Project 3: Predicting Taxi Ride Duration

Due Date: Thursday 8/13/19, 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
1a	2
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
Total	35

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]:

```
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)] 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)].

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 73.9712488)]).

Question 1a

Use a SQL query to load the `taxi` table from `taxi.db` into a Pandas DataFrame called `all_taxi`.

Only include 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)

Hint: Your solution will be shorter if you write Python code to generate the SQL query string. Try not to copy and paste code.

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

```
BEGIN QUESTION
name: q1a
points: 2
```

In [3]:

```
import sqlite3

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

# BEGIN SOLUTION
def limit(label, bounds):
    return f'({label} >= {bounds[0]} AND {label} <= {bounds[1]}) '

q = "SELECT * FROM taxi WHERE " + " AND ".join([
    limit('pickup_lon', lon_bounds),
    limit('pickup_lat', lat_bounds),
    limit('dropoff_lon', lon_bounds),
    limit('dropoff_lat', lat_bounds),
])
# END SOLUTION

all_taxi = pd.read_sql(q, conn) # SOLUTION
all_taxi.head()
```

Out[3]:

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

In [4]:

```
# TEST
all_taxi.shape
```

Out[4]:

```
(97692, 9)
```

In [5]:

```
# TEST
sum(all_taxi['duration'])
```

Out[5]:

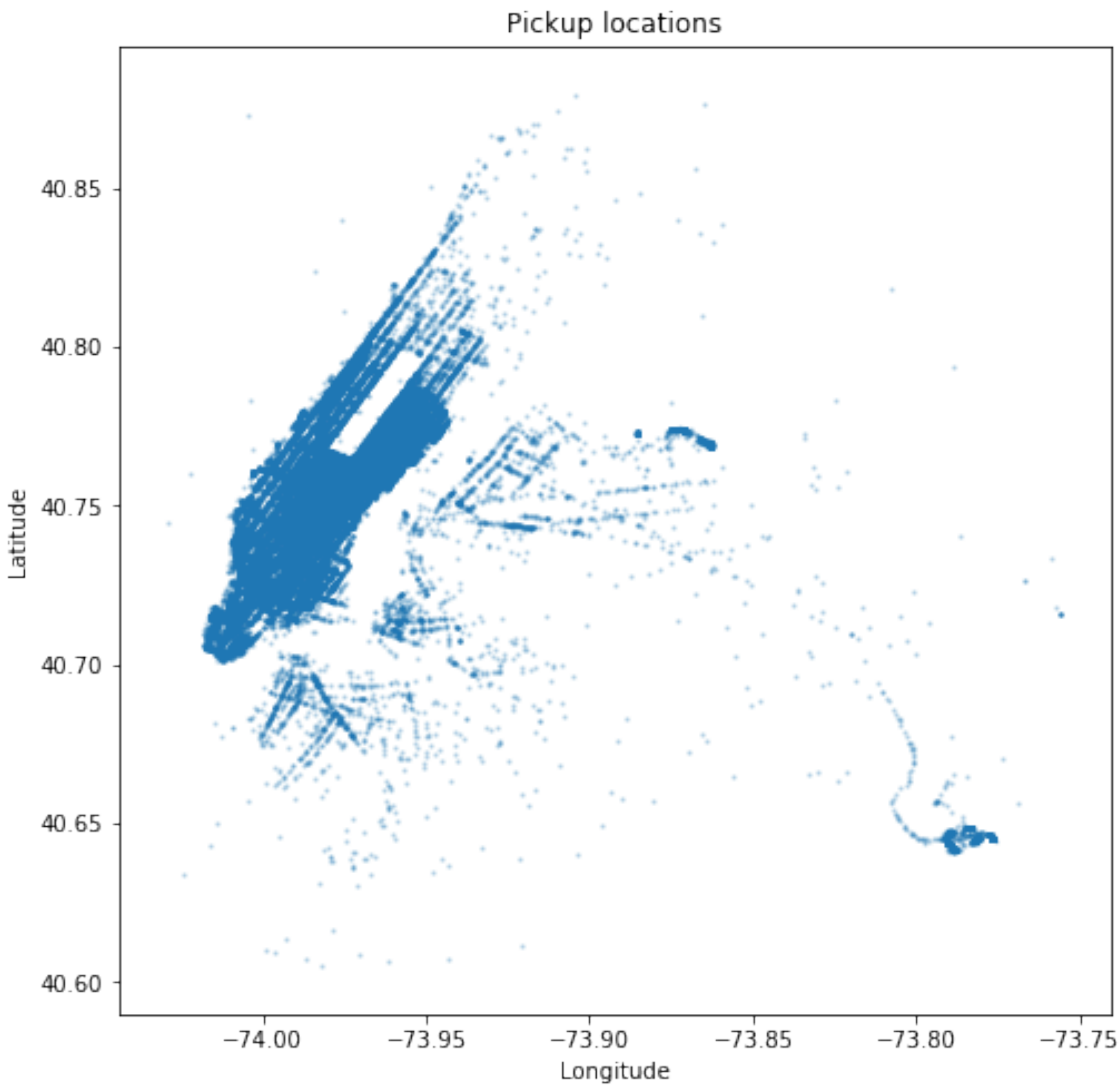
```
90553549
```

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 [6]:

```
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.

```
BEGIN QUESTION
name: q1b
points: 2
```

In [7]:

```
# BEGIN SOLUTION NO PROMPT
t = all_taxi
c = (t['passengers'] > 0)
c = c & (t['distance'] > 0)
c = c & (t['duration'] >= 60)
c = c & (t['duration'] <= 60 * 60)
c = c & (t['distance'] / t['duration'] * 60 * 60 <= 100)
# END SOLUTION
clean_taxi = t[c] # SOLUTION
```

In [8]:

```
# TEST
clean_taxi.shape
```

Out[8]:

```
(96445, 9)
```

In [9]:

```
# TEST
sum(clean_taxi['duration'])
```

Out[9]:

```
74383078
```

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
73.9712488).

