Before you turn in the homework, make sure everything runs as expected. To do so, select **Kernel** →**Restart & Run All** in the toolbar above. Remember to submit both on **DataHub** and **Gradescope**.

Please fill in your name and include a list of your collaborators below.

In [1]:
```
NAME = "Benjamin Liu"
COLLABORATORS = ""
```

Project 2: NYC Taxi Rides

Part 3: NYC Accidents Data

In the real world, data isn't always nicely bundled in one file; data can be sourced from many places with many formats. Now we will use NYC accident data to try to improve our set of features.

In this part of the project, you'll do some EDA over the combined data set. We'll do a lot of the coding work for you, but there will be a few coding subtasks for you to complete on your own, as well as many results to interpret.

Note

If your kernel dies unexpectedly, make sure you have shutdown all other notebooks. Each notebook uses valuable memory which we will need for this part of the project.

Imports

Let us start by loading the Python libraries and custom tools we will use in this part.

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import zipfile
import os
from pathlib import Path

sns.set(style="whitegrid", palette="muted")

plt.rcParams['figure.figsize'] = (12, 9)
plt.rcParams['font.size'] = 12

%matplotlib inline
```

Downloading the Data

We will use the `fetch_and_cache` utility to download the dataset.

In [3]:
```
# Download and cache urls and get the file objects.
from utils import fetch_and_cache
data_url = 'https://github.com/DS-100/fa18/raw/gh-pages/assets/datasets/collisio
file_name = 'collisions.zip'
dest_path = fetch_and_cache(data_url=data_url, file=file_name)

print(f'Located at {dest_path}')
```

```
Using version already downloaded: Sun Nov 25 04:22:44 2018
MD5 hash of file: a445b925d24f319cb60bd3ace6e4172b
Located at data/collisions.zip
```

We will store the taxi data locally before loading it.

In [4]:
```
collisions_zip = zipfile.ZipFile(dest_path, 'r')

#Extract zip files
collisions_dir = Path('data/collisions')
collisions_zip.extractall(collisions_dir)
```

Loading and Formatting Data

The following code loads the collisions data into a Pandas DataFrame.

In [5]:
```
# Run this cell to load the collisions data.
skiprows = None
collisions = pd.read_csv(collisions_dir/'collisions_2016.csv', index_col='UNIQUE
                         parse_dates={'DATETIME':["DATE","TIME"]}, skiprows=skip
collisions['TIME'] = pd.to_datetime(collisions['DATETIME']).dt.hour
collisions['DATE'] = pd.to_datetime(collisions['DATETIME']).dt.date
collisions = collisions.dropna(subset=['LATITUDE', 'LONGITUDE'])
collisions = collisions[collisions['LATITUDE'] <= 40.85]
collisions = collisions[collisions['LATITUDE'] >= 40.63]
collisions = collisions[collisions['LONGITUDE'] <= -73.65]
collisions = collisions[collisions['LONGITUDE'] >= -74.03]
collisions.info()
```

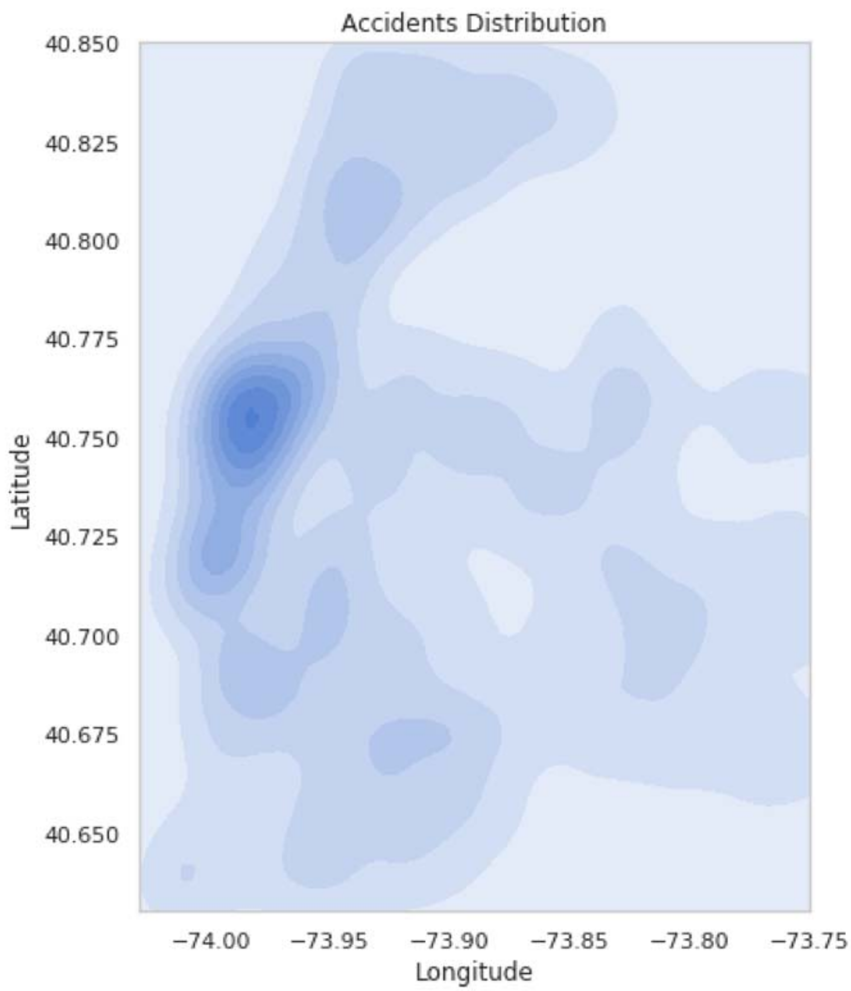
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 116691 entries, 3589202 to 3363795
Data columns (total 30 columns):
DATETIME                         116691 non-null datetime64[ns]
Unnamed: 0                       116691 non-null int64
BOROUGH                          100532 non-null object
ZIP CODE                         100513 non-null float64
LATITUDE                         116691 non-null float64
LONGITUDE                        116691 non-null float64
LOCATION                         116691 non-null object
ON STREET NAME                   95914 non-null object
CROSS STREET NAME                95757 non-null object
OFF STREET NAME                  61545 non-null object
NUMBER OF PERSONS INJURED        116691 non-null int64
NUMBER OF PERSONS KILLED         116691 non-null int64
NUMBER OF PEDESTRIANS INJURED    116691 non-null int64
NUMBER OF PEDESTRIANS KILLED     116691 non-null int64
NUMBER OF CYCLIST INJURED        116691 non-null int64
NUMBER OF CYCLIST KILLED         116691 non-null int64
NUMBER OF MOTORIST INJURED       116691 non-null int64
NUMBER OF MOTORIST KILLED        116691 non-null int64
CONTRIBUTING FACTOR VEHICLE 1    115162 non-null object
CONTRIBUTING FACTOR VEHICLE 2    101016 non-null object
CONTRIBUTING FACTOR VEHICLE 3    7772 non-null object
CONTRIBUTING FACTOR VEHICLE 4    1829 non-null object
CONTRIBUTING FACTOR VEHICLE 5    434 non-null object
VEHICLE TYPE CODE 1              115181 non-null object
VEHICLE TYPE CODE 2              92815 non-null object
VEHICLE TYPE CODE 3              7260 non-null object
VEHICLE TYPE CODE 4              1692 non-null object
VEHICLE TYPE CODE 5              403 non-null object
TIME                             116691 non-null int64
DATE                             116691 non-null object
dtypes: datetime64[ns](1), float64(3), int64(10), object(16)
memory usage: 27.6+ MB
```

1: EDA of Accidents

Let's start by plotting the latitude and longitude where accidents occur. This may give us some insight on taxi ride durations. We sample N times (given) from the collisions dataset and create a 2D KDE plot of the longitude and latitude. We make sure to set the x and y limits according to the boundaries of New York, given below.

Here is a [map of Manhattan (https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12z/da 73.9712488)](https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12z/da73.9712488) for your convenience.

In [6]:
```
# Plot lat/lon of accidents, will take a few seconds
N = 20000
city_long_border = (-74.03, -73.75)
city_lat_border = (40.63, 40.85)

sample = collisions.sample(N)
plt.figure(figsize=(6,8))
sns.kdeplot(sample["LONGITUDE"], sample["LATITUDE"], shade=True)
plt.xlim(city_long_border)
plt.ylim(city_lat_border)
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.title("Accidents Distribution")
plt.show();
```



Question 1a

What can you say about the location density of NYC collisions based on the plot above?

Hint: Here is a [page (https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12 73.9712488)](https://www.google.com/maps/place/Manhattan,+New+York,+NY/@40.7590402,-74.0394431,12) that may be useful, and [another page (https://www.6sqft.com/what-nycs-population-looks-like-day-vs-night/)](https://www.6sqft.com/what-nycs-population-looks-like-day-vs-night/) that may be useful.

In [7]:
```
q1a_answer = r"""

Since most NYC people lives in the midtown of Manhattan, traffic collisions is de

"""

# YOUR CODE HERE
# raise NotImplementedError()

print(q1a_answer)
```

```
Since most NYC people lives in the midtown of Manhattan, traffic collisions is
dense in that area. In general, it makes sense that the collosions density is h
igh in area where there are many people.
```

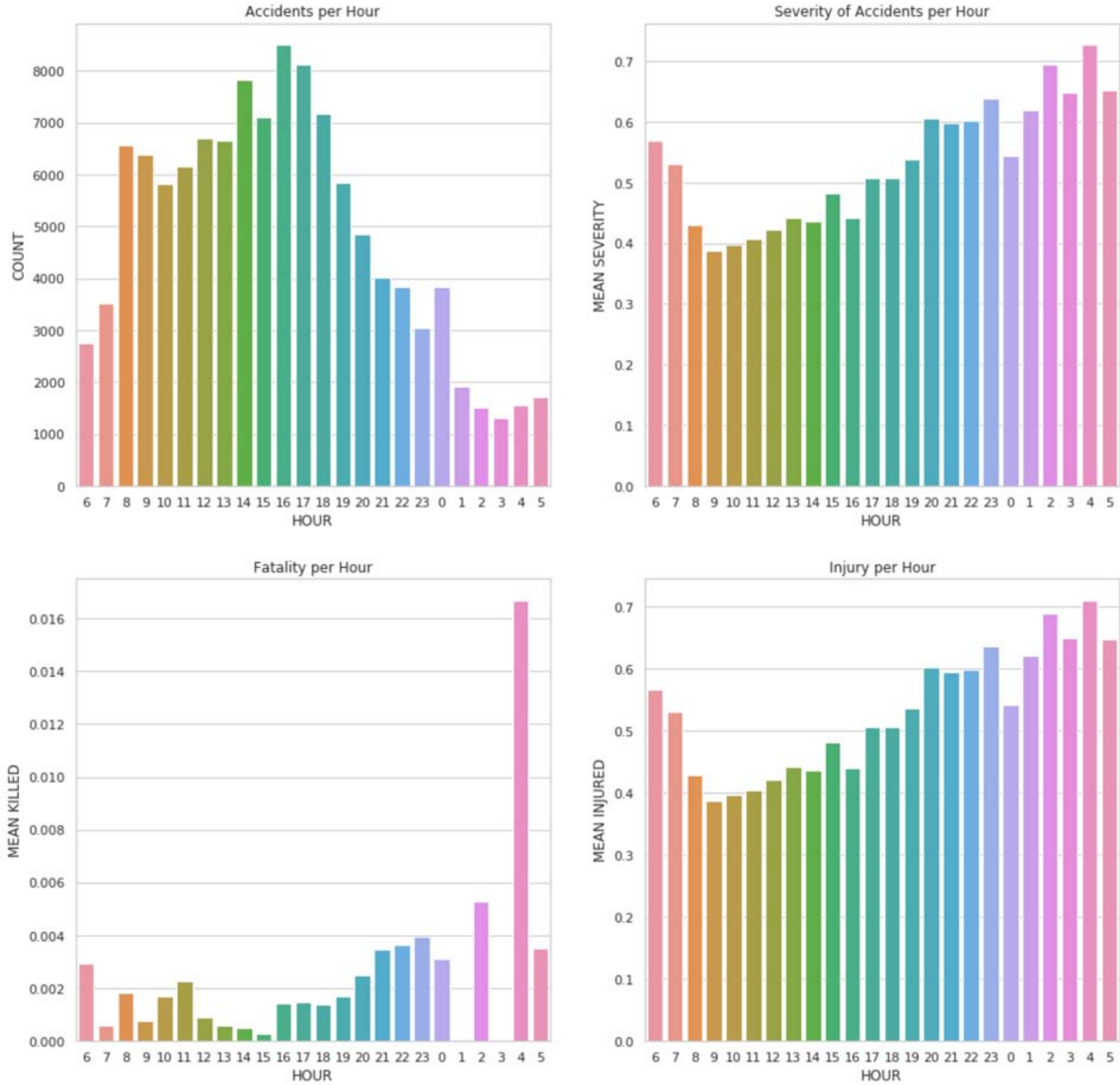
We see that an entry in accidents contains information on number of people injured/killed. Instead of using each of these columns separately, let's combine them into one column called `'SEVERITY'` . Let's also make columns `FATALITY` and `INJURY` , each aggregating the fatalities and injuries respectively.

