In [1]:
```
# Initialize autograder
# If you see an error message, you'll need to do
# pip3 install otter-grader
import otter
grader = otter.Notebook()
```

Project 3: Predicting Taxi Ride Duration

Due Date: Wednesday 3/4/20, 11:59PM

Collaboration Policy

Data science is a collaborative activity. While you may talk with others about the project, we ask that you **write your solutions individually.** If you do discuss the assignments with others please **include their names** at the top of your notebook.

Collaborators: *list collaborators here*

Score Breakdown

Question	Points
1b	2
1c	3
1d	2
2a	1
2b	2
3a	2
3b	1
3c	2
3d	2
4a	2
4b	2
4c	2
4d	2
4e	2
4f	2
4g	4
5b	7
5c	3
Total	43

This Assignment

In this project, you will use what you've learned in class to create a regression model that predicts the travel time of a taxi ride in New York. Some questions in this project are more substantial than those of past projects.

After this project, you should feel comfortable with the following:

- The data science lifecycle: data selection and cleaning, EDA, feature engineering, and model selection.
- Using `sklearn` to process data and fit linear regression models.
- Embedding linear regression as a component in a more complex model.

First, let's import:

In [2]:
```python
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
%matplotlib inline

import seaborn as sns
```

The Data

Attributes of all [yellow taxi (https://en.wikipedia.org/wiki/Taxicabs_of_New_York_City)](https://en.wikipedia.org/wiki/Taxicabs_of_New_York_City) trips in January 2016 are published by the [NYC Taxi and Limosine Commission (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page)](https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page).

The full data set takes a long time to download directly, so we've placed a simple random sample of the data into `taxi.db` , a SQLite database. You can view the code used to generate this sample in the `taxi_sample.ipynb` file included with this project (not required).

Columns of the `taxi` table in `taxi.db` include:

- `pickup_datetime` : date and time when the meter was engaged
- `dropoff_datetime` : date and time when the meter was disengaged
- `pickup_lon` : the longitude where the meter was engaged
- `pickup_lat` : the latitude where the meter was engaged
- `dropoff_lon` : the longitude where the meter was disengaged
- `dropoff_lat` : the latitude where the meter was disengaged
- `passengers` : the number of passengers in the vehicle (driver entered value)
- `distance` : trip distance
- `duration` : duration of the trip in seconds

Your goal will be to predict `duration` from the pick-up time, pick-up and drop-off locations, and distance.

Part 1: Data Selection and Cleaning

In this part, you will limit the data to trips that began and ended on Manhattan Island (map (https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12z/data=!3m1!4b1 73.9712488)).

The below cell uses a SQL query to load the `taxi` table from `taxi.db` into a Pandas DataFrame called `all_taxi`.

It only includes trips that have **both** pick-up and drop-off locations within the boundaries of New York City:

- Longitude is between -74.03 and -73.75 (inclusive of both boundaries)
- Latitude is between 40.6 and 40.88 (inclusive of both boundaries)

You don't have to change anything, just run this cell.

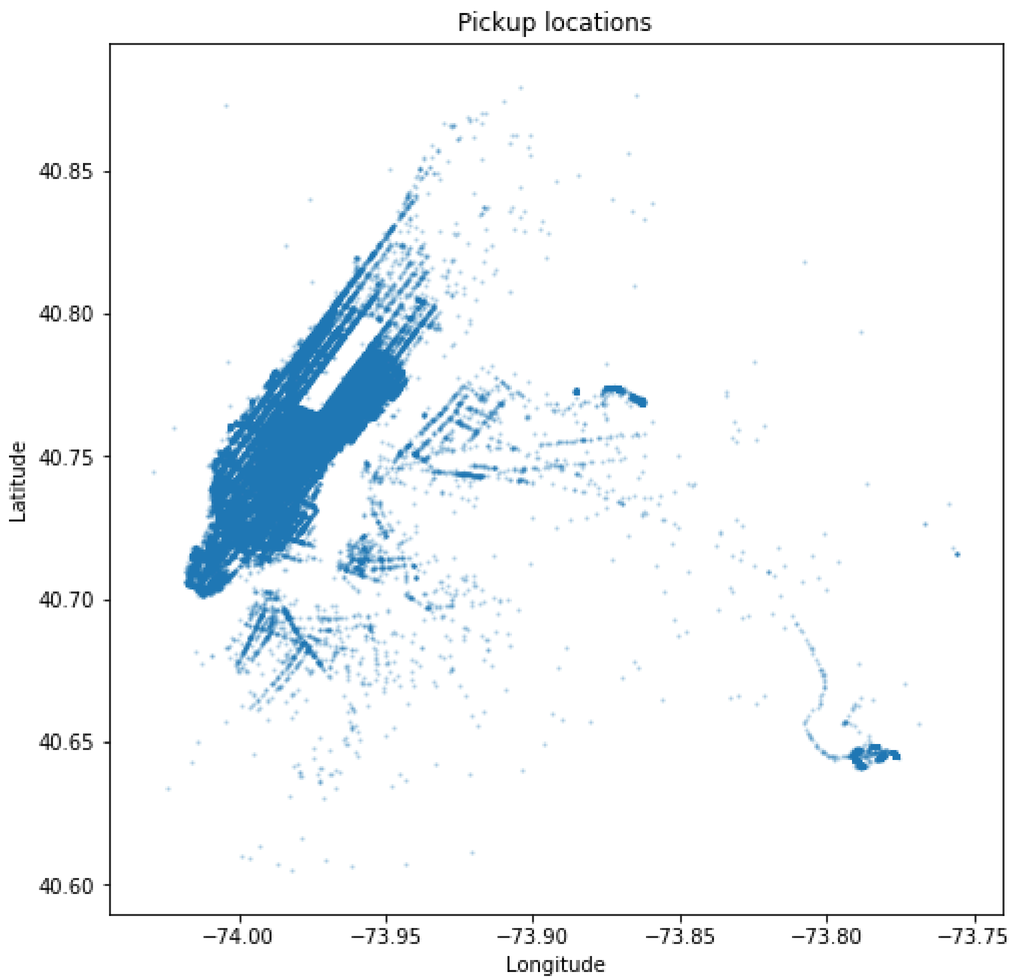
In [3]:
```
import sqlite3

conn = sqlite3.connect('taxi.db')
lon_bounds = [-74.03, -73.75]
lat_bounds = [40.6, 40.88]

c = conn.cursor()

my_string = 'SELECT * FROM taxi WHERE'

for word in ['pickup_lat', 'AND dropoff_lat']:
    my_string += ' {} BETWEEN {} AND {}'.format(word, lat_bounds[0], lat_
bounds[1])

for word in ['AND pickup_lon', 'AND dropoff_lon']:
    my_string += ' {} BETWEEN {} AND {}'.format(word, lon_bounds[0], lon_
bounds[1])

c.execute(my_string)

results = c.fetchall()

row_res = conn.execute('select * from taxi')
names = list(map(lambda x: x[0], row_res.description))


all_taxi = pd.DataFrame(results)
all_taxi.columns = names
all_taxi.head()
```

Out[3]:

	pickup_datetime	dropoff_datetime	pickup_lon	pickup_lat	dropoff_lon	dropoff_lat	passengers
0	2016-01-30 22:47:32	2016-01-30 23:03:53	-73.988251	40.743542	-74.015251	40.709808	1
1	2016-01-04 04:30:48	2016-01-04 04:36:08	-73.995888	40.760010	-73.975388	40.782200	1
2	2016-01-07 21:52:24	2016-01-07 21:57:23	-73.990440	40.730469	-73.985542	40.738510	1
3	2016-01-01 04:13:41	2016-01-01 04:19:24	-73.944725	40.714539	-73.955421	40.719173	1
4	2016-01-08 18:46:10	2016-01-08 18:54:00	-74.004494	40.706989	-74.010155	40.716751	5

A scatter plot of pickup locations shows that most of them are on the island of Manhattan. The empty white rectangle is Central Park; cars are not allowed there.

In [4]:
```
def pickup_scatter(t):
    plt.scatter(t['pickup_lon'], t['pickup_lat'], s=2, alpha=0.2)
    plt.xlabel('Longitude')
    plt.ylabel('Latitude')
    plt.title('Pickup locations')

plt.figure(figsize=(8, 8))
pickup_scatter(all_taxi)
```



The two small blobs outside of Manhattan with very high concentrations of taxi pick-ups are airports.

Question 1b

Create a DataFrame called `clean_taxi` that only includes trips with a positive passenger count, a positive distance, a duration of at least 1 minute and at most 1 hour, and an average speed of at most 100 miles per hour. Inequalities should not be strict (e.g., `<=` instead of `<`) unless comparing to 0.

