In [1]:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline
import seaborn as sns

MAX_ROWS = 10
pd.set_option('display.max_rows', MAX_ROWS)
pd.set_option('display.max_columns', 200)

sns.set_style("whitegrid")
sns.set_context("paper")
```

In [2]:

```
df = pd.read_csv('Titanic.csv')

# View
df
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050
...	...	...	...	...	...	...	...	...	...	.
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750

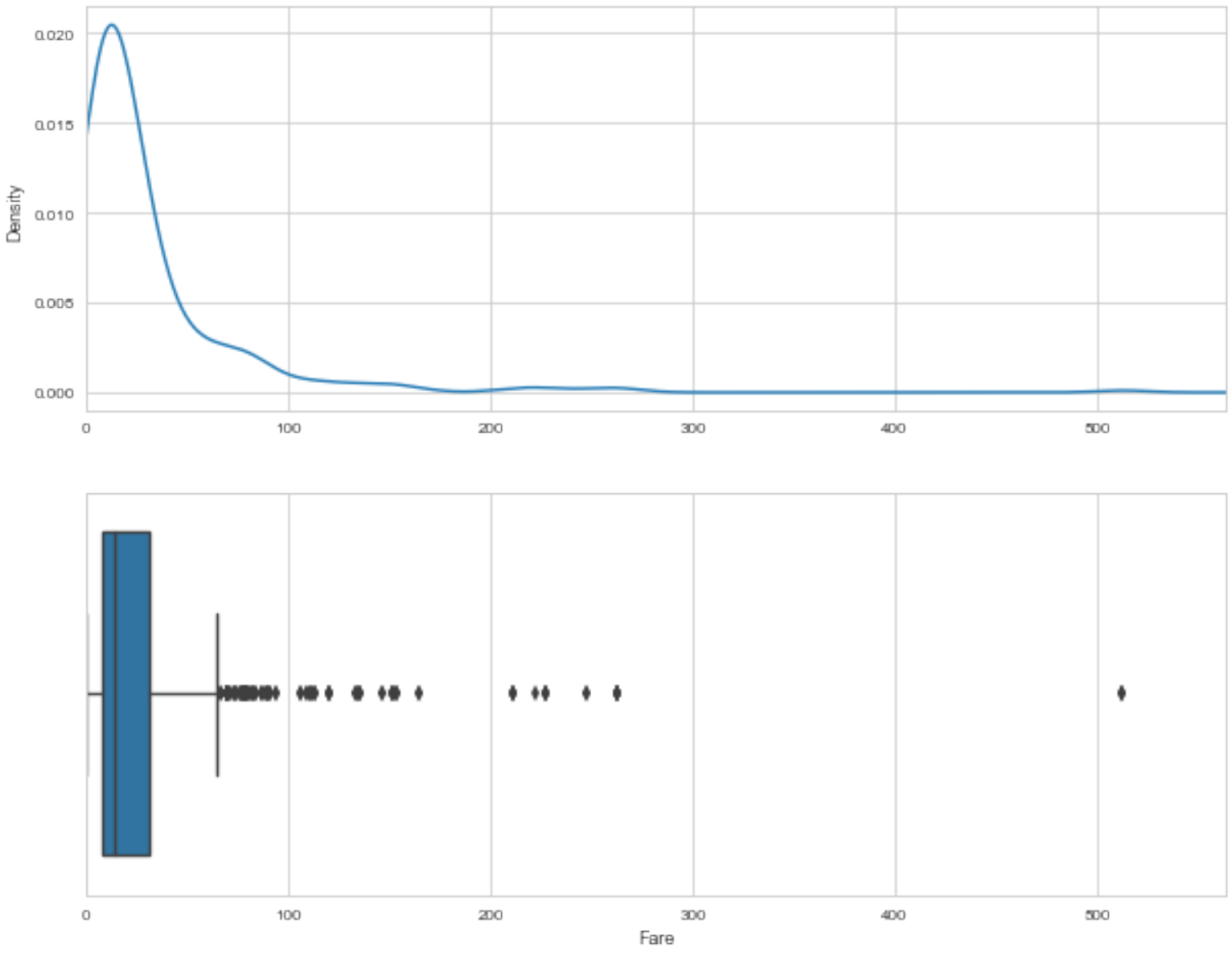
891 rows × 12 columns

In [3]:

```
i = 'Fare'

plt.figure(figsize=(10,8))
plt.subplot(211)
plt.xlim(df[i].min(), df[i].max()*1.1)

ax = df[i].plot(kind='kde')

plt.subplot(212)
plt.xlim(df[i].min(), df[i].max()*1.1)
sns.boxplot(x=df[i])
```