In [8]:
```
collisions['SEVERITY'] = collisions.filter(regex=r'NUMBER OF *').sum(axis=1)
collisions['FATALITY'] = collisions.filter(regex=r'KILLED').sum(axis=1)
collisions['INJURY'] = collisions.filter(regex=r'INJURED').sum(axis=1)
```

Now let's group by time and compare two aggregations: count vs mean. Below we plot the number of collisions and the mean severity of collisions by the hour, i.e. the `TIME` column. We visualize them side by side and set the start of our day to be 6 a.m.

Let's also take a look at the mean number of casualties per hour and the mean number of injuries per hour, plotted below.

In [9]:
```
fig, axes = plt.subplots(2, 2, figsize=(16,16))
order = np.roll(np.arange(24), -6)
ax1 = axes[0,0]
ax2 = axes[0,1]
ax3 = axes[1,0]
ax4 = axes[1,1]

collisions_count = collisions.groupby('TIME').count()
collisions_count = collisions_count.reset_index()
sns.barplot(x='TIME', y='SEVERITY', data=collisions_count, order=order, ax=ax1)
ax1.set_title("Accidents per Hour")
ax1.set_xlabel("HOUR")
ax1.set_ylabel('COUNT')


collisions_mean = collisions.groupby('TIME').mean()
collisions_mean = collisions_mean.reset_index()
sns.barplot(x='TIME', y='SEVERITY', data=collisions_mean, order=order, ax=ax2)
ax2.set_title("Severity of Accidents per Hour")
ax2.set_xlabel("HOUR")
ax2.set_ylabel('MEAN SEVERITY')

fatality_count = collisions.groupby('TIME').mean()
fatality_count = fatality_count.reset_index()
sns.barplot(x='TIME', y='FATALITY', data=fatality_count, order=order, ax=ax3)
ax3.set_title("Fatality per Hour")
ax3.set_xlabel("HOUR")
ax3.set_ylabel('MEAN KILLED')

injury_count = collisions.groupby('TIME').mean()
injury_count = injury_count.reset_index()
sns.barplot(x='TIME', y='INJURY', data=injury_count, order=order, ax=ax4)
ax4.set_title("Injury per Hour")
ax4.set_xlabel("HOUR")
ax4.set_ylabel('MEAN INJURED')

plt.show();
```



Question 1b

Based on the visualizations above, what can you say about each? Make a comparison between the accidents per hour vs the mean severity per hour. What about the number of fatalities per hour vs the number of injuries per hour? Why do we chose to have our hours start at 6 as opposed to 0?

In [10]:
```
q1b_answer = r"""

Accidents per hour vs mean severity per hour: Between 8pm to 7am, there are relat

Number of fatalities per hour vs number of injuries per hour: Except for 4am, the

Besause naturally for human drivers, a "driving day" starts at 6am and ends with

"""

# YOUR CODE HERE
# raise NotImplementedError()

print(q1b_answer)
```

```
Accidents per hour vs mean severity per hour: Between 8pm to 7am, there are rel
atively fewer accidents but the mean severity is high.

Number of fatalities per hour vs number of injuries per hour: Except for 4am, t
he fatalities per hour is tiny. The number of injuries per hour is on average h
igh.

Besause naturally for human drivers, a "driving day" starts at 6am and ends wit
h 5am of next day.
```

Let's also check the relationship between location and severity. We provide code to visualize a heat map of collisions, where the x and y coordinate are the location of the collision and the heat color is the severity of the collision. Again, we sample N points to speed up visualization.

In [11]:
```
N = 10000
sample = collisions.sample(N)

# Round / bin the latitude and longitudes
sample['lat_bin'] = np.round(sample['LATITUDE'], 3)
sample['lng_bin'] = np.round(sample['LONGITUDE'], 3)

# Average severity for regions
gby_cols = ['lat_bin', 'lng_bin']

coord_stats = (sample.groupby(gby_cols)
               .agg({'SEVERITY': 'mean'})
               .reset_index())

# Visualize the average severity per region
city_long_border = (-74.03, -73.75)
city_lat_border = (40.63, 40.85)
fig, ax = plt.subplots(ncols=1, nrows=1, figsize=(14, 10))

scatter_trips = ax.scatter(sample['LONGITUDE'].values,
                           sample['LATITUDE'].values,
                           color='grey', s=1, alpha=0.5)

scatter_cmap = ax.scatter(coord_stats['lng_bin'].values,
                          coord_stats['lat_bin'].values,
                          c=coord_stats['SEVERITY'].values,
                          cmap='viridis', s=10, alpha=0.9)

cbar = fig.colorbar(scatter_cmap)
cbar.set_label("Manhattan average severity")
ax.set_xlim(city_long_border)
ax.set_ylim(city_lat_border)
ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')
plt.title('Heatmap of Manhattan average severity')
plt.axis('off');
```

Question 1c

Do you think the location of the accident has a significant impact on the severity based on the visualization above? Additionally, identify something that could be improved in the plot above and describe how we could improve it.

In [12]:
```
q1c_answer = r"""

Based on the scatter plot above, I think the location of accident doesn't have a

Many data points actually overlap each other. We can add some tiny random noise t

"""

# YOUR CODE HERE
# raise NotImplementedError()

print(q1c_answer)
```

```
Based on the scatter plot above, I think the location of accident doesn't have
a significant impact on the collisions severity.

Many data points actually overlap each other. We can add some tiny random noise
to the longitude and latitude of each data point.
```

Question 1d

Create a plot to visualize one or more features of the `collisions` table.

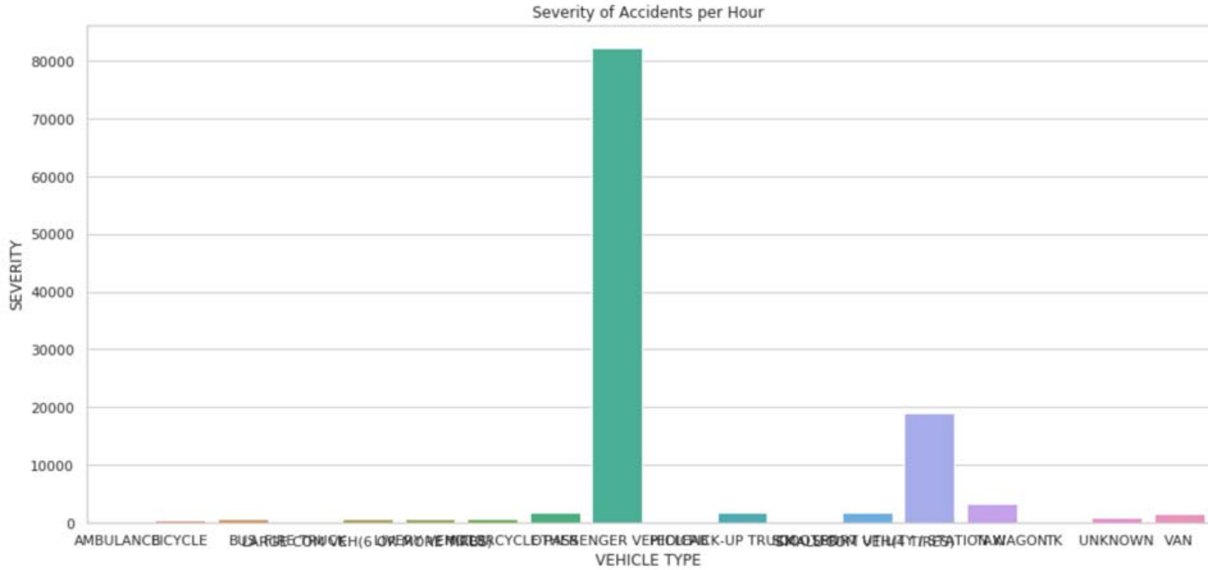
In [13]:
```
collisions.head()
```

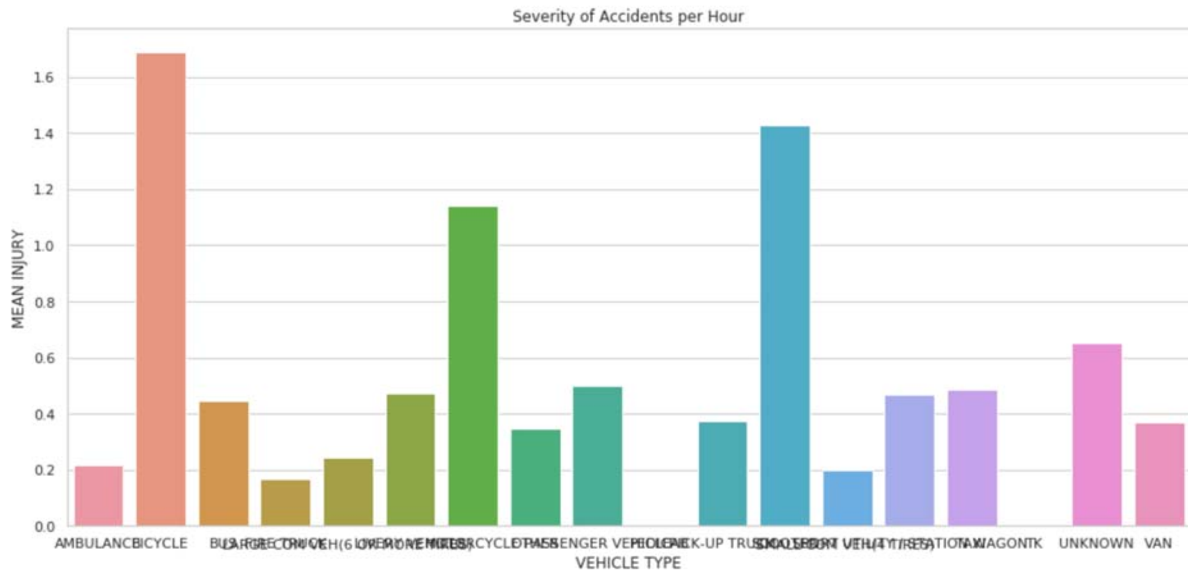
Out[13]:

	DATETIME	Unnamed: 0	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME
UNIQUE KEY								
3589202	2016-12-29 00:00:00	207836	NaN	NaN	40.844107	-73.897997	(40.8441075, -73.8979971)	NaN
3587413	2016-12-26 14:30:00	208475	NaN	NaN	40.692347	-73.881778	(40.6923473, -73.8817778)	NaN
3578151	2016-11-30 22:50:00	214339	NaN	NaN	40.755480	-73.741730	(40.75548, -73.74173)	NaN
3567096	2016-11-23 20:11:00	218291	NaN	NaN	40.771122	-73.869635	(40.7711224, -73.8696353)	NaN
3565211	2016-11-21 14:11:00	219698	NaN	NaN	40.828918	-73.838403	(40.8289179, -73.8384031)	NaN

5 rows × 33 columns

In [14]:
```
# YOUR CODE HERE
# raise NotImplementedError()
### BEGIN Solution
fig, axes = plt.subplots(2, 1, figsize=(16,16))
order = np.roll(np.arange(24), -6)
ax1 = axes[0]
ax2 = axes[1]

collisions_count = collisions.groupby('VEHICLE TYPE CODE 1').count()
collisions_count = collisions_count.reset_index()
sns.barplot(x='VEHICLE TYPE CODE 1', y='SEVERITY', data=collisions_count, ax=ax1
ax1.set_title("Severity of Accidents per Hour")
ax1.set_xlabel("VEHICLE TYPE")
ax1.set_ylabel('SEVERITY')


collisions_mean = collisions.groupby('VEHICLE TYPE CODE 1').mean()
collisions_mean = collisions_mean.reset_index()
sns.barplot(x='VEHICLE TYPE CODE 1', y='INJURY', data=collisions_mean, ax=ax2)
ax2.set_title("Severity of Accidents per Hour")
ax2.set_xlabel("VEHICLE TYPE")
ax2.set_ylabel('MEAN INJURY');

### END Solution
```





Question 1e

Answer the following questions regarding your plot in 1d.

1. What feature you're visualization
2. Why you chose this feature
3. Why you chose this visualization method

In [15]:
```
q1e_answer = r"""

I am visualizing the number of accidents catagorized by the involved vehicle type

I want to invistigate the relationship between vehicle type and accidents severit

Since the type of vehicles is catagorical data, it is good to use bar plot to vis

"""
# YOUR CODE HERE
# raise NotImplementedError()
print(q1e_answer)
```

```
I am visualizing the number of accidents catagorized by the involved vehicle ty
pe.

I want to invistigate the relationship between vehicle type and accidents sever
ity, injuries and hence I pick the vehicle type.

Since the type of vehicles is catagorical data, it is good to use bar plot to v
isualize multiple categorical data.
```

2: Combining External Datasets

It seems like accident timing and location may influence the duration of a taxi ride. Let's start to join our NYC Taxi data with our collisions data.