The provided tests check that you have constructed `clean_taxi` correctly.

In [5]:
```
clean_taxi = all_taxi[(all_taxi.passengers > 0) &
                      (all_taxi.distance > 0) &
                      (all_taxi.duration >= 60) &
                      (all_taxi.duration <= 3600) &
                      (all_taxi.distance/(all_taxi.duration/3600) <= 100
)]
```

In [6]:
```
grader.check("q1b")
```

Out[6]: All tests passed!

Question 1c (challenging)

Create a DataFrame called `manhattan_taxi` that only includes trips from `clean_taxi` that start and end within a polygon that defines the boundaries of [Manhattan Island (https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12z/data=!3m1!4b 73.9712488)](https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12z/data=!3m1!4b).

The vertices of this polygon are defined in `manhattan.csv` as (latitude, longitude) pairs, which are [published here (https://gist.github.com/baygross/5430626)](https://gist.github.com/baygross/5430626).

An efficient way to test if a point is contained within a polygon is [described on this page (http://alienryderflex.com/polygon/)](http://alienryderflex.com/polygon/). There are even implementations on that page (though not in Python). Even with an efficient approach, the process of checking each point can take several minutes. It's best to test your work on a small sample of `clean_taxi` before processing the whole thing. (To check if your code is working, draw a scatter diagram of the (lon, lat) pairs of the result; the scatter diagram should have the shape of Manhattan.)

The provided tests check that you have constructed `manhattan_taxi` correctly. It's not required that you implement the `in_manhattan` helper function, but that's recommended. If you cannot solve this problem, you can still continue with the project; see the instructions below the answer cell.

In [7]:
```
clean_taxi.columns
```

Out[7]:
```
Index(['pickup_datetime', 'dropoff_datetime', 'pickup_lon', 'pickup_la
t',
       'dropoff_lon', 'dropoff_lat', 'passengers', 'distance', 'duratio
n'],
      dtype='object')
```

In [8]:
```
manhattan = pd.read_csv('manhattan.csv')

polyX = manhattan['lon'].tolist()
polyY = manhattan['lat'].tolist()
polyCorners = len(polyX)
constant = [-1]*30
multiple = [-1]*30

def pre_calc():
    j=polyCorners-1
    for i in range(polyCorners):
        if polyY[j]==polyY[i]:
            constant[i]=polyX[i]
            multiple[i]=0
        else:
            constant[i]=polyX[i]-(polyY[i]*polyX[j])/(polyY[j]-polyY[i])+
(polyY[i]*polyX[i])/(polyY[j]-polyY[i])
            multiple[i]=(polyX[j]-polyX[i])/(polyY[j]-polyY[i])
        j=i

def in_manhattan(x, y):
    pre_calc()
    oddNodes=False
    current=polyY[polyCorners-1]>y
    previous=None
    for i in range(polyCorners):
        previous=current
        current=polyY[i]>y
        if current!=previous:
            oddNodes^=y*multiple[i]+constant[i]<x
    return oddNodes

pre_calc()

valid_data = dict()
valid_data['pickup_datetime'] = []
valid_data['dropoff_datetime'] = []
valid_data['pickup_lon'] = []
valid_data['pickup_lat'] = []
valid_data['dropoff_lon'] = []
valid_data['dropoff_lat'] = []
valid_data['passengers'] = []
valid_data['distance'] = []
valid_data['duration'] = []

for index, row in clean_taxi.iterrows():
    if in_manhattan(row['pickup_lon'], row['pickup_lat']) and in_manhatta
n(row['dropoff_lon'],row['dropoff_lat']):
        valid_data['pickup_datetime'].append(row['pickup_datetime'])
        valid_data['dropoff_datetime'].append(row['dropoff_datetime'])
        valid_data['pickup_lon'].append(row['pickup_lon'])
        valid_data['pickup_lat'].append(row['pickup_lat'])
        valid_data['dropoff_lon'].append(row['dropoff_lon'])
        valid_data['dropoff_lat'].append(row['dropoff_lat'])
        valid_data['passengers'].append(row['passengers'])
        valid_data['distance'].append(row['distance'])
```

```
        valid_data['duration'].append(row['duration'])

manhattan_taxi = pd.DataFrame.from_dict(valid_data)
```

In [9]:
```
manhattan_taxi.shape
```

Out[9]: (82800, 9)

In [10]:
```
manhattan_taxi.head(5)
```

Out[10]:

	pickup_datetime	dropoff_datetime	pickup_lon	pickup_lat	dropoff_lon	dropoff_lat	passengers
0	2016-01-30 22:47:32	2016-01-30 23:03:53	-73.988251	40.743542	-74.015251	40.709808	1
1	2016-01-04 04:30:48	2016-01-04 04:36:08	-73.995888	40.760010	-73.975388	40.782200	1
2	2016-01-07 21:52:24	2016-01-07 21:57:23	-73.990440	40.730469	-73.985542	40.738510	1
3	2016-01-08 18:46:10	2016-01-08 18:54:00	-74.004494	40.706989	-74.010155	40.716751	5
4	2016-01-02 12:39:57	2016-01-02 12:53:29	-73.958214	40.760525	-73.983360	40.760406	1

In [11]:
```
grader.check("q1c")
```

Out[11]: All tests passed!

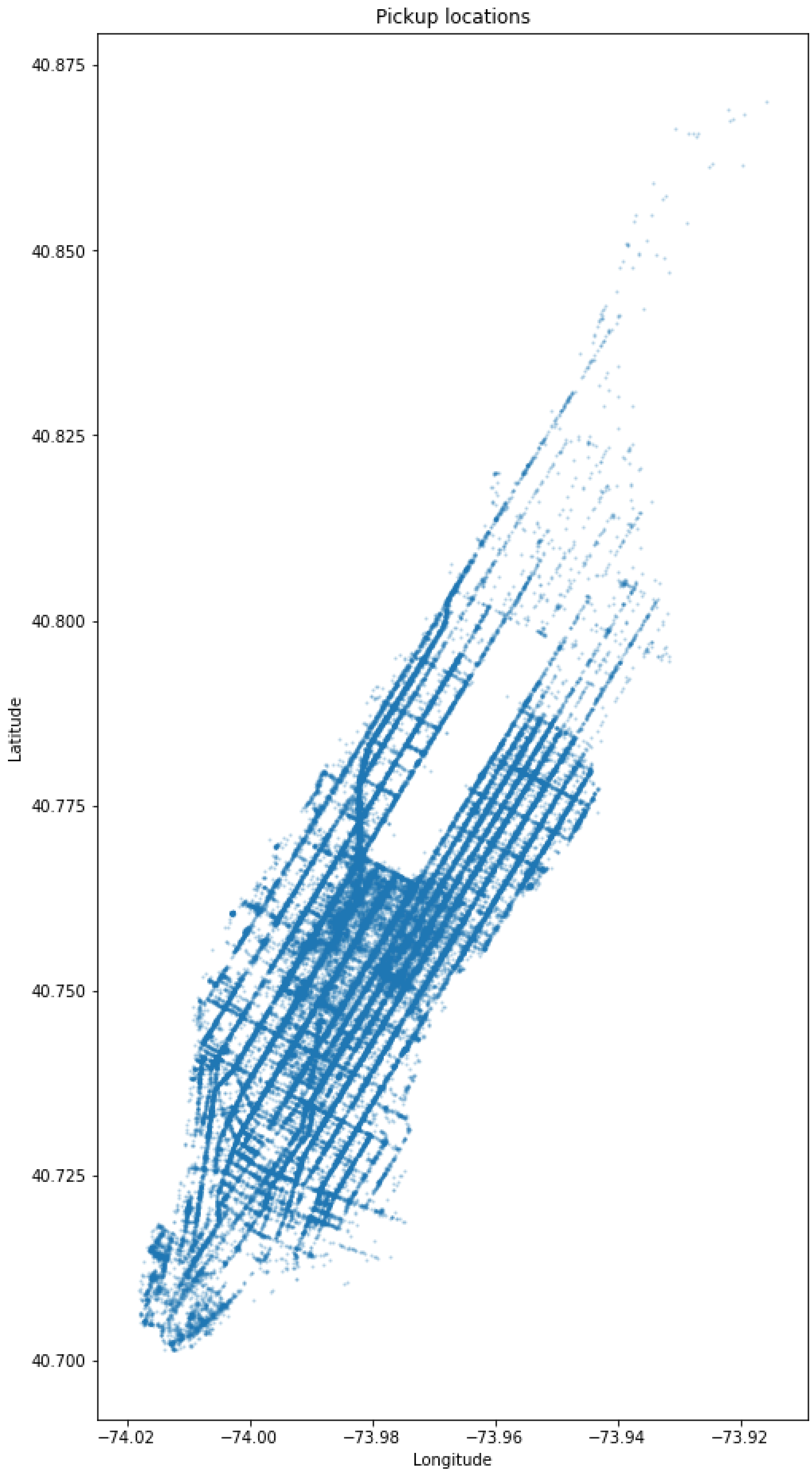
If you are unable to solve the problem above, have trouble with the tests, or want to work on the rest of the project before solving it, run the following cell to load the cleaned Manhattan data directly. (Note that you may not solve the previous problem just by loading this data file; you have to actually write the code.)