The vertices of this polygon are defined in `manhattan.csv` as (latitude, longitude) pairs, which are published here (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/). 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.

```
BEGIN QUESTION
name: q1c
points: 3
```

In [10]:

```
polygon = pd.read_csv('manhattan.csv')

# Recommended: First develop and test a function that takes a position
#              and returns whether it's in Manhattan.
def in_manhattan(x, y):
    """Whether a longitude-latitude (x, y) pair is in the Manhattan polygon."""
    # BEGIN SOLUTION
    j = polygon.shape[0]-1
    odd = False
    for i in range(polygon.shape[0]):
        xsi, xsj, ysi, ysj = xs[i], xs[j], ys[i], ys[j]
        if (ysi < y <= ysj) or (ysj < y <= ysi):
            z = (y - ysi) / (ysj - ysi) * (xsj - xsi)
            if xsi + z < x:
                odd = not odd
        j = i
    return odd
    # END SOLUTION

# Recommended: Then, apply this function to every trip to filter clean_taxi.
# BEGIN SOLUTION
import os

xs, ys = polygon['lon'], polygon['lat']

def trip_in_manhattan(row):
    pickup = in_manhattan(row['pickup_lon'], row['pickup_lat'])
    dropoff = in_manhattan(row['dropoff_lon'], row['dropoff_lat'])
    return pickup and dropoff

if os.path.exists('manhattan_taxi.csv'):
    manhattan_taxi = pd.read_csv('manhattan_taxi.csv')

if 'manhattan_taxi' not in dir():
    manhattan_taxi = clean_taxi[clean_taxi.apply(trip_in_manhattan, axis=1)]
    manhattan_taxi.to_csv('manhattan_taxi.csv', index=False)
# END SOLUTION
manhattan_taxi = manhattan_taxi # SOLUTION
```

In [11]:

```
# TEST
manhattan_taxi.shape
```

Out[11]:

```
(82800, 9)
```

In [12]:

```
# TEST
sum(manhattan_taxi['duration'])
```

Out[12]:

```
54551565
```

In [13]:

```
# TEST
manhattan_taxi.iloc[0,:]['duration']
```

Out[13]:

```
981
```

In [14]:

```
# HIDDEN TEST
manhattan_taxi.iloc[0,:]['passengers']

# NOTE: The saved manhattan_taxi.csv is watermarked by setting this value to 2 i
nstead of 1.
#       Running the solution notebook will incorrectly set this value to 2, so m
ake sure the
#       test result is set to 1 when evaluating this hidden test.
#       The reason this difference exists is to make sure that students don't so
lve the
#       problem just by loading the provided data file.
```

Out[14]:

```
2
```

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.)

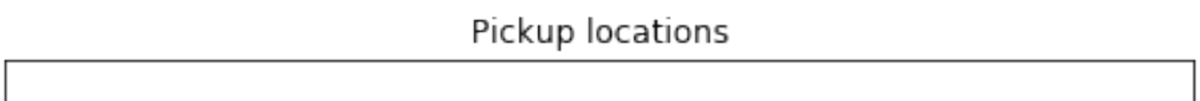
In [15]:

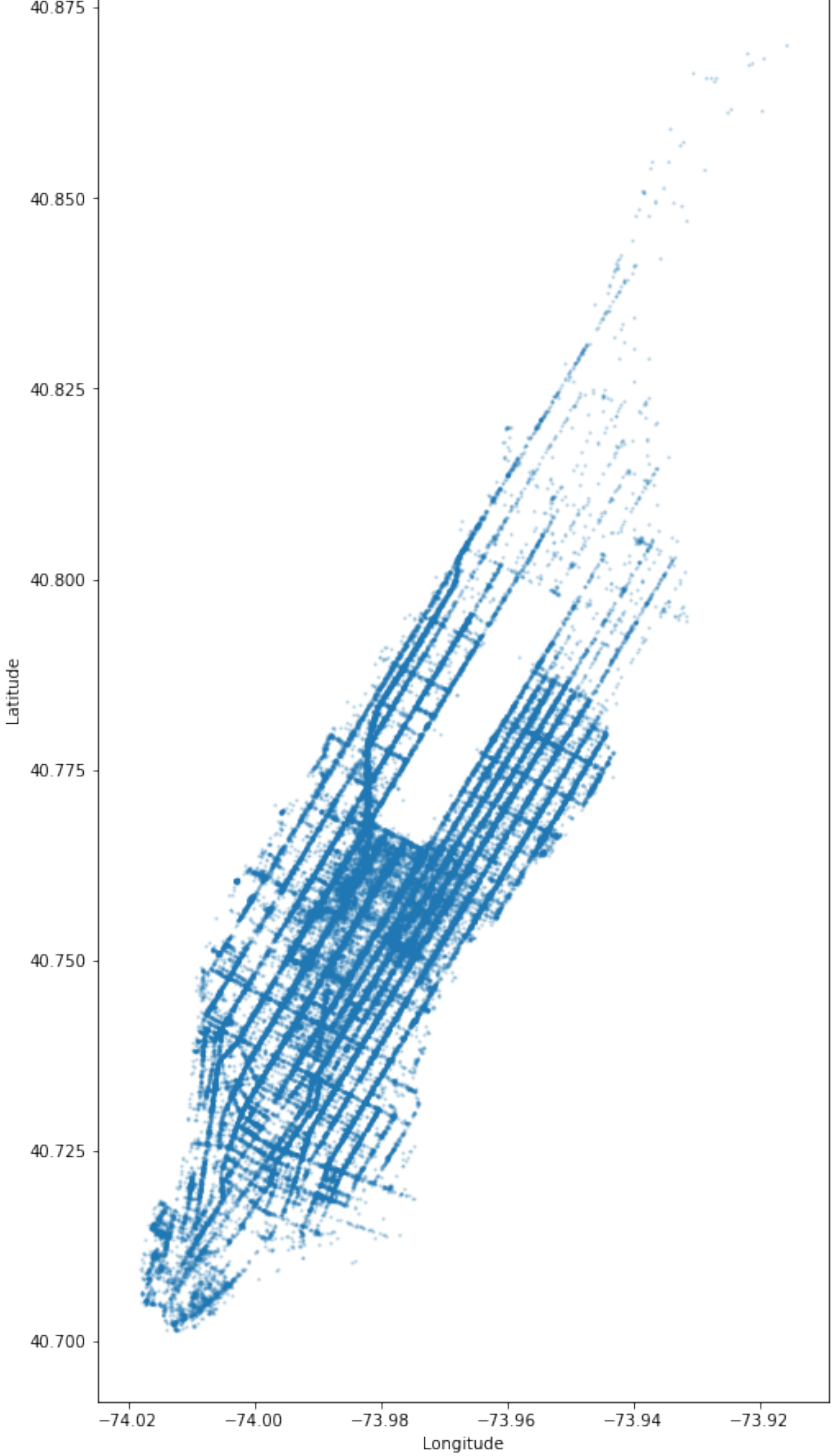
```
manhattan_taxi = pd.read_csv('manhattan_taxi.csv')
```

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

In [16]:

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




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 anomolous 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.)