Let's assume that an accident will influence traffic in the surrounding area for around 1 hour. Below, we create two columns, `START` and `END` :

- `START` : contains the recorded time of the accident
- `END` : 1 hours after `START`

Note: We chose 1 hour somewhat arbitrarily, feel free to experiment with other time intervals outside this notebook.

In [16]:
```
collisions['START'] = collisions['DATETIME']
collisions['END'] = collisions['START'] + pd.Timedelta(hours=1)
```

Question 2a

Drop all of the columns besides the following: `DATETIME` , `TIME` , `START` , `END` , `DATE` , `LATITUDE` , `LONGITUDE` , `SEVERITY` . Feel free to experiment with other subsets outside of this notebook.

In [17]:
```
collisions_subset = collisions[["DATETIME", "TIME", "START", "END", "DATE", "LATI
# YOUR CODE HERE
# raise NotImplementedError()
collisions_subset.head(5)
```

Out[17]:

UNIQUE KEY	DATETIME	TIME	START	END	DATE	LATITUDE	LONGITUDE	SEVERITY
3589202	2016-12-29 00:00:00	0	2016-12-29 00:00:00	2016-12-29 01:00:00	2016-12-29	40.844107	-73.897997	0
3587413	2016-12-26 14:30:00	14	2016-12-26 14:30:00	2016-12-26 15:30:00	2016-12-26	40.692347	-73.881778	0
3578151	2016-11-30 22:50:00	22	2016-11-30 22:50:00	2016-11-30 23:50:00	2016-11-30	40.755480	-73.741730	2
3567096	2016-11-23 20:11:00	20	2016-11-23 20:11:00	2016-11-23 21:11:00	2016-11-23	40.771122	-73.869635	0
3565211	2016-11-21 14:11:00	14	2016-11-21 14:11:00	2016-11-21 15:11:00	2016-11-21	40.828918	-73.838403	0

In [18]:
```
assert collisions_subset.shape == (116691, 8)
```

Question 2b

Now, let's merge our `collisions_subset` table with `train_df`. Start by merging with only the date. We will filter by a time window in a later question.

We should be performing a left join, where our `train_df` is the left table. This is because we want to preserve all of the taxi rides in our end result. It happens that an inner join will also work, since both tables contain data on each date.

Note that the resulting `merged` table will have multiple rows for every taxi ride row in the original `train_df` table. For example, `merged` will have 483 rows with `index` equal to 16709, because there were 483 accidents that occurred on the same date as ride #16709.

Because of memory limitation, we will select the third week of 2016 to analyze. Feel free to change to it week 1 or 2 to see if the observation is general.

In [19]:
```
data_file = Path("./", "cleaned_data.hdf")
train_df = pd.read_hdf(data_file, "train")
train_df = train_df.reset_index()
train_df = train_df[['index', 'tpep_pickup_datetime', 'pickup_longitude', 'picku
train_df['date'] = train_df['tpep_pickup_datetime'].dt.date
```

In [20]:
```
collisions_subset = collisions_subset[collisions_subset['DATETIME'].dt.weekofyea
train_df = train_df[train_df['tpep_pickup_datetime'].dt.weekofyear == 3]
```

In [21]:
```
# merge the dataframe here
merged = pd.merge(train_df, collisions_subset, how="left", left_on="date", right_

# YOUR CODE HERE
# raise NotImplementedError()

merged.head()
```

Out[21]:

	index	tpep_pickup_datetime	pickup_longitude	pickup_latitude	duration	date	DATETIME	TIM
0	16709	2016-01-21 22:28:17	-73.997986	40.741215	736.0	2016-01-21	2016-01-21 10:35:00	1
1	16709	2016-01-21 22:28:17	-73.997986	40.741215	736.0	2016-01-21	2016-01-21 13:20:00	1
2	16709	2016-01-21 22:28:17	-73.997986	40.741215	736.0	2016-01-21	2016-01-21 16:00:00	1
3	16709	2016-01-21 22:28:17	-73.997986	40.741215	736.0	2016-01-21	2016-01-21 18:30:00	1
4	16709	2016-01-21 22:28:17	-73.997986	40.741215	736.0	2016-01-21	2016-01-21 00:05:00	

In [22]:
```
assert merged.shape == (1528162, 14)
```

Question 2c

Now that our tables are merged, let's use temporal and spatial proximity to condition on the duration of the average length of a taxi ride. Let's operate under the following assumptions.

Accidents only influence the duration of a taxi ride if the following are satisfied:

1) The haversine distance between the the pickup location of the taxi ride and location of the recorded accident is within 5 (km). This is roughly 3.1 miles.

2) The start time of a taxi ride is within a 1 hour interval between the start and end of an accident.

Complete the code below to create an `'accident_close'` column in the `merged` table that indicates if an accident was close or not according to the assumptions above.

In [23]:
```
def haversine(lat1, lng1, lat2, lng2):
    """
    Compute haversine distance
    """
    lat1, lng1, lat2, lng2 = map(np.radians, (lat1, lng1, lat2, lng2))
    average_earth_radius = 6371
    lat = lat2 - lat1
    lng = lng2 - lng1
    d = np.sin(lat * 0.5) ** 2 + np.cos(lat1) * np.cos(lat2) * np.sin(lng * 0.5)
    h = 2 * average_earth_radius * np.arcsin(np.sqrt(d))
    return h

def manhattan_distance(lat1, lng1, lat2, lng2):
    """
    Compute Manhattan distance
    """
    a = haversine(lat1, lng1, lat1, lng2)
    b = haversine(lat1, lng1, lat2, lng1)
    return a + b
```

In [24]:
```
start_to_accident = haversine(merged['pickup_latitude'].values,
                              merged['pickup_longitude'].values,
                              merged['LATITUDE'].values,
                              merged['LONGITUDE'].values)
merged['start_to_accident'] = start_to_accident

# initialze accident_close column to all 0 first
merged['accident_close'] = 0

# Boolean pd.Series to select the indices for which accident_close should equal 1
# (1) record's start_to_accident <= 5
# (2) pick up time is between start and end
is_accident_close = np.array(merged['start_to_accident'] <= 5) & (np.array(merge
# YOUR CODE HERE
# raise NotImplementedError()

merged.loc[is_accident_close, 'accident_close'] = 1
```

In [25]:
```
assert merged['accident_close'].sum() > 16000
```

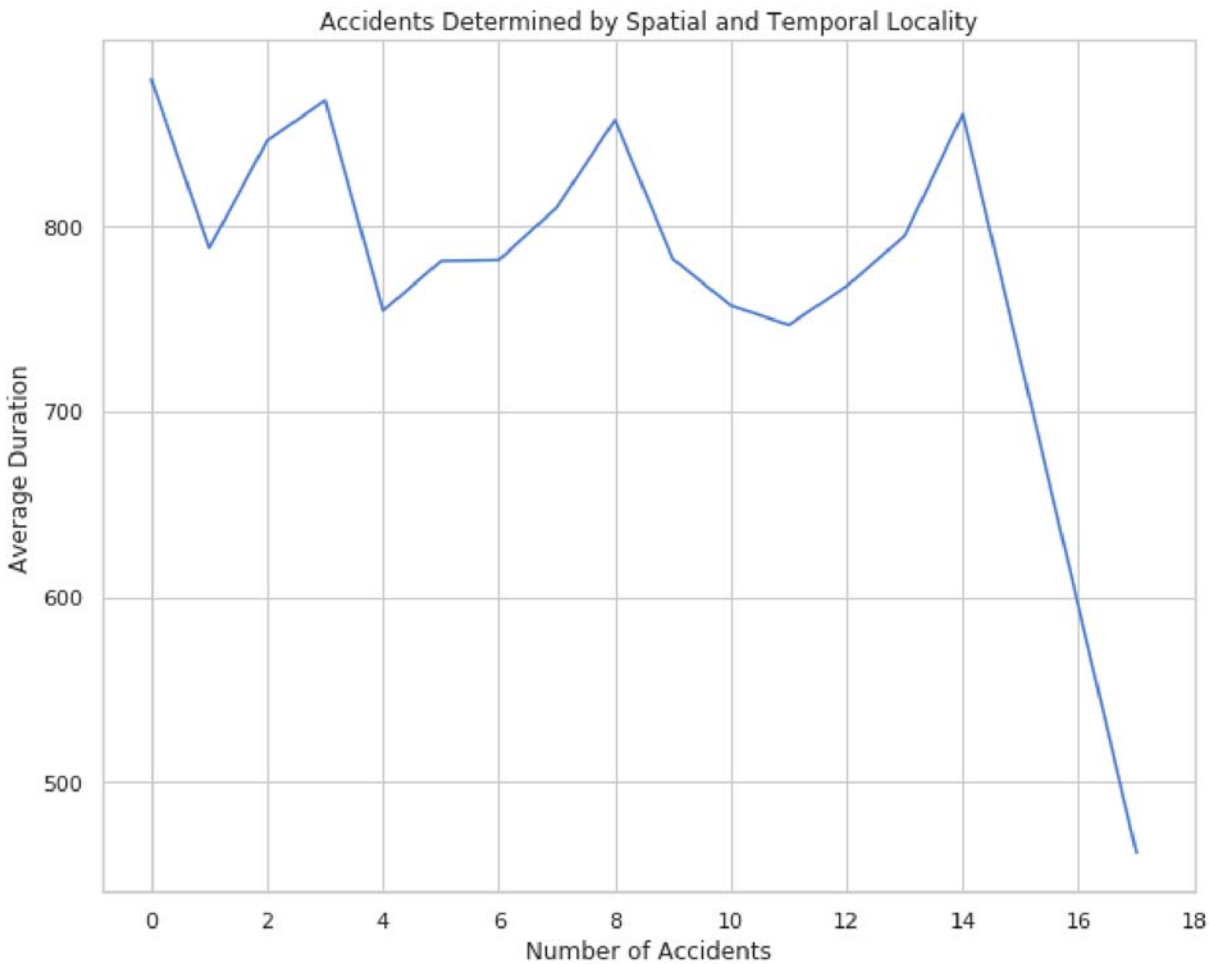
The last step is to aggregate the total number of proximal accidents. We want to count the total number of accidents that were close spatially and temporally and condition on that data.

The code below create a new data frame called `train_accidents` , which is a copy of `train_df` , but with a new column that counts the number of accidents that were close (spatially and temporally) to the pickup location/time.

In [26]:
```
train_df = train_df.set_index('index')
num_accidents = merged.groupby(['index'])['accident_close'].sum().to_frame()
train_accidents = train_df.copy()
train_accidents['num_accidents'] = num_accidents
```

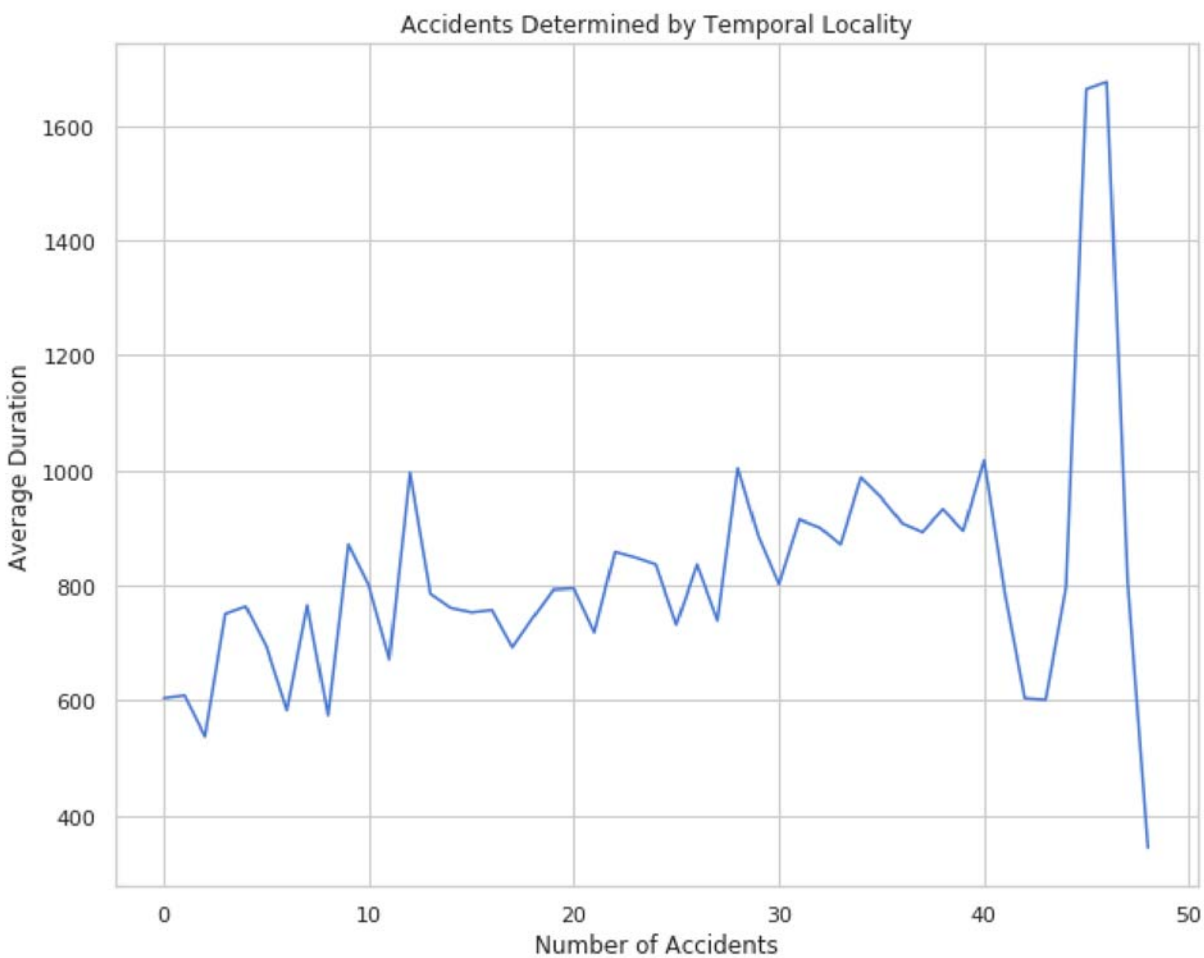
Next, for each value of `num_accidents` , we plot the average `duration` of rides with that number of accidents.