In [12]:
```
# manhattan_taxi = pd.read_csv('manhattan_taxi.csv')
```

A scatter diagram of only Manhattan taxi rides has the familiar shape of Manhattan Island.

In [13]:

```
plt.figure(figsize=(8, 16))
pickup_scatter(manhattan_taxi)
```


Pickup locations
40.875
40.850
40.825
40.800
40.775
40.750
40.725
40.700
Latitude
−74.02
−74.00
−73.98
−73.96
−73.94
−73.92
Longitude

Question 1d

Print a summary of the data selection and cleaning you performed. **Your Python code should not include any number literals, but instead should refer to the shape of `all_taxi` , `clean_taxi` , and `manhattan_taxi` .**

E.g., you should print something like: "Of the original 1000 trips, 21 anomalous trips (2.1%) were removed through data cleaning, and then the 600 trips within Manhattan were selected for further analysis."

(Note that the numbers in the example above are not accurate.)

One way to do this is with Python's f-strings. For instance,

```
name = "Joshua"
print(f"Hi {name}, how are you?")
```

prints out `Hi Joshua, how are you?`.

Please ensure that your Python code does not contain any very long lines, or we can't grade it.

Your response will be scored based on whether you generate an accurate description and do not include any number literals in your Python expression, but instead refer to the dataframes you have created.

In [14]:
```
print(f"After the SQL query, our dataframe 'all_taxi' contains taxi trip
 data from within New York. \
In particular, it has {all_taxi.shape[0]} rows and {all_taxi.shape[1]} co
lumns, where the features are: \
{all_taxi.columns.tolist()}\n")
print(f"We then filtered this data to create a new dataframe called 'clea
n_taxi', only keeping data with \
positive passenger and distance measurements, with trip durations of a mi
nimum of {int(min(clean_taxi.duration/60))} \
min and atmost {int(round(max(clean_taxi.duration/3600),0))} hour. We als
o ensured that the average speed across \
the trip was at most {round(max(clean_taxi.distance/(clean_taxi.duration/
3600)),0)} miles/hour.")
print(f"These transformations resulted in the number of rows decreasing f
rom {all_taxi.shape[0]} \
to {clean_taxi.shape[0]}.\n")
print(f"Finally, after filtering the data, we restricted the dataset to o
nly contain trips from 'clean_taxi' \
that start and end within a polygon that defines the boundaries of Manhat
tan Island by using bounding boxes.")
print(f"This resulted in the dataset shrinking from {clean_taxi.shape[0]}
 to {manhattan_taxi.shape[0]}, \
a decrease of {round(((all_taxi.shape[0]-manhattan_taxi.shape[0])/all_tax
i.shape[0])*100,2)}% from \
the original dataset.")
```

```
After the SQL query, our dataframe 'all_taxi' contains taxi trip data f
rom within New York. In particular, it has 97692 rows and 9 columns, wh
ere the features are: ['pickup_datetime', 'dropoff_datetime', 'pickup_l
on', 'pickup_lat', 'dropoff_lon', 'dropoff_lat', 'passengers', 'distanc
e', 'duration']

We then filtered this data to create a new dataframe called 'clean_tax
i', only keeping data with positive passenger and distance measurement
s, with trip durations of a minimum of 1 min and atmost 1 hour. We also
ensured that the average speed across the trip was at most 83.0 miles/h
our.
These transformations resulted in the number of rows decreasing from 97
692 to 96445.

Finally, after filtering the data, we restricted the dataset to only co
ntain trips from 'clean_taxi' that start and end within a polygon that
defines the boundaries of Manhattan Island by using bounding boxes.
This resulted in the dataset shrinking from 96445 to 82800, a decrease
of 15.24% from the original dataset.
```

Part 2: Exploratory Data Analysis

In this part, you'll choose which days to include as training data in your regression model.

Your goal is to develop a general model that could potentially be used for future taxi rides. There is no guarantee that future distributions will resemble observed distributions, but some effort to limit training data to typical examples can help ensure that the training data are representative of future observations.

January 2016 had some atypical days. New Year's Day (January 1) fell on a Friday. MLK Day was on Monday, January 18. A [historic blizzard (https://en.wikipedia.org/wiki/January_2016_United_States_blizzard)](https://en.wikipedia.org/wiki/January_2016_United_States_blizzard) passed through New York that month. Using this dataset to train a general regression model for taxi trip times must account for these unusual phenomena, and one way to account for them is to remove atypical days from the training data.

Question 2a

Add a column labeled `date` to `manhattan_taxi` that contains the date (but not the time) of pickup, formatted as a `datetime.date` value ([docs (https://docs.python.org/3/library/datetime.html#date-objects)](https://docs.python.org/3/library/datetime.html#date-objects)).

The provided tests check that you have extended `manhattan_taxi` correctly.

In [15]:
```
manhattan_taxi.info()
# We need to convert the data to a datetime object before we can extract
 date from it
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 82800 entries, 0 to 82799
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   pickup_datetime   82800 non-null  object
 1   dropoff_datetime  82800 non-null  object
 2   pickup_lon        82800 non-null  float64
 3   pickup_lat        82800 non-null  float64
 4   dropoff_lon       82800 non-null  float64
 5   dropoff_lat       82800 non-null  float64
 6   passengers        82800 non-null  int64
 7   distance          82800 non-null  float64
 8   duration          82800 non-null  int64
dtypes: float64(5), int64(2), object(2)
memory usage: 5.7+ MB
```

In [16]:
```
from datetime import datetime, date
manhattan_taxi['date'] = pd.to_datetime(manhattan_taxi['pickup_datetime'
], format="%Y/%m/%d")
manhattan_taxi['date'] = [d.date() for d in manhattan_taxi['date']]
print(type(manhattan_taxi['date'][0]))
manhattan_taxi.head()
```

```
<class 'datetime.date'>
```

Out[16]:

	pickup_datetime	dropoff_datetime	pickup_lon	pickup_lat	dropoff_lon	dropoff_lat	passengers
0	2016-01-30 22:47:32	2016-01-30 23:03:53	-73.988251	40.743542	-74.015251	40.709808	1
1	2016-01-04 04:30:48	2016-01-04 04:36:08	-73.995888	40.760010	-73.975388	40.782200	1
2	2016-01-07 21:52:24	2016-01-07 21:57:23	-73.990440	40.730469	-73.985542	40.738510	1
3	2016-01-08 18:46:10	2016-01-08 18:54:00	-74.004494	40.706989	-74.010155	40.716751	5
4	2016-01-02 12:39:57	2016-01-02 12:53:29	-73.958214	40.760525	-73.983360	40.760406	1

In [17]:
```
grader.check("q2a")
```

Out[17]: All tests passed!

Question 2b

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.

As a hint, consider how taxi usage might change on a day with a blizzard. How could you visualize/plot this?

