Exercise 7

The situation

There is a virus sweeping the globe (Ha! Like that would ever happen! But let's pretend...). You have data on the mutation rates of 4 different strains of the virus (in mutations per generation x 10e-5). You need to determine if the 4 strains mutate at generally the same rate, and can thus be treated as one in epidemiological models, or if they are different enough that they must be modeled separately.

In [2]:

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

In [3]:

```
myData = pd.read_csv("datasets/007ExerciseFile.csv")
```

In [4]:

```
display(myData)
```

	Strain	Mutation_rate
0	st_alpha	10.612005
1	st_alpha	12.586371
2	st_alpha	8.997583
3	st_alpha	11.681775
4	st_alpha	14.408237
...	...	...
163	st_delta	2.716249
164	st_delta	2.467378
165	st_delta	2.119801
166	st_delta	1.316537
167	st_delta	2.060472

168 rows × 2 columns

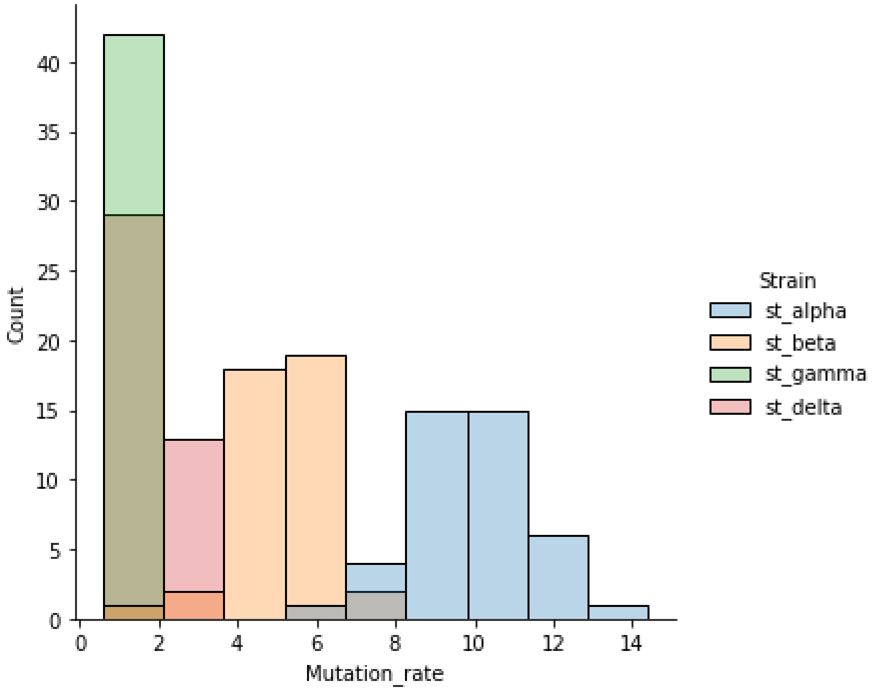
Histogram of the mutation rates of the 4 strains

In [15]:

```
sns.displot(myData, x="Mutation_rate", hue="Strain", kind="hist", alpha=0.3)
```