In [27]:
```
plt.figure(figsize=(10,8))
train_accidents.groupby('num_accidents')['duration'].mean().plot(xticks=np.arang
plt.title("Accidents Determined by Spatial and Temporal Locality")
plt.xlabel("Number of Accidents")
plt.ylabel("Average Duration")
plt.show();
```

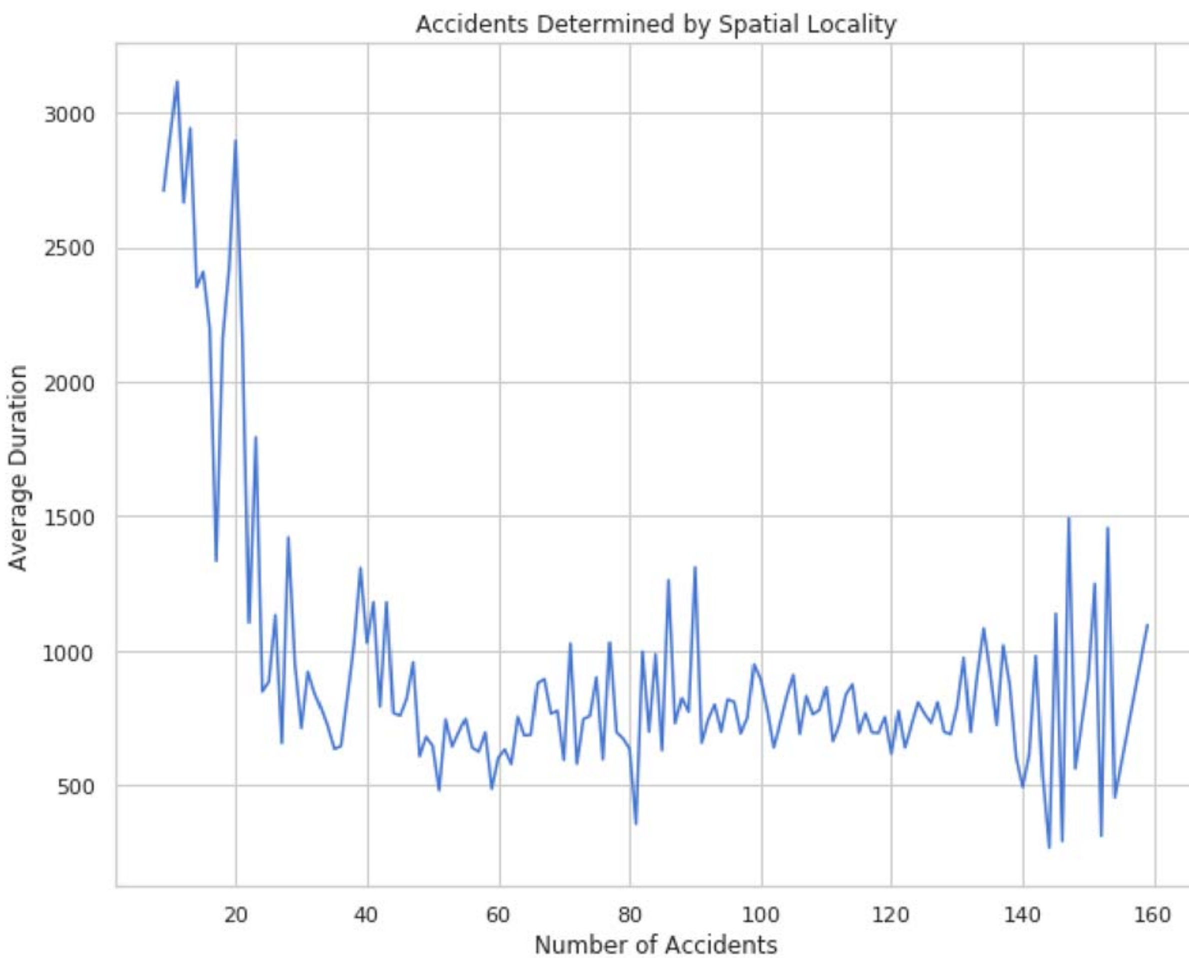


It seems that using both spatial and temporal proximity doesn't give us much insight on if collisions increase taxi ride durations. Let's try conditioning on spatial proximity and temporal proximity separately and see if there are more interesting results there.

In [28]:
```
# Temporal locality

# Condition on time
index = (((merged['tpep_pickup_datetime'] >= merged['START']) & \
         (merged['tpep_pickup_datetime'] <= merged['END'])))

# Count accidents
merged['accident_close'] = 0
merged.loc[index, 'accident_close'] = 1
num_accidents = merged.groupby(['index'])['accident_close'].sum().to_frame()
train_accidents_temporal = train_df.copy()
train_accidents_temporal['num_accidents'] = num_accidents

# Plot
plt.figure(figsize=(10,8))
train_accidents_temporal.groupby('num_accidents')['duration'].mean().plot()
plt.title("Accidents Determined by Temporal Locality")
plt.xlabel("Number of Accidents")
plt.ylabel("Average Duration")
plt.show();
```



In [29]:
```
# Spatial locality

# Condition on space
index = (merged['start_to_accident'] <= 5)

# Count accidents
merged['accident_close'] = 0
merged.loc[index, 'accident_close'] = 1
num_accidents = merged.groupby(['index'])['accident_close'].sum().to_frame()
train_accidents_spatial = train_df.copy()
train_accidents_spatial['num_accidents'] = num_accidents

# Plot
plt.figure(figsize=(10,8))
train_accidents_spatial.groupby('num_accidents')['duration'].mean().plot()
plt.title("Accidents Determined by Spatial Locality")
plt.xlabel("Number of Accidents")
plt.ylabel("Average Duration")
plt.show();
```



Question 2d

By conditioning on temporal and spatial proximity separately, we reveal different trends in average ride duration as a function of number of accidents nearby.

What can you say about the temporal and spatial proximity of accidents to taxi rides and the effect on ride duration? Think of a new hypothesis regarding accidents and taxi ride durations and explain how you would test it.

Additionally, comment on some of the assumptions being made when we condition on temporal and spatial proximity separately. What are the implications of only considering one and not the other?

In [30]:
```
q2d_answer = r"""

If the accidents are temporally closed, in a certain range, the more accidents, t
If the accidents are spatially closed, the more accidents, the shorter the ride d

Hypothesis:

When we conditioned on temporal proximity, we assume the other factors are the sa
When we conditioned on spatial proximity, we assume the other factors are the sam

"""

# YOUR CODE HERE
# raise NotImplementedError()

print(q2d_answer)
```

```
If the accidents are temporally closed, in a certain range, the more accidents,
the longger the mean ride duration.
If the accidents are spatially closed, the more accidents, the shorter the ride
duration on average.

Hypothesis:

When we conditioned on temporal proximity, we assume the other factors are the
same.
When we conditioned on spatial proximity, we assume the other factors are the s
ame. We are also assuming independency of the temporal and spatial proximity.
```

Part 3 Exports

We are not requiring you to export anything from this notebook, but you may find it useful to do so. There is a space below for you to export anything you wish.

In [31]:
```
Path("data/part3").mkdir(parents=True, exist_ok=True)
data_file = Path("data/part3", "data_part3.hdf") # Path of hdf file
...
```

Out[31]:
```
Ellipsis
```

Part 3 Conclusions

We merged the NYC Accidents dataset with our NYC Taxi dataset, conditioning on temporal and spatial locality. We explored potential features by visualizing the relationship between number of accidents and the average duration of a ride.

Please proceed to part 4 where we will be engineering more features and building our models using a processing pipeline.

Submission

You're almost done!

Before submitting this assignment, ensure that you have:

1. Restarted the Kernel (in the menubar, select Kernel→Restart & Run All)
2. Validated the notebook by clicking the "Validate" button.

Then,

1. **Submit** the assignment via the Assignments tab in **Datahub**
2. **Upload and tag** the manually reviewed portions of the assignment on **Gradescope**

In [ ]:

```
```

In [ ]:

```
```

Before you turn in the homework, make sure everything runs as expected. To do so, select **Kernel →Restart & Run All** in the toolbar above. Remember to submit both on **DataHub** and **Gradescope**.

Please fill in your name and include a list of your collaborators below.

In [1]:
```
NAME = "Benjamin Liu"
COLLABORATORS = "Victor Ding"
```

Project 2: NYC Taxi Rides

Part 4: Feature Engineering and Model Fitting

In this final part of the project, you will finally build a regression model that attempts to predict the duration of a taxi ride from all other available information.

You will build this model using a processing pipeline and submit your results to Kaggle. We will first walk you through a generic example using the data we saved from Part 1. Please carefully follow these steps as you will need to repeat this for your final model. After, we give you free reign and let you decide how you want to define your final model.

In [2]:
```
import os
import pandas as pd
import numpy as np
import sklearn.linear_model as lm
import matplotlib.pyplot as plt
import seaborn as sns
from pathlib import Path
from sqlalchemy import create_engine
from sklearn.model_selection import cross_val_score, train_test_split, GridSearcl

sns.set(style="whitegrid", palette="muted")

plt.rcParams['figure.figsize'] = (12, 9)
plt.rcParams['font.size'] = 12

%matplotlib inline
```

Training and Validation

The following code loads the training and validation data from part 1 into a Pandas DataFrame.

In [3]:
```
# Run this cell to load the data.
data_file = Path("./", "cleaned_data.hdf")
train_df = pd.read_hdf(data_file, "train")
val_df = pd.read_hdf(data_file, "val")
```

Testing

Here we load our testing data on which we will evaluate your model.

In [4]:
```
test_df = pd.read_csv("./proj2_test_data.csv")
test_df['tpep_pickup_datetime'] = pd.to_datetime(test_df['tpep_pickup_datetime']
test_df.head()
```

Out[4]:

	record_id	VendorID	tpep_pickup_datetime	passenger_count	trip_distance	pickup_longitude	pi
0	10000	1	2016-01-02 01:45:37	1	1.20	-73.982224	
1	19000	2	2016-01-02 03:05:16	1	10.90	-73.999977	
2	21000	1	2016-01-02 03:24:36	1	1.80	-73.986618	
3	23000	2	2016-01-02 03:47:38	1	5.95	-74.002922	
4	27000	1	2016-01-02 04:36:44	1	1.60	-73.986366	

In [5]:
```
test_df.describe()
```

Out[5]:

	record_id	VendorID	passenger_count	trip_distance	pickup_longitude	pickup_latitu
count	1.377400e+04	13774.000000	13774.000000	13774.000000	13774.000000	13774.0000
mean	3.465950e+07	1.536082	1.663642	2.954688	-72.953619	40.1879
std	2.015133e+07	0.498714	1.311739	3.704427	8.628431	4.7531
min	1.000000e+04	1.000000	0.000000	0.000000	-77.039436	0.0000
25%	1.719975e+07	1.000000	1.000000	1.000000	-73.992058	40.7351
50%	3.457400e+07	2.000000	1.000000	1.700000	-73.981846	40.7524
75%	5.216875e+07	2.000000	2.000000	3.157500	-73.967119	40.7672
max	6.940400e+07	2.000000	6.000000	104.800000	0.000000	40.8682

Modeling

We've finally gotten to a point where we can specify a simple model. Remember that we will be fitting our model on the training set we created in part 1. We will use our validation set to evaluate how well our model might perform on future data.

Reusable Pipeline

Throughout this assignment, you should notice that your data flows through a single processing pipeline several times. From a software engineering perspective, this should be sufficient motivation to abstract parts of our code into reusable functions/methods. We will now encapsulate our entire pipeline into a single function `process_data_gm` . gm is shorthand for "guided model".

In [6]:
```
# Copied from part 2
def haversine(lat1, lng1, lat2, lng2):
    """
    Compute haversine distance
    """
    lat1, lng1, lat2, lng2 = map(np.radians, (lat1, lng1, lat2, lng2))
    average_earth_radius = 6371
    lat = lat2 - lat1
    lng = lng2 - lng1
    d = np.sin(lat * 0.5) ** 2 + np.cos(lat1) * np.cos(lat2) * np.sin(lng * 0.5)
    h = 2 * average_earth_radius * np.arcsin(np.sqrt(d))
    return h

# Copied from part 2
def manhattan_distance(lat1, lng1, lat2, lng2):
    """
    Compute Manhattan distance
    """
    a = haversine(lat1, lng1, lat1, lng2)
    b = haversine(lat1, lng1, lat2, lng1)
    return a + b

# Copied from part 2
def bearing(lat1, lng1, lat2, lng2):
    """
    Compute the bearing, or angle, from (lat1, lng1) to (lat2, lng2).
    A bearing of 0 refers to a NORTH orientation.
    """
    lng_delta_rad = np.radians(lng2 - lng1)
    lat1, lng1, lat2, lng2 = map(np.radians, (lat1, lng1, lat2, lng2))
    y = np.sin(lng_delta_rad) * np.cos(lat2)
    x = np.cos(lat1) * np.sin(lat2) - np.sin(lat1) * np.cos(lat2) * np.cos(lng_de
    return np.degrees(np.arctan2(y, x))

# Copied from part 2
def add_time_columns(df):
    """
    Add temporal features to df
    """
    df.is_copy = False # propogate write to original dataframe
    df.loc[:, 'month'] = df['tpep_pickup_datetime'].dt.month
    df.loc[:, 'week_of_year'] = df['tpep_pickup_datetime'].dt.weekofyear
    df.loc[:, 'day_of_month'] = df['tpep_pickup_datetime'].dt.day
    df.loc[:, 'day_of_week'] = df['tpep_pickup_datetime'].dt.dayofweek
    df.loc[:, 'hour'] = df['tpep_pickup_datetime'].dt.hour
    df.loc[:, 'week_hour'] = df['tpep_pickup_datetime'].dt.weekday * 24 + df['hou
    return df

# Copied from part 2
def add_distance_columns(df):
    """
    Add distance features to df
    """
    df.is_copy = False # propogate write to original dataframe
    df.loc[:, 'manhattan'] = manhattan_distance(lat1=df['pickup_latitude'],
                                                lng1=df['pickup_longitude'],
```

```
                                                     lat2=df['dropoff_latitude'],
                                                     lng2=df['dropoff_longitude'])

    df.loc[:, 'bearing'] = bearing(lat1=df['pickup_latitude'],
                                   lng1=df['pickup_longitude'],
                                   lat2=df['dropoff_latitude'],
                                   lng2=df['dropoff_longitude'])
    df.loc[:, 'haversine'] = haversine(lat1=df['pickup_latitude'],
                                   lng1=df['pickup_longitude'],
                                   lat2=df['dropoff_latitude'],
                                   lng2=df['dropoff_longitude'])
    return df

def select_columns(data, *columns):
    return data.loc[:, columns]
```

In [7]:

```
def process_data_gm1(data, test=False):
    X = (
        data

        # Transform data
        .pipe(add_time_columns)
        .pipe(add_distance_columns)

        .pipe(select_columns,
              'pickup_longitude',
              'pickup_latitude',
              'dropoff_longitude',
              'dropoff_latitude',
              'manhattan',
             )
    )
    if test:
        y = None
    else:
        y = data['duration']

    return X, y
```

We will use our pipeline defined above to pre-process our training and test data in exactly the same way. Our functions make this relatively easy to do!

In [8]:
```
# Train
X_train, y_train = process_data_gm1(train_df)
X_val, y_val = process_data_gm1(val_df)
guided_model_1 = lm.LinearRegression(fit_intercept=True)
guided_model_1.fit(X_train, y_train)

# Predict
y_train_pred = guided_model_1.predict(X_train)
y_val_pred = guided_model_1.predict(X_val)
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```

Here, `y_val` are the correct durations for each ride, and `y_val_pred` are the predicted durations based on the 7 features above (`vendorID` , `passenger_count` , `pickup_longitude` , `pickup_latitude` , `dropoff_longitude` , `dropoff_latitude` , `manhattan`).

In [9]:
```
assert 600 <= np.median(y_train_pred) <= 700
assert 600 <= np.median(y_val_pred) <= 700
```

The resulting model really is a linear model just like we saw in class, i.e. the predictions are simply generated by the product $\Phi\theta$. For example, the line of code below generates a prediction for x_1 by computing $\phi_1^T\theta$. Here `guided_model_1.coef_` is θ and `X_train.iloc[0, :]` is ϕ_1.

Note that unlike in class, here the dummy intercept term is not included in Φ.

In [10]:
```
X_train.iloc[0, :].dot(guided_model_1.coef_) + guided_model_1.intercept_
```

Out[10]:
```
558.751330511368
```

We see that this prediction is exactly the same (except for possible floating point error) as generated by the `predict` function, which simply computes the product $\Phi\theta$, yielding predictions for every input.

In [11]:
```
y_train_pred[0]
```

Out[11]:
```
558.75133051135344
```

In this assignment, we will use Mean Absolute Error (MAE), a.k.a. mean L1 loss, to measure the quality of our models. As a reminder, this quantity is defined as:

$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Why may we want to use the MAE as a metric, as opposed to Mean Squared Error (MSE)? Using our domain knowledge that most rides are short in duration (median is roughly 600 seconds), we know that MSE is susceptible to outliers. Given that some of the outliers in our dataset are quite extreme, it is probably better to optimize for the majority of rides rather than for the outliers. You may want to remove some of these outliers later on.

In [12]:
```
def mae(actual, predicted):
    """
    Calculates MAE from actual and predicted values
    Input:
      actual (1D array-like): vector of actual values
      predicted (1D array-like): vector of predicted/fitted values
    Output:
      a float, the MAE
    """

    mae = np.mean(np.abs(actual - predicted))
    return mae
```

In [13]:
```
assert 200 <= mae(y_val_pred, y_val) <= 300
print("Validation Error: ", mae(y_val_pred, y_val))
```

```
Validation Error:  266.136130855
```

Side note: scikit-learn also has tools to compute mean absolute error (`sklearn.metrics.mean_absolute_error`). In fact, most metrics that we have discussed in this class can be found as part of the `sklearn.metrics` module (https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics). Some of these may come in handy as part of your feature engineering!

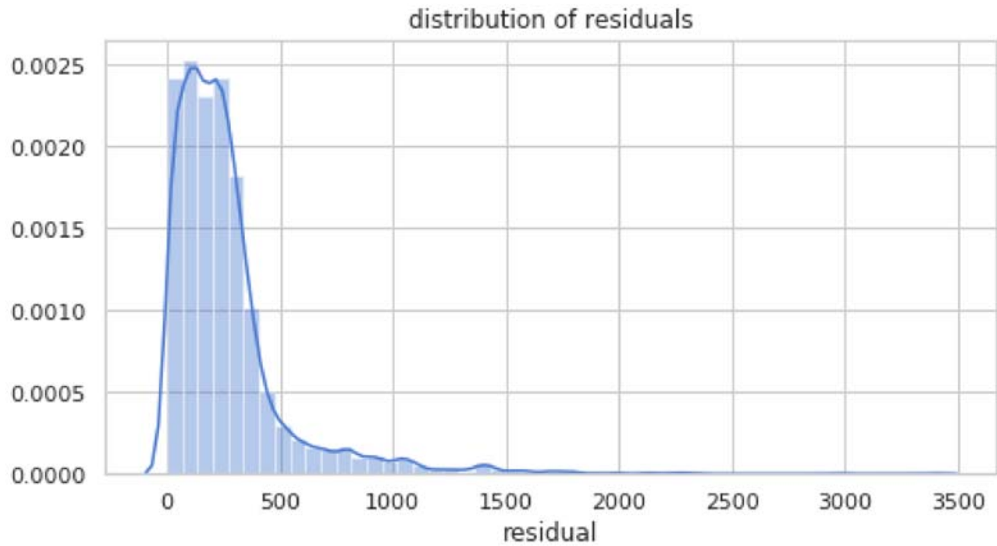
Visualizing Error

You should be getting between 200 and 300 MAE, which means your model was off by roughly 3-5 minutes on trips of average length 12 minutes. This is fairly decent performance given that our basic model uses only using the pickup/dropoff latitude and manhattan distance of the trip. 3-5 minutes may seem like a lot for a trip of 12 minutes, but keep in mind that this is the *average* error. This metric is susceptible to extreme outliers, which exist in our dataset.

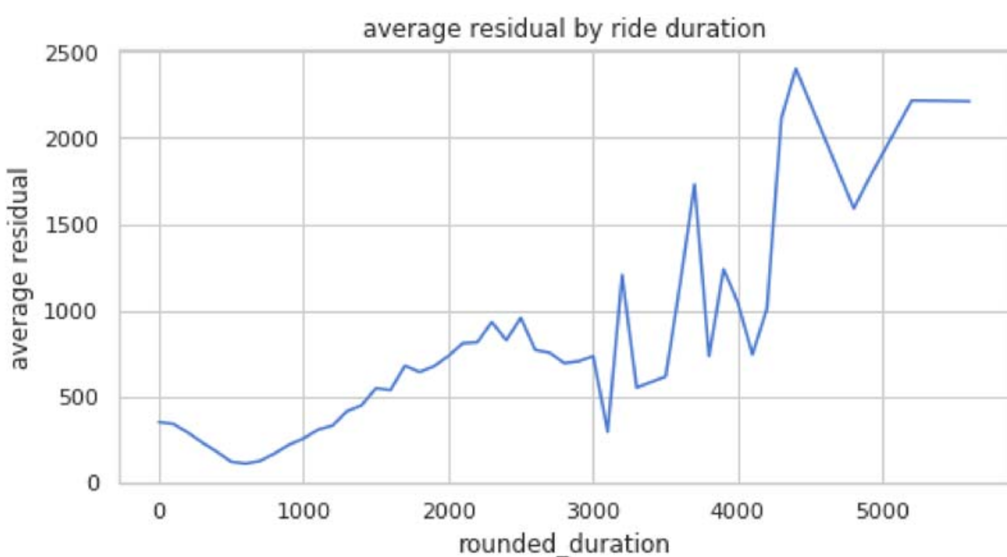
Now we will visualize the residual for the validation set. We will plot the following:

1. Distribution of residuals
2. Average residual grouping by ride duration

In [14]:
```
# Distribution of residuals
plt.figure(figsize=(8,4))
sns.distplot(np.abs(y_val - y_val_pred))
plt.xlabel('residual')
plt.title('distribution of residuals');
```



In [15]:
```
# Average residual grouping by ride duration
val_residual = X_val.copy()
val_residual['duration'] = y_val
val_residual['rounded_duration'] = np.around(y_val, -2)
val_residual['residual'] = np.abs(y_val - y_val_pred)
tmp = val_residual.groupby('rounded_duration').mean()
plt.figure(figsize=(8,4))
tmp['residual'].plot()
plt.ylabel('average residual')
plt.title('average residual by ride duration');
```



In the first visualization, we see that most of the residuals are centered around 250 seconds ~ 4 minutes. There is a minor right tail, suggesting that we are still unable to accurately fit some outliers in our data. The second visualization also suggests this, as we see the average residual increasing as a somewhat linear function of duration. But given that our average ride duration is roughly 600-700 seconds, it seems that we are indeed optimizing for the average ride because the residuals are smallest around 600-700.

Keep this in mind when creating your final model! Visualizing the error is a powerful tool and may help diagnose shortcomings of your model. Let's go ahead and submit to kaggle, although your error on the test set may be higher than 300.

Submission to Kaggle

The following code will write your predictions on the test dataset to a CSV, which you can submit to Kaggle. You may need to modify it to suit your needs, but we recommend you make a copy and preserve the original function.

Remember that if you've performed transformations or featurization on the training data, you must also perform the same transformations on the test data in order to make predictions. For example, if you've created features for the columns `pickup_datetime` or `pickup_latitude` on the training data, you must also extract the same features in order to use scikit-learn's `.predict(...)` method.

In [16]:
```
from datetime import datetime
def generate_submission(test, predictions, force=False):
    if force:
        if not os.path.isdir("submissions"):
            os.mkdir("submissions")
        submission_df = pd.DataFrame({
            "id": test_df.index.values,
            "duration": predictions,
        },
            columns=['id', 'duration'])

        timestamp = datetime.isoformat(datetime.now()).split(".")[0]

        submission_df.to_csv(f'submissions/submission_{timestamp}.csv', index=Fa

        print(f'Created a CSV file: submission_{timestamp}.csv')
        print('You may now upload this CSV file to Kaggle for scoring.')
```

In [17]:
```
X_test, _ = process_data_gm1(test_df, True)
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```

In [18]:
```
assert list(X_train.columns) == list(X_test.columns), "Different columns or diff
submission_predictions = (guided_model_1
                          .fit(X_train, y_train)
                          .predict(X_test))
submission_predictions = submission_predictions.astype(int)
submission_predictions[submission_predictions < 0] = 0
generate_submission(test_df, submission_predictions, True)
```

```
Created a CSV file: submission_2018-12-05T18:35:21.csv
You may now upload this CSV file to Kaggle for scoring.
```

In [19]:
```
# Check your submission
assert isinstance(submission_predictions, np.ndarray), "Submission not an array"
assert all(submission_predictions >= 0), "Duration must be non-negative"
assert issubclass(submission_predictions.dtype.type, np.integer), "Seconds must
```

Your Turn!

