In [10]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

In [12]:
```
taxidata = pd.read_csv("/Users/baboury/Desktop/CMP646DATA/train.csv", de
```

In [13]:
```
# let's prints the shape of the Datab
taxidata.head()
```

Out[13]:

	id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pi
0	id2875421	2	2016-03-14 17:24:55	2016-03-14 17:32:30	1	-73.982155	
1	id2377394	1	2016-06-12 00:43:35	2016-06-12 00:54:38	1	-73.980415	
2	id3858529	2	2016-01-19 11:35:24	2016-01-19 12:10:48	1	-73.979027	
3	id3504673	2	2016-04-06 19:32:31	2016-04-06 19:39:40	1	-74.010040	
4	id2181028	2	2016-03-26 13:30:55	2016-03-26 13:38:10	1	-73.973053	

In [17]:
```
#For this sub-dataset only, visualize the correlation between the aerial

index_long = (taxidata["pickup_longitude"] >= -73.82) & (taxidata["picku

data = taxidata[index_long]

index_lat = (data["pickup_latitude"] >= 40.63) & (data["pickup_latitude"
data = data[index_lat]

data.shape
```

Out[17]:
```
(25736, 11)
```

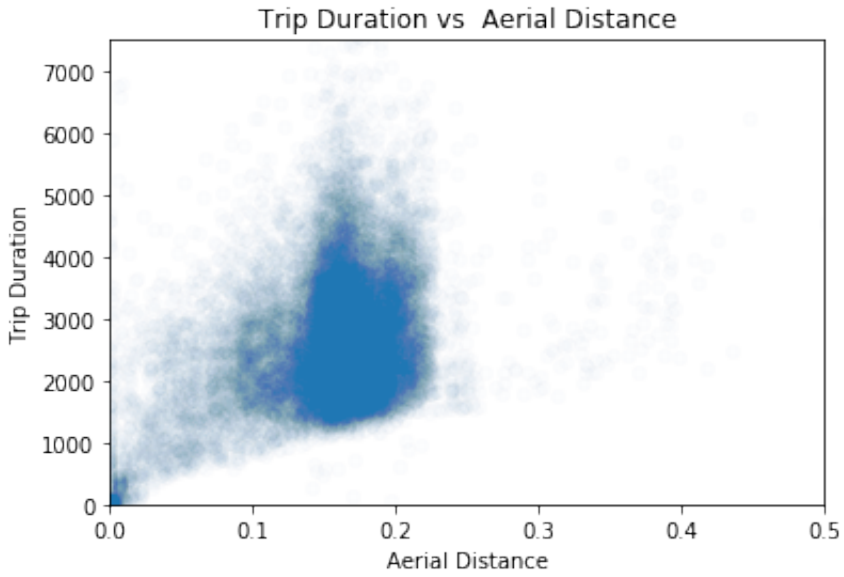
In [19]:
```
# For this sub-dataset only, visualize the drop-off location using dropo

data["aerial_distance"] = np.sqrt( ((data["pickup_longitude"] - data["dr

plt.scatter(data["aerial_distance"], data["trip_duration"], alpha = 0.01
plt.ylim(0,7500)
plt.xlim(0,0.5)
plt.xlabel("Aerial Distance")
plt.ylabel("Trip Duration")
plt.title(" Trip Duration vs  Aerial Distance ")
```

```
//anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:3: Runti
meWarning: invalid value encountered in sqrt
  This is separate from the ipykernel package so we can avoid doing im
ports until
```

Out[19]:
```
Text(0.5, 1.0, ' Trip Duration vs  Aerial Distance ')
```

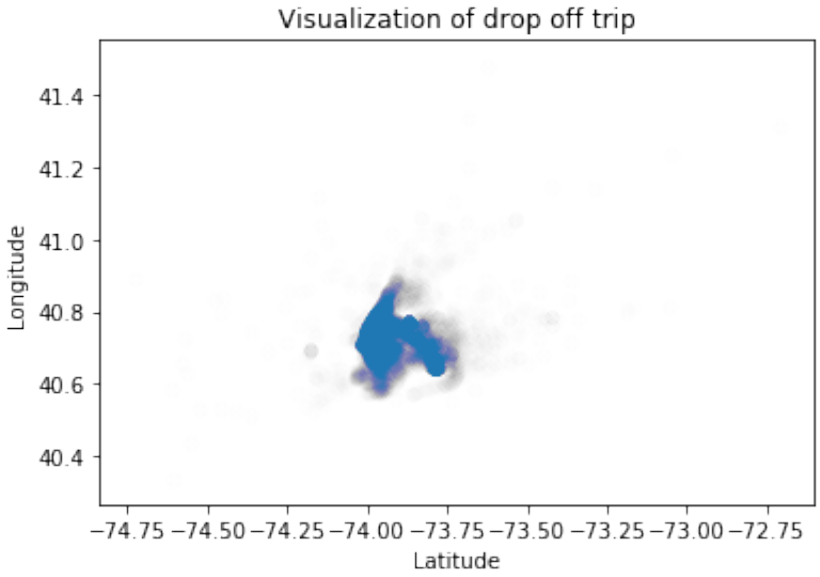


In [ ]:
There is a positive coorelation between distance and trip duration

In [ ]:
```
#For this sub-dataset only, visualize the drop-off location using dropof
```