In [18]:
```
# From the result of the below expression, we can see that the dataset on
ly contains values for the month of January.
manhattan_taxi['date'].value_counts()
```

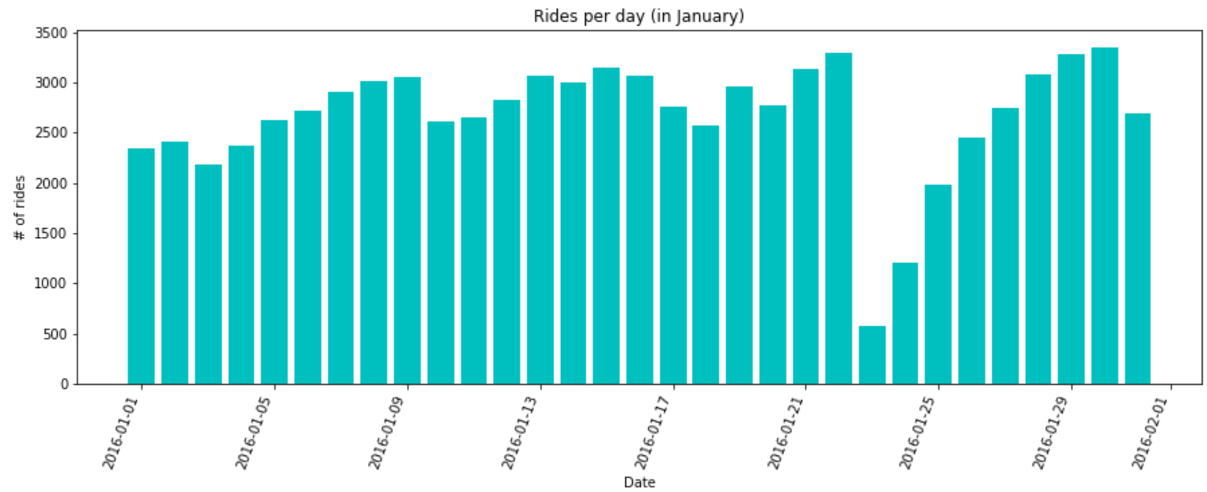
Out[18]:
```
2016-01-30    3352
2016-01-22    3291
2016-01-29    3280
2016-01-15    3139
2016-01-21    3133
2016-01-28    3083
2016-01-13    3066
2016-01-16    3059
2016-01-09    3058
2016-01-08    3010
2016-01-14    2992
2016-01-19    2963
2016-01-07    2908
2016-01-12    2829
2016-01-20    2776
2016-01-17    2753
2016-01-27    2750
2016-01-06    2721
2016-01-31    2690
2016-01-11    2645
2016-01-05    2630
2016-01-10    2605
2016-01-18    2566
2016-01-26    2445
2016-01-02    2411
2016-01-04    2368
2016-01-01    2337
2016-01-03    2177
2016-01-25    1982
2016-01-24    1203
2016-01-23     578
Name: date, dtype: int64
```

In [19]:
```
data = pd.DataFrame(manhattan_taxi['date'].value_counts().reset_index())
data.columns = ['date','count']
data.head(5)
```

Out[19]:

	date	count
0	2016-01-30	3352
1	2016-01-22	3291
2	2016-01-29	3280
3	2016-01-15	3139
4	2016-01-21	3133

In [20]:
```
fig, ax = plt.subplots(figsize=(15, 5))
ax.bar(data['date'].tolist(), data['count'].tolist(), color='c')
ax.set_title('Rides per day (in January)')
ax.set_xlabel('Date')
ax.set_ylabel('# of rides')
fig.autofmt_xdate(bottom=0.15, rotation=70, ha='right')
plt.show()
```



From the data, we can see that the the number of rides dropped dramatically on January 23rd and January 24th, and slowly starts recovering back to normal till Januar 27th, where it starts to look like previous data. This is consistent with the information presented here (https://en.wikipedia.org/wiki/January_2016_United_States_blizzard (https://en.wikipedia.org/wiki/January_2016_United_States_blizzard)), where "A travel ban was instituted for New York City and Newark, New Jersey for January 23–24."

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.)

In [21]:
```
import calendar
import re

from datetime import date

atypical = [1, 2, 3, 18, 23, 24, 25, 26]
typical_dates = [date(2016, 1, n) for n in range(1, 32) if n not in atypi
cal]
typical_dates

print('Typical dates:\n')
pat = '  [1-3]|18 | 23| 24|25 |26 '
print(re.sub(pat, '   ', calendar.month(2016, 1)))

final_taxi = manhattan_taxi[manhattan_taxi['date'].isin(typical_dates)]
```

```
Typical dates:

    January 2016
Mo Tu We Th Fr Sa Su

 4  5  6  7  8  9 10
11 12 13 14 15 16 17
   19 20 21 22
      27 28 29 30 31
```

You are welcome to perform more exploratory data analysis, but your work will not be scored. Here's a blank cell to use if you wish. In practice, further exploration would be warranted at this point, but the project is already pretty long.

In [22]:
```
# Optional: More EDA here
print(final_taxi.shape)
final_taxi.head()
```

```
(67101, 10)
```

Out[22]:

	pickup_datetime	dropoff_datetime	pickup_lon	pickup_lat	dropoff_lon	dropoff_lat	passengers
0	2016-01-30 22:47:32	2016-01-30 23:03:53	-73.988251	40.743542	-74.015251	40.709808	1
1	2016-01-04 04:30:48	2016-01-04 04:36:08	-73.995888	40.760010	-73.975388	40.782200	1
2	2016-01-07 21:52:24	2016-01-07 21:57:23	-73.990440	40.730469	-73.985542	40.738510	1
3	2016-01-08 18:46:10	2016-01-08 18:54:00	-74.004494	40.706989	-74.010155	40.716751	5
5	2016-01-17 19:21:16	2016-01-17 19:28:24	-73.978401	40.764992	-73.981018	40.762100	1

Part 3: Feature Engineering

In this part, you'll create a design matrix (i.e., feature matrix) for your linear regression model. This is analagous to the pipelines you've built already in class: you'll be adding features, removing labels, and scaling among other things.

You decide to predict trip duration from the following inputs: start location, end location, trip distance, time of day, and day of the week (*Monday, Tuesday, etc.*).

You will ensure that the process of transforming observations into a design matrix is expressed as a Python function called `design_matrix` , so that it's easy to make predictions for different samples in later parts of the project.

Because you are going to look at the data in detail in order to define features, it's best to split the data into training and test sets now, then only inspect the training set.

In [23]:
```
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)
```

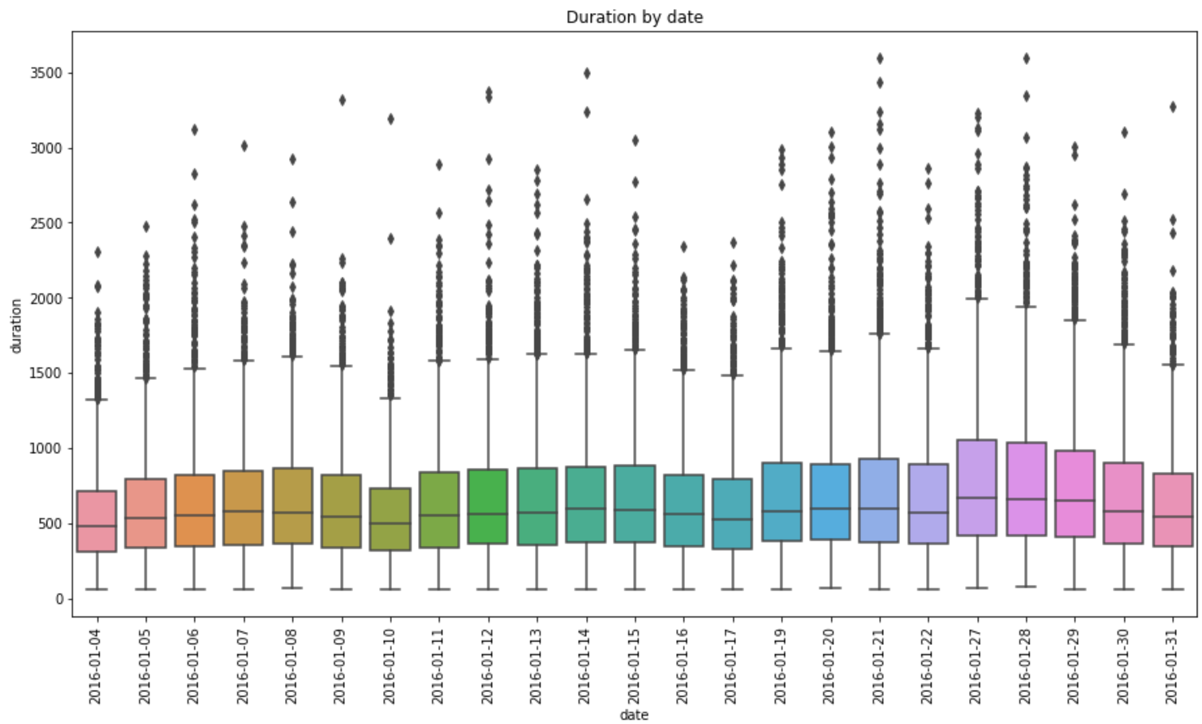
Question 3a

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 one below.