Now it's your turn! Draw upon everything you have learned this semester to find the best features to help your model accurately predict the duration of a taxi ride.

You may use whatever method you prefer in order to create features. You may use features that we created and features that you discovered yourself from any of the 2 datasets. However, we want to make it fair to students who are seeing these techniques for the first time. As such, you are only allowed regression models and their regularized forms. This means no random forest, k-nearest-neighbors, neural nets, etc.

Here are some ideas to improve your model:

- **Data selection**: January 2016 was an odd month for taxi rides due to the blizzard. Would it help to select training data differently?
- **Data cleaning**: Try cleaning your data in different ways. In particular, consider how to handle outliers.
- **Better features**: Explore the 2 datasets and find what features are most helpful. Utilize external datasets to improve your accuracy.
- **Regularization**: Try different forms of regularization to avoid fitting to the training set. Recall that `Ridge` and `Lasso` are the names of the classes in `sklearn.linear_model` that combine `LinearRegression` with regularization techniques.
- **Model selection**: You can adjust parameters of your model (e.g., the regularization parameter) to achieve higher accuracy. [GridSearchCV (http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html) may be helpful.
- **Validation**: Recall that you should use cross-validation to do feature and model selection properly! Otherwise, you will likely overfit to your training data.

There's many things you could try that could help your model. We have only suggested a few. Be creative and innovative! Please use `proj2_extras.ipynb` for all of your extraneous work. Note that you will be submitting `proj2_extras.ipynb` and we will be grading it. Please properly comment and format this notebook!

Once you are satisfied with your results, answer the questions in the Deliverables section. You may want to read this section in advance so you have an idea of what we're looking for.

Deliverables

Feature/Model Selection Process

Let's first look at selection of better features. In this following cell, describe the process of choosing good features to improve your model. You should use at least 3-4 sentences each to address the follow questions. Backup your responses with graphs supporting your claim (you can save figures and load them, no need to add the plotting code here). Use these questions to concisely summarize all of your extra work!

Question 1a

How did you find better features for your model?

In [20]:
```
q1a_answer = r"""

I use all the features from previous and train a Ridge Regression and a LASSO. At

"""
# YOUR CODE HERE
# raise NotImplementedError()
```

Question 1b

What did you try that worked / didn't work?

In [21]:
```
q1b_answer = r"""

I have tried Ridge and LASSO and it turns out that LASSO didn't work well in this

"""
# YOUR CODE HERE
# raise NotImplementedError()
```

Question 1c

What was surprising in your search for good features?

In [22]:
```
q1c_answer = r"""

I have removed the data on Jan 23 and removed the outliers, and I found that it h

"""
# YOUR CODE HERE
# raise NotImplementedError()
```

Question 2

Just as in the guided model above, you should encapsulate as much of your workflow into functions as possible. Define `process_data_fm` and `final model` in the cell below. In order to calculate your final model's MAE, we will run the code in the cell after that.

Note: You *MUST* name the model you wish to be evaluated on `final_model` . This is what we will be using to generate your predictions. We will take the state of `final_model` right after executing the cell below and run the following code:

```
# Load in test_df, solutions
X_test, _ = process_data_fm(test_df, True)
submission_predictions = final_model.predict(X_test)
# Generate score for autograding
```

We encourage you to conduct all of your exploratory work in `proj2_extras.ipynb` , which will be graded for 10 points.

In [23]:
```
from sklearn.linear_model import Ridge

def process_data_fm(data, test=False):
    X = (
        data
        # Transform data
        .pipe(add_time_columns)
        .pipe(add_distance_columns)
        .pipe(select_columns,
            'pickup_longitude',
            'pickup_latitude',
            'dropoff_longitude',
            'dropoff_latitude',
            'manhattan',
            'haversine',
            'hour',
            'trip_distance',
            'day_of_week',
            'total_amount',
            'tolls_amount',
            'tip_amount',
            'extra',
            'fare_amount'
        )
    )
    if test:
        y = None
    else:
        y = data['duration']
    return X, y

### remove outliers in training
train_df_remove = train_df[train_df['tpep_pickup_datetime'].dt.day != 23]
train_df_clean = train_df_remove[(train_df_remove['duration'] <= 4000) & (train_(

final_model = Ridge(alpha=9.82, solver="auto")
X_train_clean, y_train_clean = process_data_fm(train_df_clean)
X_val, y_val = process_data_fm(val_df)
final_model.fit(X_train_clean, y_train_clean)

# YOUR CODE HERE
# raise NotImplementedError()
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```

Out[23]:
```
Ridge(alpha=9.82, copy_X=True, fit_intercept=True, max_iter=None,
   normalize=False, random_state=None, solver='auto', tol=0.001)
```

In [24]:
```
X_train_clean.columns
```

Out[24]:
```
Index(['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'manhattan', 'haversine', 'hour', 'trip_distance',
       'day_of_week', 'total_amount', 'tolls_amount', 'tip_amount', 'extra',
       'fare_amount'],
      dtype='object')
```

In [25]:
```
test_df[['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'manhattan', 'haversine', 'hour', 'trip_distance',
       'day_of_week', 'total_amount', 'tolls_amount', 'tip_amount', 'extra',
       'fare_amount']].columns
```

Out[25]:
```
Index(['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'manhattan', 'haversine', 'hour', 'trip_distance',
       'day_of_week', 'total_amount', 'tolls_amount', 'tip_amount', 'extra',
       'fare_amount'],
      dtype='object')
```

In [26]:
```
### define a predict function
def predict(model, test_df):
    clean_index = (test_df['pickup_latitude'] <= 40.85) & (test_df['pickup_latitu
                  (test_df['dropoff_latitude'] <= 40.85) & (test_df['dropoff_la
                  (test_df['pickup_longitude'] <= -73.65) & (test_df['pickup_lo
                  (test_df['dropoff_longitude'] <= -73.65) & (test_df['dropoff_

    dirty_index = - clean_index
    if sum(dirty_index) == 0:
        return model.predict(test_df)

    clean_pred = model.predict(test_df.loc[clean_index])
    avg_duration = np.mean(clean_pred)

    pred = pd.DataFrame({
        "id": test_df.index.values,
        "duration": model.predict(test_df)
            },
                columns=["id", "duration"])

    pred.loc[dirty_index, "duration"] = avg_duration
    assert sum(clean_index) + sum(dirty_index) == len(test_df)

    return np.array(pred["duration"])
```

In [ ]:

In [ ]:

In [27]:
```
# Feel free to change this cell
# test_df_remove = test_df[test_df['tpep_pickup_datetime'].dt.day != 23]
# test_df_clean = test_df_remove[(test_df_remove['duration'] <= 4000) & (test_df_
# test_df['tpep_pickup_datetime'] = pd.to_datetime(test_df['tpep_pickup_datetime
# ### change outliers
# for i in range(len(test_df)):
#     if test_df.iloc[i, 19] < 0:
#         test_df.iloc[i, 19] = 11.300000
#     if test_df.iloc[i, 18] < 0:
#         train_df.iloc[i, 18] = 0.3
#     if test_df.iloc[i, 15] < 0:
#         train_df.iloc[i, 15] = 0.5
#     if test_df.iloc[i, 14] < 0:
#         train_df.iloc[i, 14] = 0.0
#     if test_df.iloc[i, 13] < 0:
#         train_df.iloc[i, 13] = 9.0

# X_test, _ = process_data_fm(test_df, True)

test_df = test_df[['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'manhattan', 'haversine', 'hour', 'trip_distance',
       'day_of_week', 'total_amount', 'tolls_amount', 'tip_amount', 'extra',
       'fare_amount']]
final_predictions = predict(final_model, test_df)
final_predictions = final_predictions.astype(int)
generate_submission(test_df, final_predictions, False) # Change to true to gener
```

In [28]:
```
train_df_clean.head()
```

Out[28]:

	record_id	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_dis
13242	5711100	1	2016-01-17 17:48:41	2016-01-17 17:55:53	1	
12723	4989400	1	2016-01-17 01:18:39	2016-01-17 01:21:15	1	
8508	2436400	2	2016-01-12 09:07:00	2016-01-12 09:41:17	1	
21304	10899100	2	2016-01-29 09:07:54	2016-01-29 09:18:25	1	
3817	1319400	1	2016-01-06 11:44:54	2016-01-06 11:49:55	1	

5 rows × 30 columns

In [29]:
```
test_df.head()
```

Out[29]:

	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	manhattan	haversine	hou
0	-73.982224	40.768620	-73.983765	40.779598	1.350561	1.227654	
1	-73.999977	40.738121	-73.888657	40.824364	18.968770	13.409519	
2	-73.986618	40.747379	-73.978508	40.729622	2.657731	2.089418	
3	-74.002922	40.744572	-73.942413	40.786419	9.750712	6.900764	
4	-73.986366	40.759464	-73.963081	40.760353	2.060014	1.963656	

Question 3

The following hidden cells will test your model on the test set. Please do not delete any of them if you want credit!

In [30]:
```
# NO TOUCH
```

In [31]:
```
# NOH
```

In [32]:
```
# STAHP
```

In [33]:
```
# NO MOLESTE
```

In [34]:
```
# VA-T'EN
```

In [35]:
```
# NEIN
```

In [36]:
```
# PLSNO
```

In [37]:
```
# THIS SPACE IS NOT YOURS
```

In [38]:
```
# TAWDEETAW
```

In [39]:
```
# MAU LEN
```

In [40]:
```
# ALMOST
```

In [41]:
```
# TO
```

In [42]:
```
# THE
```

In [43]:
```
# END
```

In [44]:
```
# Hmph
```

In [45]:
```
# Good riddance
```

In [46]:
```
generate_submission(test_df, submission_predictions, True)
```

```
Created a CSV file: submission_2018-12-05T18:35:22.csv
You may now upload this CSV file to Kaggle for scoring.
```

This should be the format of your CSV file.
Unix-users can verify it running `!head submission_{datetime}.csv` in a jupyter notebook cell.

```
id,duration
id3004672,965.3950873305439
id3505355,1375.0665915134596
id1217141,963.2285454171943
id2150126,1134.7680929570924
id1598245,878.5495792656438
id0668992,831.6700312449248
id1765014,993.1692116960185
id0898117,1091.1171629594755
id3905224,887.9037911118357
```

Kaggle link: [https://www.kaggle.com/t/f8b3c6acc3a045cab152060a5bc79670 (https://www.kaggle.com/t/f8b3c6acc3a045cab152060a5bc79670)](https://www.kaggle.com/t/f8b3c6acc3a045cab152060a5bc79670)

Submission

You're almost done!

Before submitting this assignment, ensure that you have:

1. Restarted the Kernel (in the menubar, select Kernel→Restart & Run All)
2. Validated the notebook by clicking the "Validate" button.

Then,

1. **Submit** the assignment via the Assignments tab in **Datahub**
2. **Upload and tag** the manually reviewed portions of the assignment on **Gradescope**

Before you turn in the homework, make sure everything runs as expected. To do so, select **Kernel →Restart & Run All** in the toolbar above. Remember to submit both on **DataHub** and **Gradescope**.

Please fill in your name and include a list of your collaborators below.

In [1]:
```
NAME = "Benjamin Liu"
COLLABORATORS = "Victor Ding"
```

Project 2: NYC Taxi Rides

Extras

Put all of your extra work in here. Feel free to save figures to use when completing Part 4.

In [2]:
```
import os
import pandas as pd
import numpy as np
import sklearn.linear_model as lm
from sklearn.model_selection import cross_val_score, train_test_split, GridSearc
import matplotlib.pyplot as plt
import seaborn as sns
from pathlib import Path
from sqlalchemy import create_engine
```

In [3]:
```
# Run this cell to load the data.
data_file = Path("./", "cleaned_data.hdf")
train_df = pd.read_hdf(data_file, "train")
val_df = pd.read_hdf(data_file, "val")
test_df = pd.read_csv("./proj2_test_data.csv")
test_df['tpep_pickup_datetime'] = pd.to_datetime(test_df['tpep_pickup_datetime']
```

Typesetting math: 100%

In [4]:
```
# get the summary of train df
train_df.describe()
```

Out[4]:

	record_id	VendorID	passenger_count	trip_distance	pickup_longitude	pickup_latitu
count	1.772400e+04	17724.000000	17724.000000	17724.000000	17724.000000	17724.0000
mean	5.320997e+06	1.535150	1.677104	2.791220	-73.973560	40.7509
std	3.158004e+06	0.498777	1.324193	3.407549	0.037279	0.0274
min	6.000000e+02	1.000000	1.000000	0.000000	-74.018150	40.6316
25%	2.604200e+06	1.000000	1.000000	1.000000	-73.991783	40.7375
50%	5.208950e+06	2.000000	1.000000	1.620000	-73.981541	40.7543
75%	8.215850e+06	2.000000	2.000000	3.000000	-73.966925	40.7684
max	1.090610e+07	2.000000	6.000000	35.430000	-73.775398	40.8471

In [5]:
```
# display the summary of test df
test_df.describe()
```

Out[5]:

	record_id	VendorID	passenger_count	trip_distance	pickup_longitude	pickup_latitu
count	1.377400e+04	13774.000000	13774.000000	13774.000000	13774.000000	13774.0000
mean	3.465950e+07	1.536082	1.663642	2.954688	-72.953619	40.1879
std	2.015133e+07	0.498714	1.311739	3.704427	8.628431	4.7531
min	1.000000e+04	1.000000	0.000000	0.000000	-77.039436	0.0000
25%	1.719975e+07	1.000000	1.000000	1.000000	-73.992058	40.7351
50%	3.457400e+07	2.000000	1.000000	1.700000	-73.981846	40.7524
75%	5.216875e+07	2.000000	2.000000	3.157500	-73.967119	40.7672
max	6.940400e+07	2.000000	6.000000	104.800000	0.000000	40.8682

In [ ]:

In [6]:
```
### Try: remove Jan 23 data, 17724 - 17603 is removed
print(train_df.shape)
train_remove = train_df[train_df['tpep_pickup_datetime'].dt.day != 23]
print(train_remove.shape)
```