Out[15]:

```
<seaborn.axisgrid.FacetGrid at 0x143927cab80>
```



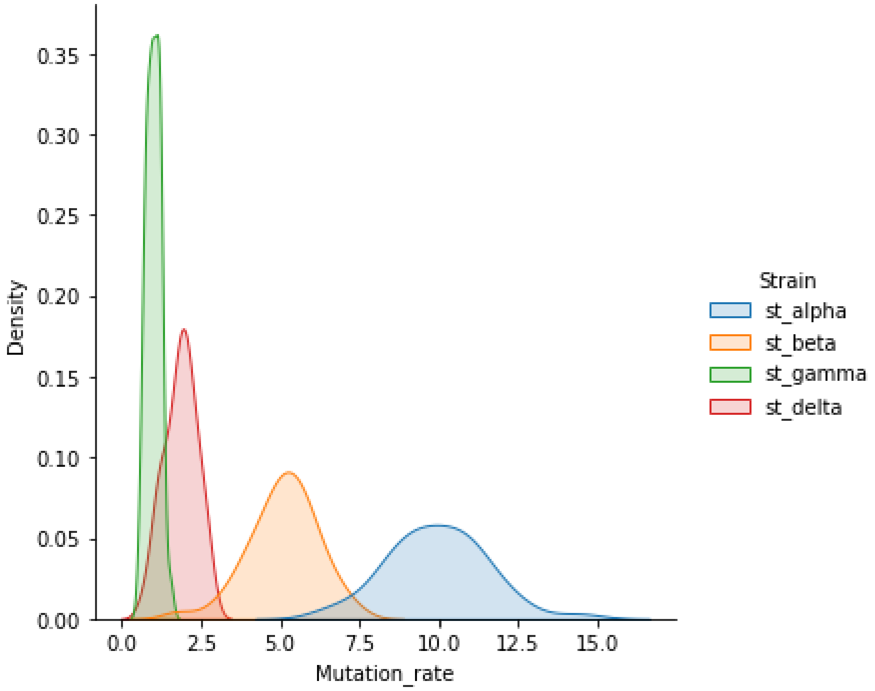
KDE plot

In [17]:

```
sns.displot(myData, x="Mutation_rate", hue="Strain", kind="kde", fill=True, alpha=0.2)
```

