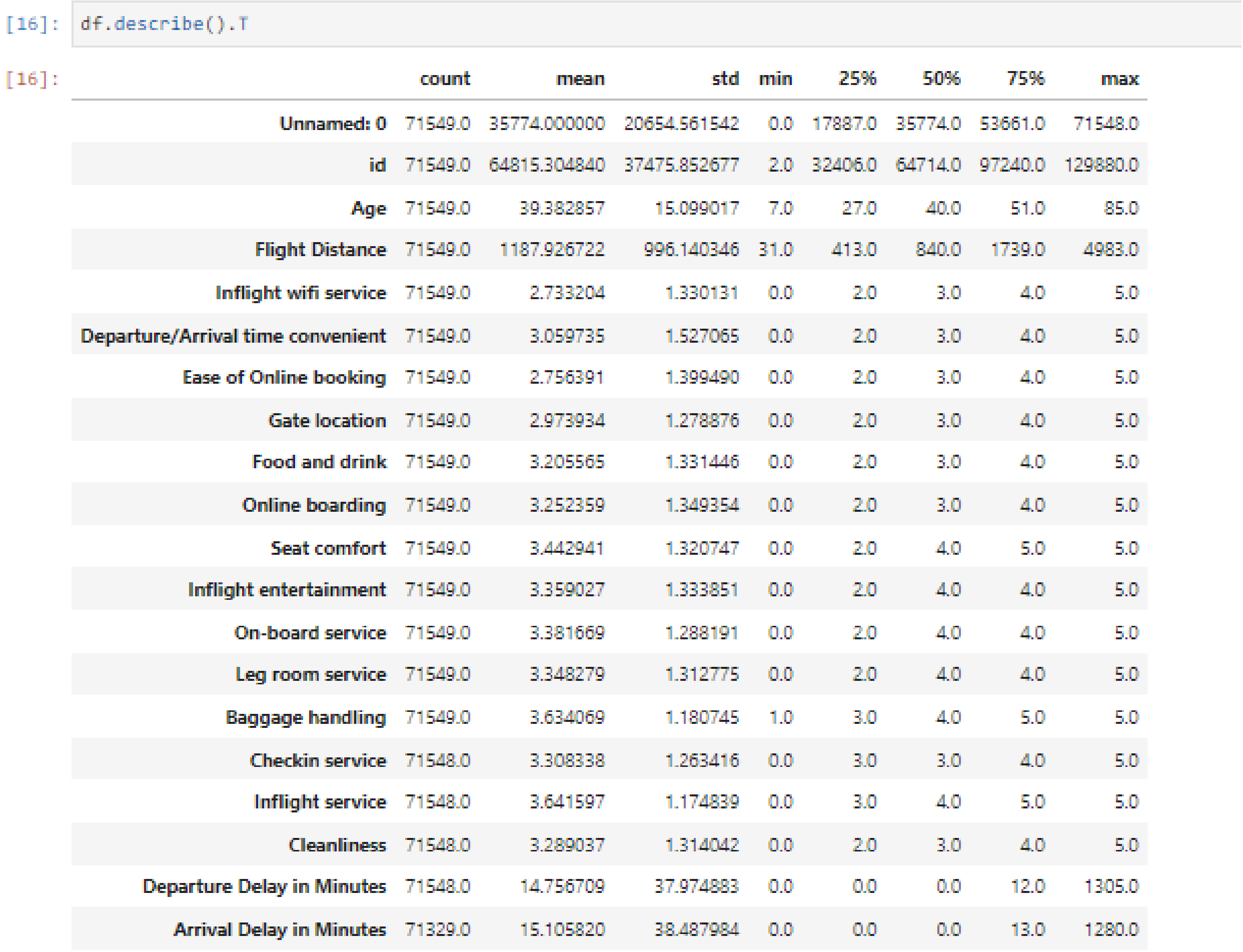
```
[14]: df.duplicated()

[14]: 0        False
      1        False
      2        False
      3        False
      4        False
               ...  
      71544    False
      71545    False
      71546    False
      71547    False
      71548    False
      Length: 71549, dtype: bool

[15]: df.duplicated().sum()

[15]: 0
```