```
(17724, 21)
(17603, 21)
```

In [7]:
```
### Try: replace outliers with the median in training data
train_df_copy = train_remove.copy()
for i in range(len(train_df_copy)):
    if train_df_copy.iloc[i, 19] < 0:
        train_df_copy.iloc[i, 19] = 11.300000
    if train_df_copy.iloc[i, 18] < 0:
        train_df_copy.iloc[i, 18] = 0.3
    if train_df_copy.iloc[i, 15] < 0:
        train_df_copy.iloc[i, 15] = 0.5
    if train_df_copy.iloc[i, 14] < 0:
        train_df_copy.iloc[i, 14] = 0.0
    if train_df_copy.iloc[i, 13] < 0:
        train_df_copy.iloc[i, 13] = 9.0
train_df_copy.describe()
```

Out[7]:

	record_id	VendorID	passenger_count	trip_distance	pickup_longitude	pickup_latitu
count	1.760300e+04	17603.000000	17603.000000	17603.000000	17603.000000	17603.0000
mean	5.294103e+06	1.534682	1.676135	2.790611	-73.973551	40.7510
std	3.152077e+06	0.498810	1.323330	3.406663	0.037278	0.0273
min	6.000000e+02	1.000000	1.000000	0.000000	-74.018150	40.6316
25%	2.585650e+06	1.000000	1.000000	1.000000	-73.991768	40.7375
50%	5.175400e+06	2.000000	1.000000	1.610000	-73.981529	40.7543
75%	8.155700e+06	2.000000	2.000000	3.000000	-73.966923	40.7684
max	1.090610e+07	2.000000	6.000000	35.430000	-73.775398	40.8471

Typesetting math: 100%

In [8]:
```
### Copy from part 2, data pre-processing
def haversine(lat1, lng1, lat2, lng2):
    """
    Compute haversine distance

    The haversine formula determines the great-circle distance between two points
    on a sphere given their longitudes and latitudes. Important in navigation, it
    is a special case of a more general formula in spherical trigonometry,
    the law of haversines, that relates the sides and angles of spherical triangl
    """
    lat1, lng1, lat2, lng2 = map(np.radians, (lat1, lng1, lat2, lng2))
    average_earth_radius = 6371
    lat = lat2 - lat1
    lng = lng2 - lng1
    d = np.sin(lat * 0.5) ** 2 + np.cos(lat1) * np.cos(lat2) * np.sin(lng * 0.5)
    h = 2 * average_earth_radius * np.arcsin(np.sqrt(d))
    return h

def manhattan_distance(lat1, lng1, lat2, lng2):
    """
    Computes Manhattan distance

    The name alludes to the grid layout of most streets on the island of Manhatta
    which causes the shortest path a car could take between two intersections in
    to have length equal to the intersections' distance in taxicab geometry.
    """
    a = haversine(lat1, lng1, lat1, lng2)
    b = haversine(lat1, lng1, lat2, lng1)
    return a + b

def bearing(lat1, lng1, lat2, lng2):
    """
    Compute the bearing, or angle, from (lat1, lng1) to (lat2, lng2).
    A bearing of 0 refers to a NORTH orientation.
    """
    lng_delta_rad = np.radians(lng2 - lng1)
    lat1, lng1, lat2, lng2 = map(np.radians, (lat1, lng1, lat2, lng2))
    y = np.sin(lng_delta_rad) * np.cos(lat2)
    x = np.cos(lat1) * np.sin(lat2) - np.sin(lat1) * np.cos(lat2) * np.cos(lng_de
    return np.degrees(np.arctan2(y, x))

def add_distance_columns(df):
    df.loc[:, 'manhattan'] = manhattan_distance(lat1=df['pickup_latitude'],
                                                lng1=df['pickup_longitude'],
                                                lat2=df['dropoff_latitude'],
                                                lng2=df['dropoff_longitude'])

    df.loc[:, 'bearing'] = bearing(lat1=df['pickup_latitude'],
                                   lng1=df['pickup_longitude'],
                                   lat2=df['dropoff_latitude'],
                                   lng2=df['dropoff_longitude'])
    df.loc[:, 'haversine'] = haversine(lat1=df['pickup_latitude'],
                                   lng1=df['pickup_longitude'],
                                   lat2=df['dropoff_latitude'],
                                   lng2=df['dropoff_longitude'])
```

Typesetting math: 100%

```
    return df

def add_time_columns(df):
    """
    Add temporal features to df
    """
    df.is_copy = False
    df.loc[:, 'month'] = df['tpep_pickup_datetime'].dt.month
    df.loc[:, 'week_of_year'] = df['tpep_pickup_datetime'].dt.weekofyear
    df.loc[:, 'day_of_month'] = df['tpep_pickup_datetime'].dt.day
    df.loc[:, 'day_of_week'] = df['tpep_pickup_datetime'].dt.dayofweek
    df.loc[:, 'hour'] = df['tpep_pickup_datetime'].dt.hour
    df.loc[:, 'week_hour'] = df['tpep_pickup_datetime'].dt.weekday * 24 + df['hou

    # No real need to return here, but we harmonize with remove_outliers for late
    return df

def select_columns(data, *columns):
    return data.loc[:, columns]
```

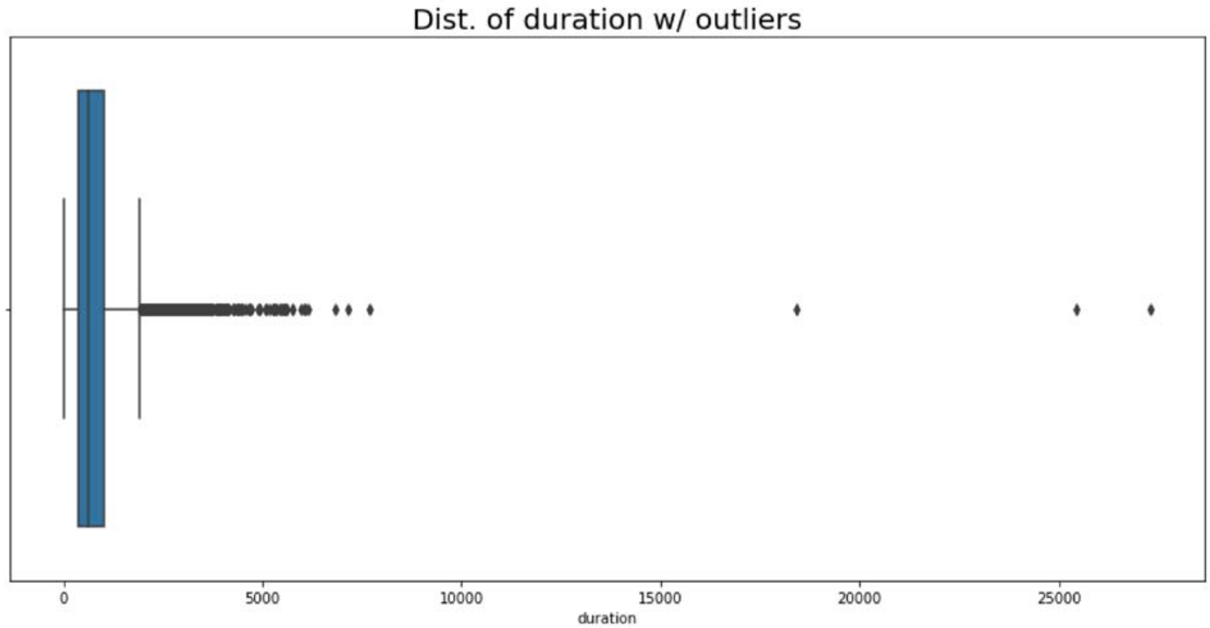
Typesetting math: 100%

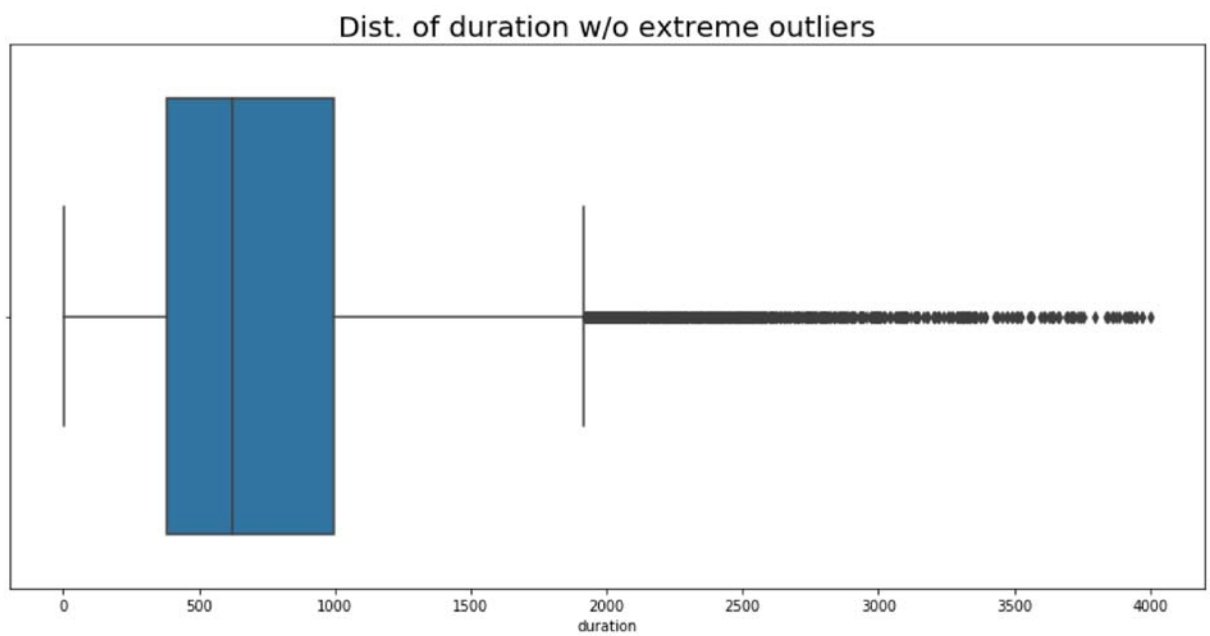
In [9]:

```
### Try: LASSO and Ridge Regression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
def process_data_fm(data, test=False):

    X = (
        data
        # Transform data
        .pipe(add_time_columns)
        .pipe(add_distance_columns)
        .pipe(select_columns,
        'pickup_longitude',
        'pickup_latitude',
        'dropoff_longitude',
        'dropoff_latitude',
        'manhattan',
        'haversine',
        'hour',
        'trip_distance',
        'day_of_week',
        'total_amount',
        'tolls_amount',
        'tip_amount',
        'extra',
        'fare_amount'
        )
    )
    if test:
        y = None
    else:
        y = data['duration']
    return X, y

def mae(actual, predicted):
    """
    mean abs error
    """
    return np.mean(np.abs(actual - predicted))
```

Typesetting math: 100%

In [10]:
```
# draw a plot to view the outliers in the duration of training data
plt.figure(figsize=(15, 7))
sns.boxplot(train_df['duration'])
plt.title('Dist. of duration w/ outliers', fontsize=20)
plt.show()
plt.figure(figsize=(15, 7))
eva_train = train_df.loc[train_df['duration'] <= 4000]
plt.title('Dist. of duration w/o extreme outliers', fontsize=20)
sns.boxplot(eva_train['duration'])
plt.show()
```





Typesetting math: 100%

In [11]:
```
train_df_clean = train_df_copy[(train_df_copy['duration'] <= 4000) & (train_df_c
X_train_new, y_train_new = process_data_fm(train_df_clean)
X_val_new, y_val_new = process_data_fm(val_df)
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```