Out[17]:

```
<seaborn.axisgrid.FacetGrid at 0x14392bc2fd0>
```



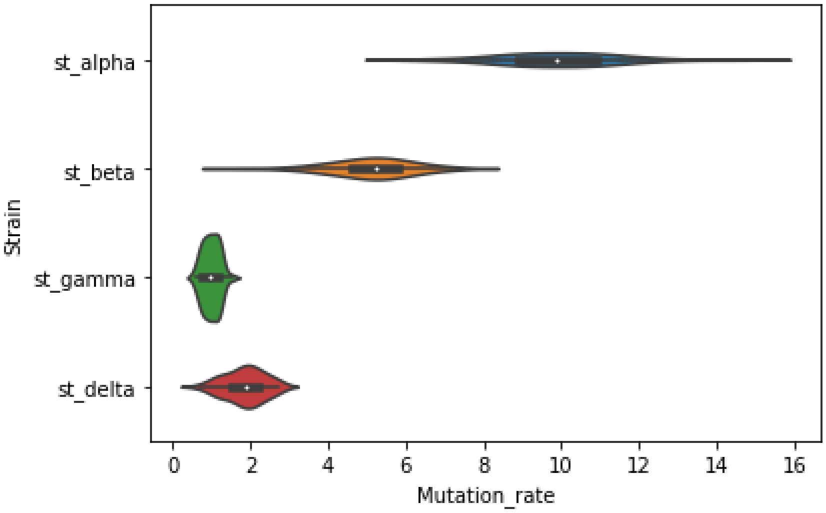
Violin plot

In [27]:

```
sns.violinplot(data=myData, x="Mutation_rate", y="Strain")
```

Out[27]:

```
<AxesSubplot:xlabel='Mutation_rate', ylabel='Strain'>
```



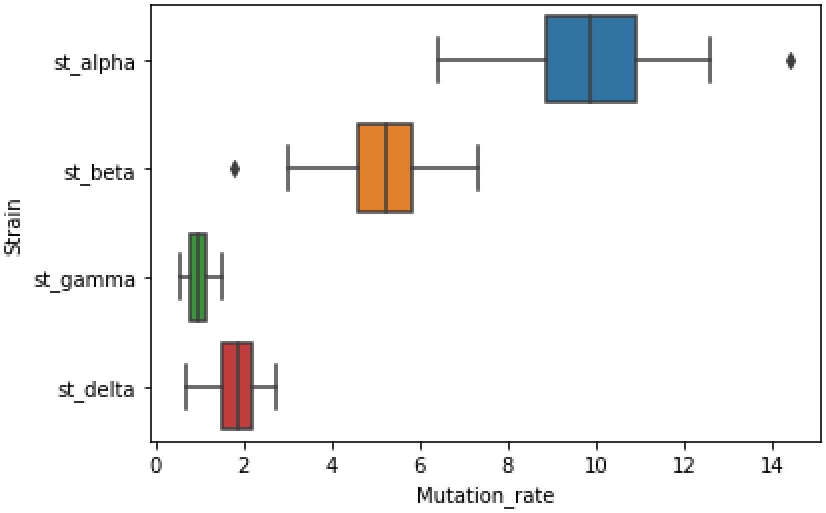
Boxplot

In [28]:

```
sns.boxplot(data=myData, x="Mutation_rate", y="Strain")
```

Out[28]:

```
<AxesSubplot:xlabel='Mutation_rate', ylabel='Strain'>
```



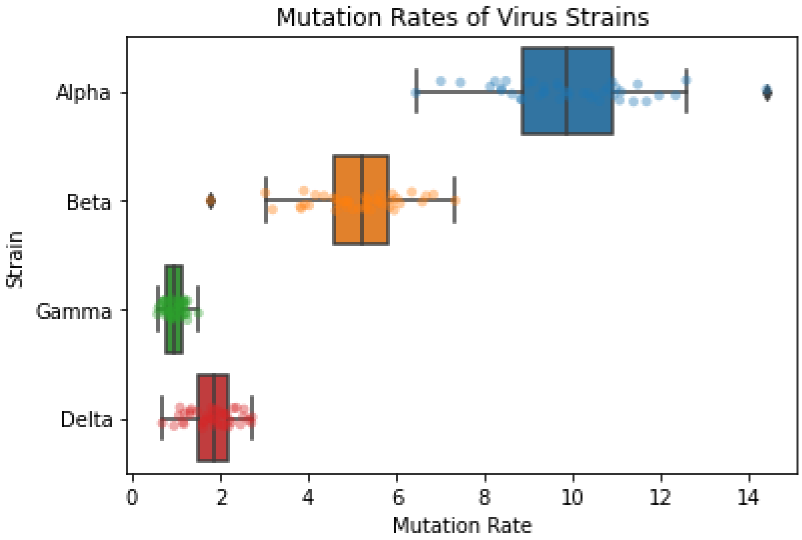
Boxplot with overlaid strip chart

In [48]:

```
plot = sns.boxplot(data=myData, x="Mutation_rate", y="Strain")
sns.stripplot(data=myData, x="Mutation_rate", y="Strain", alpha=0.4)
plot.set(xlabel="Mutation Rate")
plot.set_title("Mutation Rates of Virus Strains")
plot.set_yticklabels(["Alpha", "Beta", "Gamma", "Delta"])
```

Out[48]:

```
[Text(0, 0, 'Alpha'),
 Text(0, 1, 'Beta'),
 Text(0, 2, 'Gamma'),
 Text(0, 3, 'Delta')]
```



Summary statistics

In [22]:

```
myData.groupby("Strain").describe()
```

Out[22]:

	Mutation_rate							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Strain								
st_alpha	42.0	9.938535	1.563850	6.437518	8.884385	9.871964	10.926445	14.408237
st_beta	42.0	5.074079	1.089212	1.783542	4.591669	5.225036	5.824638	7.337440
st_delta	42.0	1.839395	0.508917	0.688354	1.513231	1.868183	2.180814	2.730632
st_gamma	42.0	0.974370	0.214968	0.577240	0.775191	0.962160	1.159783	1.510792

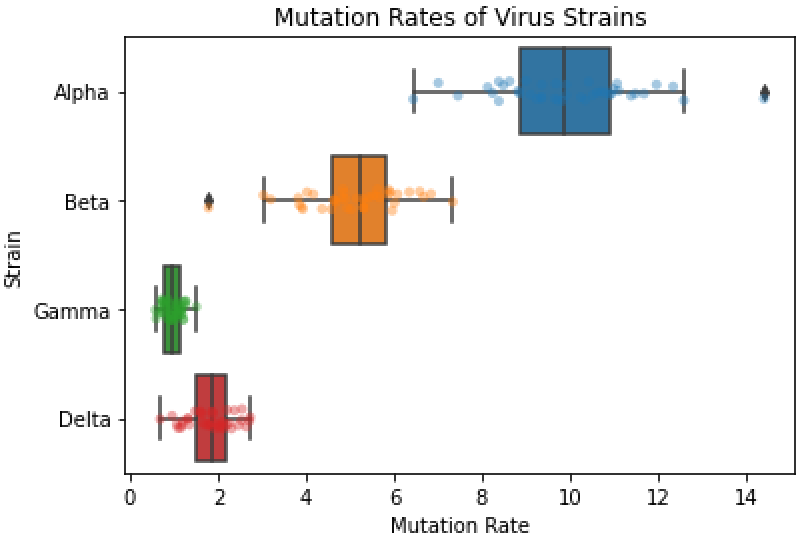
Boxplot to show Fauci

In [49]:

```
plot = sns.boxplot(data=myData, x="Mutation_rate", y="Strain")
sns.stripplot(data=myData, x="Mutation_rate", y="Strain", alpha=0.4)
plot.set(xlabel="Mutation Rate")
plot.set_title("Mutation Rates of Virus Strains")
plot.set_yticklabels(["Alpha", "Beta", "Gamma", "Delta"])
```

Out[49]:

```
[Text(0, 0, 'Alpha'),
 Text(0, 1, 'Beta'),
 Text(0, 2, 'Gamma'),
 Text(0, 3, 'Delta')]
```



Each of the strains seems to mutate slightly different, so I would not recommend treating the strains as one epidemiological model. Looking at the boxplot above, the Alpha strain mutates the fastest. The Beta strain was the second fastest to mutate. Delta has the 3rd fastest mutation rate, and the Gamma strain has the slowest mutation.