- Summary

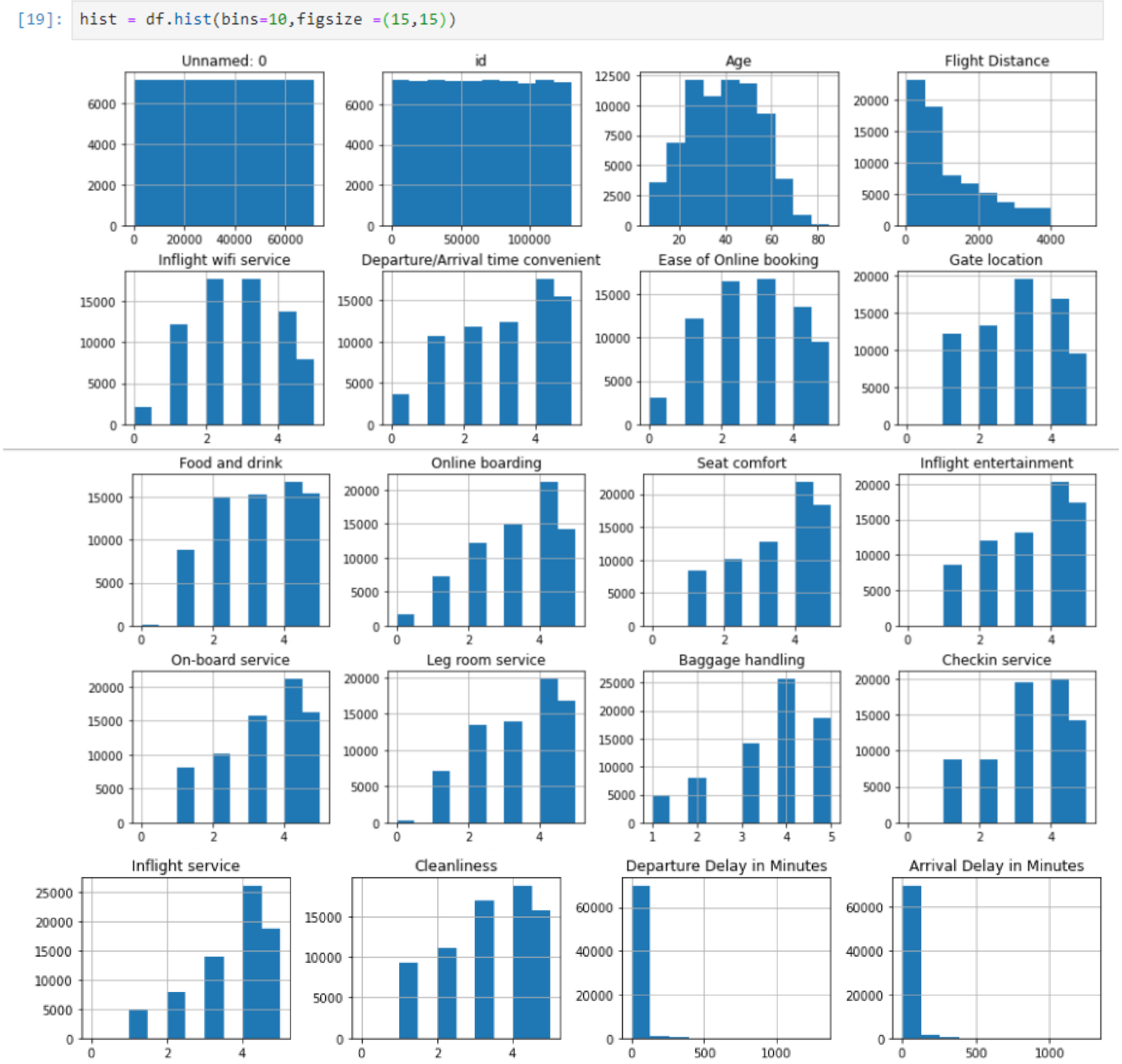


```
[16]: df.describe().T

[16]:
```

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	71549.0	35774.000000	20654.561542	0.0	17887.0	35774.0	53661.0	71548.0
id	71549.0	64815.304840	37475.852677	2.0	32406.0	64714.0	97240.0	129880.0
Age	71549.0	39.382857	15.099017	7.0	27.0	40.0	51.0	85.0
Flight Distance	71549.0	1187.926722	996.140346	31.0	413.0	840.0	1739.0	4983.0
Inflight wifi service	71549.0	2.733204	1.330131	0.0	2.0	3.0	4.0	5.0
Departure/Arrival time convenient	71549.0	3.059735	1.527065	0.0	2.0	3.0	4.0	5.0
Ease of Online booking	71549.0	2.756391	1.399490	0.0	2.0	3.0	4.0	5.0
Gate location	71549.0	2.973934	1.278876	0.0	2.0	3.0	4.0	5.0
Food and drink	71549.0	3.205565	1.331446	0.0	2.0	3.0	4.0	5.0
Online boarding	71549.0	3.252359	1.349354	0.0	2.0	3.0	4.0	5.0
Seat comfort	71549.0	3.442941	1.320747	0.0	2.0	4.0	5.0	5.0