In [23]:
```
plt.scatter(data["dropoff_longitude"], data["dropoff_latitude"], alpha =
plt.title("Visualization of drop off trip")
plt.xlabel("Latitude")
plt.ylabel("Longitude")
```

Out[23]:
```
Text(0, 0.5, 'Longitude')
```



In [26]:
```
#There was a snow storm on Jan 23, 2016. Is the distribution of trip_dur
#different from the rest of the year?

taxidata["pickup_datetime"] = taxidata["pickup_datetime"].astype(np.date
```

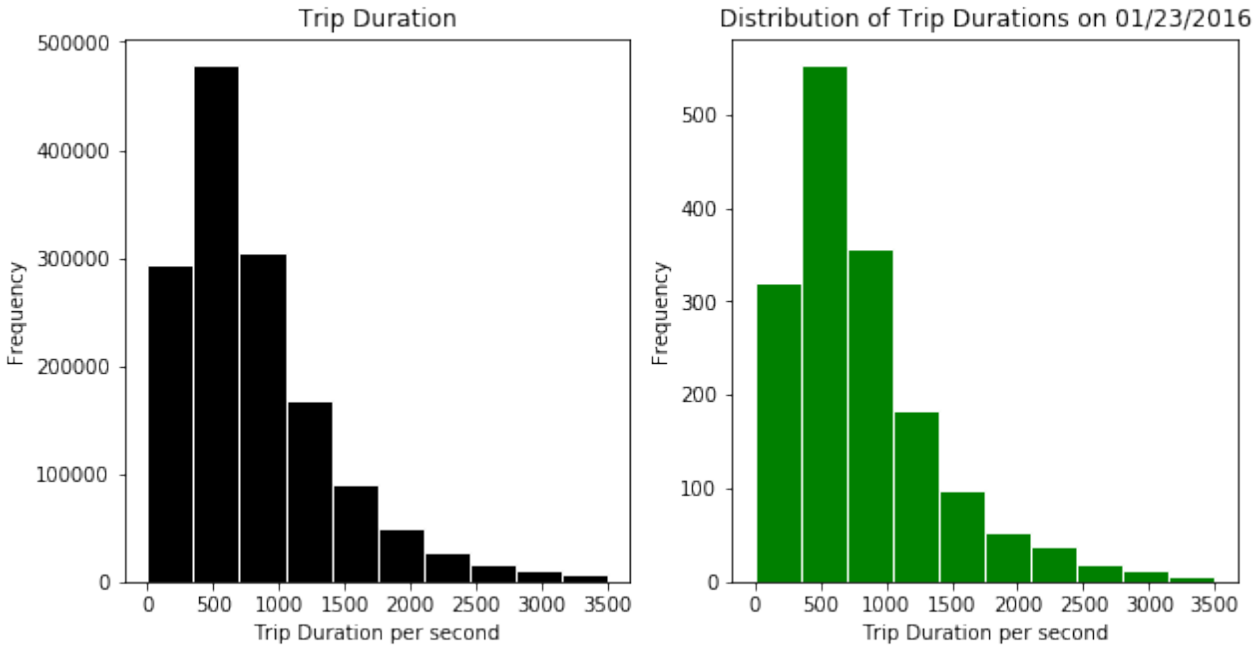
In [29]:
```
snow_storm = (taxidata["pickup_datetime"].dt.year == 2016) & (taxidata["
snow_storm = taxidata[snow_storm]
snow_storm.shape
```

Out[29]:
```
(1648, 11)
```

In [40]:
```
fig = plt.figure(figsize = (10, 5))
ax1 = fig.add_subplot(1,2,1)
ax2 = fig.add_subplot(1,2,2)

ax1.hist(taxidata["trip_duration"], range = (0,3500), edgecolor = "white
ax1.set_title("Trip Duration")
ax1.set_xlabel("Trip Duration per second")
ax1.set_ylabel("Frequency")
ax2.hist(snowstorm_taxidata["trip_duration"], range = (0,3500), edgecolo
ax2.set_title("Distribution of Trip Durations on 01/23/2016")
ax2.set_xlabel("Trip Duration per second")
ax2.set_ylabel("Frequency")
```

Out[40]:
```
Text(0, 0.5, 'Frequency')
```



In [ ]:
```
The trip distrubition is pretty much the same as in January
```