You can generate this type of plot using `sns.boxplot`

In [24]:
```
order = sorted(train.date.unique())
fig, ax = plt.subplots(figsize=(15,8))
sns.boxplot(ax=ax, x=train['date'],y=train['duration'], order=order)
plt.xticks(rotation='vertical')
ax.set_title('Duration by date')
```

Out[24]:
```
Text(0.5, 1.0, 'Duration by date')
```



Question 3b

In one or two sentences, describe the assocation between the day of the week and the duration of a taxi trip. Your answer should be supported by your boxplot above.

Note: The end of Part 2 showed a calendar for these dates and their corresponding days of the week.

Response to 3b

Boxplots describe minimum, 1st quartile, median, 3rd quartile and maximum (non-unusual observation), along with any outliers. Based on the above plot, it appears that Sundays (01/10, 01/17) have shorter trip durations, as see from having a lower quartile, median, 3rd quartile and non-unusual maximum observation, which are lower than all the other days for that particular week. Apart from this, the median ride duration across days is approximately 500seconds, or about 8.5 minutes, while all days appear to have approximately same minimum duration of 5-10 seconds (this might be an error or cancelled rides). The maximum non-unusual observation on Fridays seems to be higher than those on other days of the week. 01/27 appears to be an abberation in terms of ride durations, and it is probably because it was the first day after the travel ban was lifted from the storm + roads cleared up so trips could take place.

Below, the provided `augment` function adds various columns to a taxi ride dataframe.

- `hour` : The integer hour of the pickup time. E.g., a 3:45pm taxi ride would have `15` as the hour. A 12:20am ride would have `0` .
- `day` : The day of the week with Monday=0, Sunday=6.
- `weekend` : 1 if and only if the `day` is Saturday or Sunday.
- `period` : 1 for early morning (12am-6am), 2 for daytime (6am-6pm), and 3 for night (6pm-12pm).
- `speed` : Average speed in miles per hour.

No changes are required; just run this cell.

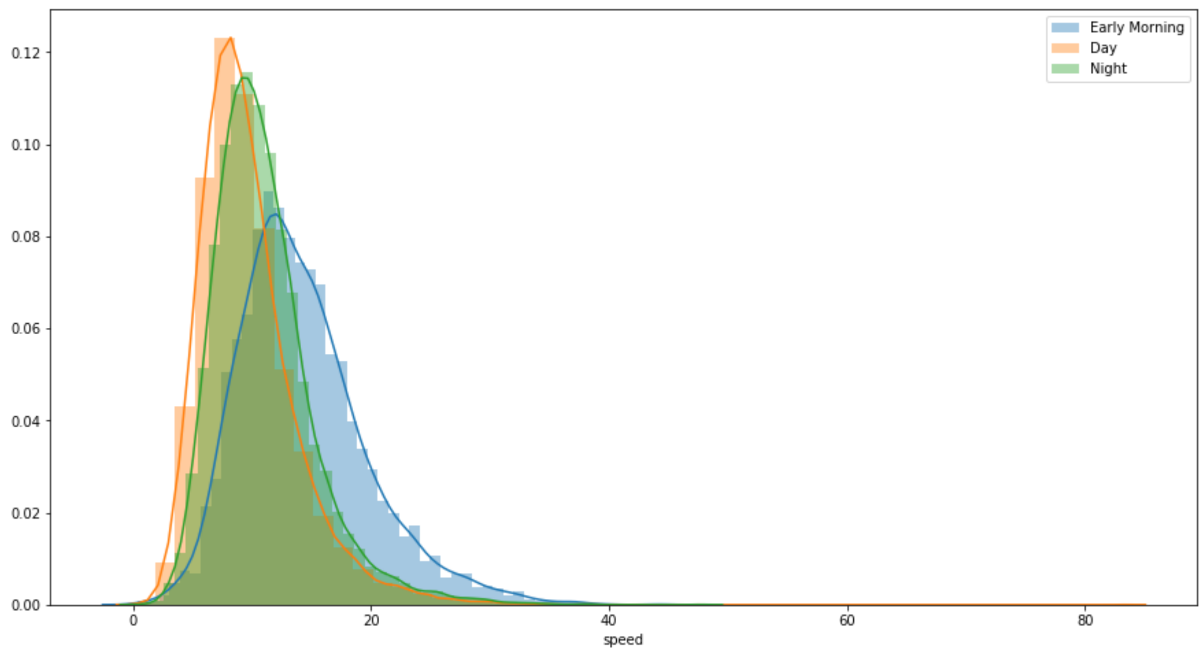
In [25]:
```
def speed(t):
    """Return a column of speeds in miles per hour."""
    return t['distance'] / t['duration'] * 60 * 60

def augment(t):
    """Augment a dataframe t with additional columns."""
    u = t.copy()
    pickup_time = pd.to_datetime(t['pickup_datetime'])
    u.loc[:, 'hour'] = pickup_time.dt.hour
    u.loc[:, 'day'] = pickup_time.dt.weekday
    u.loc[:, 'weekend'] = (pickup_time.dt.weekday >= 5).astype(int)
    u.loc[:, 'period'] = np.digitize(pickup_time.dt.hour, [0, 6, 18])
    u.loc[:, 'speed'] = speed(t)
    return u

train = augment(train)
test = augment(test)
train.iloc[0,:] # An example row
```

Out[25]:
```
pickup_datetime     2016-01-21 18:02:20
dropoff_datetime    2016-01-21 18:27:54
pickup_lon                     -73.9942
pickup_lat                       40.751
dropoff_lon                    -73.9637
dropoff_lat                     40.7711
passengers                            1
distance                           2.77
duration                           1534
date                         2016-01-21
hour                                 18
day                                   3
weekend                               0
period                                3
speed                           6.50065
Name: 14043, dtype: object
```

Question 3c

Use `sns.distplot` to create an overlaid histogram comparing the distribution of average speeds for taxi rides that start in the early morning (12am-6am), day (6am-6pm; 12 hours), and night (6pm-12am; 6 hours). Your plot should look like this:

In [26]:
```
early_morning = train[train.period == 1]
day = train[train.period == 2]
night = train[train.period == 3]
fig, ax = plt.subplots(figsize=(15,8))
sns.distplot(ax=ax, a=early_morning['speed'], label='Early Morning')
sns.distplot(ax=ax, a=day['speed'], label='Day')
sns.distplot(ax=ax, a=night['speed'], label='Night')
ax.legend()
plt.show()
```



It looks like the time of day is associated with the average speed of a taxi ride.

Question 3d

Manhattan can roughly be divided into Lower, Midtown, and Upper regions. Instead of studying a map, let's approximate by finding the first principal component of the pick-up location (latitude and longitude).

[Principal component analysis (https://en.wikipedia.org/wiki/Principal_component_analysis)](https://en.wikipedia.org/wiki/Principal_component_analysis) (PCA) is a technique that finds new axes as linear combinations of your current axes. These axes are found such that the first returned axis (the first principal component) explains the most variation in values, the 2nd the second most, etc.

Add a `region` column to `train` that categorizes each pick-up location as 0, 1, or 2 based on the value of each point's first principal component, such that an equal number of points fall into each region.

Read the documentation of [`pd.qcut` (https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.qcut.html)](https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.qcut.html), which categorizes points in a distribution into equal-frequency bins.

You don't need to add any lines to this solution. Just fill in the assignment statements to complete the implementation.

Before implementing PCA, it is important to scale and shift your values. The line with `np.linalg.svd` will return your transformation matrix, among other things. You can then use this matrix to convert points in (lat, lon) space into (PC1, PC2) space.

Hint: If you are failing the tests, try visualizing your processed data to understand what your code might be doing wrong.

The provided tests ensure that you have answered the question correctly.