Inflight entertainment	71549.0	3.359027	1.333851	0.0	2.0	4.0	4.0	5.0
On-board service	71549.0	3.381669	1.288191	0.0	2.0	4.0	4.0	5.0
Leg room service	71549.0	3.348279	1.312775	0.0	2.0	4.0	4.0	5.0
Baggage handling	71549.0	3.634069	1.180745	1.0	3.0	4.0	5.0	5.0
Checkin service	71548.0	3.308338	1.263416	0.0	3.0	3.0	4.0	5.0
Inflight service	71548.0	3.641597	1.174839	0.0	3.0	4.0	5.0	5.0
Cleanliness	71548.0	3.289037	1.314042	0.0	2.0	3.0	4.0	5.0
Departure Delay in Minutes	71548.0	14.756709	37.974883	0.0	0.0	0.0	12.0	1305.0
Arrival Delay in Minutes	71329.0	15.105820	38.487984	0.0	0.0	0.0	13.0	1280.0

- Histogram

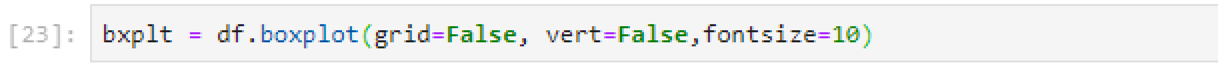


```
[19]: hist = df.hist(bins=10,figsize =(15,15))
```

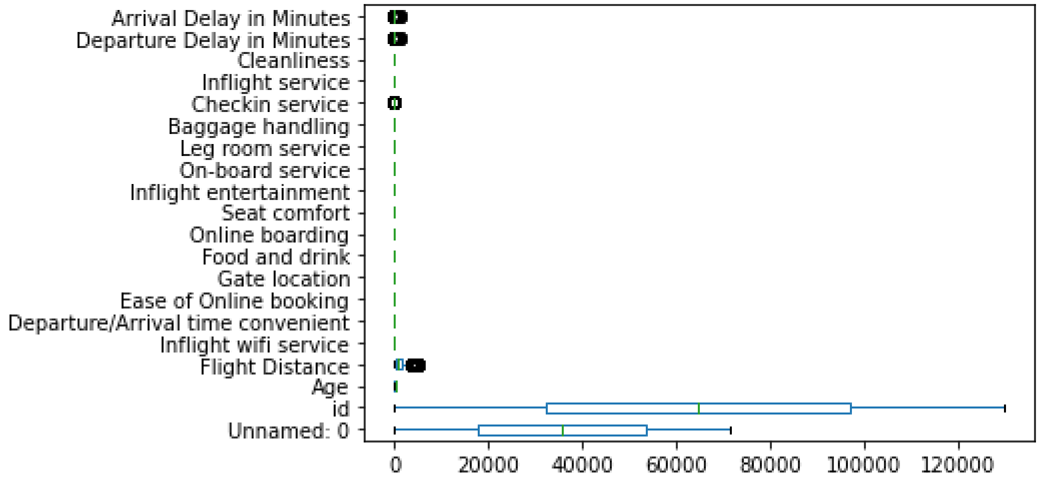
Select categorical column only ,We can see in above histogram, review the total number of each unique value per column, and compare them to each other.