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.

```
BEGIN QUESTION
name: q1d
points: 2
manual: True
```

In [17]:

```
# BEGIN SOLUTION
all_rows, clean_rows, man_rows = [t.shape[0] for t in [all_taxi, clean_taxi, man
hattan_taxi]]
cleaned = all_rows - clean_rows
cleaned_percentage = str((cleaned / all_rows) * 100)[:3]
print(f"Of the original {all_rows} trips, {cleaned} anomolous trips ({cleaned_pe
rcentage}%) "
      f"were removed through data cleaning, and then the {man_rows} trips within
Manhattan "
       "were selected for further analysis.")
# END SOLUTION
```

```
Of the original 97692 trips, 1247 anomolous trips (1.2%) were remove
d through data cleaning, and then the 82800 trips within Manhattan w
ere selected for further analysis.
```

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 Years 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) 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)).

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

```
BEGIN QUESTION
name: q2a
points: 1
```

In [18]:

```
manhattan_taxi.loc[:, 'date'] = pd.to_datetime(manhattan_taxi['pickup_datetime']
).dt.date # SOLUTION
manhattan_taxi.head()
```

Out[18]:

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

In [19]:

```
# TEST
manhattan_taxi.shape
```

Out[19]:

```
(82800, 10)
```

In [20]:

```
# TEST
list(manhattan_taxi.groupby('date').size())[:8]
```

Out[20]:

```
[2337, 2411, 2177, 2368, 2630, 2721, 2908, 3010]
```

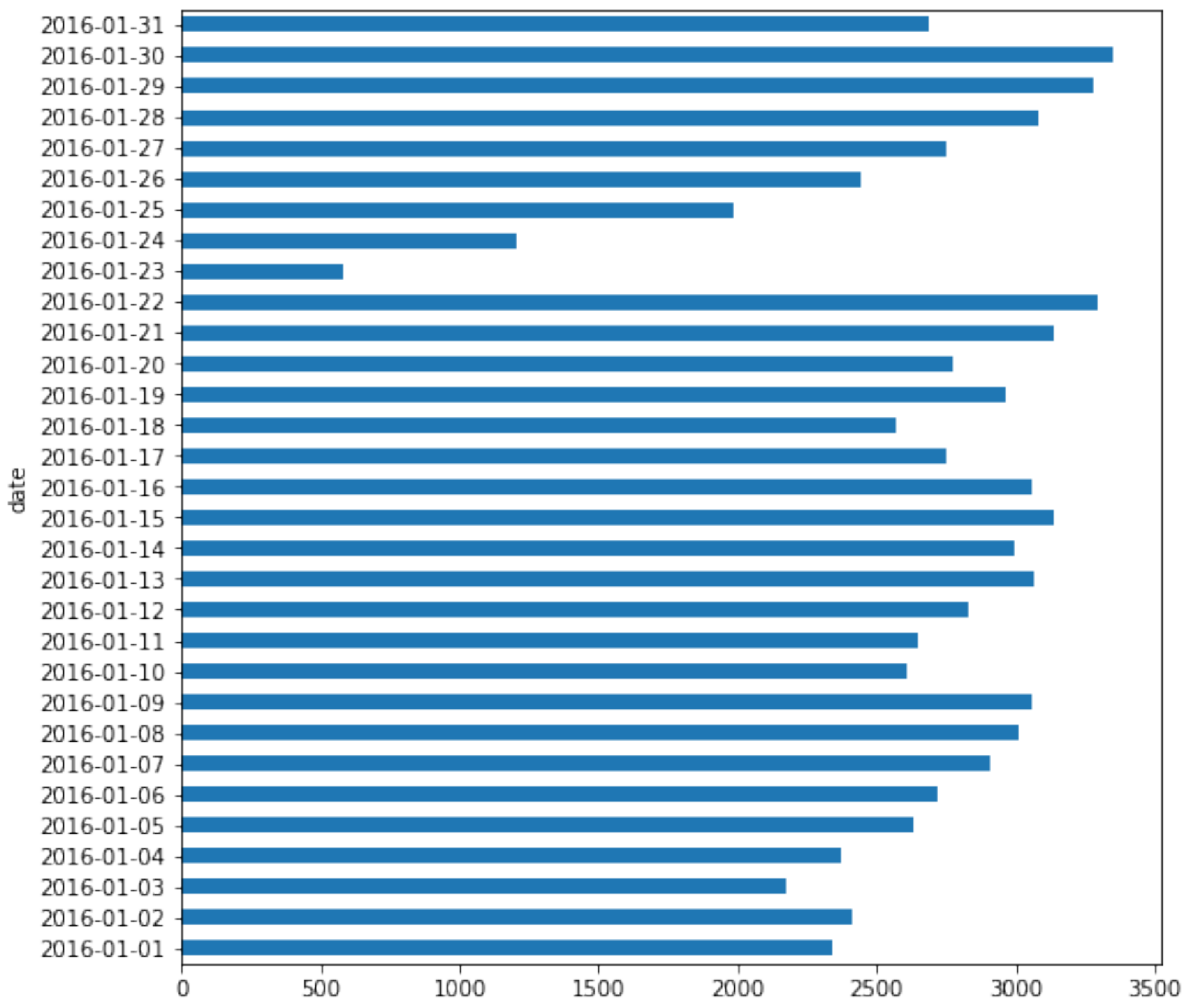
Question 2b

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.

```
BEGIN QUESTION
name: q2b
points: 2
manual: True
format: image
```

In [21]:

```
# BEGIN SOLUTION
manhattan_taxi.groupby('date').size().plot(kind='barh', figsize=(8, 8));
# END SOLUTION
```



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 [22]:

```
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 atypical]
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 [23]:

```
# Optional: More EDA here
```

Part 3: Feature Engineering

In this part, you'll create a design matrix (i.e., feature matrix) for your linear regression model. 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 [24]:

```
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)
```

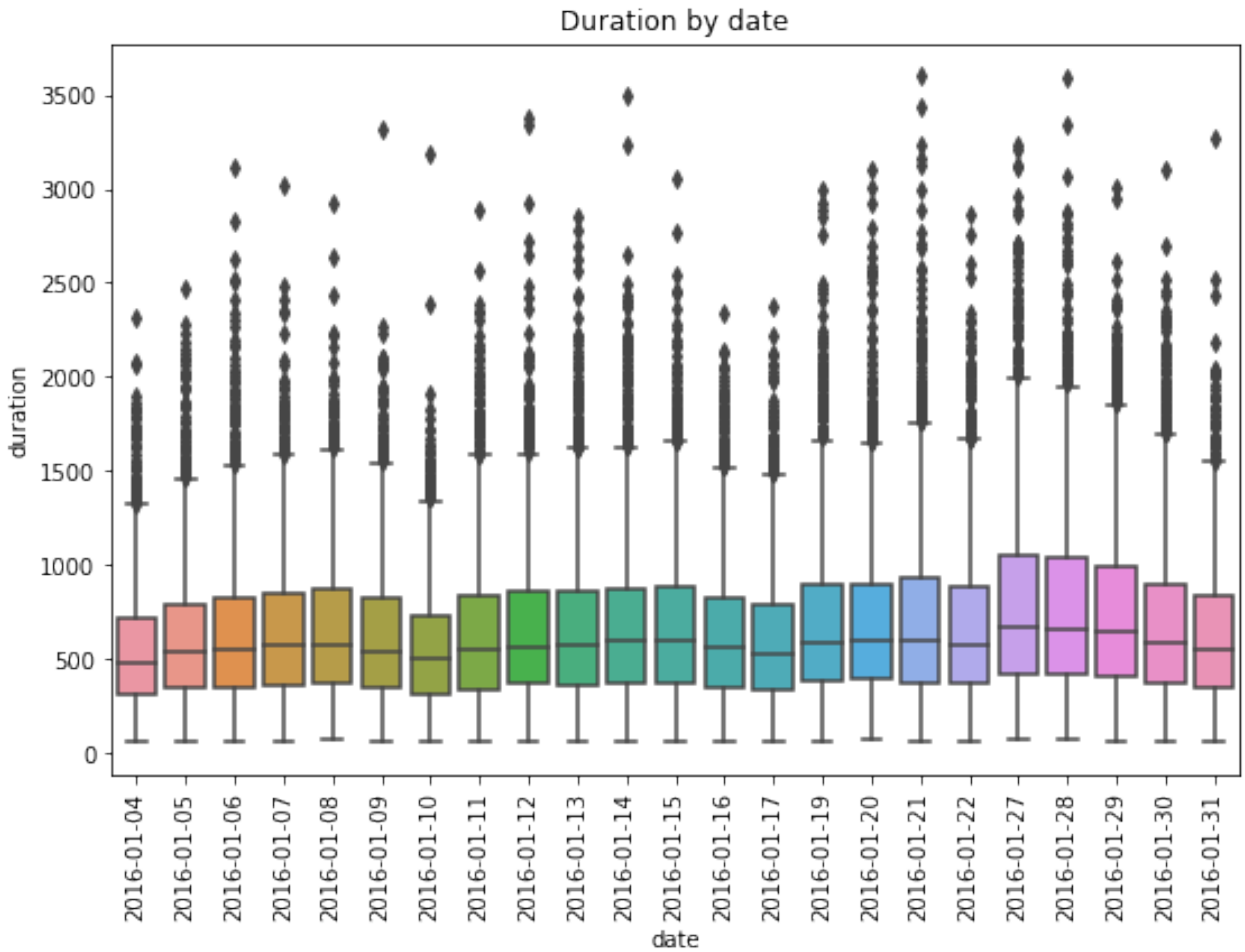
Question 3a

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 this:

```
BEGIN QUESTION
name: q3a
points: 2
manual: True
format: image
```

In [25]:

```
# BEGIN SOLUTION
data = train.sort_values('date')
plt.figure(figsize=(9, 6))
sns.boxplot('date', 'duration', data=data);
plt.xticks(rotation=90);
plt.title('Duration by date');
# END SOLUTION
```



Question 3b

In one or two sentences, describe the assocation between the day of the week and the duration of a taxi trip.

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

```
BEGIN QUESTION
name: q3b
points: 1
manual: True
```

SOLUTION:

At least in the first two weeks, the duration upper quartile increases throughout the week from Monday to Friday, dips on Saturday, dips a lot on Sunday, then increases for the next Monday. The last week doesn't follow this pattern as clearly, perhaps because of residual effects of the blizzard.

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 [26]:

```
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[26]:

```
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
```

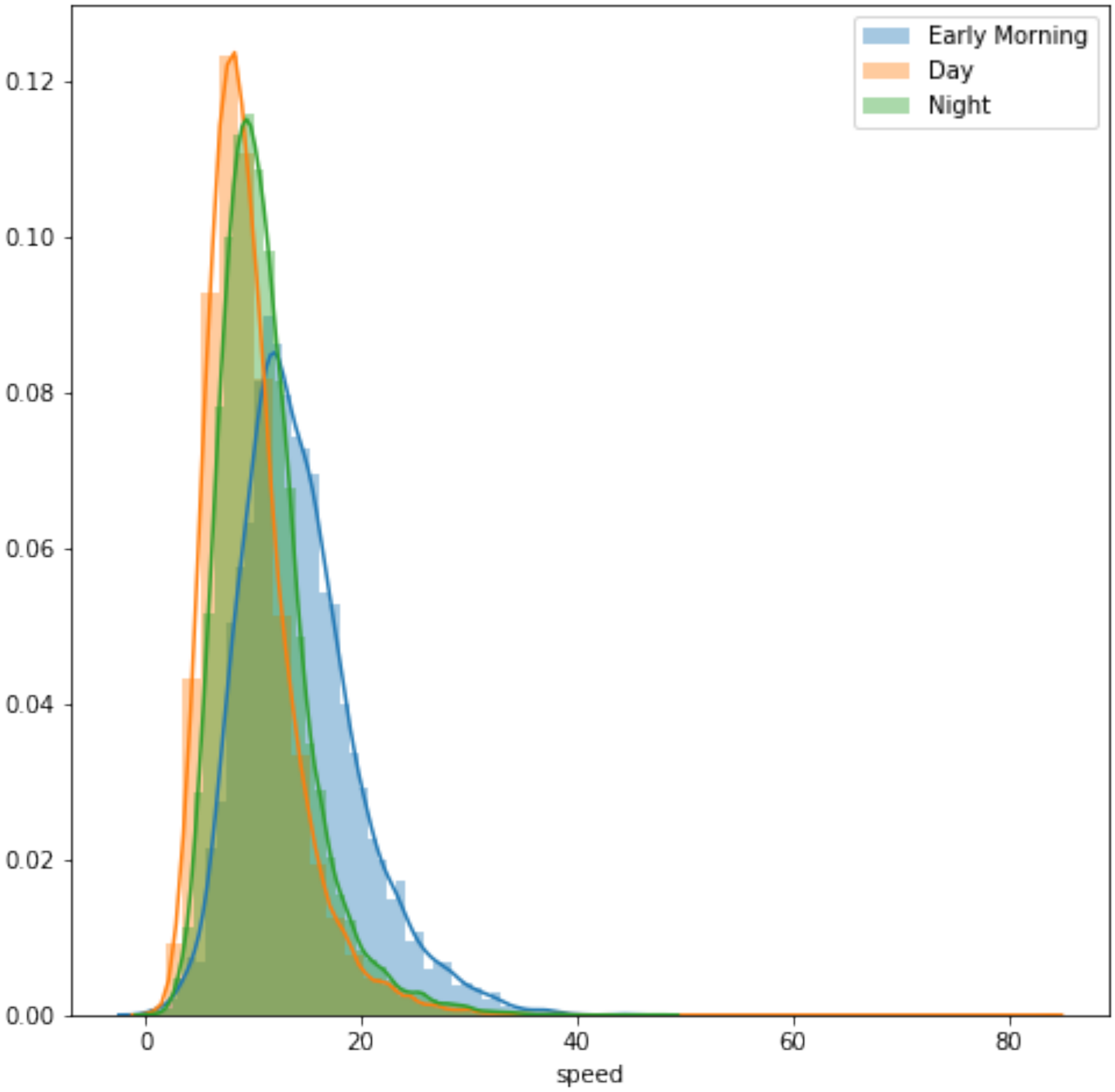
Question 3c

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:

```
BEGIN QUESTION
name: q3c
points: 2
manual: True
format: image
```

In [27]:

```
# BEGIN SOLUTION
plt.figure(figsize=(8, 8))
for i, s in enumerate(['Early Morning', 'Day', 'Night']):
    sns.distplot(train[train['period'] == i+1]['speed'], label=s)
plt.legend();
# END SOLUTION
```



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).

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.

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

```
BEGIN QUESTION
name: q3d
points: 2
```

In [28]:

```python
# Find the first principle component
D = train[['pickup_lon', 'pickup_lat']].values # SOLUTION
pca_n = D.shape[0] # SOLUTION
pca_means = np.mean(D, axis=0) # SOLUTION
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']].values # SOLUTION
    assert D.shape[0] == t.shape[0], 'You set D using the incorrect table'
    # Always use the same data transformation used to compute vt
    X = (D - pca_means) / np.sqrt(pca_n)
    first_pc = X @ vt.T[:,0] # SOLUTION
    t.loc[:,'region'] = pd.qcut(first_pc, 3, labels=[0, 1, 2])

add_region(train)
add_region(test)
```

In [29]:

```
# TEST
np.isclose(s[0], 0.02514825, 1e-3)
```

Out[29]:

```
True
```

In [30]:

```
# TEST
train.shape
```

Out[30]:

```
(53680, 16)
```

In [31]:

```
# TEST
test.shape
```

Out[31]:

```
(13421, 16)
```

In [32]:

```
# TEST
list(train['region'][:8])
```

Out[32]:

```
[1, 1, 0, 1, 2, 1, 1, 0]
```

In [33]:

```
# TEST
list(test['region'][:8])
```

Out[33]:

```
[2, 1, 2, 0, 1, 0, 1, 2]
```

In [34]:

```
# TEST
sum(train[train['region']==1]['duration'])
```

Out[34]:

```
11666210
```

In [35]:

```
# TEST
sum(test[test['region']==1]['duration'])
```

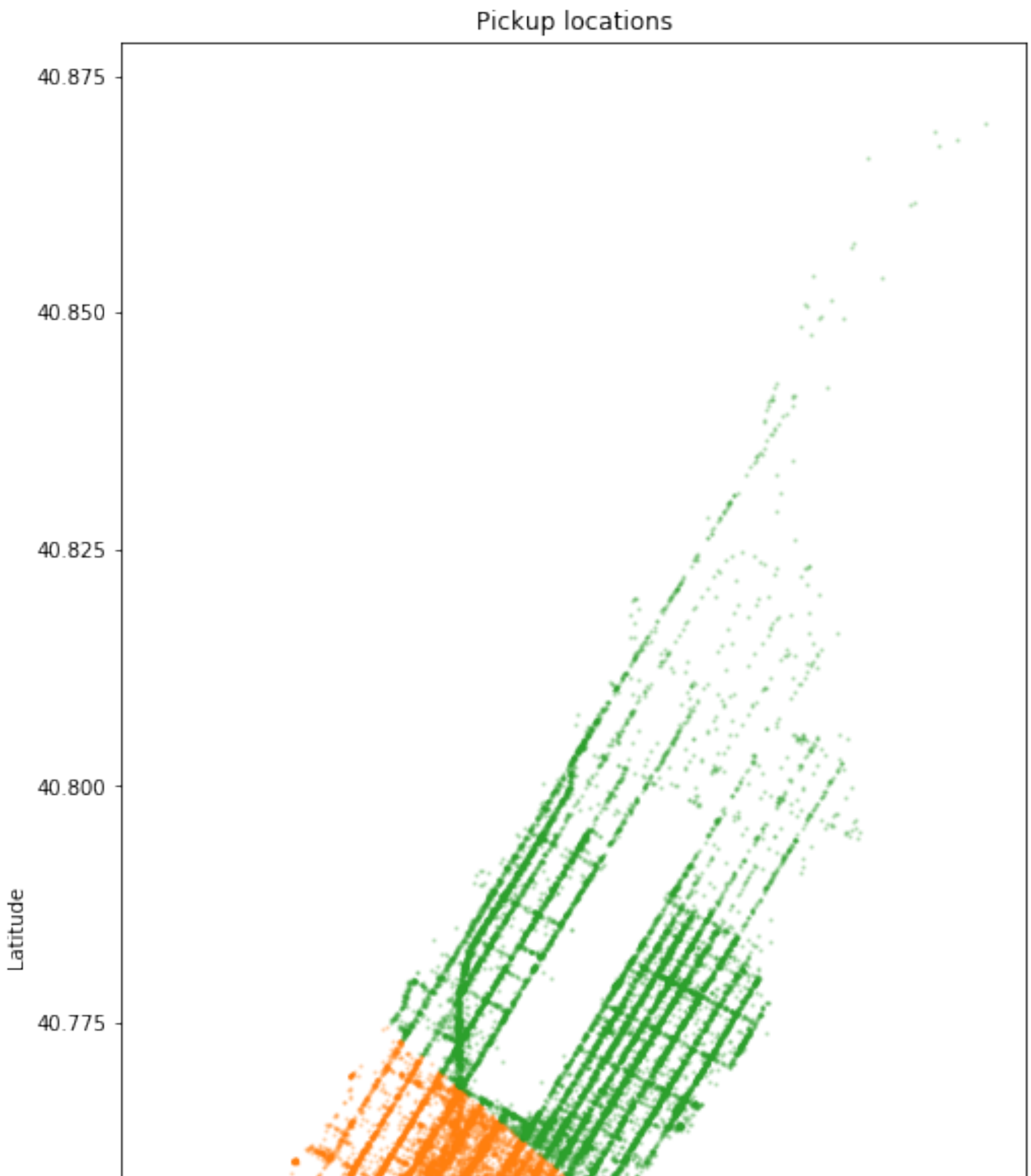
Out[35]:

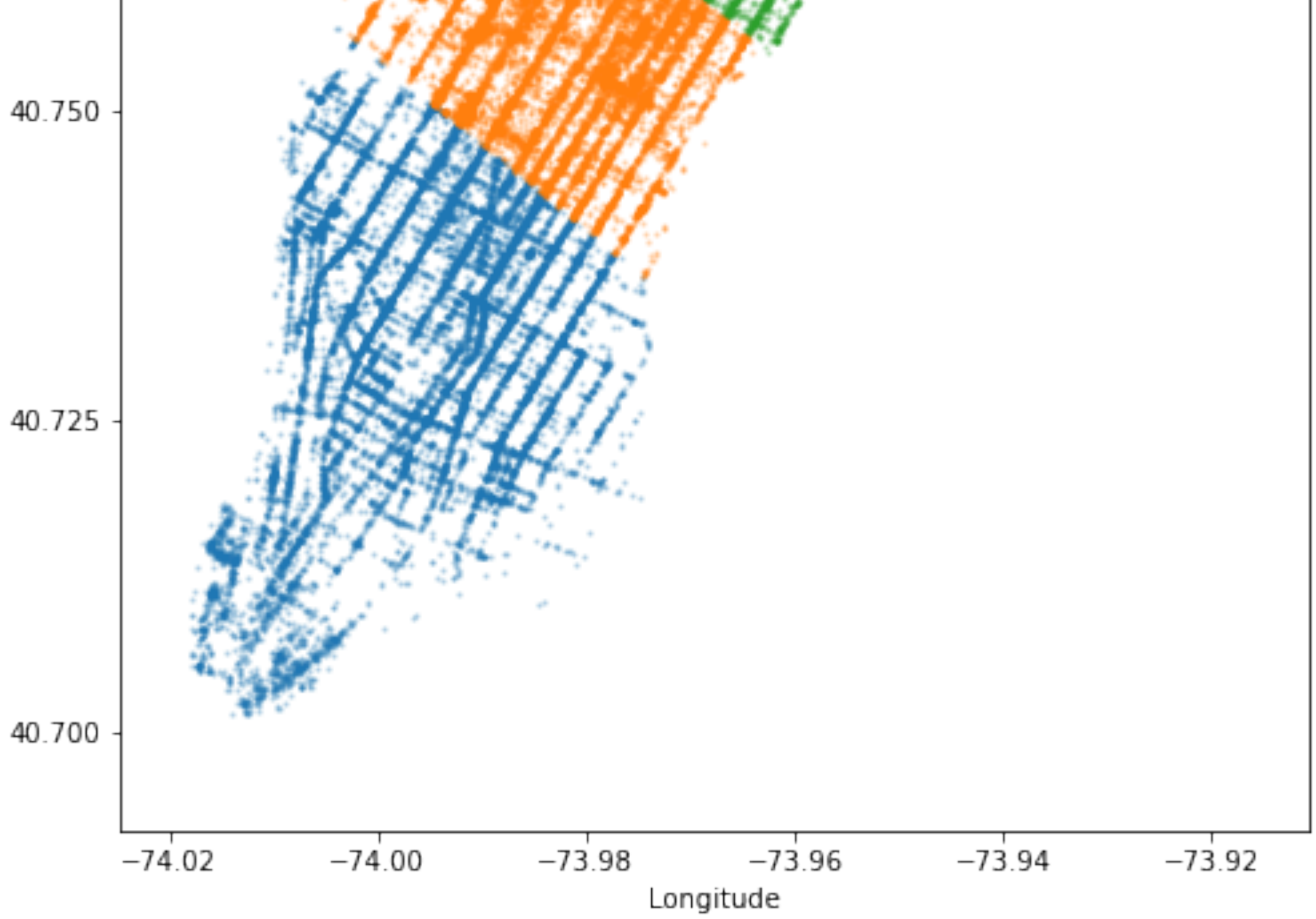
```
2897696
```

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 [36]:

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





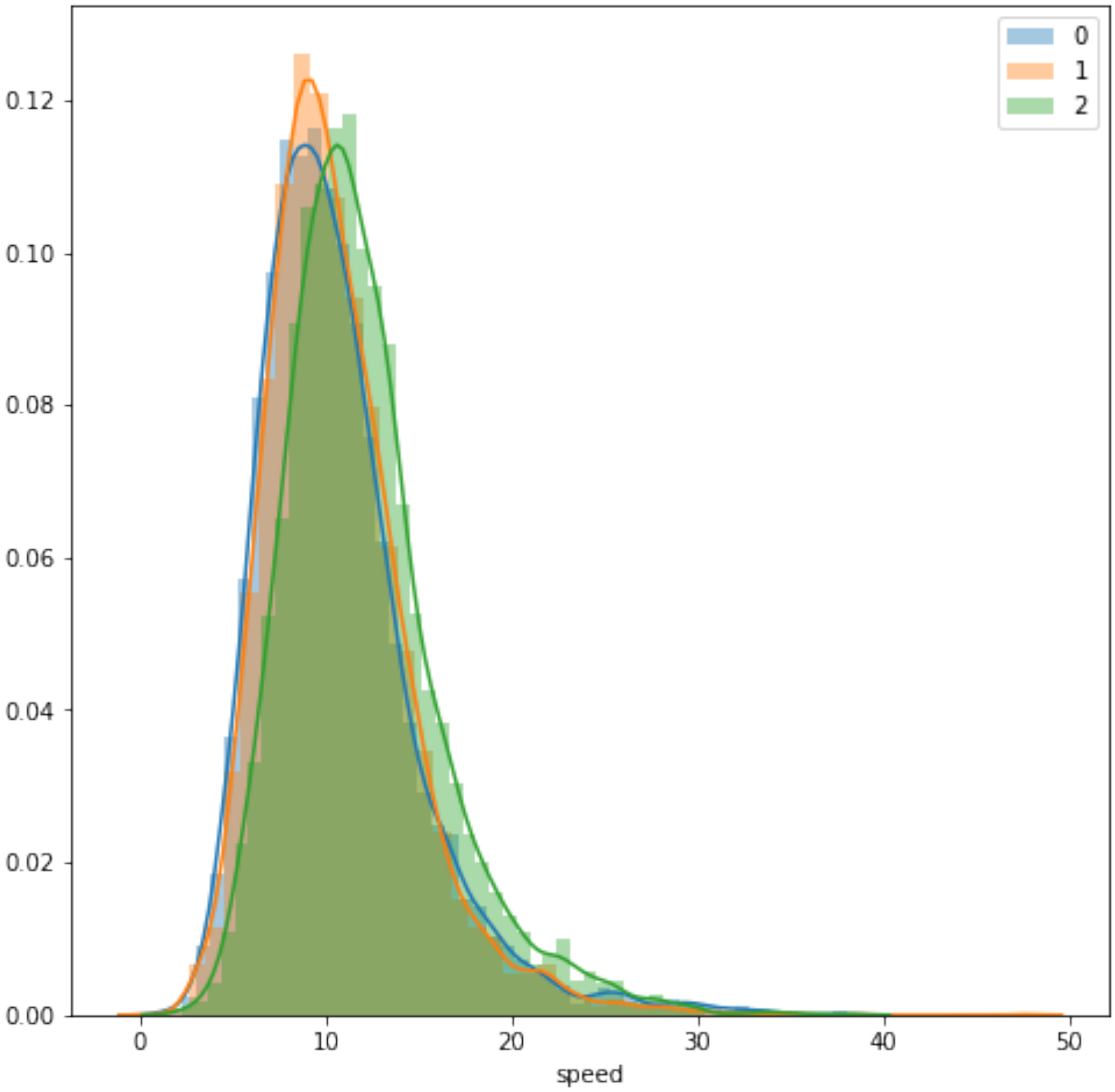
Questoin 3e (ungraded)

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 [37]:

```
# BEGIN SOLUTION
plt.figure(figsize=(8, 8))
t = train[train['period'] == 3]

for i in range(3):
    sns.distplot(t[t['region'] == i]['speed'], label=str(i))
plt.legend();
# END SOLUTION
```



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 [38]:

```
from sklearn.preprocessing import StandardScaler

num_vars = ['pickup_lon', 'pickup_lat', 'dropoff_lon', 'dropoff_lat', 'distance'
]
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 units
    categoricals = [pd.get_dummies(t[s], prefix=s, drop_first=True) for s in cat
_vars]
    return pd.concat([scaled] + categoricals, axis=1)

design_matrix(train).iloc[0,:]
```

Out[38]:

```
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.

```
BEGIN QUESTION
name: q4a
points: 2
```

In [39]:

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

constant_rmse = rmse(np.mean(train['duration']) - test['duration']) # SOLUTION
constant_rmse
```

Out[39]:

```
399.1437572352677
```

In [40]:

```
# TEST
350 <= constant_rmse <= 450
```

Out[40]:

```
True
```

In [41]:

```
# HIDDEN TEST
np.isclose(constant_rmse, 399.14376, 1e-4)
```

Out[41]:

```
True
```

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.

```
BEGIN QUESTION
name: q4b
points: 2
```

In [42]:

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
# BEGIN SOLUTION
model.fit(train[['distance']], train['duration'])
predictions = model.predict(test[['distance']])
errors = predictions - test['duration']
# END SOLUTION
simple_rmse = rmse(errors) # SOLUTION
simple_rmse
```

Out[42]:

```
276.78411050003365
```

In [43]:

```
# TEST
260 <= simple_rmse <= 300
```

Out[43]:

```
True
```

In [44]:

```
# HIDDEN TEST
np.isclose(simple_rmse, 276.78411, 1e-4)
```

Out[44]:

```
True
```

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.

```
BEGIN QUESTION
name: q4c
points: 2
```

In [45]:

```
model = LinearRegression()
# BEGIN SOLUTION
model.fit(design_matrix(train), train['duration'])
predictions = model.predict(design_matrix(test))
errors = predictions - test['duration']
# END SOLUTION
linear_rmse = rmse(errors) # SOLUTION
linear_rmse
```

Out[45]:

```
255.19146631882776
```

In [46]:

```
# TEST
list(design_matrix(test).sum())[10:15]
```

Out[46]:

```
[290.0, 511.0, 699.0, 687.0, 683.0]
```

In [47]:

```
# TEST
250 <= linear_rmse <= 260
```

Out[47]:

```
True
```

In [48]:

```
# TEST
np.isclose(linear_rmse, 255.19147, 1e-4)
```

Out[48]:

True

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.

```
BEGIN QUESTION
name: q4d
points: 2
```

In [49]:

```
model = LinearRegression()
errors = []

for v in np.unique(train['period']):
    # BEGIN SOLUTION
    v_train = train[train['period'] == v]
    v_test = test[test['period'] == v]
    model.fit(design_matrix(v_train), v_train['duration'])
    predictions = model.predict(design_matrix(v_test))
    errors.extend(predictions - v_test['duration'])
    # END SOLUTION

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

Out[49]:

246.62868831165173

In [50]:

```
# TEST
240 <= period_rmse <= 255
```

Out[50]:

True

In [51]:

```
# HIDDEN TEST
np.isclose(period_rmse, 246.628688, 1e-4)
```

Out[51]:

```
True
```

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 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.

```
BEGIN QUESTION
name: q4e
points: 2
manual: True
```

SOLUTION: Linear regression can only model linear associations. There must be an assocation between period, other covariates, and duration that is non-linear. An example is that speed (which is the slope of distance) varies by period, but a linear model with the provided design matrix cannot express this interaction.

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.

```
BEGIN QUESTION
name: q4f
points: 2
```

In [52]:

```
model = LinearRegression()
# BEGIN SOLUTION
model.fit(design_matrix(train), train['speed'])
speed_predictions = model.predict(design_matrix(test))
duration_predictions = test['distance'] / speed_predictions * 60 * 60
errors = duration_predictions - test['duration']
# END SOLUTION
speed_rmse = rmse(errors) # SOLUTION
speed_rmse
```

Out[52]:

```
243.01798368514974
```

In [53]:

```
# TEST
240 <= speed_rmse <= 255
```

Out[53]:

```
True
```

In [54]:

```
# HIDDEN TEST
np.isclose(speed_rmse, 243.01798, 1e-4)
```

Out[54]:

```
True
```

Optional: Explain why predicting speed leads to a more accurate regression model than predicting duration directly.