In [27]:
```
# Find the first principle component
D = train[['pickup_lon','pickup_lat']]
pca_n = train.shape[0]
pca_means = np.mean(D)
X = (D - pca_means) / np.sqrt(pca_n)
u, s, vt = np.linalg.svd(X, full_matrices=False)

def add_region(t):
    """Add a region column to t based on vt above."""
    D = t[['pickup_lon', 'pickup_lat']]
    assert D.shape[0] == t.shape[0], 'You set D using the incorrect tabl
e'
    # Always use the same data transformation used to compute vt
    pca_n = D.shape[0]
    X = (D - pca_means) / np.sqrt(pca_n)
    first_pc = X @ vt.transpose()[0]
    t.loc[:,'region'] = pd.qcut(first_pc, 3, labels=[0, 1, 2])

add_region(train)
add_region(test)
```

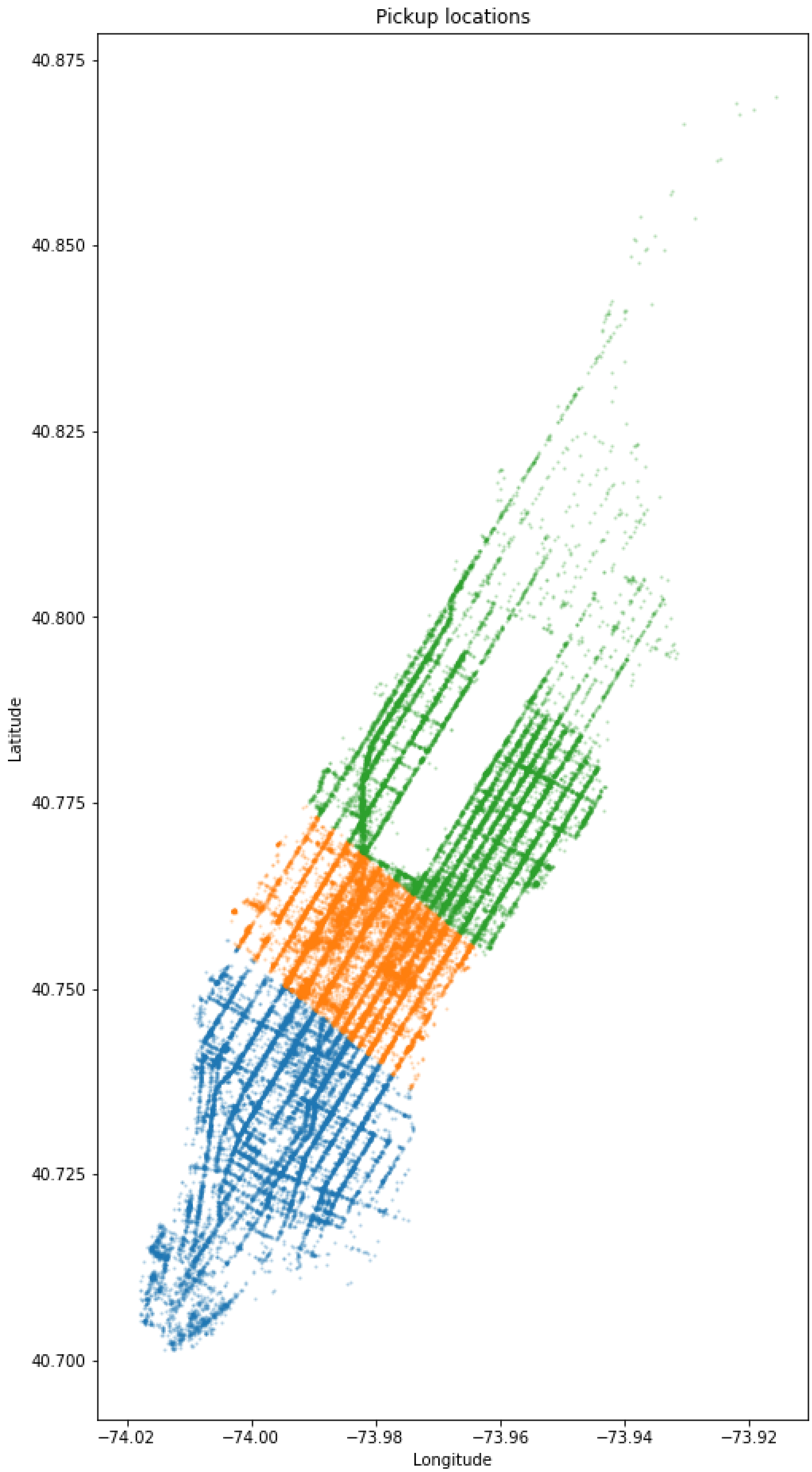
In [28]:
```
grader.check("q3d")
```

Out[28]: All tests passed!

Let's see how PCA divided the trips into three groups. These regions do roughly correspond to Lower Manhattan (below 14th street), Midtown Manhattan (between 14th and the park), and Upper Manhattan (bordering Central Park). No prior knowledge of New York geography was required!

In [29]:

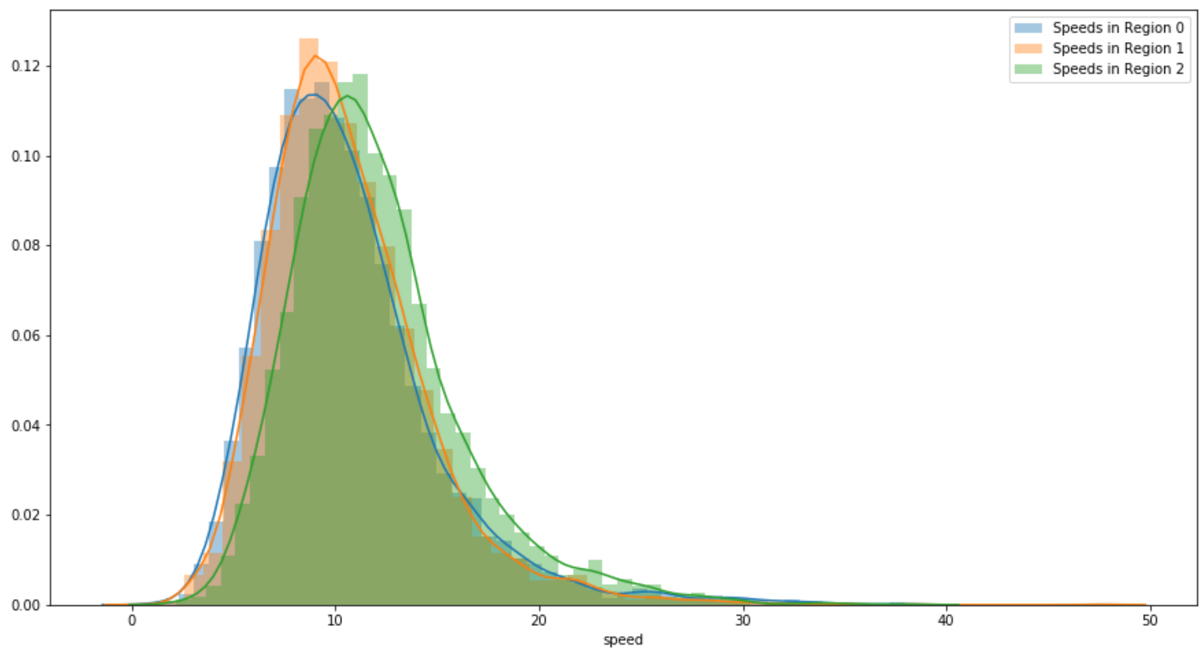
```
plt.figure(figsize=(8, 16))
for i in [0, 1, 2]:
    pickup_scatter(train[train['region'] == i])
```


Pickup locations
Latitude
Longitude
40.875
40.850
40.825
40.800
40.775
40.750
40.725
40.700
−74.02
−74.00
−73.98
−73.96
−73.94
−73.92

Question 3e (ungraded)

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 [30]:
```
region_1 = train[(train['region'] == 0) & (train.period == 3)]
region_2 = train[(train['region'] == 1) & (train.period == 3)]
region_3 = train[(train['region'] == 2) & (train.period == 3)]
fig, ax = plt.subplots(figsize=(15,8))
sns.distplot(ax=ax, a=region_1['speed'], label='Speeds in Region 0')
sns.distplot(ax=ax, a=region_2['speed'], label='Speeds in Region 1')
sns.distplot(ax=ax, a=region_3['speed'], label='Speeds in Region 2')
ax.legend()
plt.show()
```



Finally, we create a design matrix that includes many of these features. Quantitative features are converted to standard units, while categorical features are converted to dummy variables using one-hot encoding. The `period` is not included because it is a linear combination of the `hour`. The `weekend` variable is not included because it is a linear combination of the `day`. The `speed` is not included because it was computed from the `duration`; it's impossible to know the speed without knowing the duration, given that you know the distance.

In [31]:

```
from sklearn.preprocessing import StandardScaler

num_vars = ['pickup_lon', 'pickup_lat', 'dropoff_lon', 'dropoff_lat', 'di
stance']
cat_vars = ['hour', 'day', 'region']

scaler = StandardScaler()
scaler.fit(train[num_vars])

def design_matrix(t):
    """Create a design matrix from taxi ride dataframe t."""
    scaled = t[num_vars].copy()
    scaled.iloc[:,:] = scaler.transform(scaled) # Convert to standard uni
ts
    categoricals = [pd.get_dummies(t[s], prefix=s, drop_first=True) for s
 in cat_vars]
    return pd.concat([scaled] + categoricals, axis=1)

# This processes the full train set, then gives us the first item
# Use this function to get a processed copy of the dataframe passed in
# for training / evaluation
design_matrix(train).iloc[0,:]
```

```
Out[31]: pickup_lon    -0.805821
         pickup_lat    -0.171761
         dropoff_lon    0.954062
         dropoff_lat    0.624203
         distance       0.626326
         hour_1         0.000000
         hour_2         0.000000
         hour_3         0.000000
         hour_4         0.000000
         hour_5         0.000000
         hour_6         0.000000
         hour_7         0.000000
         hour_8         0.000000
         hour_9         0.000000
         hour_10        0.000000
         hour_11        0.000000
         hour_12        0.000000
         hour_13        0.000000
         hour_14        0.000000
         hour_15        0.000000
         hour_16        0.000000
         hour_17        0.000000
         hour_18        1.000000
         hour_19        0.000000
         hour_20        0.000000
         hour_21        0.000000
         hour_22        0.000000
         hour_23        0.000000
         day_1          0.000000
         day_2          0.000000
         day_3          1.000000
         day_4          0.000000
         day_5          0.000000
         day_6          0.000000
         region_1       1.000000
         region_2       0.000000
         Name: 14043, dtype: float64
```

Part 4: Model Selection

In this part, you will select a regression model to predict the duration of a taxi ride.

Important: *Tests in this part do not confirm that you have answered correctly. Instead, they check that you're somewhat close in order to detect major errors. It is up to you to calculate the results correctly based on the question descriptions.*

Question 4a

Assign `constant_rmse` to the root mean squared error on the **test** set for a constant model that always predicts the mean duration of all **training set** taxi rides.

In [32]:
```
def rmse(errors):
    """Return the root mean squared error."""
    return np.sqrt(np.mean(errors ** 2))
```

In [33]:
```
mean_duration = np.mean(train['duration'])
constant_rmse = rmse(test.duration - mean_duration)
constant_rmse
```

Out[33]: 399.1437572352666

In [34]:
```
grader.check("q4a")
```

Out[34]: All tests passed!

Question 4b

Assign `simple_rmse` to the root mean squared error on the test set for a simple linear regression model that uses only the distance of the taxi ride as a feature (and includes an intercept).

Terminology Note: Simple linear regression means that there is only one covariate. Multiple linear regression means that there is more than one. In either case, you can use the `LinearRegression` model from `sklearn` to fit the parameters to data.

In [35]:
```
from sklearn.linear_model import LinearRegression

model = LinearRegression(fit_intercept=True)
model.fit(np.array(train['distance']).reshape(-1, 1), np.array(train['dur
ation']).reshape(-1, 1))
predictions = model.predict(np.array(test['distance']).reshape(-1, 1))
simple_rmse = rmse(predictions - np.array(test['duration']).reshape(-1,1
))
simple_rmse
```

Out[35]: 276.7841105000342

In [36]:
```
grader.check("q4b")
```

Out[36]: All tests passed!

Question 4c

Assign `linear_rmse` to the root mean squared error on the test set for a linear regression model fitted to the training set without regularization, using the design matrix defined by the `design_matrix` function from Part 3.

The provided tests check that you have answered the question correctly and that your `design_matrix` function is working as intended.

In [37]:
```
design_train = design_matrix(train)
design_test = design_matrix(test)

model = LinearRegression()
model.fit(design_train, train['duration'])
predictions = model.predict(design_test)
linear_rmse = rmse(predictions - test['duration'])
linear_rmse
```

Out[37]: 255.19146631882757

In [38]:
```
grader.check("q4c")
```

Out[38]: All tests passed!

Question 4d

For each possible value of `period`, fit an unregularized linear regression model to the subset of the training set in that `period`. Assign `period_rmse` to the root mean squared error on the test set for a model that first chooses linear regression parameters based on the observed period of the taxi ride, then predicts the duration using those parameters. Again, fit to the training set and use the `design_matrix` function for features.

In [39]:
```
model = LinearRegression()
errors = []

for v in np.unique(train['period']):
    train_subset = design_matrix(train[train.period == v])
    test_subset = design_matrix(test[test.period == v])
    model.fit(train_subset, train[train.period == v]['duration'])
    predictions = model.predict(test_subset)
    error = predictions - test[test.period == v]['duration']
    for err in error:
        errors.append(err)

period_rmse = rmse(np.array(errors))
period_rmse
```

Out[39]: 246.62868831165176

In [40]:
```
grader.check("q4d")
```

Out[40]: All tests passed!

This approach is a simple form of decision tree regression, where a different regression function is estimated for each possible choice among a collection of choices. In this case, the depth of the tree is only 1.

Question 4e

In one or two sentences, explain how the `period` regression model above could possibly outperform linear regression when the design matrix for linear regression already includes one feature for each possible hour, which can be combined linearly to determine the `period` value.

There are two main reasons this model works better:

1. The period regression model above almost works like a random forest model, where we build three different trees using subsets of data (each period becomes its own subset), and measure the prediction quality across all 3 models. Because of the subsets of data, each model is slightly more generalized, and therefore the "collective" model is able to generalize well.
2. Although hour could be combined together to form the period value in some linear combination because of the way linear regression works, the data would still have incredibly high variance because it is a linear combination and the model would need to optimize 6 different theta values. That being said, binning them collectively into a single feature reduces this variance tremendously, as it gets summarized under one feature. While this does potentially result in a loss of information because it essentially also works as a form of dimensionality reduction, in this scenario, it instead seems to work in linear regression's favour.

Question 4f

Instead of predicting duration directly, an alternative is to predict the average *speed* of the taxi ride using linear regression, then compute an estimate of the duration from the predicted speed and observed distance for each ride.

Assign `speed_rmse` to the root mean squared error in the **duration** predicted by a model that first predicts speed as a linear combination of features from the `design_matrix` function, fitted on the training set, then predicts duration from the predicted speed and observed distance.

Hint: Speed is in miles per hour, but duration is measured in seconds. You'll need the fact that there are 60 * 60 = 3,600 seconds in an hour.

In [41]:
```python
# Converting speed from miles per hour to miles per second
train_speed = train['speed']/3600

design_train = design_matrix(train)
design_test = design_matrix(test)

model = LinearRegression()
model.fit(design_train, train_speed)
predictions = model.predict(design_test)

predicted_durations = []
for idx, prediction in enumerate(predictions):
    predicted_durations.append(test['distance'].iloc[idx]/prediction)

speed_rmse = rmse(predicted_durations - test['duration'])
speed_rmse
```

Out[41]: 243.0179836851495

In [42]:
```python
grader.check("q4f")
```

Out[42]: All tests passed!

In [43]:
```python
train[['duration','speed']].describe()
```

Out[43]:

	duration	speed
count	53680.000000	53680.000000
mean	661.777310	10.567878
std	406.675522	4.502984
min	60.000000	0.201399
25%	361.000000	7.504225
50%	571.000000	9.759036
75%	867.000000	12.689808
max	3598.000000	83.478261

Reasoning: I think this happens because there much lesser variance in speed, but a large variation in trip duration (as see in the subset of the table.describe(). Even after scaling it, the variance is still captured by the values in the range [-1,1] for duration, and this makes it harder for the model to predict accurately.

Question 4g

Finally, complete the function `tree_regression_errors` (and helper function `speed_error`) that combines the ideas from the two previous models and generalizes to multiple categorical variables.

The `tree_regression_errors` should:

- Find a different linear regression model for each possible combination of the variables in `choices`;
- Fit to the specified `outcome` (on train) and predict that `outcome` (on test) for each combination (`outcome` will be `'duration'` or `'speed'`);
- Use the specified `error_fn` (either `duration_error` or `speed_error`) to compute the error in predicted duration using the predicted outcome;
- Aggregate those errors over the whole test set and return them.

You should find that including each of `period`, `region`, and `weekend` improves prediction accuracy, and that predicting speed rather than duration leads to more accurate duration predictions.

If you're stuck, try putting print statements in the skeleton code to see what it's doing.

In [44]:
```python
choices = ['period', 'region', 'weekend']
print(train.groupby(choices).size())
```

```
period  region  weekend
1       0       0           980
                1          1604
        1       0           792
                1           658
        2       0           453
                1           381
2       0       0          6508
                1          2179
        1       0          7728
                1          2199
        2       0          9283
                1          2694
3       0       0          4905
                1          1718
        1       0          5135
                1          1381
        2       0          3900
                1          1182
dtype: int64
```

In [45]:
```
model = LinearRegression()
choices = ['period', 'region', 'weekend']

def duration_error(predictions, observations):
    """Error between duration predictions (array) and observations (data
 frame)"""
    return predictions - observations['duration']

def speed_error(predictions, observations):
    """Duration error between speed predictions and duration observation
s"""
    return ((observations['distance']/(predictions/3600)) - observations[
'duration'])

def tree_regression_errors(outcome='duration', error_fn=duration_error):
    """Return errors for all examples in test using a tree regression mod
el."""
    errors = []
    for vs in train.groupby(choices).size().index:
        v_train, v_test = train, test
        for v, c in zip(vs, choices):
            v_train = v_train[v_train[c] == v]
            v_test = v_test[v_test[c] == v]
        model.fit(design_matrix(v_train), v_train[outcome])
        predictions = model.predict(design_matrix(v_test))
        errors.extend(error_fn(predictions, v_test))
    return errors

errors = tree_regression_errors()
errors_via_speed = tree_regression_errors('speed', speed_error)
tree_rmse = rmse(np.array(errors))
tree_speed_rmse = rmse(np.array(errors_via_speed))
print('Duration:', tree_rmse, '\nSpeed:', tree_speed_rmse)
```

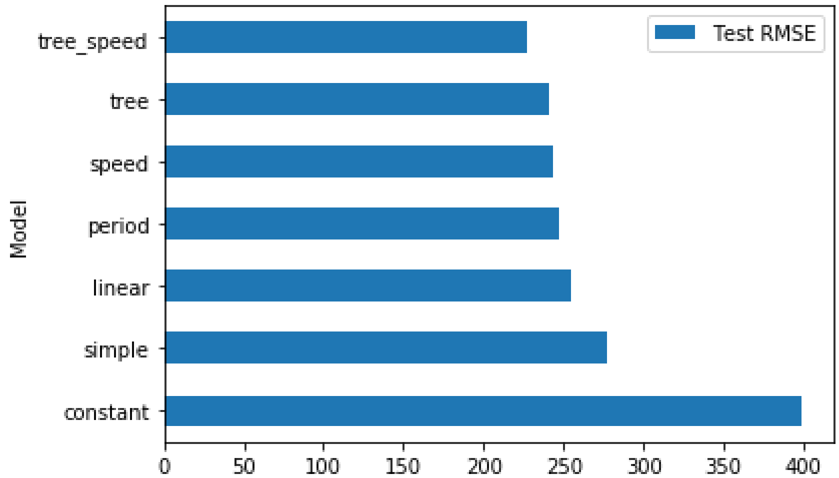
```
Duration: 240.3395219270353
Speed: 226.90793945018308
```

In [46]:
```
grader.check("q4g")
```

Out[46]: All tests passed!

Here's a summary of your results:

In [47]:
```
models = ['constant', 'simple', 'linear', 'period', 'speed', 'tree', 'tre
e_speed']
pd.DataFrame.from_dict({
    'Model': models,
    'Test RMSE': [eval(m + '_rmse') for m in models]
}).set_index('Model').plot(kind='barh');
```



Part 5: Building on your own

In this part you'll build a regression model of your own design, with the goal of achieving even higher performance than you've seen already. You will be graded on your performance relative to others in the class, with higher performance (lower RMSE) receiving more points.

Question 5a

In the below cell (feel free to add your own additional cells), train a regression model of your choice on the same train dataset split used above. The model can incorporate anything you've learned from the class so far.

The model you train will be used for questions 5b and 5c

In [48]:
```
design_train = design_matrix(train)
design_test = design_matrix(test)
```

In [49]:
```python
from sklearn.ensemble import GradientBoostingRegressor
regr = GradientBoostingRegressor(min_samples_split=8,
                                 max_depth=8,
                                 min_samples_leaf=4,
                                 loss='ls',
                                 n_estimators=150,
                                 random_state=42)
regr.fit(design_train, train['duration'])
```

Out[49]:
```
GradientBoostingRegressor(alpha=0.9, ccp_alpha=0.0, criterion='friedman
_mse',
                          init=None, learning_rate=0.1, loss='ls', max_
depth=8,
                          max_features=None, max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split
=None,
                          min_samples_leaf=4, min_samples_split=8,
                          min_weight_fraction_leaf=0.0, n_estimators=15
0,
                          n_iter_no_change=None, presort='deprecated',
                          random_state=42, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_star
t=False)
```

Question 5b

Print a summary of your model's performance. You **must** include the RMSE on the train and test sets. Do not hardcode any values or you won't receive credit.

Don't include any long lines or we won't be able to grade your response.

In [50]:
```python
predictions = regr.predict(design_train)
train_rmse = rmse(predictions - train['duration'])
print("Train RMSE:",train_rmse)
```

```
Train RMSE: 155.21067955386602
```

In [51]:
```python
predictions = regr.predict(design_test)
test_rmse = rmse(predictions - test['duration'])
print("Test RMSE:",test_rmse)
```

```
Test RMSE: 196.72180791541697
```

Question 5c

Describe why you selected the model you did and what you did to try and improve performance over the models in section 4.

Responses should be at most a few sentences.

My intuition said that it would be interesting to try a model that can map non-linear relationships. On that note, I decided to try 3 different models - XGBoost, RandomForest and GradientBoosting. XGBoost, which is supposed to really great for both classification and regression tasks (especially classification), actually resulted in a Test MSE of 775, which was extremely high. This was true even after playing around with a variety of hyperparameters. I wasn't able to figure out why this was this case, but it didn't prove to be promising enough.

I then tried Random Forest, and after a lot of effort, I was able to bring down my test RMSE to about 216, where I played around with `max-depth` , `min_samples_leaf` , `min_samples_split` and a few other hyperaparameters. This looked promising and made sense, since our tree based approach (in 4d and 4g) previously also worked out quite well. I also set n_estimators to 50, which gave me a decent test RMSE, but proved to be slow. I knew that increasing n_estimators would give me a much better test RMSE, but it would not work for the sake of this project. So, I decided to play around with GradientBoosting, which also had roughly the same parameters. I was able to bring down my test RMSE to 196.7, which was a significant improvement to all the other models I had tried as well as our results from 4f.

As a reflection, gradient boosting converted weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. Each tree created improves on the previous tree created, thereby creating a large number of strong trees. Because of this, even with a fairly shallow max-depth, we were able to achieve a low Test RMSE.

Congratulations! You've carried out the entire data science lifecycle for a challenging regression problem.

In Part 1 on data selection, you solved a domain-specific programming problem relevant to the analysis when choosing only those taxi rides that started and ended in Manhattan.

In Part 2 on EDA, you used the data to assess the impact of a historical event---the 2016 blizzard---and filtered the data accordingly.

In Part 3 on feature engineering, you used PCA to divide up the map of Manhattan into regions that roughly corresponded to the standard geographic description of the island.

In Part 4 on model selection, you found that using linear regression in practice can involve more than just choosing a design matrix. Tree regression made better use of categorical variables than linear regression. The domain knowledge that duration is a simple function of distance and speed allowed you to predict duration more accurately by first predicting speed.

In Part 5, you made your own model using techniques you've learned throughout the course.

Hopefully, it is apparent that all of these steps are required to reach a reliable conclusion about what inputs and model structure are helpful in predicting the duration of a taxi ride in Manhattan.

Future Work

Here are some questions to ponder:

- The regression model would have been more accurate if we had used the date itself as a feature instead of just the day of the week. Why didn't we do that?
- Does collecting this information about every taxi ride introduce a privacy risk? The original data also included the total fare; how could someone use this information combined with an individual's credit card records to determine their location?
- Why did we treat `hour` as a categorical variable instead of a quantitative variable? Would a similar treatment be beneficial for latitude and longitude?
- Why are Google Maps estimates of ride time much more accurate than our estimates?

Here are some possible extensions to the project:

- An alternative to throwing out atypical days is to condition on a feature that makes them atypical, such as the weather or holiday calendar. How would you do that?
- Training a different linear regression model for every possible combination of categorical variables can overfit. How would you select which variables to include in a decision tree instead of just using them all?
- Your models use the observed distance as an input, but the distance is only observed after the ride is over. How could you estimate the distance from the pick-up and drop-off locations?
- How would you incorporate traffic data into the model?

In [54]:

```
# Save your notebook first, then run this cell to generate a PDF.
# Note, the download link will likely not work.
# Find the pdf in the same directory as your proj3.ipynb
grader.export("proj3.ipynb", filtering=False)
```

Your file has been exported. Download it here (proj3.pdf)!

In [ ]:

```

```