In [12]:
```
### Try: compare Lasso and Ridge
model = Lasso(alpha=1)
model.fit(X_train_new, y_train_new)
y_train_pred_new = model.predict(X_train_new)
y_val_pred_new = model.predict(X_val_new)
print(mae(y_train_pred_new, y_train_new))
print(mae(y_val_pred_new, y_val_new))
```

```
91.6189609903
110.372183238
```

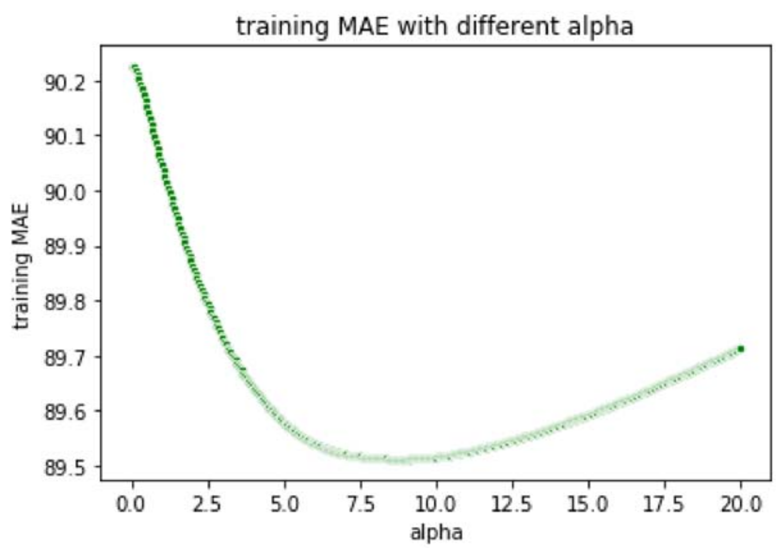
```
/srv/conda/envs/data100/lib/python3.6/site-packages/sklearn/linear_model/coordi
nate_descent.py:491: ConvergenceWarning: Objective did not converge. You might
want to increase the number of iterations. Fitting data with very small alpha m
ay cause precision problems.
  ConvergenceWarning)
```

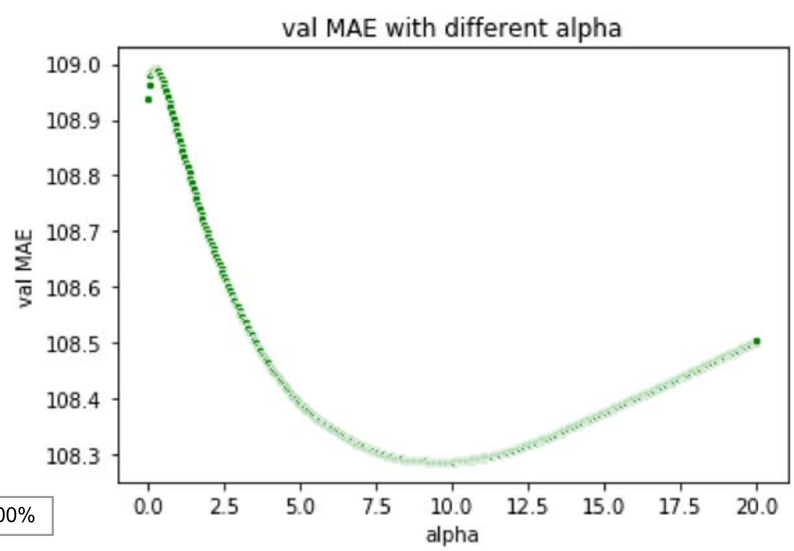
In [13]:
```
### Try: compare Lasso and Ridge
model = Ridge(alpha=1)
model.fit(X_train_new, y_train_new)
y_train_pred_new = model.predict(X_train_new)
y_val_pred_new = model.predict(X_val_new)
print(mae(y_train_pred_new, y_train_new))
print(mae(y_val_pred_new, y_val_new))
```

```
90.0460977754
108.872421536
```

Typesetting math: 100%

In [14]:
```
# Find the best hyperparam regualrization alpha for Ridge Regression
penalty = np.linspace(0, 20, 400)
mae_train = []
mae_val = []
for i in penalty:
    model = Ridge(alpha=i)
    model.fit(X_train_new, y_train_new)
    y_train_pred = model.predict(X_train_new)
    y_val_pred = model.predict(X_val_new)
    mae_train.append(mae(y_train_pred, y_train_new))
    mae_val.append(mae(y_val_pred, y_val_new))

sns.scatterplot(x=penalty, y=mae_train, s=20, color='g')
plt.xlabel('alpha')
plt.ylabel('training MAE')
plt.title('training MAE with different alpha')
plt.show()
sns.scatterplot(x=penalty, y=mae_val, s=20, color='g')
plt.xlabel('alpha')
plt.ylabel('val MAE')
plt.title('val MAE with different alpha')
plt.show()
```



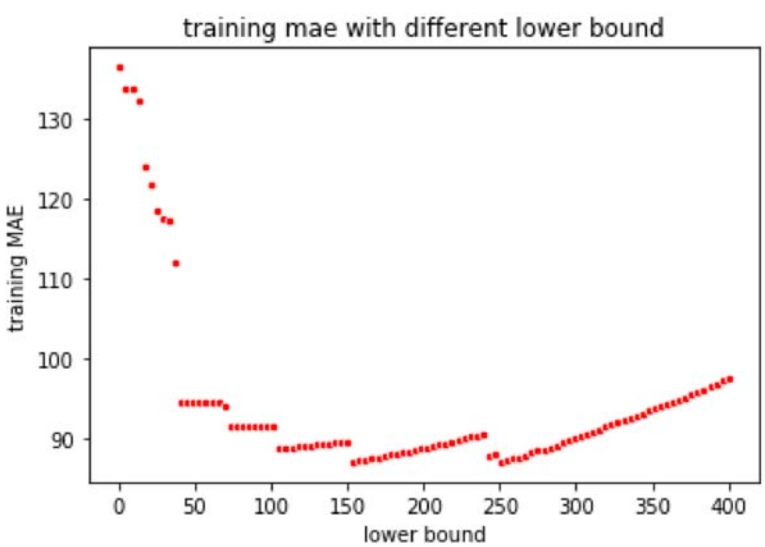


Typesetting math: 100%

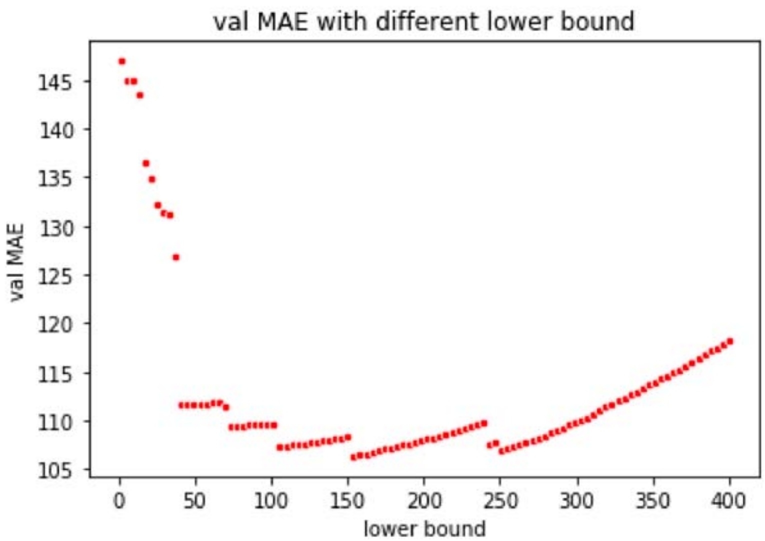
Typesetting math: 100%

In [15]:
```
### Remove outliers in training set
lower = np.linspace(1, 400, 100)
mae_train_1 = []
mae_val_1 = []
for i in lower:
    train_df_clean = train_df[(train_df['duration'] <= 4000) & (train_df['durati
    X_train_new, y_train_new = process_data_fm(train_df_clean)
    X_val_new, y_val_new = process_data_fm(val_df)
    model = Ridge(alpha=9.82)
    model.fit(X_train_new, y_train_new)
    y_train_pred_new = model.predict(X_train_new)
    y_val_pred_new = model.predict(X_val_new)
    mae_train_1.append(mae(y_train_pred_new, y_train_new))
    mae_val_1.append(mae(y_val_pred_new, y_val_new))

sns.scatterplot(x=lower, y=mae_train_1, s=20, color='r')
plt.xlabel('lower bound')
plt.ylabel('training MAE')
plt.title('training mae with different lower bound')
plt.show()
sns.scatterplot(x=lower, y=mae_val_1, s=20, color='r')
plt.xlabel('lower bound')
plt.ylabel('val MAE')
plt.title('val MAE with different lower bound')
plt.show()
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```



Typesetting math: 100%



In [16]:

```
### Find the best solver for Ridge Regression
### auto solver, alpha = 9.82
train_df_clean = train_df_copy[(train_df_copy['duration'] <= 4000) & (train_df_c
X_train_new, y_train_new = process_data_fm(train_df_clean)
X_val_new, y_val_new = process_data_fm(val_df)

grid_params = {'solver':['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag',
'saga']}
model = Ridge(alpha=9.82)
final_model = GridSearchCV(estimator=model, param_grid=grid_params,cv=5)
final_model.fit(X_train_new, y_train_new)

best_param = final_model.best_params_
print(best_param)

y_train_pred_new = final_model.predict(X_train_new)
y_val_pred_new = final_model.predict(X_val_new)
print(mae(y_train_pred_new, y_train_new))
print(mae(y_val_pred_new, y_val_new))
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)

{'solver': 'auto'}
89.5152021478
108.286184627
```

Typesetting math: 100%

In [17]:
```
### define a predict function
def predict(model, test_df):
    clean_index = (test_df['pickup_latitude'] <= 40.85) & (test_df['pickup_latitu
                   (test_df['dropoff_latitude'] <= 40.85) & (test_df['dropoff_la
                   (test_df['pickup_longitude'] <= -73.65) & (test_df['pickup_lo
                   (test_df['dropoff_longitude'] <= -73.65) & (test_df['dropoff_

    dirty_index = - clean_index
    if sum(dirty_index) == 0:
        return model.predict(test_df)

    clean_pred = model.predict(test_df.loc[clean_index])
    avg_duration = np.mean(clean_pred)

    pred = pd.DataFrame({
        "id": test_df.index.values,
        "duration": model.predict(test_df)
            },
                columns=["id", "duration"])

    pred.loc[dirty_index, "duration"] = avg_duration
    assert sum(clean_index) + sum(dirty_index) == len(test_df)

    return np.array(pred["duration"])
```

In [18]:
```
from datetime import datetime
def generate_submission(test, predictions, force=False):
    if force:
        if not os.path.isdir("submissions"):
            os.mkdir("submissions")
        submission_df = pd.DataFrame({
            "id": test_df.index.values,
            "duration": predictions,
        },
            columns=['id', 'duration'])

        timestamp = datetime.isoformat(datetime.now()).split(".")[0]

        submission_df.to_csv(f'submissions/submission_{timestamp}.csv', index=Fa

        print(f'Created a CSV file: submission_{timestamp}.csv')
        print('You may now upload this CSV file to Kaggle for scoring.')
```

Typesetting math: 100%

In [19]:
```
test_df.head()
```

Out[19]:

	record_id	VendorID	tpep_pickup_datetime	passenger_count	trip_distance	pickup_longitude	pi
0	10000	1	2016-01-02 01:45:37	1	1.20	-73.982224	
1	19000	2	2016-01-02 03:05:16	1	10.90	-73.999977	
2	21000	1	2016-01-02 03:24:36	1	1.80	-73.986618	
3	23000	2	2016-01-02 03:47:38	1	5.95	-74.002922	
4	27000	1	2016-01-02 04:36:44	1	1.60	-73.986366	

In [20]:
```
_, _ = process_data_fm(test_df, True)
test_df = test_df[['pickup_longitude', 'pickup_latitude', 'dropoff_longitude',
       'dropoff_latitude', 'manhattan', 'haversine', 'hour', 'trip_distance',
       'day_of_week', 'total_amount', 'tolls_amount', 'tip_amount', 'extra',
       'fare_amount']]
final_predictions = predict(final_model, test_df)
final_predictions = final_predictions.astype(int)
generate_submission(test_df, final_predictions, True)
```

```
Created a CSV file: submission_2018-12-05T19:00:22.csv
You may now upload this CSV file to Kaggle for scoring.
```

```
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
8: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  object.__getattribute__(self, name)
/srv/conda/envs/data100/lib/python3.6/site-packages/pandas/core/generic.py:438
9: FutureWarning: Attribute 'is_copy' is deprecated and will be removed in a fu
ture version.
  return object.__setattr__(self, name, value)
```

Submission

You're almost done!

Before submitting this assignment, ensure that you have:

1. Restarted the Kernel (in the menubar, select Kernel$\rightarrow$Restart & Run All)
2. Validated the notebook by clicking the "Validate" button.

Then,

1. **Submit** the assignment via the Assignments tab in **Datahub**
2. **Upload and tag** the manually reviewed portions of the assignment on **Gradescope**

In [ ]:

Typesetting math: 100%

```
In [ ]:
```

Typesetting math: 100%