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.

```
BEGIN QUESTION
name: q4g
points: 4
```

In [65]:

```
for vs in train.groupby(choices).size().index:
    print(vs)
```

```
(1, 0, 0)
(1, 0, 1)
(1, 1, 0)
(1, 1, 1)
(1, 2, 0)
(1, 2, 1)
(2, 0, 0)
(2, 0, 1)
(2, 1, 0)
(2, 1, 1)
(2, 2, 0)
(2, 2, 1)
(3, 0, 0)
(3, 0, 1)
(3, 1, 0)
(3, 1, 1)
(3, 2, 0)
(3, 2, 1)
```

In [71]:

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

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

def speed_error(predictions, observations):
    """Duration error between speed predictions and duration observations"""
    # BEGIN SOLUTION
    return duration_error(observations['distance'] / predictions * 60 * 60, obse
rvations)
    # END SOLUTION

def tree_regression_errors(outcome='duration', error_fn=duration_error):
    """Return errors for all examples in test using a tree regression model."""
    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] # SOLUTION
            v_test = v_test[v_test[c] == v] # SOLUTION
            print(v_train.shape, v_test.shape)
        # BEGIN SOLUTION
        model.fit(design_matrix(v_train), v_train[outcome])
        predictions = model.predict(design_matrix(v_test))
        errors.extend(error_fn(predictions, v_test))
        print()
        # END SOLUTION
    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)
```

```
(4868, 16) (1264, 16)
(2584, 16) (665, 16)
(980, 16) (250, 16)

(4868, 16) (1264, 16)
(2584, 16) (665, 16)
(1604, 16) (415, 16)

(4868, 16) (1264, 16)
(1450, 16) (373, 16)
(792, 16) (201, 16)

(4868, 16) (1264, 16)
(1450, 16) (373, 16)
```

```
(658, 16) (172, 16)

(4868, 16) (1264, 16)
(834, 16) (226, 16)
(453, 16) (121, 16)

(4868, 16) (1264, 16)
(834, 16) (226, 16)
(381, 16) (105, 16)

(30591, 16) (7585, 16)
(8687, 16) (2165, 16)
(6508, 16) (1613, 16)

(30591, 16) (7585, 16)
(8687, 16) (2165, 16)
(2179, 16) (552, 16)

(30591, 16) (7585, 16)
(9927, 16) (2472, 16)
(7728, 16) (1942, 16)

(30591, 16) (7585, 16)
(9927, 16) (2472, 16)
(2199, 16) (530, 16)

(30591, 16) (7585, 16)
(11977, 16) (2948, 16)
(9283, 16) (2258, 16)

(30591, 16) (7585, 16)
(11977, 16) (2948, 16)
(2694, 16) (690, 16)

(18221, 16) (4572, 16)
(6623, 16) (1644, 16)
(4905, 16) (1224, 16)

(18221, 16) (4572, 16)
(6623, 16) (1644, 16)
(1718, 16) (420, 16)

(18221, 16) (4572, 16)
(6516, 16) (1628, 16)
(5135, 16) (1285, 16)

(18221, 16) (4572, 16)
(6516, 16) (1628, 16)
(1381, 16) (343, 16)

(18221, 16) (4572, 16)
(5082, 16) (1300, 16)
(3900, 16) (1006, 16)
```

```
(18221, 16) (4572, 16)
(5082, 16) (1300, 16)
(1182, 16) (294, 16)
```

In [74]:

```
test.shape
```

Out[74]:

```
(13421, 16)
```

In [73]:

```
len(errors)
```

Out[73]:

```
13421
```

In [56]:

```
# TEST
240 <= tree_rmse <= 245
```

Out[56]:

```
True
```

In [57]:

```
# TEST
225 <= tree_speed_rmse <= 240
```

Out[57]:

```
True
```

In [58]:

```
# HIDDEN TEST
np.isclose(tree_rmse, 240.33952, 1e-4)
```

Out[58]:

```
True
```

In [59]:

```
# HIDDEN TEST
np.isclose(tree_speed_rmse, 226.907939, 1e-4)
```

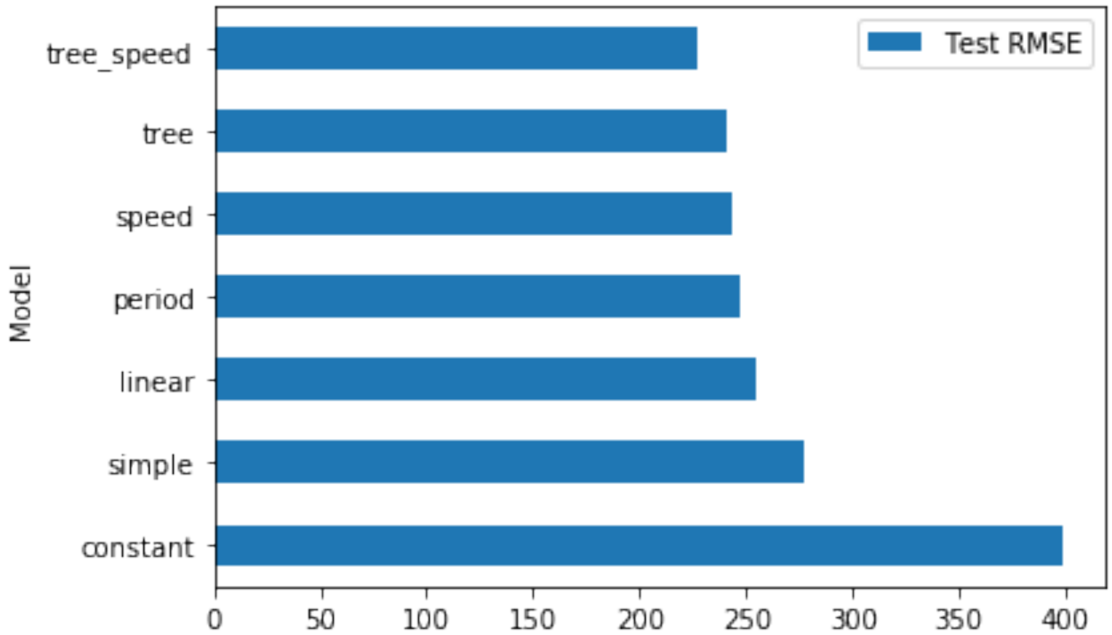
Out[59]:

True

Here's a summary of your results:

In [60]:

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



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.

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?