- Review Outliers

  Create box plot for all columns

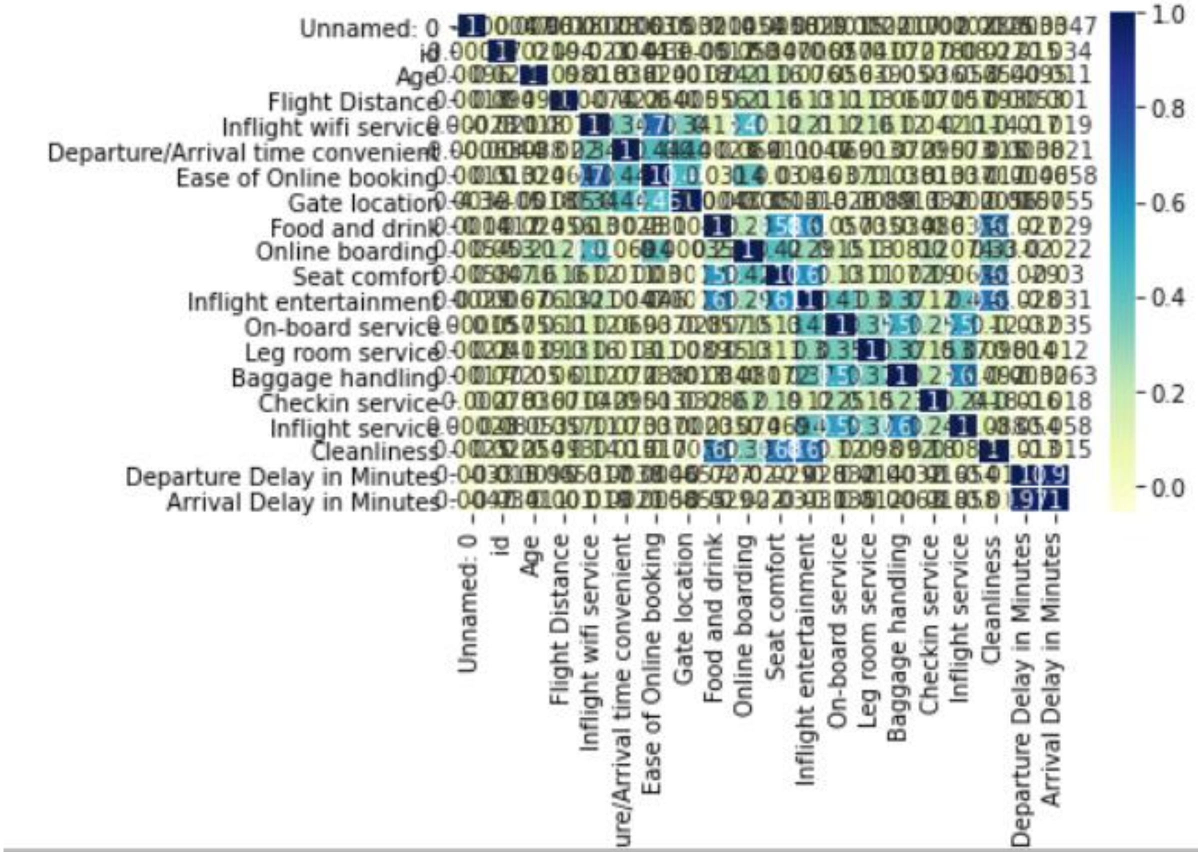


```
[23]: bxplt = df.boxplot(grid=False, vert=False,fontsize=10)
```



Identify outliers . the column at left side.

- Data Relation



It gives the relationship between the columns .

- Pair plots , Evaluate the column distribution against each other columns