Out[3]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11536c240>
```



In [4]:

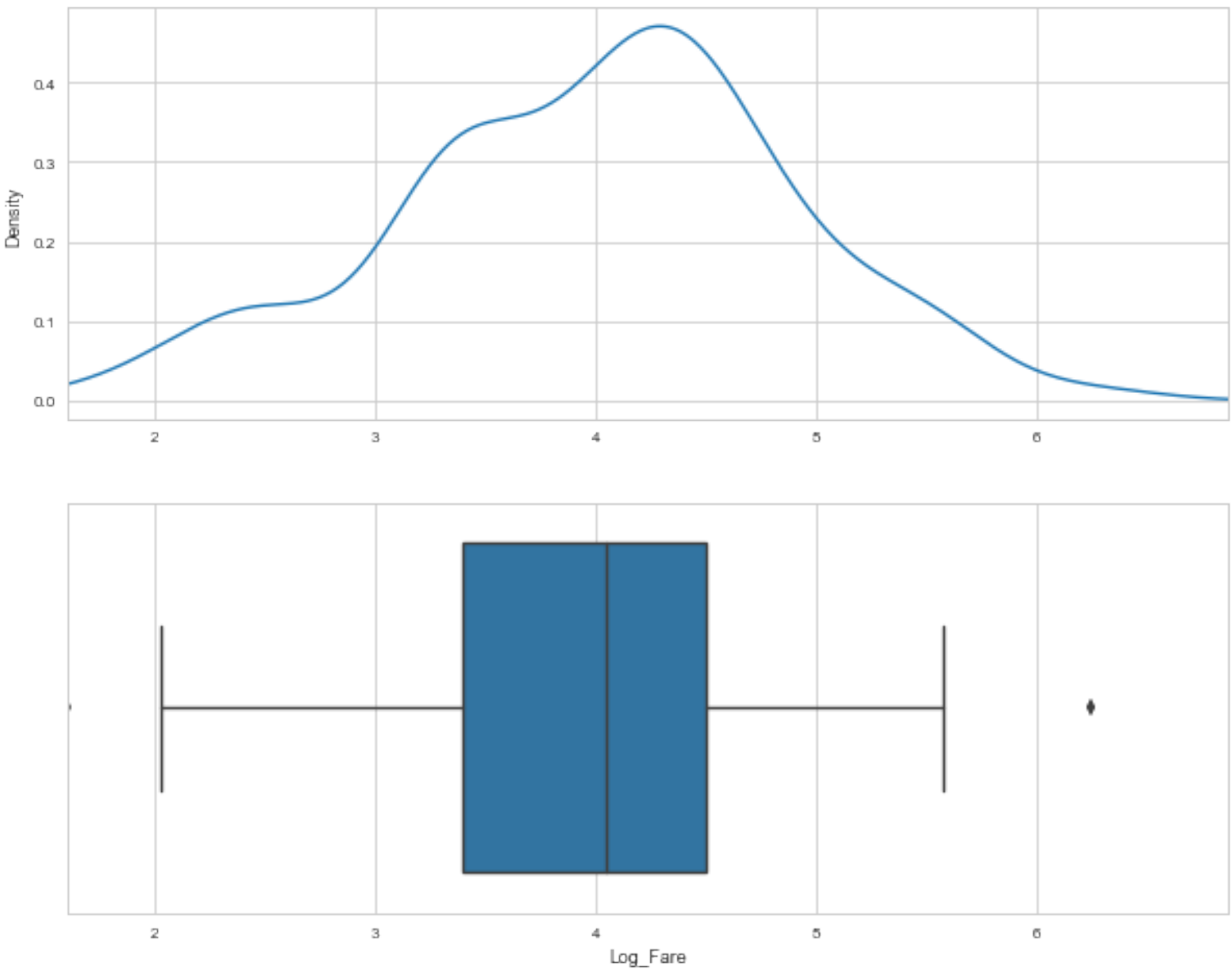
```
# Remove any zeros (otherwise we get (-inf)
df.loc[df.Fare == 0, 'Fare'] = np.nan

# Drop NA
df.dropna(inplace=True)

# Log Transform
df['Log_' + i] = np.log(df[i])
```

In [5]:

```
i = 'Log_Fare'

plt.figure(figsize=(10,8))
plt.subplot(211)
plt.xlim(df[i].min(), df[i].max()*1.1)

ax = df[i].plot(kind='kde')

plt.subplot(212)
plt.xlim(df[i].min(), df[i].max()*1.1)
sns.boxplot(x=df[i])
```

Out[5]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11542c5c0>
```



In [6]:

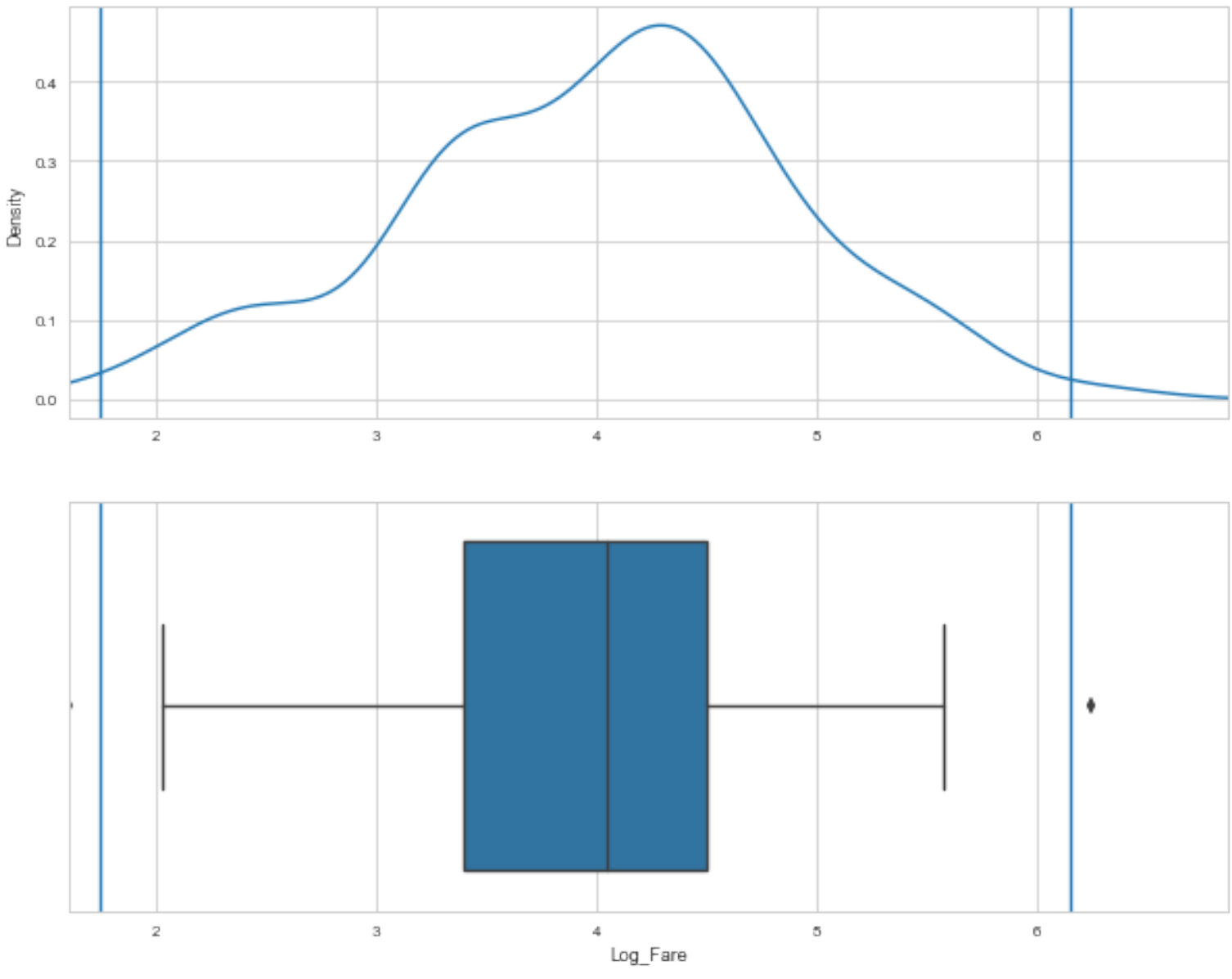
```
q75, q25 = np.percentile(df.Log_Fare.dropna(), [75 ,25])
iqr = q75 - q25

min = q25 - (iqr*1.5)
max = q75 + (iqr*1.5)
```

In [7]:

```
i = 'Log_Fare'

plt.figure(figsize=(10,8))
plt.subplot(211)
plt.xlim(df[i].min(), df[i].max()*1.1)
plt.axvline(x=min)
plt.axvline(x=max)

ax = df[i].plot(kind='kde')

plt.subplot(212)
plt.xlim(df[i].min(), df[i].max()*1.1)
sns.boxplot(x=df[i])
plt.axvline(x=min)
plt.axvline(x=max)
```

Out[7]:

```
<matplotlib.lines.Line2D at 0x115a190b8>
```



In [8]:

```
df['Outlier'] = 0

df.loc[df[i] < min, 'Outlier'] = 1
df.loc[df[i] > max, 'Outlier'] = 1
```

In [9]:

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 181 entries, 1 to 889
Data columns (total 14 columns):
PassengerId    181 non-null int64
Survived       181 non-null int64
Pclass         181 non-null int64
Name           181 non-null object
Sex            181 non-null object
Age            181 non-null float64
SibSp          181 non-null int64
Parch          181 non-null int64
Ticket         181 non-null object
Fare           181 non-null float64
Cabin          181 non-null object
Embarked       181 non-null object
Log_Fare       181 non-null float64
Outlier        181 non-null int64
dtypes: float64(3), int64(6), object(5)
memory usage: 21.2+ KB
None
```

In [ ]: