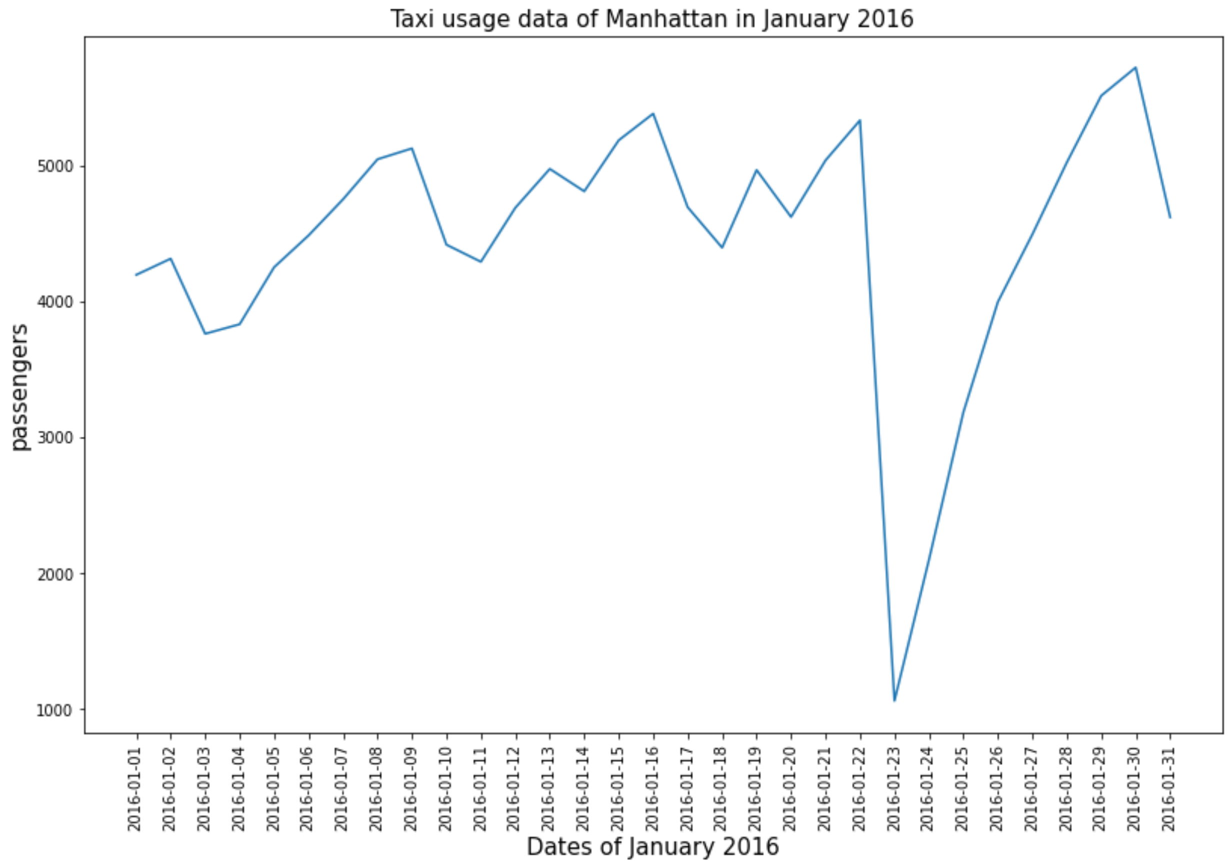
Create a data visualization that allows you to identify which dates were affected by the historic blizzard of January 2016. Make sure that the visualization type is appropriate for the visualized data.

Hint: How do you expect taxi usage to differ on blizzard days?

In [13]:
```
my_df = manhattan_taxi.loc[:, ['date', 'passengers']].groupby('date').agg(n

plt.figure(figsize=(13.5, 8.5))
ax = sns.lineplot(x=my_df.index, y=my_df['passengers'])
ax.set(xticks=my_df.index)
ax.set_xticklabels(labels=my_df.index, rotation=90)
ax.set_xlabel('Dates of January 2016', fontsize = 15)
ax.set_ylabel('passengers', fontsize = 15)
ax.set_title('Taxi usage data of Manhattan in January 2016', fontsize = 15)
```

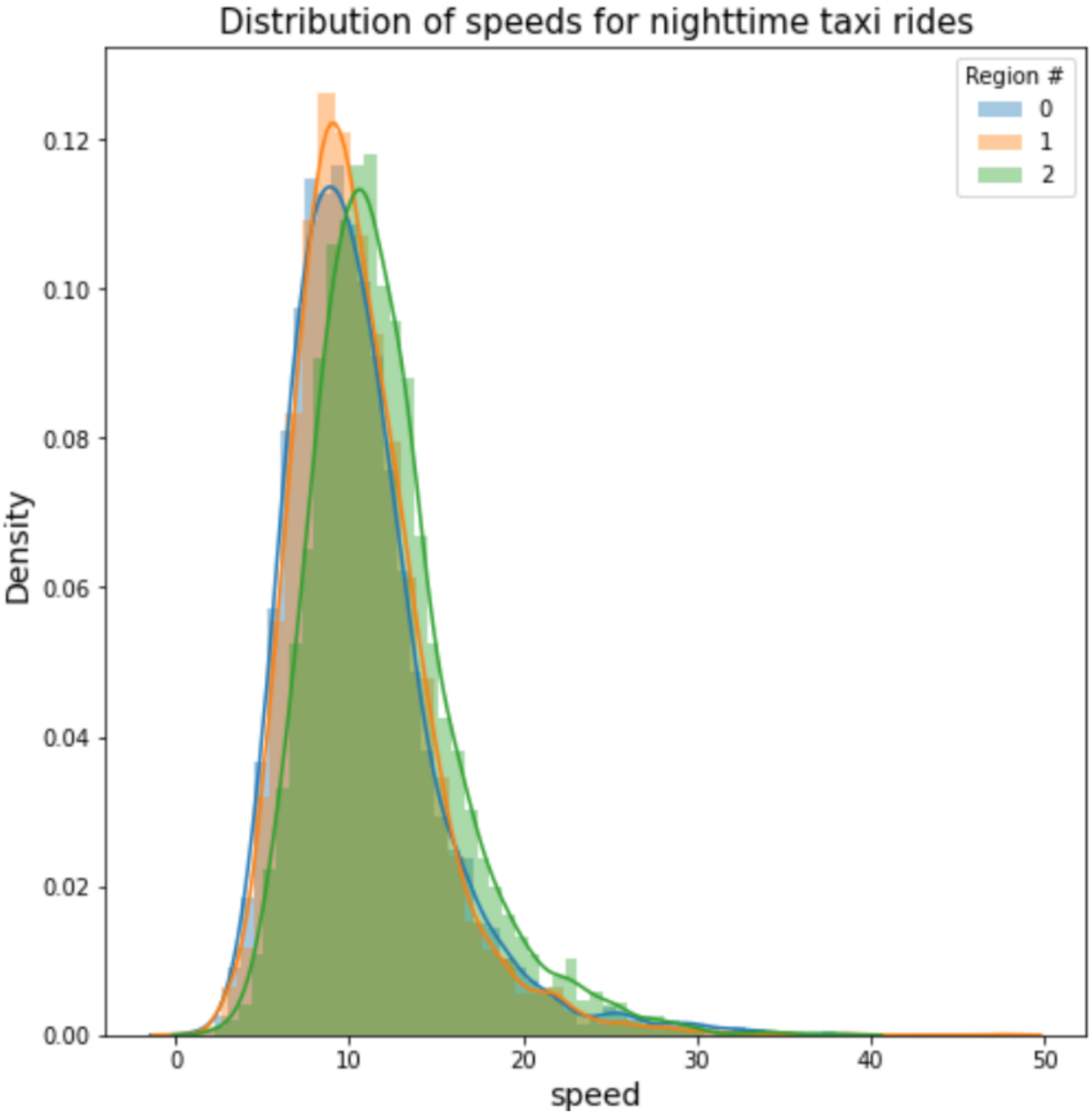


Finally, we have generated a list of dates that should have a fairly typical distribution of taxi rides, which excludes holidays and blizzards. The cell below assigns `final_taxi` to the subset of `manhattan_taxi` that is on these days. (No changes are needed; just run this cell.)

Use `sns.distplot` to create an overlaid histogram comparing the distribution of speeds for nighttime taxi rides (6pm-12am) in the three different regions defined above. Does it appear that there is an association between region and average speed during the night?

In [24]:
```
region_0 = train[train['region'] == 0]
region_0_night = region_0[(region_0['hour'] >= 18) & (region_0['hour'] < 24
region_1 = train[train['region'] == 1]
region_1_night = region_1[(region_1['hour'] >= 18) & (region_1['hour'] < 24
region_2 = train[train['region'] == 2]
region_2_night = region_2[(region_2['hour'] >= 18) & (region_2['hour'] < 24

plt.figure(figsize=(8, 8.3))
ax = sns.distplot(region_0_night['speed'], label='0')
ax = sns.distplot(region_1_night['speed'], label='1')
ax = sns.distplot(region_2_night['speed'], label='2')
ax.legend().set_title('Region #')
ax.set_xlabel('speed', fontsize=14)
ax.set_ylabel('Density', fontsize=14)
ax.set_title('Distribution of speeds for nighttime taxi rides', fontsize=15
```



In [16]:
```
import sklearn.model_selection

train, test = sklearn.model_selection.train_test_split(
    final_taxi, train_size=0.8, test_size=0.2, random_state=42)
print('Train:', train.shape, 'Test:', test.shape)
```

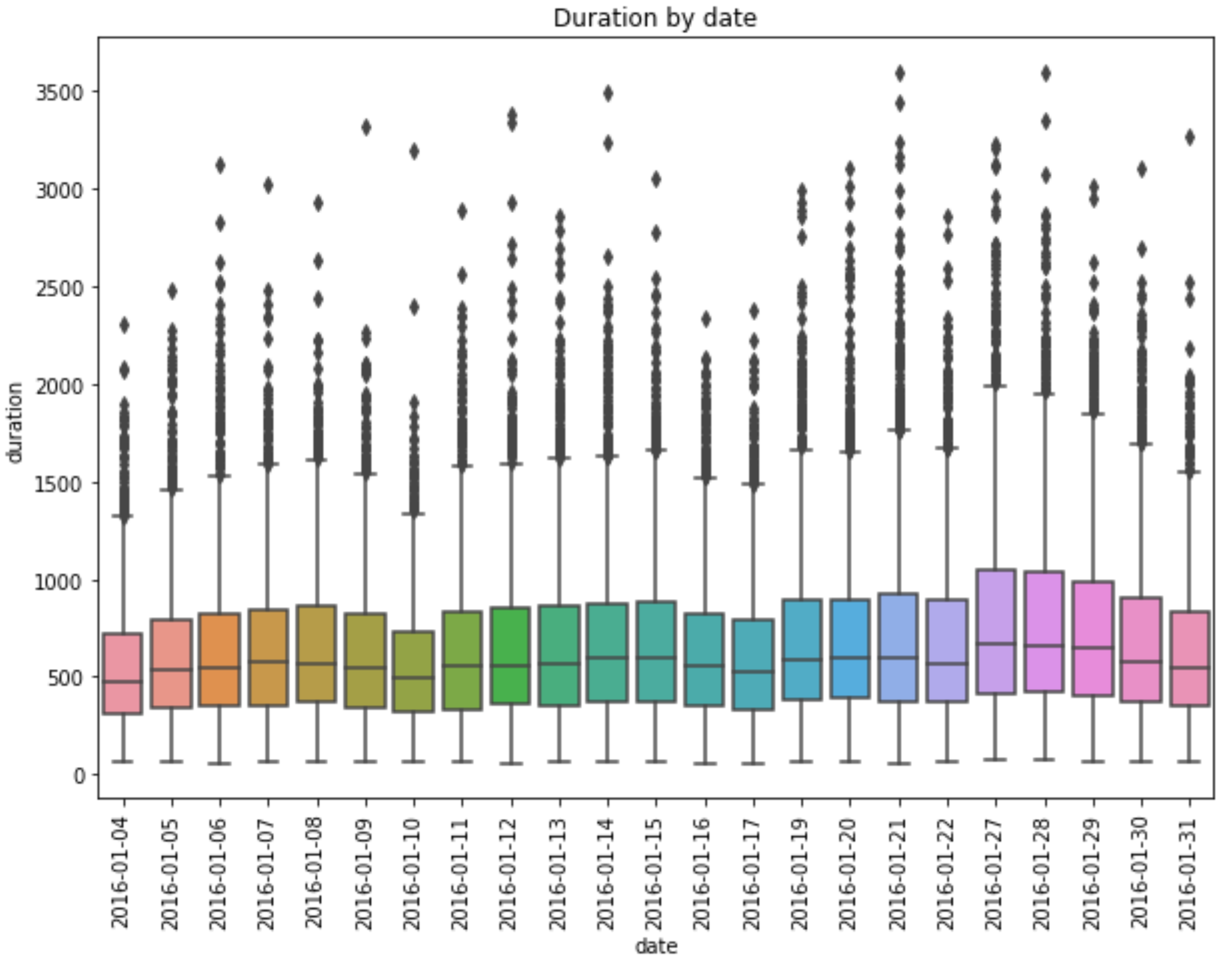
```
Train: (53680, 10) Test: (13421, 10)
```

Question 3a

Create a box plot that compares the distributions of taxi trip durations for each day **using `train` only**. Individual dates shoud appear on the horizontal axis, and duration values should appear on the vertical axis. Your plot should look like the following.

Hint: Use `sns.boxplot`.

In [17]:
```
plt.figure(figsize=(10, 7))
my_train = train.sort_values('date', ascending=True)
ax = sns.boxplot(x=my_train['date'], y=my_train['duration'])
ax.set_title('Duration by date')
ax.set_xticklabels(ax.get_xticklabels(), rotation=90);
```


Duration by date
duration
date